



Design of Rapid Warning System for Detection of Pollution and Anomalies in High Dimensional Process by Using Adaptive Network Fuzzy Inference System

Monireh Khayat¹, Rassoul Noorossana^{2*}, Paria Soleimani³, Sadigh Rasissi⁴

¹ PhD Student, Department of Industrial Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran

² Professor, School of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran

³ Associate Professor, Department of Industrial Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran

⁴ Professor, Department of Industrial Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran

Abstract

Intelligent water quality monitoring and the early identification of pollution events have become major challenges in water resources management because of the nonlinear, noisy, and uncertain nature of environmental data. This study proposes a hybrid framework based on wavelet transformation and an adaptive neuro-fuzzy inference system to develop an early warning system. First, total phosphorus (TP) concentration was estimated using a surrogate relationship. The resulting time series was then preprocessed using wavelet transformation to reduce noise while preserving the meaningful signal components and was subsequently used as input to the ANFIS model. Following model training, residual errors were calculated as the differences between the observed and predicted values, and the anomaly-detection thresholds were established by fitting several statistical distributions. The Akaike and Bayesian information criteria and the Kolmogorov–Smirnov test indicated that the logistic distribution provided a better fit to the residual errors than the normal and Laplace distributions. To evaluate the proposed framework, two pollution scenarios with different intensities were defined, and performance indicators derived from the confusion matrix were calculated. Under the second scenario, the Wavelet–ANFIS model achieved the best performance, with an accuracy of 98.07%, recall of 98.98%, specificity of 99.04%, and an F1-score of 99.08%. The findings further demonstrated that wavelet preprocessing reduced prediction errors, improved model stability, and enhanced the detection of actual pollution events. Overall, the proposed framework can serve as an effective tool for real-time water quality monitoring and early pollution warning in environments characterized by nonlinear, noisy, and uncertain data.

Keywords: water pollution, anomaly, high dimensional process, adaptive network fuzzy inference system, wavelet

* Corresponding Author: Rassoul Noorossana
Email: rassoul@iust.ac.ir



مقاله پژوهشی

تاریخ پذیرش: ۱۴۰۵/۵/۱۴

تاریخ دریافت: ۱۴۰۵/۳/۸

طراحی سامانه‌ی هشدار سریع برای تشخیص آلودگی و ناهنجاری‌ها در فرایندهای با ابعاد بالا با استفاده از سیستم استنتاج فازی عصبی تطبیقی

منیره خیاط^۱، رسول نورالسنّا^{۲*}، پریا سلیمانی^۳، صدیق رئیسی^۴

^۱ دانشجوی دکتری گروه مهندسی صنایع، دانشگاه آزاد اسلامی واحد تهران جنوب، تهران، ایران

^۲ استاد دانشکده مهندسی صنایع، دانشگاه علم و صنعت ایران، تهران، ایران

^۳ دانشیار، گروه مهندسی صنایع، واحد تهران جنوب، دانشگاه آزاد اسلامی، تهران، ایران

^۴ استاد، گروه مهندسی صنایع، واحد تهران جنوب، دانشگاه آزاد اسلامی، تهران، ایران

چکیده

پایش هوشمند کیفیت آب و شناسایی زودهنگام رخدادهای آلودگی، به دلیل ماهیت غیرخطی، نویزی و نامطمئن داده‌های محیطی، از چالش‌های اساسی مدیریت منابع آب به شمار می‌رود. این پژوهش با هدف توسعه یک سامانه هشدار سریع، چارچوبی ترکیبی مبتنی بر تبدیل موجک و سیستم استنتاج فازی-عصبی تطبیقی ارائه می‌کند. در مرحله نخست، غلظت فسفر کل (TP) با استفاده از یک رابطه جانشین برآورد شد. سپس سری زمانی حاصل به منظور کاهش نویز و حفظ مؤلفه‌های معنادار سیگنال، با استفاده از تبدیل موجک پیش‌پردازش و به عنوان ورودی مدل ANFIS به کار گرفته شد. پس از آموزش مدل، خطای باقیمانده از اختلاف میان مقادیر واقعی و پیش‌بینی شده محاسبه شد و آستانه تشخیص ناهنجاری از طریق برازش توزیع‌های آماری مختلف تعیین گردید. نتایج معیارهای اطلاعاتی آکائیک و بیزی و آزمون کولموگوروف-اسمیرنوف نشان داد که توزیع لجستیک، در مقایسه با توزیع‌های نرمال و لاپلاس، برازش مناسب‌تری برای مدل‌سازی خطاهای باقیمانده دارد. برای ارزیابی عملکرد چارچوب پیشنهادی، دو سناریوی آلودگی با شدت‌های متفاوت تعریف و شاخص‌های استخراج‌شده از ماتریس درهم‌ریختگی محاسبه شد. مدل Wavelet-ANFIS در سناریوی دوم، با دستیابی به صحت ۹۸،۰۷ درصد، فراخوانی ۹۸،۹۸ درصد، ویژگی ۹۹،۰۴ درصد و امتیاز F1 برابر با ۹۹،۰۸ درصد، بهترین عملکرد را نشان داد. یافته‌ها همچنین بیانگر آن است که پیش‌پردازش موجکی موجب کاهش خطای پیش‌بینی، افزایش پایداری مدل و بهبود قابلیت شناسایی رخدادهای واقعی آلودگی می‌شود. در مجموع، چارچوب پیشنهادی می‌تواند به عنوان ابزاری مؤثر برای پایش برخط کیفیت آب و صدور هشدار سریع در محیط‌های دارای داده‌های غیرخطی، نویزی و همراه با عدم قطعیت مورد استفاده قرار گیرد.

کلمات کلیدی: آلودگی آب، ناهنجاری، فرایند با ابعاد بالا، سیستم استنتاج فازی عصبی تطبیقی، تبدیل موجک

۱-مقدمه

آب آشامیدنی حیاتی‌ترین نیاز بشر است، اما تنها حدود سه درصد از منابع آبی کره‌ی زمین را تشکیل می‌دهد و از همین مقدار اندک نیز تنها بخش محدودی به‌صورت آب‌های سطحی و زیرزمینی در دسترس است. در سال‌های اخیر، رشد روزافزون جمعیت، گسترش سریع شهرنشینی، توسعه‌ی فعالیت‌های صنعتی و کشاورزی، و دفع یا رهاسازی غیراصولی فاضلاب در حوضه‌های آبریز و آبخوان‌ها، از مهم‌ترین نگرانی‌های دولت‌ها و تصمیم‌گیرندگان حوزه‌ی مدیریت منابع آب به شمار می‌رود. در راستای اهداف توسعه‌ی پایدار، تدوین راهبردها و سیاست‌گذاری‌های سیاسی و اقتصادی مرتبط، اهمیتی ویژه یافته است. [۱] شاخص‌های کیفیتی متعددی وجود دارند که تأثیر قابل توجهی بر سلامت آب‌های سطحی دارند. این شاخص‌ها متأثر از عواملی چون ناهمگونی شرایط طبیعی، ویژگی‌های متفاوت لایه‌های خاک در مسیر جریان، وجود منابع معدنی در مجاورت آبخوان‌ها، نوع کاربری اراضی اطراف رودخانه‌ها، بارش‌های موسمی و انتقال مواد معدنی توسط رواناب‌ها، نشت یا خرابی سامانه‌های فاضلاب شهری و صنعتی هستند [۲]. این عوامل در کنار وابستگی شدید به متغیرهای محیطی مانند نور، بارش، دما و pH، موجب بروز عدم قطعیت ذاتی در کیفیت آب می‌شوند. مسئله‌ی عدم قطعیت و احتمال فراتر رفتن شاخص‌های کیفیت از حدود مجاز، یکی از تهدیدهای جدی برای سلامت انسان، گیاهان، جانوران و گونه‌های آبی به‌شمار می‌رود [۳].

بسیاری از شاخص‌های کلیدی کیفیت آب، مانند کل نیتروژن (TN)، کل فسفر (TP) و کل جامدات معلق (TSS)، را نمی‌توان با وضوح زمانی بالا پایش کرد و برخی از آن‌ها نیز توسط حسگرها به‌صورت برخط قابل اندازه‌گیری نیستند [۴]. بنابراین لازم است این شاخص‌ها به کمک داده‌های پایش برخط و با استفاده از روابط رگرسیونی ساده بیان شوند. روابط جانشین می‌توانند برای تخمین غلظت شاخص‌های کیفی با وضوح زمانی بسیار بالا به کار روند. در پژوهشی، رابطه‌ی میان کدورت با TP و TSS بررسی شد تا بتوان برآوردهایی با بسامد بالا از این پارامترها فراهم آورد؛ شاخص‌هایی که به روش‌های سنتی قابل سنجش دقیق نیستند [۵]. این روابط جانشین در مطالعات مختلف برای اندازه‌گیری بار آلودگی در آب [۶]، هشدار شکوفایی جلبک‌ها [۷] و اعتبارسنجی مدل‌های هیدرولوژیکی [۸] به کار رفته‌اند. بنابراین، پایش، پیش‌بینی و ارزیابی منابع آبی از منظر وجود آلاینده‌ها اهمیت ویژه‌ای دارد. در سال‌های اخیر، پژوهشگران پیشرفت‌های قابل توجهی در توسعه و بهبود روش‌های تشخیص ناهنجاری حاصل کرده‌اند؛ با در نظر گرفتن عواملی مانند زمان هشدار، دقت در پیش‌بینی شاخص‌های مورد مطالعه، و توانایی در تشخیص و طبقه‌بندی انواع آلودگی‌ها؛ این پیشرفت‌ها در ادامه مورد بررسی قرار خواهند گرفت.

تشخیص ناهنجاری‌ها در فرآیندهای با ابعاد بالا به معنای شناسایی الگوهایی است که از رفتار طبیعی مشاهدات انحراف دارند. رشد سریع فناوری‌های مرتبط با نمونه‌برداری آزمایشگاهی، سامانه‌های ماهواره‌ای و شبکه‌های انتقال داده‌ی بی‌سیم منجر به تولید حجم عظیمی از داده‌های کیفی و هیدرولیکی شده است [۹]. از این‌رو، ارائه‌ی الگوریتم‌های تحلیل‌گر و مدل‌های پیش‌بینی‌کننده برای تشخیص آلودگی یا برآورد کیفیت منابع آب آشامیدنی یکی از چالش‌های مهم در حوزه‌ی سیستم‌های پایش کیفیت آب است. روش‌های تشخیص آلودگی و ناهنجاری‌های کیفیت آب^۱ (WQAD) به‌طور کلی در سه دسته‌ی اصلی جای می‌گیرند [۱۰]:

۱. روش‌های سنتی و آزمایشگاهی

۲. روش‌های مبتنی بر تحلیل‌های آماری و استنباطی

۳. روش‌های هوشمند مبتنی بر داده

آب‌های سطحی بخشی از فرآیندهای با ابعاد بالا محسوب می‌شوند، زیرا متغیرهای کیفیت آب تحت تأثیر عوامل متعددی قرار دارند. استفاده از مدل‌های مکانیکی سنتی برای توصیف تغییرات با بسامد بالای کیفیت آب دشوار است، چرا که داده‌ها اغلب دارای عدم قطعیت، غیرخطی و چندبُعدی هستند. در مقابل، مدل‌های داده‌محور مانند سیستم‌های فازی عصبی تطبیقی (ANFIS^۲) و سیستم‌های موزج ANFIS به محققین اجازه می‌دهند تغییرات بلندمدت کیفیت آب را بررسی کنند. روش‌های مبتنی بر ANFIS از جمله مدل‌های محبوب در تحلیل داده‌های غیرخطی هستند و می‌توانند با یادگیری حافظه و فرایندهای خودتطبیقی، روابط میان ورودی‌ها و خروجی‌ها را شناسایی کنند. [۱۱] پژوهش‌های متعددی از شبکه‌های عصبی با تأخیر زمانی برای پیش‌بینی خطاهای

¹ Water Quality Anomaly Detection

² Adaptive Neuro-Fuzzy Inference System

مدل‌های هیدرولوژیکی و بهینه‌سازی ساختار شبکه (وزن‌های سیناپسی، تعداد نرون‌های پنهان و دوره‌های آموزش) استفاده کرده‌اند [۱۲].

در مطالعات بعدی، استفاده از مدل‌های موجک ANFIS نتایج قابل توجهی در تحلیل سری‌های زمانی بلندمدت کیفیت آب داشته است. مدل‌های ANFIS برای پیش‌بینی غلظت آلاینده‌ها در آب‌های سطحی به کار رفته‌اند و نتایج نشان می‌دهد که این مدل‌ها ابزارهایی کارآمد و کاربردی برای مهندسان و مدیران منابع آب هستند [۱۳-۱۴]. در پژوهش‌های اخیر، ترکیب موجک با ANFIS موجب بهبود چشمگیر دقت پیش‌بینی در مقایسه با مدل‌های شبکه عصبی کلاسیک^۱ (ANN) شده است [۱۵]. در مطالعه‌ی پیشین [۱۶] از روش RBF استفاده شد که نسبت به ANFIS دقت پایین‌تری داشت. از آن‌جا که آب یکی از حیاتی‌ترین نیازهای بشر است، شناسایی دقیق آلودگی آن ضروری است؛ از این‌رو در این پژوهش از روش ANFIS استفاده شده که توانایی بالایی در تحلیل داده‌های با ابعاد زیاد، غیرخطی و نامطمئن دارد.

این پژوهش تلاشی است برای ارائه‌ی یک چارچوب داده‌محور با هدف تشخیص ناهنجاری‌های کیفیت آب با بهره‌گیری از روش‌های یادگیری ماشین. در ادامه‌ی مقاله، ابتدا مطالعه‌ی موردی معرفی می‌شود. در بخش سوم، روش‌شناسی تحقیق توضیح داده شده است. بخش چهارم شامل نتایج شبیه‌سازی، تحلیل‌ها و مقایسه‌ی عملکرد مدل پیشنهادی با روش‌های پیشین است. در بخش پایانی، نتیجه‌گیری و پیشنهادهایی برای پژوهش‌های آینده ارائه شده است.

۲- مواد و روش‌ها

۲-۱- منطقه مورد مطالعه- ساختار داده‌ها

حوضه رودخانه Potomac River به عنوان منطقه مورد مطالعه در این پژوهش انتخاب شد. رودخانه پوتوماک با طول تقریبی ۶۵۲ کیلومتر و مساحت حوضه آبریز حدود ۳۸,۰۰۰ کیلومتر مربع، یکی از مهم‌ترین رودخانه‌های ساحل شرقی ایالات متحده آمریکا است. این رودخانه از رشته‌کوه‌های آپالچی سرچشمه گرفته و پس از عبور از چندین ایالت، به خلیج چساپیک تخلیه می‌شود. تنوع کاربری اراضی، فعالیت‌های انسانی و رواناب‌های شهری و کشاورزی موجب شده است که این حوضه در معرض تغییرات کیفی و رخدادهای آلودگی قرار گیرد؛ از این‌رو، رودخانه پوتوماک یکی از مناطق مناسب برای مطالعات پایش کیفیت آب و تشخیص ناهنجاری‌های زیست‌محیطی محسوب می‌شود [18].



شکل ۱- موقعیت جغرافیائی و توزیع جمعیتی مجاور رودخانه

۲-۱-۱- ایستگاه‌های پایش

به منظور افزایش اعتبار نتایج و کاهش سوگیری مکانی، دو ایستگاه پایش واقع در نواحی هیدرولوژیکی متفاوت انتخاب شدند:

- ایستگاه شماره 016329000 (واقع در جریان بالادستی رودخانه ، North Fork of the Shenandoah River)^۲

^۱ Artificial Neural Network

^۲. اطلاعات ایستگاه ۰۱۶۳۹۰۰۰

• ایستگاه شماره 01646305 (واقع در جریان پایین دست رودخانه در نزدیکی منطقه شهری McLean)^۱

در این پژوهش، ایستگاه بالادست (USGS-01632900) برای ساخت الگوی پایه و آموزش مدل مورد استفاده قرار گرفته است، زیرا این ایستگاه در شاخه‌ای با پیچیدگی هیدرولوژیکی کمتر و تأثیرات انسانی محدودتر واقع شده و امکان استخراج رفتار مرجع و نسبتاً پایدار سیستم را فراهم می‌کند. در مقابل، ایستگاه پایین دست (USGS-01646305) که در ناحیه‌ای با شرایط هیدرولوژیکی متفاوت و تأثیرپذیری بیشتر از فعالیت‌های انسانی قرار دارد، برای اعتبارسنجی مستقل مدل به کار گرفته شده است.

انتخاب ایستگاه‌هایی در موقعیت‌های جغرافیایی و هیدرولوژیکی متفاوت با اهداف زیر انجام شده است:

۱. در نظر گرفتن ناهمگونی مکانی و تفاوت‌های هیدرولوژیکی در سطح حوضه
۲. ارزیابی قابلیت تعمیم‌پذیری مدل در شرایط مکانی مستقل از داده‌های آموزش
۳. کاهش بیش‌برازش مکانی^۲ از طریق جداسازی کامل محل آموزش و آزمون
۴. شبیه‌سازی شرایط عملی یک سامانه هشدار سریع که باید در محیطی متفاوت از داده‌های آموزشی نیز عملکرد قابل قبول داشته باشد.

همچنین هیچ‌گونه هم‌پوشانی مکانی یا زمانی بین داده‌های مورد استفاده برای آموزش و داده‌های آزمون وجود نداشته است، که این امر از بروز نشت داده^۳ جلوگیری می‌کند.

۲-۱-۲- ساختار تقسیم‌بندی داده‌ها

در این پژوهش، داده‌های پایش پیوسته با فاصله نمونه‌برداری ۱۵ دقیقه‌ای (معادل ۹۶ مشاهده در روز برای هر متغیر) در بازه زمانی ۱ ژانویه ۲۰۲۲ تا ۲ مارس ۲۰۲۲ گردآوری و تحلیل شده‌اند. این بازه زمانی معادل ۶۰ روز بوده و در مجموع برای هر متغیر ۵۷۶۰ مشاهده (60×96) جمع‌آوری شده است.

به منظور شبیه‌سازی شرایط واقعی یک سامانه هشدار سریع و جلوگیری از نشت اطلاعات زمانی، تقسیم‌بندی داده‌ها به صورت ترتیبی و بدون هم‌پوشانی انجام شد. در این طراحی، ۴۰ روز نخست به آموزش مدل اختصاص یافته و ۲۰ روز پایانی صرفاً برای ارزیابی نهایی عملکرد مدل مورد استفاده قرار گرفت. پارامترهای مورد استفاده شامل کدورت، دمای آب و هدایت الکتریکی ویژه بوده‌اند که با فاصله نمونه‌برداری ۱۵ دقیقه‌ای استخراج شده‌اند.

مجموعه آموزش^۴

- ✓ بازه زمانی: ۱ ژانویه ۲۰۲۲ تا ۹ فوریه ۲۰۲۲ (۴۰ روز)
- ✓ تعداد داده‌ها: ۳۸۴۰ مشاهده برای هر متغیر (کدورت، ضریب هدایت و دما)
- ✓ کاربرد: آموزش مدل مرجع، کالیبراسیون پارامترها و تعیین آستانه‌های تصمیم‌گیری.

مجموعه آزمون^۵

- ✓ بازه زمانی: ۱۰ فوریه ۲۰۲۲ تا ۲ مارس ۲۰۲۲ (۲۰ روز)
- ✓ تعداد داده‌ها: ۱۹۲۰ مشاهده برای هر متغیر
- ✓ کاربرد: ارزیابی مستقل عملکرد مدل بدون استفاده در فرآیند آموزش که نمای این تقسیم‌بندی در شکل (۲) مصور شده است.

<https://waterdata.usgs.gov/monitoring-location/USGS-01632900/#dataTypeId=continuous-00065-0&period=P7D&showFieldMeasurements=true>

^۱ اطلاعات ایستگاه ۰۱۶۴۶۳۰۵

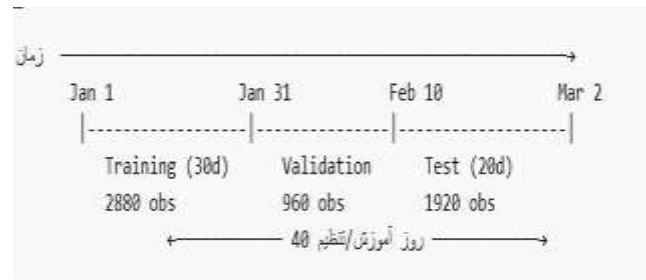
<https://waterdata.usgs.gov/monitoring-location/USGS-01646305/#dataTypeId=continuous-00065-0&period=P7D&showFieldMeasurements=true>

^۲ Spatial Overfitting

^۳ Data Leakage

^۴ Training Set

^۵ Test Set



شکل ۲- ساختار داده های آموزش و تست

جدول ۱- متغیرهای کیفی و هیدرولیکی مورد استفاده در پژوهش و مشخصات اندازه گیری آنها

نام متغیر	نماد	واحد اندازه گیری	توضیح
کدورت	Turbidity (T)	NTU	شاخص میزان ذرات معلق و کلوئیدی در آب
دمای آب	Temperature (Temp)	°C	دمای لحظه ای آب ثبت شده توسط حسگر
هدایت الکتریکی ویژه	Specific Conductance (SC)	μS/cm (at 25°C)	شاخص غلظت یون های محلول در آب
دبی جریان	Discharge (Q)	ft ³ /s یا m ³ /s	حجم جریان عبوری از مقطع رودخانه در واحد زمان

۲-۲- ساخت مدل جایگزین در تخمین شاخص های کیفی (Surrogate)

مدل های جانشین در واقع به کشف و توصیف روابط میان متغیرهای ورودی و خروجی در داده های مشاهده شده با رویکرد آماری می پردازند. در بسیاری از موارد، شاخص های کیفی یا کمی در حوزه های آبریز را نمی توان به صورت مستقیم اندازه گیری یا پایش کرد، زیرا ماهیتی پویا دارند یا محدودیت هایی در ابزارهای سنجش وجود دارد. در چنین شرایطی، مدل های جانشین با انتخاب مؤلفه هایی که بیشترین تأثیر را بر متغیر پاسخ دارند، امکان دستیابی به مقادیر شاخص ها را با فرکانس بالا فراهم می سازند [۲۰]. به طور معمول، متغیرهای کیفی آب مانند کل فسفر، کل نیتروژن و اکسیژن خواهی شیمیایی به صورت مستقیم از طریق داده های برخط قابل اندازه گیری نیستند. بنابراین به کارگیری مدل های جانشین نقش مهمی در ارتقای دقت و کارایی مدل های زیست محیطی ایفا می کند. [۲۱].

برای مثال، مواد جامد معلق از مهم ترین عوامل مؤثر بر مقادیر TN و TP در آب های سطحی طبیعی به شمار می روند [۲۲]. کدورت^۱ که میزان شفافیت نسبی آب را نشان می دهد، متأثر از مقدار مواد معلق است و معمولاً به عنوان شاخصی جانشین برای تخمین غلظت مواد معلق، TP و TN استفاده می شود [۲۳]. از سوی دیگر، رسانایی ویژه^۲ نیز با غلظت های TN و TP ارتباط مستقیم دارد و افزایش آن می تواند بر فرآیندهای رسوب گذاری و جذب نیتروژن و فسفر تأثیر بگذارد. همچنین، نیتروژن و فسفر محلول به طور نزدیکی با دما و pH آب مرتبط اند؛ بنابراین، غلظت TN و TP رابطه ای معنادار با دمای آب دارند [۲۴]. معادله ۱ یک مدل ساده از متغیرهای جایگزین با عدم قطعیت را نشان می دهد. در این معادله، $\hat{y}_i$ متغیر وابسته یا پاسخ است و ضرایب b_0 و b_1 به ترتیب مقدار ثابت و شیب خط هستند.

$$\hat{y}_i = mx_i + b + e_i \quad i = 1, 2, \dots, n \quad (1)$$

$$e_i = y_i - \hat{y}_i \quad (2)$$

در معادله ۱، y_i مشاهده i ام و m شیب معادله پیش بینی کننده و e_i خطای بین مقدار واقعی مشاهده i ام و مقدار پیش بینی شده است. پارامترها و کیفیت مدل بر اساس شاخص های آماری مختلفی مانند حداقل مربعات خطا MSE، ضریب تعیین R^2 ، میانگین مربعات خطا و RMAE تخمین زده می شوند.

¹ Turbidity

² Specific Conductance

$$MSE = \sqrt{\frac{\sum_i^m [y_i - \hat{y}_i]^2}{n - 2}} \quad (۳)$$

$$R^2 = 1 - sse/ssy \quad (۴)$$

$$sse = \sum_i^m [y_i - \hat{y}_i]^2, ssy = \sum_i^m (y_i - \bar{y}_i)^2 \quad (۵)$$

$$RMAE = \frac{1/n \cdot \sum_i^m |y_i - \hat{y}_i|}{\bar{y}_i} \cdot 100 \quad (۶)$$

در این پژوهش، به منظور افزایش دقت برآورد و اتکاپذیری نتایج از روابط رگرسیونی ارائه شده توسط کریستنسن و همکاران (۲۰۰۰) استفاده شده است. نتایج جدول شماره (۲) نشان می دهد که هر سه متغیر مستقل کدورت، هدایت الکتریکی و دمای آب از نظر آماری در سطح معناداری بالا هستند ($p < 0.001$) این امر بیانگر آن است که اثر هریک از این متغیرها بر تغییرات غلظت کل فسفر به صورت آماری قابل اتکاست. مقدار ضریب تعیین $R^2 = 0.96$ نشان می دهد که مدل قادر است حدود 96 درصد از تغییرات غلظت فسفات را بر اساس سه پارامتر مستقل دیگر شناسائی و توصیف کند. مقدار خطای میانگین مربعات ($MSE=0.00179$) نیز حاکی از دقت بالای پیش بینی است.

$$TP = 0.00103TUR + 0.0057WT - 0.227\log_{10} SC + 0.776 \quad (۷)$$

جدول ۲- معادله و شاخص های ارزیابی مدل رگرسیونی جانشین برای برآورد فسفر کل

شاخص شیمیایی کیفیت آب	معادله رگرسیون	سطح معناداری	میانگین مربعات خطا	ضریب تعیین
Total phosphorus	$TP = 0.00103TUR + 0.0057WT - 0.227\log_{10} SC + 0.776$	<0.001	0.00179	0.960

۲-۳- حذف نویز سری های زمانی^۱

روش های مبتنی بر تبدیل موجک در دهه های اخیر به طور گسترده در مدل سازی داده محور سیستم های هیدرولوژیکی و محیط زیستی مورد استفاده قرار گرفته اند [۲۵]. دلیل این استقبال، توانایی موجک ها در تحلیل همزمان سیگنال در دو حوزه زمان و فرکانس و قابلیت آن ها در تجزیه چندرزولوشنی سری های زمانی نایستا است؛ ویژگی ای که در بسیاری از فرآیندهای طبیعی و محیطی اهمیت اساسی دارد.

تبدیل موجک یک نگاشت خطی مبتنی بر ضرب داخلی بین سیگنال و نسخه های مقیاس یافته و انتقال یافته موجک مادر است که با اعمال تابع موجک مادر بر سیگنال اصلی در مقیاس ها و انتقال های مختلف، بازنمایی دوبعدی زمان-مقیاس از داده فراهم می کند. این بازنمایی امکان استخراج ویژگی های پنهان، تغییرات موضعی و رفتارهای گذرای سیگنال را فراهم می سازد. در روند تجزیه سیگنال توسط موجک، سری زمانی به دو دسته مؤلفه اصلی تفکیک می شود:

۱. مؤلفه های فرکانس پایین که بیانگر روندهای بلندمدت و ساختار کلی پدیده هستند.

۲. مؤلفه های فرکانس بالا که شامل نوسانات کوتاه مدت، تغییرات سریع و نویزهای ناشی از خطاهای اندازه گیری یا آشفتگی های محیطی هستند.

این تفکیک، امکان حذف نویز بدون تخریب روند فیزیکی حاکم بر پدیده را فراهم می کند.

^۱ Wavelet Denoising -Preprocessing

در این مطالعه از موجک دابشیز^۱ مرتبه ۴ (db4) استفاده شده است. انتخاب این موجک به دلیل پشتیبانی فشرده، قابلیت محلی سازی مناسب در حوزه زمان-فرکانس و کاربرد گسترده در تحلیل سری‌های زمانی هیدرولوژیکی صورت گرفته است. وجود پشتیبانی فشرده در موجک db4 باعث می‌شود که خروجی موجک تنها به یک بازه زمانی محدود از سیگنال وابسته باشد و اثر هر رویداد به کل دامنه زمانی منتشر نشود. در داده‌های پایش برخط کیفیت آب، بسیاری از تغییرات به صورت موضعی و گذرا رخ می‌دهند؛ برای مثال افزایش ناگهانی کدورت یا فسفر کل در اثر یک رویداد بارشی، ورود رواناب کشاورزی یا تخلیه مقطعی آلاینده که معمولاً چنین رخدادهایی معمولاً در یک بازه زمانی کوتاه ظاهر شده و سپس فروکش می‌کنند.

در این حالت اثر یک جهش کوتاه مدت می‌تواند در کل سری زمانی پخش شود و موجب اعوجاج در روند کلی سیگنال گردد. در حالیکه تمرکز و فشرده‌گی موجک دابشیز سبب می‌شود انرژی ناشی از یک رویداد کوتاه مدت تنها در همان بازه زمانی متمرکز باقی بماند. به بیان دیگر، یک پیک لحظه‌ای در شاخص کدورت یا فسفر تنها در ضرایب موجک مربوط به همان بازه منعکس می‌شود و به سایر بخش‌های سری زمانی نشت نمی‌کند. در چارچوب روابط ریاضی این پژوهش، سری زمانی حاصل از پایش برخط شاخص‌های کیفیت آب از دو مؤلفه اصلی تشکیل شده است:

$$S(i) = f(t) + \sigma \times e(t) \quad t = 0, \dots, n-1 \quad (8)$$

این رابطه بیان می‌کند که مقدار مشاهده شده شاخص کیفیت آب در زمان t برابر است با جمع سیگنال واقعی سیستم $f(t)$ و یک مؤلفه نویزی تصادفی که شدت آن با ضریب 6 مقیاس می‌شود. هدف از پایش پردازش موجک، برآورد و تقریب واقعی سیگنال مورد پایش و حذف مؤلفه نویزی بدون تخریب ساختار دینامیکی سیگنال است. تابع موجک گسسته را می‌توان به شکل زیر بیان کرد:

$$\psi_{j,k}(t) = a_0^{-\frac{j}{2}} \psi(a_0^{-j} t - kb_0) \quad (9)$$

که در آن a_0^j پارامتر مقیاس و b_0 پارامتر انتقال زمانی پنجره موجک و حرکت آن بر روی سیگنال اصلی است، t زمان است و j شاخص سطح تجزیه است. ضرایب تبدیل موجک گسسته از طریق ضرب داخلی سیگنال با موجک‌های مقیاس یافته و انتقال یافته محاسبه می‌شوند که به صورت زیر تعریف می‌گردد:

$$C_{j,k} = \int_{-\infty}^{\infty} f(t) \psi_{j,k}^*(t) dt \quad (10)$$

که در آن $\psi_{j,k}^*$ مزدوج مختلط موجک پایه در مقیاس t و انتقال زمانی k است. ضرایب $C_{j,k}$ میزان شباهت موضعی سیگنال با موجک در یک مقیاس و موقعیت زمانی مشخص را نشان می‌دهند. از آنجا که تبدیل موجک یک تبدیل خطی و معکوس پذیر است، سیگنال اصلی را می‌توان از طریق جمع ضرایب موجک و توابع پایه بازسازی نمود. و بازسازی سری اصلی نیز از طریق رابطه‌ی زیر انجام می‌شود:

$$f(t) = C \sum_{-\infty}^{\infty} \sum_{-\infty}^{\infty} C_{j,k} \psi_{j,k}(t) \quad (11)$$

که در آن C یک ضریب نرمال سازی است. این رابطه نشان می‌دهد که سیگنال اصلی به صورت ترکیبی از موجک‌های مقیاس یافته و انتقال یافته بازسازی می‌شود.

¹ Daubechies

در نتیجه، با حذف یا تضعیف ضرایب موجک متناظر با مؤلفه‌های نویزی و بازسازی سیگنال از ضرایب اصلاح‌شده، سری زمانی پالایش‌شده‌ای حاصل می‌شود که نمایانگر رفتار واقعی سیستم بوده و به‌عنوان ورودی پایدار و قابل اعتماد برای مدل ANFIS در مرحله مدل‌سازی مورد استفاده قرار می‌گیرد.

۲-۳-۱- روش آستانه‌گذاری^۱

آستانه‌گذاری یکی از مراحل مهم در حذف نویز مبتنی بر تبدیل موجک است که با هدف تفکیک مؤلفه‌های معنادار سیگنال از مؤلفه‌های نویزی انجام می‌شود. پس از تجزیه سیگنال به ضرایب موجک در مقیاس‌های مختلف، ضرایب با دامنه بزرگ معمولاً نمایانگر ساختار واقعی سیگنال هستند، در حالی که ضرایب کوچک‌تر غالباً ناشی از نویز اندازه‌گیری هستند. در این روش با تعریف یک مقدار آستانه، ضرایب کوچک‌تر از آستانه حذف یا تضعیف شده و ضرایب مهم حفظ می‌شوند در این پژوهش، مقدار آستانه بر اساس قاعده جهانی Donohoe تعیین شده است که به صورت $T = \sigma\sqrt{2\ln N}$ تعریف می‌شود؛ در این رابطه N تعداد ضرایب موجک و σ برآورد انحراف معیار نویز است. اعمال این آستانه بر ضرایب موجک موجب حذف مؤلفه‌های فرکانس بالا و نویزهای تصادفی می‌شود، در حالی که ساختار اصلی سیگنال حفظ می‌گردد. در نتیجه، سیگنال بازسازی‌شده نمایانگر رفتار واقعی سیستم بوده و داده‌ای پایدارتر و قابل اعتمادتر برای مراحل بعدی مدل‌سازی، از جمله آموزش مدل‌های یادگیری ماشین و سامانه‌های تشخیص ناهنجاری، فراهم می‌کند [40].

۲-۳-۲- اثر لبه^۲

یکی از چالش‌های رایج در تحلیل موجک سری‌های زمانی محدود، بروز اعوجاج در نواحی ابتدایی و انتهایی داده‌های سری زمانی است که به عنوان اثر لبه‌ها شناخته می‌شود. این پدیده زمانی رخ می‌دهد که تابع موجک در نزدیکی مرزهای سری زمانی قرار گرفته و بخشی از آن خارج از دامنه داده‌ها قرار گیرد. برای مشخص کردن نواحی تحت تأثیر این پدیده در طیف موجک، مفهوم مخروط تأثیر^۳ معرفی می‌شود که محدوده‌ای را مشخص می‌کند که در آن ضرایب موجک ممکن است تحت تأثیر کمبود داده در مرزها قرار گیرند. در این پژوهش به منظور ارزیابی حساسیت نتایج تحلیل موجک نسبت به اثر لبه‌ها، دو سناریوی مختلف برای مدیریت مرزهای سری زمانی مورد بررسی قرار گرفت. در سناریوی نخست، تحلیل موجک مستقیماً بر روی سری زمانی اصلی و بدون اعمال هیچ‌گونه گسترش در داده‌ها انجام شد. در سناریوی دوم، پیش از اجرای تبدیل موجک، سری زمانی با استفاده از روش امتداد متقارن^۴ در دو انتهای بازه زمانی گسترش داده شد. در این روش مقادیر سری زمانی در نزدیکی مرزها به صورت انعکاسی بازسازی می‌شوند تا از ایجاد ناپیوستگی‌های مصنوعی در سیگنال جلوگیری شود. شکل ۴ نمودار طیف توان موجک پیوسته مورلت^۵ برای سری زمانی فسفات در طول ۶۰ روز از ابتدای ژانویه ۲۰۲۲ تا ۲۰۲۲ مارچ ۲۰۲۲ را نشان می‌دهد. محور افقی بازه زمانی ۶۰ روزه و محور عمودی مقیاس موجک را نمایش می‌دهد. بررسی الگوی انرژی در این طیف نشان می‌دهد تغییرات فسفات در طول دوره مطالعه به صورت رویداد های ناپیوسته و مقطعی ظاهر شده اند. به ویژه در روز های ۱۰، ۳۵، ۱۸ و ۵۰ ام که بیانگر اختلالات گذرا در میزان مشاهدات هستند. در راستای بررسی منشأ جهش های فسفات در دوره ۶۰ روزه، سری زمانی شاخص دبی در همان ایستگاه ۰۱۶۳۲۹۰۰، در شکل ۴ نمایش داده شده است. همانطور که مشاهده می‌شود متناظر با پیک های مقادیر فسفات، تغییرات دبی آب نیز که بدلیل بارش و جاری شدن روان آب های سطحی و حمل املاح اراضی مجاور بوده است، قابل مشاهده است.

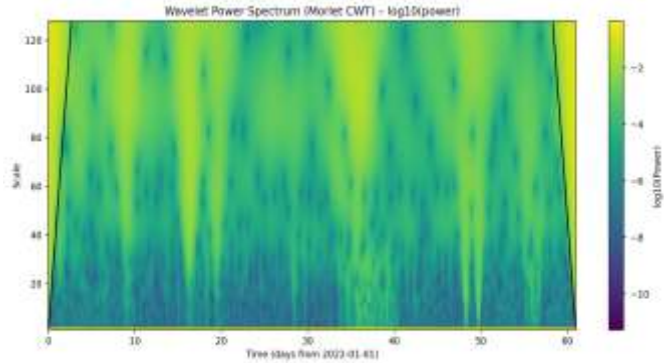
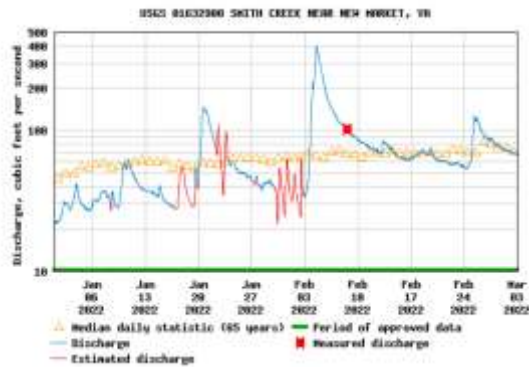
¹ Thresholding

² Boundary Effects

³ Cone of Influence

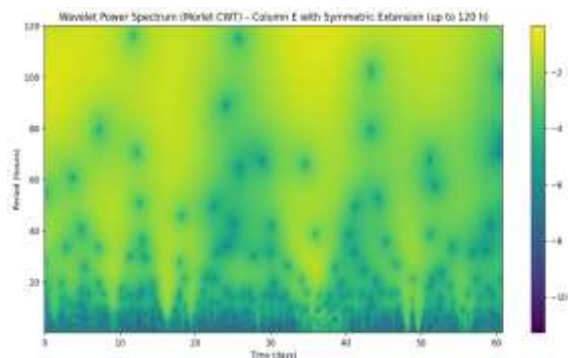
⁴ Symmetric Extension

⁵ Morlet Wavele



شکل ۳- طیف توان موجک پیوسته موجک - سری زمانی فسفات بازه زمانی ۶۰ روزه
شکل ۴- تغییرات زمانی دبی رودخانه در ایستگاه ۰۱۶۳۲۹۰۰ بازه زمانی ۶۰ روزه

در ادامه و به منظور ارزیابی و تحلیل حساسیت اثر لبه ها در سری زمانی مشاهدات، زمانی با استفاده از روش امتداد متقارن در دو انتهای بازه زمانی گسترش داده شد. در این روش مقادیر سری زمانی در نزدیکی مرزها به صورت انعکاسی بازسازی می شوند تا از ایجاد ناپیوستگی های مصنوعی در سیگنال جلوگیری شود. (شکل ۵)



شکل ۵- طیف توان موجک سری زمانی ۶۰ روزه فسفات پس از اعمال گسترش متقارن آینه ای جهت کاهش اثر لبه ها

۳-۳-۲- تحلیل طیف توان - تبدیل مقیاس موجک^۱

در تحلیل موجک، محور عمودی طیف توان در ابتدا بر اساس **مقیاس موجک**^۲ تعریف می شود. با این حال، برای تفسیر فیزیکی بهتر نتایج، این مقیاس ها به **دوره زمانی واقعی**^۳ تبدیل می شوند. در این پژوهش فاصله نمونه برداری داده ها ۱۵ دقیقه (معادل ۰٫۲۵ ساعت) بوده و تحلیل با استفاده از موجک Morlet انجام شده است. در این شرایط، رابطه تقریبی بین مقیاس در موجک مورلت و چرخه زمانی به صورت رابطه زیر بیان می شود:

$$Period \approx 1.03 * Scale * \Delta t \quad (12)$$

که در این مطالعه $\Delta t = 0.25$ است. مقیاس های Scale نیز طبق یک دنباله لگاریتمی به صورت نمایی بزرگ می شود. این دنباله بر اساس رابطه روبرو شکل می یابد. $s_j = s_0 * 2^{js_j}$ در جدول ۳ نمونه های از تبدیل مقیاس موجک و تفسیر چرخه زمانی معادل هر نوسان بر اساس ساعت نمایش داده شده است.

¹ Power Spectrum Analysis

² Scale

³ Period

جدول ۳- تبدیل مقیاس‌های موجک به دوره زمانی معادل و تفسیر چرخه‌های زمانی

گام های مقیاس (J)	مقیاس	چرخه زمانی	تفسیر زمانی
۰	۲	۰,۵۲ ساعت	تغییرات کمتر از یک ساعت
۱	۲,۳۸	۰,۶۱ ساعت	نوسانات کوتاه مدت
۴	۴	۱,۰۳ ساعت	نوسانات حدود یک ساعت
۱۷	۴۰	۱۰,۳ ساعت	نوسانات نیم روزه

با استفاده از روابط فوق، مقیاس‌های موجک به دوره‌های زمانی واقعی بر حسب ساعت تبدیل شده‌اند تا امکان تفسیر چرخه‌های زمانی موجود در سری زمانی فراهم گردد. عنوان مثال، مقیاس ۴۰ تقریباً معادل چرخه‌ای در حدود ۱۰ ساعت، مقیاس ۹۶ معادل چرخه‌ای نزدیک به ۲۴ ساعت و مقیاس ۱۲۰ معادل دوره‌ای حدود ۳۱ ساعت است. برای نمونه، در حوالی روز ۳۶ دوره مطالعه یک ناحیه با توان بالا در محدوده مقیاس‌های نزدیک به ۲۴ ساعت مشاهده می‌شود که بیانگر وقوع یک پیک فسفات با رفتار نوسانی تقریباً روزانه است؛ به این معنا که افزایش و کاهش غلظت فسفات در این رویداد در بازه‌ای حدود یک شبانه‌روز تداوم داشته و چرخه کامل تغییرات آن تقریباً ۲۴ ساعت طول کشیده است.

به منظور ارزیابی میزان تأثیر لبه‌ها در تحلیل موجک، طیف توان سری زمانی فسفات در دو حالت شامل داده‌های اصلی و داده‌های گسترش‌یافته با استفاده از روش گسترش متقارن آینه‌ای^۱ محاسبه و با یکدیگر مقایسه گردید. شکل (۳) و (۵) نتایج نشان داد که الگوی کلی توزیع انرژی در هر دو طیف توان بسیار مشابه بوده و چرخه‌های غالب در مقیاس‌های زمانی یکسانی ظاهر می‌شوند. این موضوع نشان می‌دهد که چرخه‌های شناسایی شده ناشی از رفتار واقعی سری زمانی بوده و به اعوجاج ناشی از مرزهای داده مرتبط نیستند. همچنین عدم مشاهده تغییرات قابل توجه در ساختار طیف توان پس از اعمال گسترش متقارن بیانگر آن است که تحلیل موجک نسبت به اثرات مرزی حساسیت اندکی دارد. با توجه به طول مناسب سری زمانی، ۵۷۶۰ نمونه با تراکم بالای نمونه‌برداری (فاصله زمانی ۱۵ دقیقه)، و پایداری الگوی انرژی در مقیاس‌های مختلف، می‌توان نتیجه گرفت که اثر لبه‌ها در تحلیل حاضر ناچیز بوده و نتایج به‌دست‌آمده از تبدیل موجک از اعتبار و قابلیت اطمینان بالایی برخوردار است.

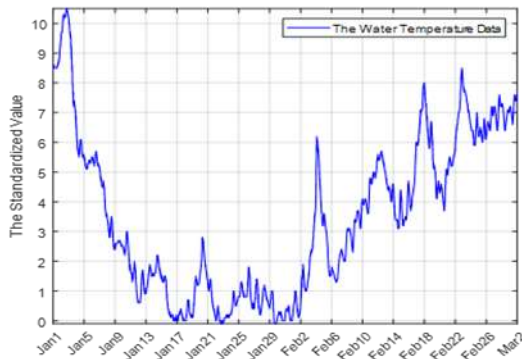
۲-۳-۴- استخراج ویژگی^۲

در این مرحله، به‌منظور بهبود دقت تشخیص ناهنجاری، ویژگی‌های معنادار از داده‌های پیش‌پردازش‌شده استخراج شدند. پس از اعمال تبدیل موجک و حذف نویز، سری‌های زمانی پالایش‌شده به‌عنوان ورودی این مرحله مورد استفاده قرار گرفتند. ویژگی‌ها در دو دسته اصلی تعریف شدند: (۱) متغیرهای مشاهداتی مستقیم شامل کدورت، دمای آب و هدایت الکتریکی و (۲) متغیر جانشین شامل غلظت فسفات برآورد شده که از روابط رگرسیونی استخراج شد. همچنین، مؤلفه‌های فرکانس پایین حاصل از تبدیل موجک به‌منظور حفظ روند اصلی سیگنال و کاهش اثر نویز به مجموعه ویژگی‌ها افزوده شدند.

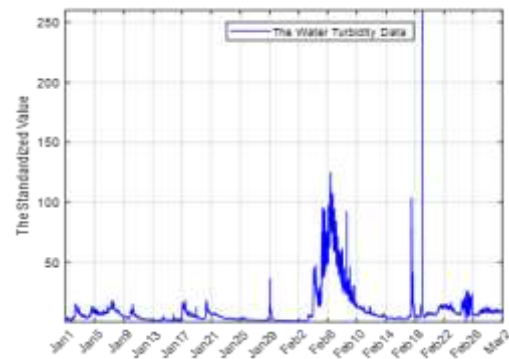
در مجموع، ترکیب ویژگی‌های فیزیکی، شیمیایی و فرکانسی استخراج‌شده، یک نمایش جامع از رفتار سیستم کیفیت آب فراهم می‌کند که به مدل ANFIS امکان می‌دهد روابط غیرخطی و پیچیده میان متغیرها را با دقت بالاتری شناسایی کرده و ناهنجاری‌ها را به‌صورت مؤثر تشخیص دهد. همان‌طور که در شکل (۶) مشاهده می‌شود، شاخص کدورت دو نقطه‌ای اوج دارد: یکی در مشاهده‌ی شماره‌ی ۳۵۴۹ مربوط به تاریخ ۶ فوریه و دیگری در مشاهده‌ی شماره‌ی ۳۷۶۹ مربوط به ۹ فوریه. در همین تاریخ، دبی رودخانه به بیشترین مقدار خود در کل دوره، یعنی حدود ۴۳۰۰۰ فوت مکعب بر ثانیه (معادل ۷۳۰۰۵۷ مترمکعب در دقیقه) رسیده است. بارندگی یکی از عوامل اصلی افزایش حجم جریان رودخانه است و با شست‌وشوی املاح خاک در اراضی مجاور، شاخص‌هایی مانند کدورت نیز افزایش می‌یابد که بر تغییرات فسفات تأثیر مستقیم دارند.

^۱ Symmetric Mirror Extension

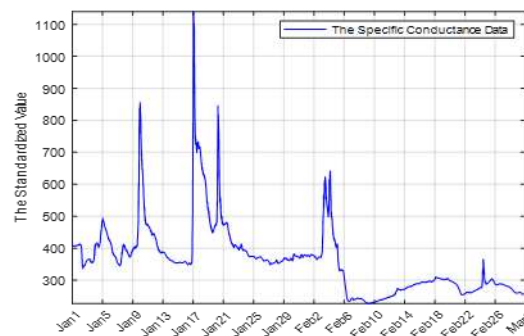
^۲ Feature Extraction



شکل ۷- داده های دمای آب



شکل ۶- داده های کدورت آب



شکل ۸- داده های رسانایی ویژه

بررسی دقیق تر نمودار فسفات نشان می دهد که تقریباً در تمامی موارد، افزایش فسفات با افزایش دبی جریان سطحی همراه بوده است. در مشاهده ی شماره ی ۲۷۳۹ که مربوط به ۲۹ ژانویه ساعت ۹:۴۵ صبح است، دبی رودخانه حدود ۵۰۷۷۰ فوت مکعب بر ثانیه (معادل ۹۸۰۳،۲ مترمکعب در دقیقه) و کدورت حدود ۲،۵ واحد فنو (FNU) گزارش شده است. این وضعیت احتمالاً بیانگر وقوع یک رویداد ناهنجار در بخشی از محدوده ی تحت پایش است. در پژوهش های آتی تلاش می شود تا ماهیت این نوع رخدادها تفکیک شود تا امکان پیشگیری سریع از آلودگی های محیطی فراهم گردد.

۲-۴- مدل سازی سری زمانی

مدل سازی سری های زمانی به تحلیل دنباله ای از مشاهدات یک پدیده در فواصل زمانی منظم با هدف شناسایی الگوهای وابستگی، ساختار نوسانات و پیش بینی رفتار آینده اطلاق می شود [۲۷]. در چارچوب این پژوهش، با توجه به ماهیت غیرایستا، غیرخطی و دارای عدم قطعیت داده های کیفیت آب، تحلیل سری زمانی به عنوان یک گام میانی برای درک رفتار دینامیکی سیستم و آماده سازی داده ها برای مدل سازی نهایی مورد استفاده قرار گرفته است. ورودی این مرحله، سری های زمانی پالایش شده حاصل از پیش پردازش موجه و مجموعه ویژگی های استخراج شده در بخش قبل است.

در انتخاب نوع مدل پیش بینی سری زمانی، توجه به ویژگی ها و پراکندگی داده ها، وابستگی متغیرها و عوامل محیطی ناشناخته بسیار مهم است [۳۲]. در حوزه ی مدیریت منابع آب، ناپایداری و عدم قطعیت متغیرها از مسائل اساسی محسوب می شوند. بنابراین، استفاده از مدل های خودرگرسیون با واریانس شرطی ناپایدار مانند ARCH و مدل های توسعه یافته ی آن مانند GARCH می تواند بسیار مؤثر باشد [۳۳]. از سوی دیگر، شرایط غیرخطی و پیچیدگی ذاتی متغیرهای کیفی باعث می شود تکیه بر یک مدل منفرد برای پیش بینی کافی نباشد. به همین دلیل، در پژوهش حاضر کارایی مدل های ARMA، ARCH و GARCH بررسی و مقایسه شده است [۳۴-۳۶]. مدل $ARMA^1$ به عنوان یک مدل خطی پایه، برای شناسایی وابستگی های زمانی بین مشاهدات و خطاهای گذشته

¹ Autoregressive Moving Average

به کار گرفته شد. این مدل امکان بررسی ساختار خودرگرسیون و میانگین متحرک داده‌ها را فراهم می‌سازد، با این حال، فرض واریانس ثابت در این مدل، کاربرد آن را در داده‌های با نوسانات شدید محدود می‌کند.

به منظور غلبه بر این محدودیت، مدل ARCH^۱ برای مدل‌سازی واریانس شرطی و بررسی تغییرات زمانی در پراکندگی داده‌ها مورد استفاده قرار گرفت. این مدل قادر است وابستگی واریانس به مشاهدات گذشته را شناسایی کرده و رفتار ناپایدار سری زمانی را بهتر توصیف کند. در ادامه، مدل GARCH^۲ به عنوان توسعه یافته مدل ARCH، برای مدل‌سازی دقیق‌تر نوسانات واریانس به کار گرفته شد. این مدل با در نظر گرفتن هم‌زمان اثرات گذشته واریانس و خطاها، توانایی بالاتری در شبیه‌سازی رفتار نوسانی و غیرخطی سری‌های زمانی محیطی دارد و در بسیاری از مطالعات هیدرولوژیکی عملکرد بهتری نسبت به مدل‌های ساده‌تر نشان داده است [33]. هدف از به کارگیری این مدل‌ها در چارچوب پژوهش حاضر، صرفاً پیش‌بینی مستقیم سری زمانی نبوده، بلکه بررسی و مقایسه توانایی آن‌ها در بازنمایی ساختار دینامیکی داده‌ها و فراهم‌سازی یک مبنای مقایسه برای ارزیابی عملکرد مدل پیشنهادی مبتنی بر ANFIS است. در این راستا، خروجی بهترین مدل سری زمانی به عنوان یکی از ورودی‌های کمکی در فرآیند تحلیل نهایی و مقایسه عملکرد مورد استفاده قرار گرفته است. در مجموع، تحلیل سری زمانی در این پژوهش نقش یک لایه تحلیلی میانی را ایفا می‌کند که با استخراج اطلاعات ساختاری از داده‌ها، درک بهتری از رفتار سیستم فراهم کرده و زمینه را برای مدل‌سازی دقیق‌تر و تشخیص مؤثرتر ناهنجاری‌ها در مرحله بعدی فراهم می‌سازد.

۲-۴-۱-۱ مدل ARMA

خود رگرسیون میانگین متحرک نوعی مدل خطی برای سری‌های زمانی است که نخستین بار توسط باکس و جنکینز معرفی شد [۳۶]. این مدل شامل دو بخش اصلی است: مشاهدات گذشته و خطاهای پیش‌بینی گذشته که در معادله ۱۰ نشان داده شده است. فرمول کلی مدل به صورت زیر است:

$$ARMA(p, q): Y_t = c + \sum_{i=1}^p \varphi_i \cdot Y_{t-i} + \sum_{i=1}^q \theta_i \cdot \epsilon_{t-i} + \epsilon_t \quad (13)$$

در این رابطه، در این رابطه، Y_t متغیری است که باید بر اساس مشاهدات قبلی تخمین زده شود و ضرایب φ_i ، θ_i و c پارامترهای مدل و p ، q ترتیب بازگشت مشاهدات به مراحل قبل هستند. ϵ_t یک دنباله تصادفی از جملات خطا با میانگین صفر و واریانس ثابت است. یکی از شرایط استفاده از مدل‌های خطی مانند ARMA، ایستایی داده‌ها^۳ است. نایستایی به این معناست که ویژگی‌های آماری مشاهدات در طول زمان تغییر می‌کند و این امر بر دقت مدل تأثیر می‌گذارد. برای ایستار کردن داده‌ها، معمولاً از عملگر تفاضلی استفاده می‌شود تا روند زمانی حذف شود و میانگین و واریانس به مقدار ثابت نزدیک گردد. معمولاً با چندین مرحله مشتق‌گیری از یک سری ناپایدار، خواص آماری مشاهدات مستقل از زمان می‌شوند یا به عبارت دیگر، مدل ایستا می‌شود. در معادله (۱۳)، ∇x_t عملگر رگرسیون است که به صورت یک معادله دیفرانسیل، مشاهدات را یک یا چند مرحله به عقب می‌برد و در نتیجه، یک سری جدید در سمت راست معادله تشکیل می‌شود که میانگین آن نزدیک به صفر و واریانس آن ثابت خواهد بود.

$$\begin{aligned} \nabla y_t &= y_t - y_{t-1} \\ y_t &= \log y_t - \log y_{t-1} \end{aligned} \quad (13)$$

۲-۴-۲ مدل ARCH

مدل ARCH یا مدل خودرگرسیون با واریانس شرطی ناپایدار اولین بار توسط انگل در سال ۱۹۸۲ معرفی شد [۳۷]. می‌توان گفت که y_t در رابطه (۱۲) یک فرآیند ARCH از نوع (p) است اگر واریانس داده‌ها یک ترکیب خطی وزن‌دار از جملات توان دوم مشاهدات در مراحل گذشته باشد و $b_i > 0$ و $a_i > 0$ باشد.

^۱ Autoregressive Conditional Heteroskedasticity

^۲ Generalized ARCH

^۳ Stationarity

$$y_t = a_0 + \sum_{i=1}^p a_i \cdot y_{t-i} + \epsilon_t \quad (14)$$

در دنیای واقعی، پدیده‌های غیرخطی مانند متغیرهای کیفیت آب و تغییرات و وابستگی شدید این شاخص‌ها به سایر عوامل محیطی و غیرمحیطی، عموماً مسائلی مانند نوسانات و تغییرات لحظه‌ای در واریانس مشاهدات را مطرح می‌کنند. در مطالعات کیفی یا هیدرولیکی، که در آن‌ها اثر ARCH در داده‌ها مشاهده می‌شود، مدل‌های خودرگرسیون شرطی عملکرد مؤثری از خود نشان داده‌اند [۳۸، ۳۹].

۳-۴-۲- مدل GARCH

تفاوت این مدل با ARCH در مدل‌سازی واریانس‌های شرطی و ناپایدار است. به عبارت دیگر، می‌توان گفت مشاهدات در یک پدیده غیرخطی زمانی از مدل GARCH پیروی می‌کنند که واریانس آن‌ها در معادله (۱۵) اعمال شود.

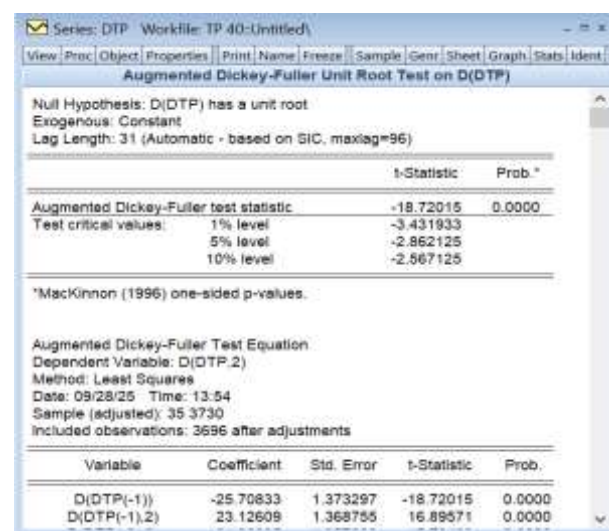
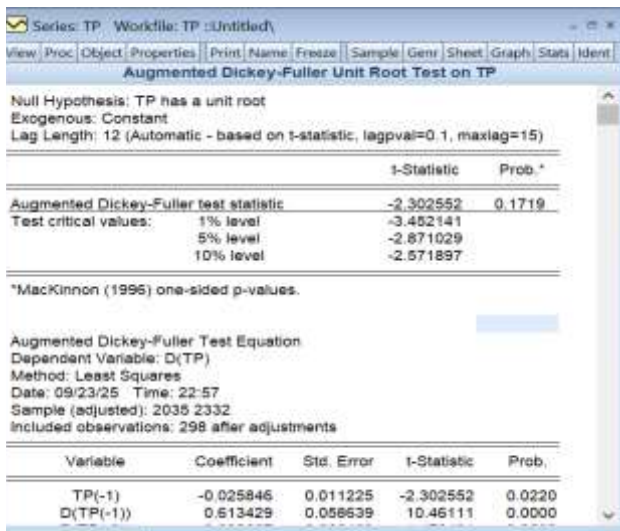
$$\sigma_t^2 = a + \sum_{i=1}^p b_i y_{t-1}^2 + \dots + \sum_{j=1}^q c_j \sigma_{t-j}^2 \quad \text{if } a > 0, b_i \geq 0, c_j \geq 0 \quad (15)$$

این مدل قادر است پدیده‌های ناپایدار و متغیر را با دقت بالاتری مدل‌سازی کند و معمولاً در پیش‌بینی رفتارهای غیرخطی داده‌های محیطی مانند کیفیت آب مؤثرتر از مدل‌های ساده‌تر است. در پژوهش حاضر، مدل‌های ARCH، ARMA و GARCH روی داده‌های حاصل از معادلات جانشین شاخص‌های TP و TN اعمال شدند. در نهایت، سری زمانی با بهترین عملکرد برای هر متغیر انتخاب و به عنوان ورودی مدل اصلی ANFIS مورد استفاده قرار گرفت. داده‌های آموزش و آزمون براساس مدل GARCH در شکل‌های ۹ نمایش داده شده‌اند.

۲-۴-۴- تنظیمات مدل و برآورد پارامترها (Eviews.8)

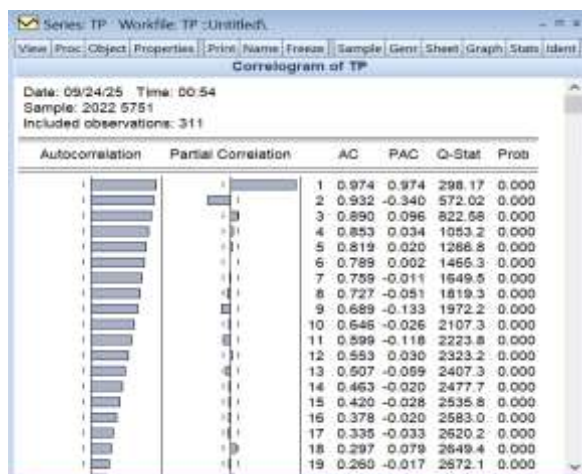
دربخش شبیه‌سازی مربوط به سری زمانی، از نرم افزار Eviews.8 استفاده شده است. سری زمانی فسفات با گام زمانی ۱۵ دقیقه و در بازه ۴۰ روزه مورد تحلیل قرار گرفت. بررسی اولیه روند با استفاده از مدل خطی (عرض از مبدأ و روند) نشان داد که این مدل توان توضیح‌دهی بسیار محدودی دارد ($R^2 \approx 0.025$) و باقیمانده‌ها دارای خودهمبستگی شدید هستند که بیانگر ناکارآمدی مدل‌های صرفاً روندی است. آزمون ریشه واحد دیکی-فولر تعمیم‌یافته (ADF) شکل (۹) نشان داد که سری زمانی نامانا است ($p\text{-value} = 0.17$)، لذا یک مرحله تفاضل‌گیری برای ماناسازی انجام شد. پس از تفاضل‌گیری، مقدار آماره آزمون به حدود -18.72 کاهش یافت. بنابراین فرض صفر نامانایی سری زمانی به‌طور قاطع رد شد، که نشان‌دهنده مناسب بودن سری برای مدل‌سازی است.

شکل ۹- خروجی تفاضل گیری، ضرایب و شاخص های آماری قبل از تفاضل گیری

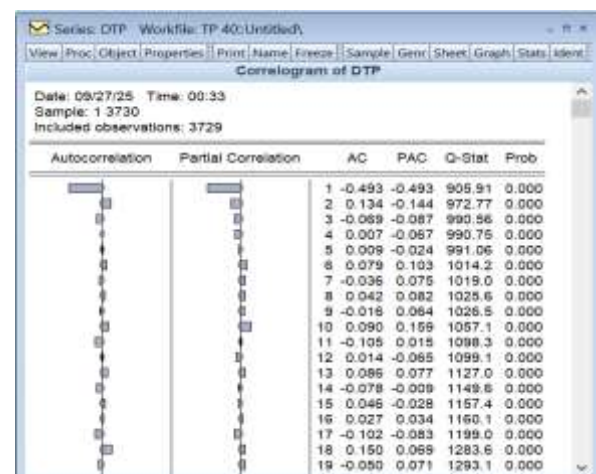


همانطور که در شکل (۱۰) نیز مشاهده می شود ضرایب خود همبستگی ACF در لگ های متعدد (حتی تا لگ های بالا) بزرگ و به صورت آهسته کاهش می یابند؛ این الگو نشانه ی وجود روند و ناپایداری است. همچنین مقدار Q-Stat بالا و $Prob \approx 0.000$ در اکثر لگ ها بیانگر رد فرضیه نویز سفید و وجود خودهمبستگی معنادار در کل سری است. در PACF نیز باقی ماندن چند لگ معنادار تأییدکننده وابستگی ساختاری سری است. پس از تفاضل گیری، ضرایب ACF به سرعت به اطراف صفر نوسان می کنند، شکل (۱۰) و اغلب داخل باند اطمینان قرار می گیرند؛ تنها یکی دو لگ ابتدایی ممکن است معنادار باشند که نشان دهنده حذف روند و ایستاسازی سری است. کاهش شدید مقادیر خودهمبستگی و رفتار نوسانی اطراف صفر، نشانه نزدیک شدن سری به یک فرآیند ایستا با وابستگی کوتاه مدت است.

شکل ۱۰- کورلوگرام سری زمانی قبل و بعد تفاضل گیری



(بعد تفاضل گیری)



(قبل تفاضل گیری)

از نظر آماری، مقایسه این دو حالت نشان می دهد که تفاضل گیری باعث کاهش خودهمبستگی بلندمدت، بهبود ویژگی های تصادفی سری و فراهم شدن شرایط لازم برای برازش مدل های AR/MA شده است؛ به طوری که پس از ایستاسازی، ساختار مدل را می توان با تکیه بر لگ های معنادار محدود در ACF/PAC تعیین کرد. انتخاب ساختار مدل بر اساس نمودارهای خودهمبستگی (ACF) و خودهمبستگی جزئی (PACF) انجام شد. کاهش تدریجی ACF و قطع PACF در وقفه دوم، لزوم استفاده از مدل های

ترکیبی را نشان می دهد. براین اساس، مدل های $AR(1)$ ، $MA(1)$ ، $ARMA(1,1)$ برای مدلسازی رفتار خودبازگشتی و میانگین متحرک داده ها در نظر گرفته شدند.

شکل ۱۱- برازش های مختلف برای بررسی رفتار خودبازگشتی و میانگین متحرک داده ها



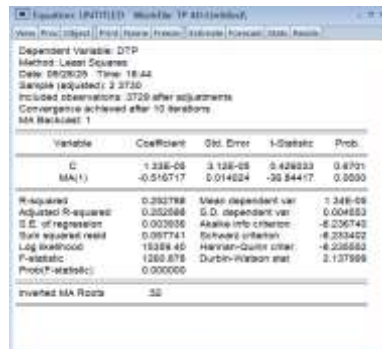
Dependent Variable: DTP
Method: Least Squares
Date: 09/28/25 Time: 18:24
Sample (adjusted): 3 3730
Included observations: 3728 after adjustments
Convergence achieved after 3 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.32E-08	4.38E-08	0.305477	0.7619
AR(1)	-0.492716	0.014236	-34.58076	0.0000

R-squared: 0.242752 Mean dependent var: 1.34E-05
Adjusted R-squared: 0.242948 S.D. dependent var: 0.004553
S.E. of regression: 0.022963 Akaike info criterion: -8.223129
Sum squared resid: 0.038516 Schwarz criterion: -8.219760
Log likelihood: 15329.91 Hannan-Quinn criter.: -8.221941
F-statistic: 1194.447 Durbin-Watson stat.: 2.141893
Prob(F-statistic): 0.000000

Inverted AR Roots: -.49

برازش $AR(1)$



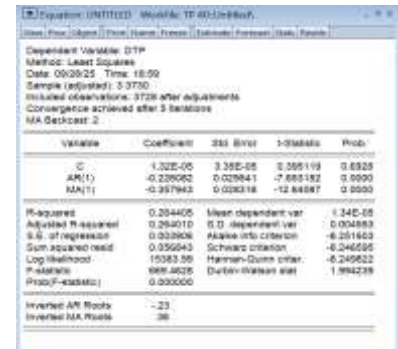
Dependent Variable: DTP
Method: Least Squares
Date: 09/28/25 Time: 18:44
Sample (adjusted): 3 3730
Included observations: 3728 after adjustments
Convergence achieved after 10 iterations
MA Backcast: 1

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.33E-08	3.13E-08	0.426333	0.6701
MA(1)	-0.516717	0.014424	-35.84477	0.0000

R-squared: 0.252768 Mean dependent var: 1.34E-05
Adjusted R-squared: 0.252568 S.D. dependent var: 0.004553
S.E. of regression: 0.022963 Akaike info criterion: -8.236740
Sum squared resid: 0.038516 Schwarz criterion: -8.233422
Log likelihood: 15329.45 Hannan-Quinn criter.: -8.235562
F-statistic: 1265.876 Durbin-Watson stat.: 2.137898
Prob(F-statistic): 0.000000

Inverted MA Roots: .52

برازش $MA(1,1)$



Dependent Variable: DTP
Method: Least Squares
Date: 09/28/25 Time: 18:59
Sample (adjusted): 3 3730
Included observations: 3728 after adjustments
Convergence achieved after 3 iterations
MA Backcast: 2

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.32E-08	3.39E-08	0.395199	0.6938
AR(1)	-0.239002	0.029844	-7.885182	0.0000
MA(1)	-0.357943	0.026318	-12.84887	0.0000

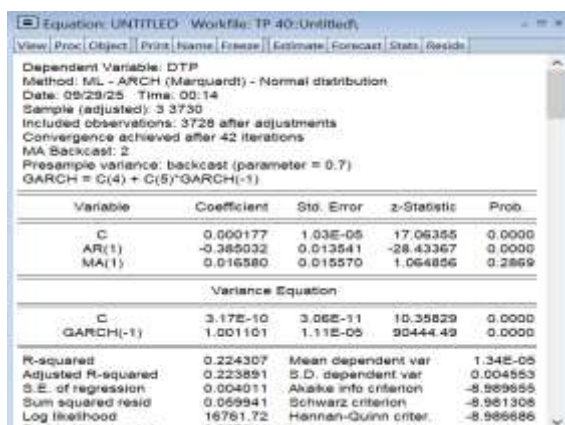
R-squared: 0.294405 Mean dependent var: 1.34E-05
Adjusted R-squared: 0.294010 S.D. dependent var: 0.004553
S.E. of regression: 0.022963 Akaike info criterion: -8.251603
Sum squared resid: 0.035043 Schwarz criterion: -8.246556
Log likelihood: 15351.59 Hannan-Quinn criter.: -8.249522
F-statistic: 988.8035 Durbin-Watson stat.: 1.994238
Prob(F-statistic): 0.000000

Inverted AR Roots: -.23
Inverted MA Roots: .38

برازش $ARMA(1,1)$

در ادامه، به منظور بررسی کیفیت سایر مدل ها و با توجه به تأیید وجود اثر ARCH در باقیمانده های مدل $ARMA(p < 0.01)$ ، وجود ناهمسانی واریانس نیز بر اساس نتایج ارائه شده در شکل (۱۲) تأیید شد. بر این اساس، مدل های $GARCH(1,1)$ و در نهایت $GARCH(1,1)$ برای مدل سازی واریانس شرطی داده ها به کار گرفته شدند. تمامی مدل ها با استفاده از روش حداکثر درست نمایی در نرم افزار EViews 8 برآورد شدند و حداکثر طول وقفه برابر با ۹۶ مشاهده، معادل یک شبانه روز، در نظر گرفته شد. همان گونه که در شکل (۱۲) مشاهده می شود، مقدار تابع درست نمایی حدود ۱۶۷۶۱ و مقادیر معیارهای اطلاعاتی آکائیک و شوارتز به ترتیب برابر با $-۸/۹۸۹$ و $-۸/۹۸۱$ به دست آمدند. همچنین، معناداری آماری ضرایب مدل، به ویژه ضریب مربوط به مؤلفه $GARCH$ ، نشان می دهد که مدل انتخاب شده توانسته است ناهمسانی واریانس موجود در داده ها را به طور مناسبی بازنمایی کند.

شکل ۱۲- تأیید وجود اثر آرج در داده ها و برازش مدل گارچ



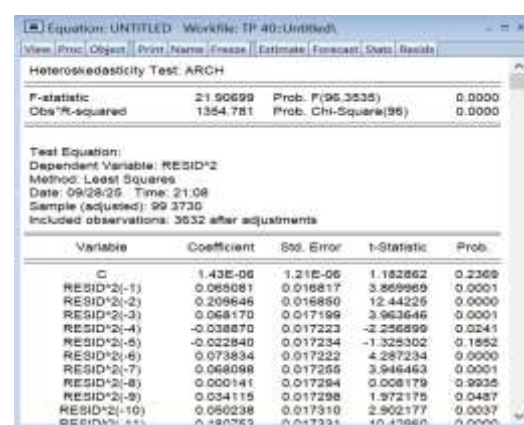
Dependent Variable: DTP
Method: ML - ARCH (Marquardt) - Normal distribution
Date: 09/28/25 Time: 00:14
Sample (adjusted): 3 3730
Included observations: 3728 after adjustments
Convergence achieved after 42 iterations
MA Backcast: 2
Presample variance: backcast (parameter = 0.7)
 $GARCH = C(4) + C(5)*GARCH(1)$

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	0.000177	1.03E-05	17.06355	0.0000
AR(1)	-0.385032	0.013641	-28.43367	0.0000
MA(1)	0.016580	0.015570	1.064856	0.2869

Variance Equation

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3.17E-10	3.06E-11	10.35829	0.0000
GARCH(1)	1.001101	1.11E-05	90444.49	0.0000

R-squared: 0.224307 Mean dependent var: 1.34E-05
Adjusted R-squared: 0.223891 S.D. dependent var: 0.004553
S.E. of regression: 0.004011 Akaike info criterion: -8.989655
Sum squared resid: 0.059941 Schwarz criterion: -8.981308
Log likelihood: 16761.72 Hannan-Quinn criter.: -8.986886
Durbin-Watson stat.: 3.987344



Heteroskedasticity Test: ARCH

	F-statistic	Prob. F(95.3535)	Obs*R-squared	Prob. Chi-Square(96)
	21.90699	0.0000	1354.781	0.0000

Test Equation:

Dependent Variable: RESID*2
Method: Least Squares
Date: 09/28/25 Time: 21:08
Sample (adjusted): 99 3730
Included observations: 3632 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.43E-06	1.21E-06	1.182862	0.2369
RESID*2(-1)	0.065081	0.016817	3.869969	0.0001
RESID*2(-2)	0.020846	0.016880	1.244225	0.0000
RESID*2(-3)	0.068170	0.017199	3.963646	0.0001
RESID*2(-4)	-0.038870	0.017223	-2.256599	0.0241
RESID*2(-5)	-0.032840	0.017234	-1.925302	0.1552
RESID*2(-6)	0.073834	0.017232	4.287234	0.0000
RESID*2(-7)	0.068098	0.017255	3.946463	0.0001
RESID*2(-8)	0.000141	0.017294	0.008179	0.9935
RESID*2(-9)	0.034115	0.017298	1.972175	0.0487
RESID*2(-10)	0.050238	0.017310	2.902177	0.0037
RESID*2(-11)	0.007953	0.017331	0.458868	0.6500

۲-۴-۵- ارزیابی و انتخاب مدل بهینه سری زمانی^۱

مقایسه تجمیعی مدل ها نشان می دهد که اگرچه مدل های AR ، MA و $ARMA$ قادر به مدل سازی ساختار میانگین هستند، اما تنها مدل های دارای واریانس شرطی می توانند دینامیک واقعی سری را بازنمایی کنند. در این میان، مدل $ARMA(1,1)$ و $GARCH(1,1)$ به دلیل بهبود همزمان در معیارهای اطلاعاتی، کاهش خطای استاندارد و حذف خودهمبستگی در باقیمانده ها،

^۱ Time Series Model Selection

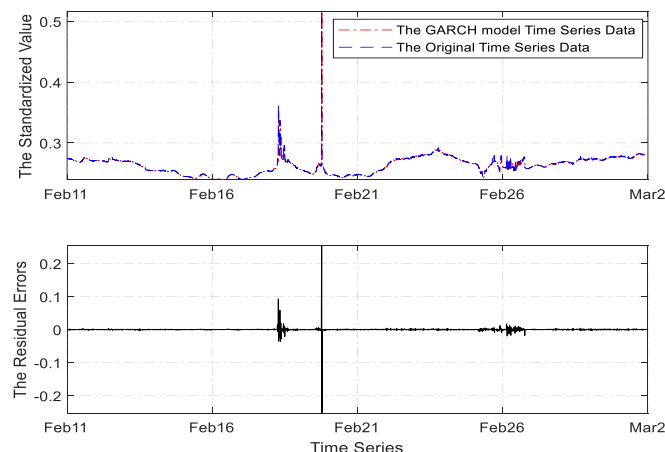
به عنوان مدل بهینه انتخاب شد. در این نوع از مدل (آرما-گارچ) همزمان وابستگی خطی در میانگین سری و ناهمسانی واریانس در نظر گرفته شده است. در نتیجه بازتولید رفتار کلی داده ها با دقت بالایی بازنمایی می شود و تطابق مناسبی بین داده های واقعی و مقادیر برآورد شده برقرار نماید. نتایج برآورد نشان داد که در میان مدل های خطی، مدل $ARMA(1,1)$ نسبت به مدل های AR و MA دارای عملکرد بهتری است و مقادیر کمتری از شاخص های آکائیک (AIC) و شوارتز (BIC) را ارائه می دهد. با این حال، مقدار پایین ضریب تعیین پائین در شکل (۱۱) (حدود ۰.۲۵)، بیانگر آن است که مدل های خطی به تنهایی قادر به توضیح کامل رفتار سری نیستند. با در نظر گرفتن ناهمسانی واریانس، مدل های $ARCH$ و $GARCH$ بهبود قابل توجهی در شاخص های برازش ایجاد کردند. به ویژه، مدل $ARMA-GARCH$ با کمترین مقدار شاخص آکائیک (-10.80) و شوارتز و همچنین آماره دوربین-واتسن نزدیک به ۲، بهترین عملکرد را در میان تمامی مدل ها نشان داد. البته، اگرچه مدل $ARMA(1,1)$ کمترین مقدار میانگین مربعات خطا را نشان می دهد، اما مدل $ARMA-GARCH$ با داشتن بیشترین مقدار حداکثر درست نمایی و کمترین مقدار آکائیک از نظر برازش کلی و لحاظ کردن ناهمسانی واریانس شرطی، عملکرد مناسب تری ارائه می دهد.

جدول ۴- نتایج مقایسه مدل های سری زمانی

مدل	log - likelihood	AIC	BIC	SE	BIC	DW	MSE
AR(1)	16329.91	-8.22313	-8.21979	0.003963	-8.21979	2.141	0,00001571
MA(1)	15359.4	-8.23674	-8.2334	0.003963	-8.2334	2.137	0,00001549
ARMA(1,1)	15383.99	-8.2516	-8.2466	0.003906	-8.2466	1.994	0,00001526
GARCH(1)	16761.72	-9.59606	-9.59068	0.003045	-9.59068	2.977	0,00001609
ARMA(1,1)-GARCH(1,1)	20175.9	-10.82022	-10.80854	0.004049	-10.80854	2.52	0,00001639

در ادامه، نتایج مربوط به داده های آزمون فاصله زمانی ۱۱ فوریه تا ۲ مارچ در شکل (۱۳) نشان می دهد که مدل $GARCH$ در توصیف رفتار کلی سری زمانی عملکرد قابل قبولی دارد، اما در مواجهه با پیک های ناگهانی و نوسانات شدید، دقت آن کاهش می یابد. به ویژه در حوالی رخداد های اوج، میزان خطای باقیمانده افزایش یافته و توزیع آن از حالت یکنواخت فاصله می گیرد. این موضوع بیانگر آن است که اگرچه مدل های مبتنی بر واریانس شرطی مانند $GARCH$ قادر به مدل سازی بخشی از ناهمسانی واریانس هستند، اما در بازنمایی کامل رفتار های غیرخطی و پیچیده سیستم های کیفیت آب محدودیت دارند. بر همین اساس، استفاده از مدل های داده محور پیشرفته تر مانند $ANFIS$ در ادامه پژوهش، به عنوان یک رویکرد مکمل برای بهبود دقت پیش بینی و تشخیص ناهنجاری ها ضروری به نظر می رسد.

شکل ۱۳- آموزش سری های زمانی مدل $GARCH$

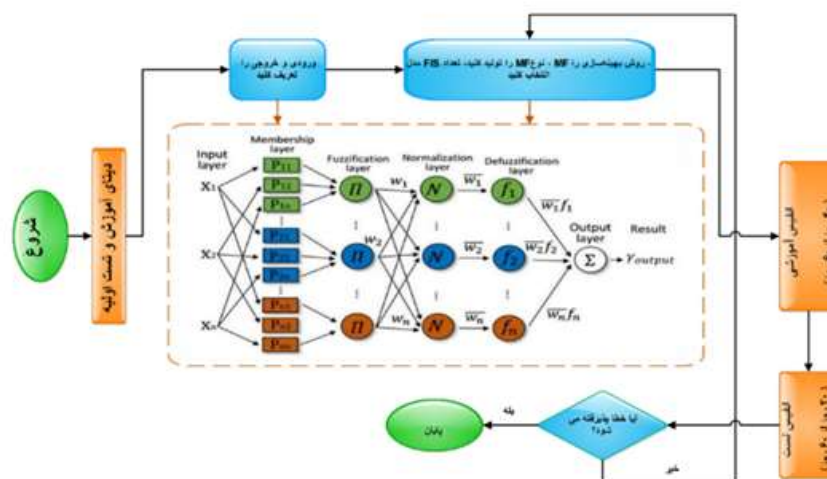


۵-۲- سیستم استنتاج فازی-عصبی تطبیقی^۱

مدل سیستم استنتاج فازی-عصبی تطبیقی به عنوان یک چارچوب ترکیبی، توانایی تلفیق قابلیت تفسیرپذیری منطق فازی با قدرت یادگیری شبکه‌های عصبی را فراهم می‌سازد. این مدل معمولاً بر پایه ساختار سوگنو و در قالب یک معماری چندلایه (اغلب پنج‌لایه) پیاده‌سازی می‌شود که در آن، توابع عضویت ورودی‌ها، قواعد فازی و پارامترهای بخش نتیجه به صورت همزمان و داده‌محور تنظیم می‌گردند. در ANFIS، ابتدا فضای ورودی از طریق توابع عضویت مناسب (مانند گاوسی یا مثلثی) فازی‌سازی شده و سپس با استفاده از مجموعه‌ای از قواعد اگر-آنگاه، نگاشت غیرخطی بین متغیرهای ورودی و خروجی برقرار می‌شود.

در این پژوهش، به منظور مدل‌سازی غیرخطی سری زمانی کیفیت آب، از مدل (ANFIS) مبتنی بر ساختار سوگنو مرتبه اول استفاده شد. ورودی‌های مدل شامل مؤلفه‌های فرکانس پایین حاصل از تبدیل موجک دابشیز (۴db) برای سه متغیر کدورت (TURB)، دما (WT) و هدایت الکتریکی (SC) بوده و خروجی مدل، مؤلفه فرکانس پایین پارامتر هدف (TP) در نظر گرفته شد. داده‌ها به دو بخش آموزش (۳۰ روز اول، ۲۸۸۰ نمونه) و آزمون (۱۱ روز پایانی) تقسیم شدند. شکل (۱۴) نمای گرافیکی از بخش انفیس مدل و اتصال آن به مرحله قبل (پیش پردازش) و آماده سازی سری زمانی آموزش دیده به منظور تخمین آستانه آلودگی را نشان می‌دهد.

شکل ۱۴- نمای شماتیک مدل انفیس



۵-۲-۱- تنظیم پارامترهای مدل^۲

برای تولید سیستم فازی اولیه، سه رویکرد شامل شبکه‌بندی^۳، خوشه‌بندی کاهشی^۴ و خوشه‌بندی FCM وجود دارد که در این مقاله، ساختار مبتنی بر خوشه‌بندی کاهشی انتخاب گردید. این انتخاب به دلیل توانایی این روش در تعیین خودکار تعداد خوشه‌ها و مراکز قوانین فازی بر اساس توزیع داده‌ها، بدون نیاز به تعریف اولیه تعداد خوشه‌ها صورت گرفت. در مقایسه با روش شبکه‌بندی^۵ که افزایش ابعاد ورودی منجر به رشد نمایی تعداد قواعد می‌شود، خوشه‌بندی کاهشی از پیچیدگی محاسباتی پایین‌تری برخوردار است. همچنین، در مقایسه با روش FCM که به مقداردهی اولیه حساس بوده و نیازمند تعیین تعداد خوشه‌ها از پیش است، این روش پایداری بیشتری در استخراج ساختار اولیه مدل ارائه می‌دهد.

¹ ANFIS

² ANFIS Model Parameter Tuning

³ Grid Partitioning

⁴ Subtractive Clustering

⁵ Grid Partitioning

بنابراین، با توجه به ماهیت غیرخطی و ابعاد نسبتاً بالای داده‌های کیفیت آب، استفاده از خوشه‌بندی کاهشی منجر به ساختار بهینه‌تر و کارآمدتر برای مدل ANFIS شد در این روش، پارامتر شعاع تأثیر خوشه‌ها^۱ برای هر ورودی برابر با ۰,۳ در نظر گرفته شد که منجر به استخراج قواعد فازی به صورت داده‌محور گردید. توابع عضویت ورودی‌ها از نوع گاوسی^۲ انتخاب شدند و تعداد آن‌ها به صورت خودکار بر اساس ساختار خوشه‌ها تعیین شد (برخلاف روش Grid که در آن به صورت ثابت ۸ تابع عضویت برای هر ورودی قابل تعریف بود. در جداول (۵) و (۶) دسته پارامترهای این بخش بر اساس تعلق به زیربخش FIS و یا بخش آموزش و اعتبار سنجی به تفکیک بیان شده است.

جدول ۵- پارامترهای سیستم فازی (FIS)

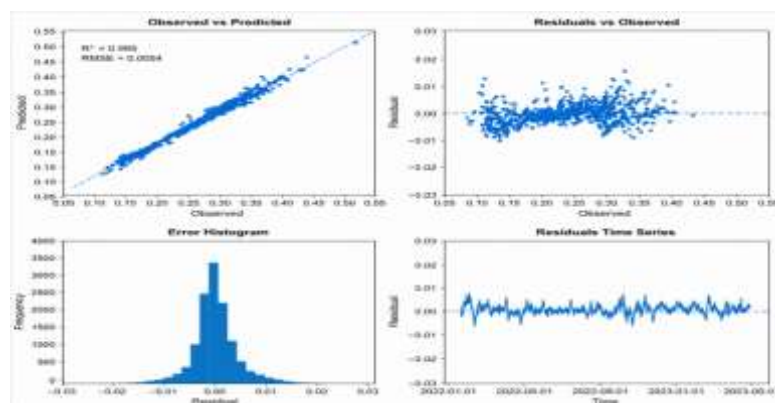
الگوریتم یادگیری	LSE+GradientDescent
تعداد تکرار (epoch)	100
نرخ یادگیری اولیه	0.01
نرخ کاهش گام	0.9
افزایش گام	0.1

جدول ۶- پارامترهای آموزش

روش خوشه بندی	Subtractive Clustring
شعاع خوشه	۰,۳
تابع عضویت	گوسین
تعداد قواعد	خودکار

در بخش اعتبارسنجی نیز از ۳۸۴۰ داده متعلق به ۴۰ روز از فاصله زمانی ۱ ژانویه ۲۰۲۲ لغایت ۹ فوریه ۲۰۲۲ به مدت ۴۰ روز و داده های تست ۲۰ روز انتهایی بازه زمانی از ۱۰ فوریه لغایت ۲ مارس به مدت ۲۰ روز و با رویکرد Cross validation استفاده شد. به منظور ارزیابی دقت پیش بینی مدل در بخش انفیس نمودارهای تشخیص در شکل (۱۵) و (۱۶) ارائه شده اند که نشان می‌دهند مدل پیشنهادی از دقت پیش‌بینی بالایی برخوردار است. تمرکز نقاط حول خط ۴۵ درجه و مقدار بالای ضریب تعیین (۰,۹۸۵) همراه با RMSE بسیار پایین، بیانگر توانایی مدل در بازتولید دقیق رفتار واقعی سیستم است. همچنین، توزیع باقیمانده‌ها در هیستوگرام خطا به صورت تقریباً نرمال و متمرکز حول صفر است که تأیید می‌کند خطاها به طور تصادفی توزیع شده‌اند و مدل دچار بیش‌برازش نشده است.

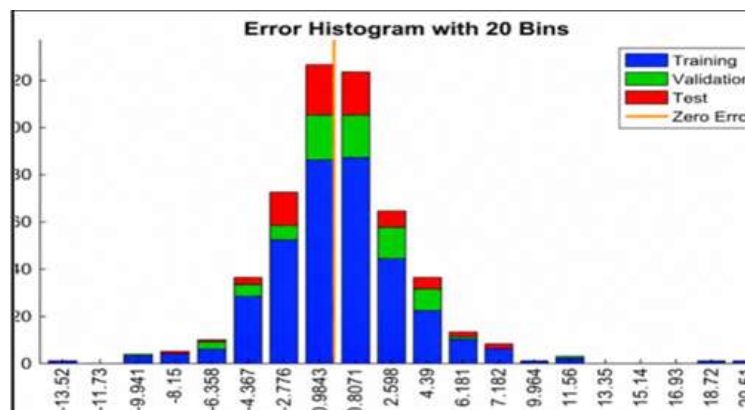
شکل ۱۵- ارزیابی عملکرد مدل



¹ Cluster InfluenceRange

² Gaussmf

شکل ۱۶- هیستوگرام خطای باقیمانده ها



۳- شبیه سازی و نتایج عددی

۳-۱- چارچوب فرایند مدل - ساخت مرجع رفتار نرمال

در این پژوهش، هدف اصلی مدل پیشنهادی صرفاً پیش‌بینی مقادیر فسفات نبوده، بلکه ایجاد یک خط پایه رفتاری برای شناسایی شرایط نرمال سیستم و آشکارسازی انحراف از آن بوده است. به همین منظور، فرایند مدل‌سازی در قالب چند گام متوالی شامل جمع‌آوری داده‌ها، استخراج غلظت فسفات با استفاده از روابط جایگزین، حذف نویز و پالایش سری زمانی با تبدیل موجک، و آموزش مدل ANFIS برای یادگیری رفتار نرمال سیستم طراحی گردید. مجموعه مراحل ارائه‌شده در این بخش در واقع فرایند ایجاد مرجع تشخیص ناهنجاری و ساخت خط پایه رفتار نرمال سیستم را تشکیل می‌دهند که به ترتیب بیان شده اند.

➤ گام اول: شرح نمونه‌برداری و جمع‌آوری داده‌ها

در مطالعه پیش رو، جمع آوری مشاهدات در فواصل زمانی ۱۵ دقیقه‌ای، درایستگاه های سنجش شماره ۰۱۶۴۶۳۰۵ و ۰۱۶۳۲۹۰۰۰ صورت پذیرفته است. با توجه به اینکه تمرکز این مطالعه بر پایش و تشخیص ناهنجاری از منظر فسفات در آب‌های سطحی است، تنها عوامل کلیدی مانند کدورت، رسانایی ویژه و دما از محل برداشت می‌شوند.

➤ گام دوم: استخراج داده‌های غلظت فسفات بر اساس رابطه جایگزین

در محیط‌های واقعی و هنگام مواجهه با تجزیه و تحلیل مسائل کیفی، به ویژه هنگامی که وابستگی و ابعاد مشخصه مشاهدات زیاد است، روش تخمین جایگزین می‌تواند به طور قابل توجهی به توسعه و انعطاف‌پذیری مدل تخمین‌گر کمک کند. غلظت شاخص فسفات به عنوان یک متغیر بسیار متغیر و با همبستگی بالا که نمی‌توان آن را مستقیماً اندازه‌گیری کرد، رابطه معنی‌داری با ضریب کدورت، دما و هدایت الکتریکی ویژه دارد.

$$TP = 0.00103TUR + 0.0057WT - 0.227\log_{10}SC + 0.776 \quad (16)$$

تفسیر معادله ۱۶ نشان می‌دهد که ضرایب پیش‌بینی تغییرات فسفات با ضریب اطمینان ۹۶٪، R^2 ، متغیرهای صحیح و معناداری بودند.

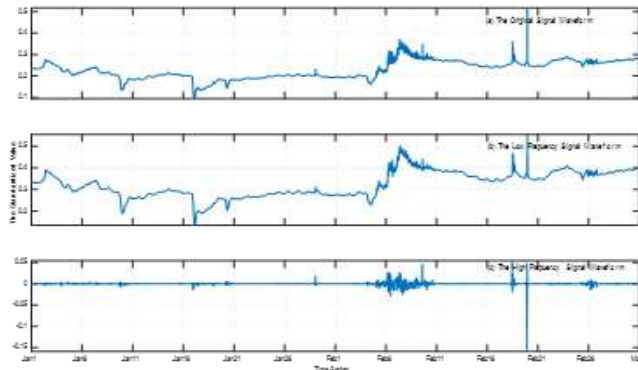
➤ گام سوم: موجک و تقریب سری زمانی و حذف مولفه های نویز

به منظور کاهش نویز و حذف نوسانات غیرضروری موجود در سری زمانی فسفات، پیش‌پردازش داده‌ها با استفاده از تبدیل موجک^۱ انجام شد. شکل (۱۷) فرایند اعمال موجک دابشیز در ۴ سطح بر روی سری زمانی فسفات را نشان می‌دهد. استفاده از این روش موجب استخراج مؤلفه‌های اصلی سیگنال و کاهش اثر نویزهای محیطی و نوسانات تصادفی گردید. داده‌های پالایش‌شده حاصل از این

^۱ Wavelet Transform

مرحله، به عنوان ورودی مدل ANFIS مورد استفاده قرار گرفتند تا فرآیند یادگیری رفتار نرمال سیستم با دقت بالاتری انجام شود. کاهش نویز در این مرحله نقش مهمی در بهبود کیفیت پیش‌بینی مدل و افزایش دقت تشخیص ناهنجاری‌ها ایفا می‌کند.

شکل ۱۷- تجزیه سری زمانی فسفات با موجک دابشیز

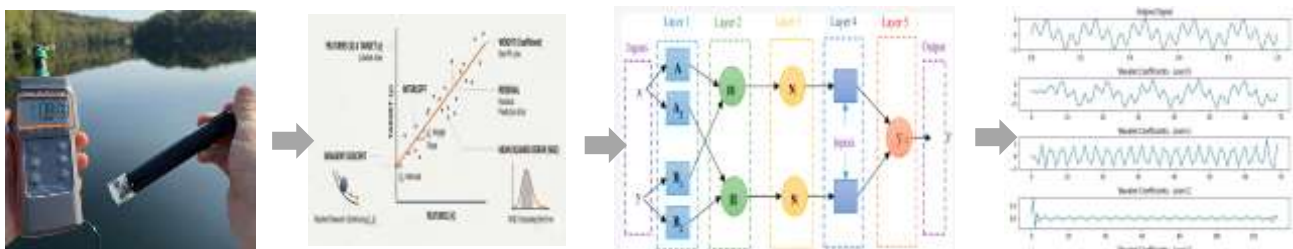


➤ گام چهارم: مدل‌سازی سری زمانی

اندازه‌گیری و جمع‌آوری مشاهدات مربوط به تاریخ ۲۶ ژانویه ۲۰۲۲ تا ۲۳ مارس ۲۰۲۲ و برای هر روز در فواصل ۱۵ دقیقه‌ای، ۹۶ مشاهده برای هر متغیر مانند کدورت، هدایت الکتریکی، دما و غیره انجام شده است. داده‌ها به دو بخش آموزش و تست تقسیم شده‌اند. برای بخش داده‌های آموزشی، از ۲۶ ژانویه تا ۲ مارس و برای بخش تست، از ۳ مارس تا ۲۳ مارس استفاده شد. برای هر روز از این بازه، ۹۶ مشاهده از هر شاخص اندازه‌گیری شده است. پس از محاسبه شاخص غلظت فسفات بر اساس رابطه جایگزین ذکر شده در معادله ۱۵، بر اساس تعداد کل مشاهدات جمع‌آوری شده، سه مدل سری زمانی ذکر شده در میحث مدل‌سازی برای داده‌های غلظت فسفات ARCH، GARCH و ARMA اعمال شد. (مجموعه داده‌های آموزشی ۳۸۹۲ فسفات است که در فاصله ۱۵ دقیقه از ایستگاه شماره ۰۱۶۴۶۳۰۵ رودخانه نمونه‌برداری شده‌اند).

این بخش از الگوریتم، مهم‌ترین بخش ساختار مدل پیشنهادی است. این بخش به مدل این قابلیت را می‌دهد که با آموزش کافی از داده‌های ورودی، ناهنجاری‌ها را در داده‌های آنلاین تشخیص دهد. پاپ لاین (۲۴) رابطه کل ساختار مدل پیشنهادی را در مدل نشان می‌دهد. برای ایجاد مرجعی برای تشخیص رفتار در الگوریتم‌های داده‌محور، ابتدا باید داده‌های جمع‌آوری شده برای آموزش، پاکسازی و نویزدایی شوند. برای این منظور، از تبدیل ویولت دابشیز با سطح تجزیه تا ۴ مرحله استفاده شده است. با تجزیه سری زمانی اصلی به دو بخش فرکانس بالا و فرکانس پایین، تابع موجک، اجزای فرکانس پایین را برای ورود به ANFIS آماده می‌کند. در این بخش، سطح آستانه تشخیص ناهنجاری (آستانه ناهنجاری) با مقایسه سری زمانی آموزش دیده و سری زمانی واقعی استخراج و محاسبه شده و معیار تصمیم‌گیری تعیین شده است.

شکل ۱۸- نمای گرافیکی بخش مرجع تشخیص ناهنجاری مدل پیشنهادی



۲-۳ - تحلیل خطای باقیمانده

در روش پیشنهادی، تشخیص آلودگی مستقیماً بر اساس مقدار خام فسفات انجام نشده است، بلکه از اختلاف بین سری زمانی واقعی و سری زمانی پیش‌بینی‌شده توسط مدل ANFIS استفاده شده است. دلیل اصلی این موضوع آن است که در سامانه‌های محیطی، مقادیر کیفیت آب همواره دارای نوسانات طبیعی، نویز محیطی، تغییرات فصلی و رفتارهای غیرخطی هستند. بنابراین، اگر تنها مقدار خام فسفات به عنوان معیار تصمیم‌گیری استفاده شود، ممکن است بسیاری از تغییرات طبیعی سیستم به اشتباه به عنوان آلودگی شناسایی شوند. به همین دلیل، در این تحقیق رفتار «مورد انتظار» سیستم ابتدا توسط مدل ANFIS یاد گرفته شد و سپس اختلاف بین رفتار واقعی و رفتار پیش‌بینی‌شده به عنوان شاخص تشخیص ناهنجاری در نظر گرفته شد.

مدل ANFIS در واقع الگوی نرمال سیستم را بر اساس داده‌های تاریخی و روابط غیرخطی بین متغیرهای کیفیت آب یاد می‌گیرد. هنگامی که سیستم در شرایط عادی قرار دارد، مقدار واقعی و مقدار پیش‌بینی‌شده اختلاف کمی با یکدیگر دارند و در نتیجه مقدار خطای باقی‌مانده کوچک باقی می‌ماند. اما زمانی که یک رخداد غیرعادی مانند افزایش ناگهانی آلودگی رخ دهد، رفتار واقعی سیستم از الگوی نرمال فاصله می‌گیرد و مدل دیگر قادر نیست مقدار جدید را با دقت بالا پیش‌بینی کند. در این حالت، اختلاف بین مقدار واقعی و مقدار پیش‌بینی‌شده افزایش یافته و خطای باقی‌مانده به طور ناگهانی بزرگ می‌شود. بنابراین خطای باقی‌مانده می‌تواند به عنوان یک شاخص حساس برای آشکارسازی رخدادها و غیرعادی مورد استفاده قرار گیرد.

از دیدگاه آماری، سری زمانی خطاها نمایانگر فاصله بین رفتار واقعی سیستم و رفتار مورد انتظار مدل است. هنگامی که این فاصله در محدوده نرمال قرار داشته باشد، سیستم در شرایط عادی تلقی می‌شود. اما اگر مقدار خطا از حدود آماری تعریف‌شده عبور کند، نشان‌دهنده آن است که رفتار سیستم دیگر با الگوی پایه سازگار نیست و احتمال وقوع آلودگی یا رخداد غیرعادی وجود دارد. به همین دلیل، در این تحقیق آستانه تشخیص آلودگی نه بر اساس مقدار مستقیم فسفات، بلکه بر اساس توزیع آماری خطاهای باقی‌مانده تعیین گردید. این رویکرد موجب می‌شود که سامانه تشخیص ناهنجاری نسبت به تغییرات واقعی و غیرعادی حساس‌تر بوده و در عین حال اثر نوسانات طبیعی و نویز محیطی کاهش یابد.

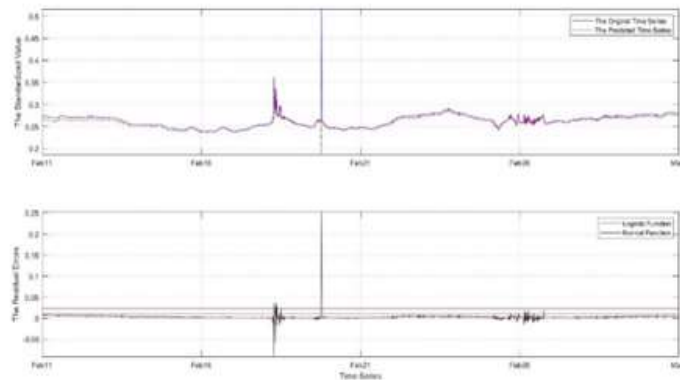
در واقع، مدل ANFIS به عنوان یک مدل پیش‌بینی‌کننده رفتار نرمال عمل می‌کند و خطای باقی‌مانده به عنوان معیار سنجش میزان انحراف سیستم از شرایط عادی مورد استفاده قرار می‌گیرد. هرچه اختلاف بین مقدار واقعی و مقدار پیش‌بینی‌شده بیشتر شود، احتمال وجود رخداد آلودگی نیز بیشتر خواهد بود. به همین دلیل، در سناریوهای دارای آلودگی شدیدتر، مقدار خطا سریع‌تر از حدود آستانه عبور کرده و مدل توانسته است رخدادها و غیرعادی را با دقت و حساسیت بالاتری شناسایی نماید.

از تفاوت دو سری زمانی داده‌های اصلی و سری زمانی-فرکانسی آموزش دیده از ANFIS، سری زمانی-فرکانسی باقیمانده خطا که در شکل (۱۷) نشان داده شده است، بدست می‌آید. همانطور که مشاهده می‌شود، تابع توزیع لجستیک برازش بهتری با خطای کمتر در مقایسه با تابع توزیع نرمال روی باقیمانده‌های خطا دارد.

شکل (۱۹) سری زمانی استانداردشده فسفات و همچنین تغییرات خطاهای باقی‌مانده را در دوره آزمون ۲۰ روزه نشان می‌دهد. همان‌گونه که مشاهده می‌شود، در شرایط عادی مقدار خطا در محدوده نزدیک به صفر باقی مانده است، در حالی که در رخدادها و غیرعادی، مقدار خطای باقی‌مانده به طور ناگهانی افزایش یافته و از حدود آستانه آماری عبور می‌کند. این موضوع نشان می‌دهد که انحراف رفتار واقعی سیستم از الگوی نرمال یادگرفته‌شده توسط مدل، به‌خوبی از طریق خطاهای باقی‌مانده قابل آشکارسازی است.

از دیدگاه آماری، سری زمانی خطاها نمایانگر فاصله بین رفتار واقعی سیستم و رفتار مورد انتظار مدل است. هنگامی که این فاصله در محدوده نرمال قرار داشته باشد، سیستم در وضعیت عادی تلقی می‌شود، اما عبور خطا از حدود آماری تعریف‌شده می‌تواند نشان‌دهنده وقوع رخدادها و غیرعادی یا آلودگی باشد. به همین دلیل، در این پژوهش آستانه تشخیص آلودگی نه بر اساس مقدار مستقیم فسفات، بلکه بر اساس توزیع آماری خطاهای باقی‌مانده تعیین گردید.

نتایج برازش توزیع‌ها بر روی سری زمانی خطاها نشان داد که توزیع لجستیک نسبت به توزیع نرمال تطابق مناسب‌تری با رفتار باقیمانده‌ها دارد. در این مطالعه، پارامترهای توزیع برابر با $\mu=0.0001$ و $s=0.015$ به‌دست آمدند که بر این اساس، حد پایین آستانه آلودگی برابر با -0.098 و حد بالای آن برابر با 0.100 تعیین گردید. در این مطالعه، $\mu=0.0001$ و $s=0.015$ است. بنابراین، حد پایین آستانه آلودگی $Thresh_{inf} = -0.098$ و حد بالای آستانه $Thresh_{sup} = 0.100$ است.



شکل ۱۹- سری زمانی استاندارد شده TP و سری خطاهای باقیمانده در طول دوره آزمایش (۲۰ روز)

۳-۳- برازش توزیع آماری

در این مرحله، به منظور تعیین مناسب‌ترین توزیع آماری برای مدل‌سازی خطاهای باقی‌مانده، مقادیر خطای باقی‌مانده استخراج شد. این مقادیر بیانگر اختلاف میان مقدار واقعی و مقدار پیش‌بینی‌شده توسط مدل هستند. سپس سه توزیع آماری نرمال، لجستیک و لاپلاس بر روی داده‌های خطا برازش داده شدند. در جدول (۷) شاخص‌های ارزیابی برازش هر سه تابع توزیع ارائه شده است.

جدول ۷- مقایسه برازش توزیع‌های آماری مختلف بر خطاهای باقیمانده مدل بر اساس معیارهای نیکویی برازش

تابع توزیع	میانگین	پارامتر مقیاس	شاخص آکائیک	شاخص بیزین	آماره آزمون KS کولموگروف-اسمیرنوف	حداکثر درست نمایی
نرمال	-6.688e-05	0.0053	-44491.61	-44478.26	0.96	22247.80
لجستیک	9.26e-05	0.002628	-45842.43	-45829.08	0.046	22923.217
لاپلاس	3.87e-05	0.003735	-45516.63	-45503.29	0.066	22760.318

برای ارزیابی کیفیت برازش هر توزیع، شاخص آماری شامل حداکثر درست نمایی، معیار اطلاعات آکائیک AIC، معیار اطلاعات BIC، آماره کولموگروف-اسمیرنوف و مقدار احتمال آن محاسبه شد. نتایج نشان داد که توزیع لجستیک با بیشترین مقدار شاخص درست نمایی و کمترین مقدار آماره آکائیک بهترین برازش را نسبت به دو توزیع دیگر ارائه داده است. در مقایسه، توزیع نرمال دارای آماره KS برابر با ۰,۰۹۶ و توزیع لاپلاس دارای آماره KS برابر با ۰,۰۶۶ بود. بنابراین، با توجه به بالاتر بودن درست نمایی و پایین‌تر بودن AIC، BIC و KS، توزیع لجستیک به عنوان مناسب‌ترین توزیع برای مدل‌سازی خطاهای باقی‌مانده و مبنای محاسبه آستانه آلودگی، انتخاب شد.

بر اساس پارامترهای برازش‌شده برای توزیع لجستیک، مقدار پارامتر مکان برابر با $\mu = 9.26 \times 10^{-5}$ و پارامتر مقیاس برابر با $s = 0.002628$ به دست آمد. انتخاب توزیع لجستیک برای ادامه تحلیل از این جهت اهمیت دارد که این توزیع نسبت به نرمال توانایی بهتری در نمایش دنباله‌های سنگین‌تر و نوسانات شدیدتر دارد؛ ویژگی‌ای که برای تحلیل داده‌های محیطی و تشخیص ناهنجاری‌های کیفیت آب بسیار مهم است.

۳-۴- تعیین حدود آستانه بر اساس تابع توزیع لجستیک^۱

به منظور تعیین حدود تشخیص آلودگی، پس از برازش تابع توزیع لجستیک بر روی سری زمانی خطاهای باقی‌مانده از تابع کوانتایل این توزیع برای محاسبه بازه اطمینان آماری استفاده شد.

^۱ Threshold Boundary Determination

$$f(x, \mu, s) = \frac{e^{-\frac{x-\mu}{s}}}{s(1 + e^{-\frac{x-\mu}{s}})} \quad (۱۷)$$

$$Thresh_{inf} = \mu + \ln\left(\frac{p_{inf}}{1 - p_{inf}}\right)s = \mu - 5.29s \quad (۱۸)$$

$$Thresh_{sup} = \mu + \ln\left(\frac{p_{sup}}{1 - p_{sup}}\right)s = \mu + 5.29s \quad (۱۹)$$

که در آن μ پارامتر مکان یا میانگین توزیع خطاها و پارامتر S مقیاس یا میزان پراکندگی داده‌ها است. با استفاده از این مقادیر عددی جدول (۷)، حدود بالا و پایین بازه اطمینان برای سطوح مختلف اطمینان مانند ۹۵٪ و ۹۹٪ محاسبه گردید. بر این اساس، انتظار می‌رود بخش عمده رفتار نرمال سیستم در داخل این بازه قرار گیرد، و داده‌های خارج از حدود فوق به عنوان رخدادهای غیرعادی یا آلودگی احتمالی شناسایی شوند. از آنجا که توزیع لجستیک دارای دنباله‌های ضخیم‌تری نسبت به توزیع نرمال است، این روش توانایی بیشتری در آشکارسازی رخدادهای شدید و نادر آلودگی در سری زمانی کیفیت آب فراهم می‌کند.

$$Q_p = \mu + s \cdot \ln \frac{p}{1 - p} \quad (۲۰)$$

$$Q_p = 9.26 * 10^{-5} \mp 0.0026 \cdot \ln \frac{95}{1 - 95} \quad (۲۱)$$

۵-۳- ارزیابی عملکرد مدل - ماتریس درهم ریختگی^۱

به منظور ارزیابی عملکرد مدل‌های پیشنهادی در تشخیص رخدادهای غیرعادی کیفیت آب، از ماتریس درهم ریختگی^۲ استفاده شد. این ماتریس یکی از متداول‌ترین ابزارهای ارزیابی در مسائل طبقه‌بندی و آشکارسازی ناهنجاری است که امکان تحلیل عملکرد مدل را در تفکیک صحیح شرایط نرمال و آلوده فراهم می‌کند. در این روش، خروجی مدل با برچسب واقعی داده‌ها مقایسه شده و نتایج در قالب چهار حالت شامل مثبت صحیح (TP)، منفی صحیح (TN)، مثبت کاذب (FP) و منفی کاذب (FN) بیان می‌شود. در این میان، TP و TN نشان‌دهنده تشخیص صحیح نمونه‌ها بوده، در حالی که FP و FN بیانگر خطاهای طبقه‌بندی مدل هستند. استفاده از ماتریس درهم ریختگی این امکان را فراهم می‌کند که علاوه بر دقت کلی مدل، نوع و ماهیت خطاهای رخ داده نیز مورد بررسی قرار گیرد. این موضوع در سامانه‌های پایش کیفیت آب اهمیت ویژه‌ای دارد، زیرا عدم شناسایی رخداد واقعی آلودگی (FN) می‌تواند منجر به پیامدهای زیست‌محیطی و مدیریتی جدی شود، در حالی که افزایش هشدارهای کاذب (FP) موجب کاهش قابلیت اعتماد سامانه و افزایش هزینه‌های عملیاتی خواهد شد. بر اساس مقادیر حاصل از ماتریس درهم ریختگی، شاخص‌های ارزیابی شامل Accuracy، Misclassification Rate، Sensitivity Recall و Specificity محاسبه شدند تا عملکرد مدل‌های ANFIS و Wavelet در سناریوهای مختلف آلودگی مورد مقایسه قرار گیرد. جدول (۸) ساختار کلی ماتریس درهم ریختگی را نشان می‌دهد.

جدول ۸- ماتریس درهم ریختگی

Predicted Label			
Negative	Positive		
FN	TP	Positive	
TN	FP	Negative	

^۱ Model Performance Evaluation Using Confusion Matrix

^۲ Confusion Matrix

دقت: نسبت کل نمونه‌های درست طبقه‌بندی شده به کل داده‌ها را نشان می‌دهد و به صورت زیر محاسبه می‌شود:

$$\text{Accuracy} = (TP + TN) / (TP + FN + FP + TN) \quad (22)$$

نرخ طبقه‌بندی نادرست^۱: بیانگر نسبت کل خطاهای مدل به تعداد کل نمونه‌ها می‌باشد. فرمول محاسبه آن به شرح زیر است:

$$\text{Mis Classification Rate} = (FP + FN) / (TP + FN + FP + TN) \quad (23)$$

حساسیت: توانایی مدل در شناسایی صحیح رخدادهای مثبت واقعی را نشان می‌دهد:

$$\text{Sensitivity (TPR)} = TP / (TP + FN) \quad (24)$$

ویژگی^۲: نیز بیانگر توانایی مدل در تشخیص صحیح نمونه‌های منفی واقعی می‌باشد. فرمول محاسبه آن به شکل معادله زیر است:

$$\text{Specificity} = TN / (FP + TN) \quad (25)$$

۳-۶- طراحی سناریوهای آلودگی و ماتریس‌های درهم ریختگی

در این پژوهش، به منظور ارزیابی اثربخشی الگوریتم پیشنهادی در تشخیص رخدادهای غیرعادی، دو سناریوی آلودگی مصنوعی در بازه زمانی ساعت ۱۰ تا ۱۲، طی دوره ۱۱ فوریه تا ۲ مارس، تعریف و اجرا شد. در سناریوی نخست (I)، غلظت شاخص‌های مورد بررسی در بازه مذکور به دو برابر مقادیر سری زمانی اصلی افزایش یافت. در سناریوی دوم (II) نیز، با حفظ همان بازه زمانی، غلظت شاخص‌ها به سه برابر مقادیر اولیه افزایش داده شد. سپس عملکرد مدل ANFIS و مدل ترکیبی Wavelet-ANFIS در تشخیص این رخدادها ارزیابی شد. برای تشکیل ماتریس‌های درهم‌ریختگی، مقادیر واقعی فسر کل با مقادیر پیش‌بینی شده توسط مدل‌های ANFIS و Wavelet-ANFIS مقایسه و اختلاف میان آن‌ها به عنوان خطای باقیمانده محاسبه شد. در ادامه، با استفاده از حدود بالایی و پایینی تعیین شده بر اساس تابع توزیع لجستیک، هر مشاهده در یکی از دو طبقه «نرمال» یا «آلوده/ناهنجار» قرار گرفت. بدین ترتیب، مشاهده‌ای که خطای باقیمانده آن در محدوده آستانه قرار داشت، به عنوان وضعیت نرمال طبقه‌بندی شد؛ در مقابل، عبور خطای باقیمانده از حد بالایی یا پایینی آستانه به منزله وقوع آلودگی یا ناهنجاری تلقی و آن مشاهده علامت‌گذاری شد.

۳-۷- مقایسه عملکرد مدل در سناریوهای متفاوت^۳

پس از تعیین وضعیت واقعی و وضعیت پیش‌بینی شده هر مشاهده، تعداد داده‌های قرارگرفته در چهار حالت اصلی ماتریس درهم‌ریختگی محاسبه گردید. اگر یک رخداد آلودگی به درستی توسط مدل شناسایی شده باشد، به عنوان مثبت صحیح (TP) ثبت می‌شود. در مقابل، اگر مدل یک شرایط نرمال را به اشتباه آلوده تشخیص دهد، در دسته مثبت کاذب (FP) قرار می‌گیرد. همچنین، اگر شرایط نرمال به درستی تشخیص داده شود، به عنوان منفی صحیح (TN) و اگر رخداد واقعی آلودگی توسط مدل شناسایی نشود، به عنوان منفی کاذب (FN) در نظر گرفته می‌شود. در واقع، اعداد داخل هر ماتریس نشان‌دهنده تعداد دفعاتی هستند که مدل توانسته وضعیت واقعی سیستم را به درستی یا نادرستی تشخیص دهد. در این تحقیق، چهار ماتریس درهم‌ریختگی برای دو مدل ANFIS و Wavelet-ANFIS و در دو سناریوی مختلف شدت آلودگی تشکیل شد. به این ترتیب، عملکرد مدل‌ها نه تنها بر اساس دقت کلی، بلکه بر اساس توانایی آن‌ها در شناسایی رخدادهای واقعی آلودگی و کاهش هشدارهای کاذب مورد ارزیابی قرار

¹ Misclassification Rate

² Specificity

³ Comparative Performance Under Different Scenarios

گرفت. این روش امکان تحلیل دقیق حساسیت مدل نسبت به تغییرات کیفیت آب و بررسی قابلیت اطمینان سامانه تشخیص ناهنجاری را فراهم می‌کند.

جدول ۹- نتایج پارامترهای ارزیابی با دو سناریوی مختلف

ANFIS-Sen.I				Wavelet-ANFIS-Sen.I				ANFIS-Sen.II				Wavelet-ANFIS-Sen.II			
		Predicted				Predicted				Predicted				Predicted	
		P	N			P	N			P	N			P	N
	P	۱۶۴۶	۱۳۶		P	۱۷۲۰	۵۴		P	۱۶۹۰	۸۵		P	۱۷۶۴	۲۰
	N	۳۰	۱۰۸		N	۲۶	۱۲۰		N	۱۵	۱۳۰		N	۱۷	۱۱۹

پیش بینی اشتباه

پیش بینی صحیح

- مقایسه شاخص های دقت، نرخ طبقه بندی، حساسیت، ویژگی در سناریوهای آلودگی

	سناریو ۱		سناریو ۲	
	ANFIS RBF	Wavelet-ANFIS Wavelet-RBF	ANFIS RBF	Wavelet-ANFIS Wavelet-RBF
دقت (Accuracy)	91.35%	95.83%	94.79%	98.07%
نرخ طبقه بندی نادرست (Misclassification Rate)	91%	98%	93%	93.4%
	8.65%	4.17%	5.21%	1.93%
حساسیت (Recall)	8%	1%	6%	6%
	92.37%	96.97%	95.21%	98.98%
ویژگی (Specifity)	92%	98%	93%	99.7%
	78.26%	82.19%	89.66%	99.04%
F1-Score	89%	97%	93%	99.1%
	95.2%	97.7%	93.8%	97.1%
	91.9%	98.2%		99.08%
				99.5%

بررسی و مقایسه شاخص دقت در دو ستون مربوط به مدل انفیس به تنهایی و مدل Wavelet-ANFIS در هر دو سناریو دارای بالاترین نرخ طبقه بندی صحیح بوده است. در سناریوی اول، دقت مدل Wavelet-ANFIS برابر با ۹۵,۸۳٪ و در سناریوی دوم برابر با ۹۸,۰۷٪ به دست آمد، در حالی که مدل ANFIS بدون اعمال مویک به ترتیب مقادیر ۹۱,۳۵٪ و ۹۴,۷۹٪ را نشان داد. این نتایج بیانگر آن است که حذف مؤلفه های نویزی پیش از مرحله یادگیری، نقش بسیار مهمی در بهبود توانایی مدل برای تشخیص شرایط نرمال و غیرنرمال داشته است.

نتایج حاصل از ماتریس های درهم ریختگی نشان می دهد که مدل پیشنهادی Wavelet-ANFIS عملکرد بهتری نسبت به مدل ANFIS منفرد و همچنین روش مبتنی بر شبکه عصبی شعاعی ارائه شده در پژوهش پیشین [۱۶] داشته است. در تحقیق قبلی، از شبکه عصبی شعاعی و مویک برای تشخیص ناهنجاری های کیفیت آب استفاده شده بود، در حالی که در پژوهش حاضر، سیستم استنتاج فازی-عصبی ANFIS به منظور افزایش پایداری مدل و کاهش اثر نویزهای فرکانس بالا به کار گرفته شد. مقایسه شاخص های ارزیابی نشان داد که استفاده از تبدیل مویک پیش از مرحله یادگیری موجب افزایش دقت تشخیص، کاهش نرخ طبقه بندی نادرست و بهبود حساسیت مدل در شناسایی رخدادهای واقعی آلودگی شده است.

با وجود عملکرد مطلوب مدل پیشنهادی، در برخی موارد مقدار Specificity نسبت به روش مبتنی بر شبکه عصبی شعاعی اندکی کمتر مشاهده شد. این موضوع نشان می دهد که مدل ANFIS به دلیل حساسیت بالاتر به تغییرات غیرخطی و الگوهای مرزی، تمایل بیشتری به شناسایی تغییرات کوچک به عنوان رخداد آلودگی دارد. در نتیجه، مدل تلاش می کند از حذف رخدادهای واقعی آلودگی جلوگیری نماید؛ بنابراین مقدار True Positive افزایش یافته و False Negative کاهش می یابد که در نهایت منجر به افزایش Recall می شود. با این حال، این رفتار می تواند موجب افزایش نسبی False Positive و کاهش جزئی Specificity گردد.

به عبارت دیگر، مدل رویکردی محافظه کارانه اتخاذ می‌کند و ترجیح می‌دهد هشدار اضافی تولید کند تا اینکه یک رخداد واقعی آلودگی را از دست بدهد. در سامانه‌های پایش محیط‌زیستی، از دست دادن یک آلودگی واقعی می‌تواند پیامدهای جدی‌تری نسبت به هشدار کاذب داشته باشد؛ بنابراین دستیابی به Recall بالا و حساسیت بیشتر در این پژوهش اهمیت بالاتری نسبت به بیشینه‌سازی Specificity داشته است.

۴- نتیجه‌گیری

نتایج این پژوهش نشان داد که ترکیب تبدیل موجک با سیستم استنتاج فازی-عصبی تطبیقی (Wavelet-ANFIS) توانسته است عملکرد بسیار مؤثری در تشخیص ناهنجاری‌های کیفیت آب ارائه دهد. در این تحقیق، ابتدا با استفاده از داده‌های کیفی آب شامل کدورت، هدایت الکتریکی ویژه و دمای آب، مقادیر فسفات توسط مدل‌های ANFIS و Wavelet-ANFIS پیش‌بینی گردید. سپس اختلاف بین مقادیر واقعی و مقادیر پیش‌بینی‌شده به صورت خطای باقی‌مانده محاسبه شد و به عنوان شاخص اصلی تشخیص ناهنجاری مورد استفاده قرار گرفت. این رویکرد موجب شد که فرآیند آشکارسازی آلودگی بر پایه میزان انحراف رفتار واقعی سیستم از الگوی نرمال پیش‌بینی‌شده انجام شود، و فرایند تشخیص صرفاً بر اساس مقدار خام غلظت فسفات این سنجش صورت نپذیرد. به منظور تعیین حدود تشخیص آلودگی، توزیع آماری خطاهای باقی‌مانده مورد تحلیل قرار گرفت و سه تابع توزیع نرمال، لاپلاس و لجستیک بر روی داده‌ها برازش داده شدند. نتایج شاخص‌های برازش نشان داد که توزیع لجستیک بهترین برازش را برای مدل‌سازی رفتار باقی‌مانده‌ها ارائه می‌دهد. دلیل این موضوع، توانایی بیشتر توزیع لجستیک در نمایش دنباله‌های سنگین‌تر و رخدادهای نادر محیطی نسبت به توزیع نرمال است. بر اساس پارامترهای استخراج‌شده از توزیع لجستیک، حدود بالا و پایین آستانه آلودگی با استفاده از تابع کوانتایل محاسبه گردید و هر مشاهده‌ای که مقدار خطای آن خارج از بازه اطمینان قرار می‌گرفت، به عنوان رخداد غیرعادی یا آلودگی احتمالی شناسایی شد.

بررسی شاخص F1-score نیز نتایج به‌دست‌آمده از Accuracy و Recall را تأیید می‌کند. از آنجا که میانگین هماهنگ Precision و Recall است، مقدار بالاتر آن نشان‌دهنده تعادل بهتر مدل بین شناسایی صحیح رخدادهای آلودگی و کاهش تشخیص‌های نادرست است. نتایج نشان داد که مدل‌های مبتنی بر Wavelet در هر دو سناریو دارای F1-score بالاتری نسبت به مدل‌های منفرد هستند. به‌ویژه، مدل Wavelet-ANFIS در سناریوی دوم با F1-score برابر با ۹۹٫۰۸٪ بهترین عملکرد را نشان داد. این نتیجه بیانگر آن است که مدل پیشنهادی علاوه بر افزایش حساسیت در تشخیص آلودگی، توانسته پایداری و تعادل طبقه‌بندی را نیز حفظ کند. همچنین در سناریوی دوم، مدل Wavelet-ANFIS با دستیابی به Accuracy برابر با ۹۸٫۰۷٪، Recall برابر با ۹۸٫۸۸، Precision برابر با ۹۹٫۰۴٪ و F1-score برابر با ۹۹٪ بهترین عملکرد را در میان تمامی مدل‌ها نشان داد. بالاتر بودن F1-score در مدل RBF نسبت به ANFIS در برخی موارد نشان می‌دهد که شبکه شعاعی تعادل بهتری میان Precision و Recall برقرار کرده است. در مقابل، مدل ANFIS به دلیل حساسیت بیشتر به نوسانات غیرخطی و الگوهای مرزی، توانایی بالاتری در شناسایی رخدادهای واقعی آلودگی نشان داده است. با این حال، این حساسیت بیشتر می‌تواند با افزایش نسبی هشدارهای کاذب همراه باشد. بنابراین، کاهش جزئی F1-score در مدل ANFIS لزوماً به معنی ضعف مدل نیست، بلکه بیانگر نوعی مبادله میان افزایش حساسیت تشخیص و کنترل خطاهای مثبت کاذب است. در کاربردهای پایش کیفیت آب، کاهش رخدادهای از دست‌رفته آلودگی اهمیت ویژه‌ای دارد؛ از این رو، Recall بالا همچنان یکی از معیارهای کلیدی ارزیابی مدل محسوب می‌شود. این نتایج بیانگر آن است که حذف مؤلفه‌های نویزی فرکانس بالا توسط تبدیل موجک موجب کاهش پراکندگی خطاها، افزایش تمرکز Residual Errorها در اطراف میانگین و بهبود پایداری فرایند طبقه‌بندی شده است.

نتایج همچنین نشان داد که با افزایش شدت آلودگی، عملکرد مدل در تشخیص ناهنجاری‌ها بهبود یافته است. در سناریوهای دارای آلودگی شدیدتر، فاصله بین رفتار واقعی سیستم و الگوی پایه پیش‌بینی‌شده افزایش یافته و در نتیجه مقدار Residual Error با وضوح بیشتری از حدود آستانه عبور کرده است. این موضوع سبب شد که مدل بتواند رخدادهای شدید آلودگی را با حساسیت و دقت بالاتری شناسایی نماید. با این حال، تشخیص آلودگی‌های کم‌دامنه و تغییرات تدریجی همچنان چالش‌برانگیزتر از رخدادهای شدید است، زیرا این نوع تغییرات ممکن است همپوشانی بیشتری نسبت به نوسانات طبیعی داشته باشد و تفکیک آن‌ها از رفتار عادی سیستم دشوارتر باشد.

مقایسه نتایج این پژوهش با مطالعه پیشین مبتنی بر شبکه عصبی شعاعی [۱۶] نشان داد که استفاده از Wavelet-ANFIS موجب بهبود قابل توجه عملکرد سامانه تشخیص ناهنجاری شده است. این موضوع بیانگر آن است که ترکیب روش‌های کاهش نویز، استخراج ویژگی و یادگیری غیرخطی می‌تواند قابلیت اطمینان سامانه‌های پایش هوشمند کیفیت آب را افزایش دهد. در مجموع، نتایج این پژوهش نشان می‌دهد که مدل Wavelet-ANFIS دارای توانایی بالایی در شناسایی رخدادهای غیرعادی کیفیت آب بوده و می‌تواند به عنوان ابزاری مؤثر در سامانه‌های هشدار سریع آلودگی و مدیریت هوشمند منابع آب مورد استفاده قرار گیرد.

اعلامیه‌ها

حمایت مالی:

این پژوهش هیچ‌گونه حمایت مالی از سازمان‌های دولتی، خصوصی یا غیرانتفاعی دریافت نکرده است و به‌صورت مستقل توسط نویسندگان انجام شده است.

تضاد منافع:

نویسندگان اعلام می‌دارند که هیچ‌گونه تضاد منافع مرتبط با این پژوهش وجود ندارد.

تأییدیه اخلاقی:

این مطالعه شامل انجام پژوهش بر روی انسان یا حیوان نبوده و بنابراین نیاز به اخذ تأییدیه اخلاقی و کد اخلاق پژوهش نداشته است.

رضایت آگاهانه:

مطالعه حاضر نیازمند اخذ رضایت آگاهانه نبوده است.

دسترسی به داده‌ها:

داده‌های مورد استفاده در این پژوهش بنا به درخواست منطقی از نویسنده مسئول قابل دسترسی است.

سهام نویسندگان:

تمامی نویسندگان در جمع‌آوری داده‌ها، تحلیل و تفسیر نتایج، نگارش مقاله و بازبینی و تأیید نسخه نهایی مقاله مشارکت داشته‌اند. تمامی نویسندگان نسخه نهایی مقاله را مطالعه و تأیید کرده‌اند.

۵-منابع و مراجع

1. Akhtar, N., Syakir Ishak, M. I., Bhawani, S. A., and Umar, K. (2021). "Various natural and anthropogenic factors responsible for water quality degradation: A review." *Water*, 13(19), 2660. <https://doi.org/10.3390/w13192660>
2. Xu, G., Fan, H., Oliver, D. M., Dai, Y., Li, H., Shi, Y., Long, H., Xiong, K., and Zhao, Z. (2022). "Decoding river pollution trends and their landscape determinants in an ecologically fragile karst basin using a machine learning model." *Environmental Research*, 214(Part 4), 113843. <https://doi.org/10.1016/j.envres.2022.113843>
3. Wali, S. U. (2021). "The need for a multi-pollutant approach to model the movement of pollutants in surface-water: A review of status and future challenges." *International Journal of Agriculture and Animal Production*, 1(1), 26–58. <https://doi.org/10.55529/ijaap.11.26.58>
4. Bieroza, M. Z., and Heathwaite, A. L. (2015). "Seasonal variation in phosphorus concentration–discharge hysteresis inferred from high-frequency in situ monitoring." *Journal of Hydrology*, 524, 333–347. <https://doi.org/10.1016/j.jhydrol.2015.02.036>

5. Kusari, L. (2022). "Turbidity as a surrogate for the determination of total phosphorus, using relationship based on sub-sampling techniques." *Ecological Engineering & Environmental Technology*, 23(4), 88–93. <https://doi.org/10.12912/27197050/150233>
6. Bowes, M. J., Loewenthal, M., Read, D. S., Hutchins, M. G., Prudhomme, C., Armstrong, L. K., Harman, S. A., Wickham, H. D., Gozzard, E., and Carvalho, L. (2016). "Identifying multiple stressor controls on phytoplankton dynamics in the River Thames (UK) using high-frequency water quality data." *Science of the Total Environment*, 569–570, 1489–1499. <https://doi.org/10.1016/j.scitotenv.2016.06.239>
7. Mozo, A., Morón-López, J., Vakaruk, S., Pompa-Pernía, A. G., González-Prieto, A., Pascual Aguilar, J. A., Gómez-Canaval, S., and Ortiz, J. M. (2022). "Chlorophyll soft-sensor based on machine learning models for algal bloom predictions." *Scientific Reports*, 12, 13529. <https://doi.org/10.1038/s41598-022-17299-5>
8. Yoon, H. N., Marshall, L., Sharma, A., and Kim, S. (2022). "Bayesian model calibration using surrogate streamflow in ungauged catchments." *Water Resources Research*, 58(1), e2021WR031287. <https://doi.org/10.1029/2021WR031287>
9. Liu, P., Wang, J., Sangaiah, A. K., Xie, Y., and Yin, X. (2019). "Analysis and prediction of water quality using LSTM deep neural networks in IoT environment." *Sustainability*, 11(7), 2058. <https://doi.org/10.3390/su11072058>
10. Dogo, E. M., Nwulu, N. I., Twala, B., and Aigbavboa, C. (2019). "A survey of machine learning methods applied to anomaly detection on drinking-water quality data." *Urban Water Journal*, 16(3), 235–248. <https://doi.org/10.1080/1573062X.2019.1637002>
11. Mola, M., and Amiri-Ahouee, R. (2021). "ANFIS model based on fuzzy C-mean, grid partitioning and subtractive clustering to detection of stator winding inter-turn fault for PM synchronous motor." *International Transactions on Electrical Energy Systems*, 31(3), e12770. <https://doi.org/10.1002/2050-7038.12770>
12. Sun, Y., Babovic, V., and Chan, E. S. (2010). "Multi-step-ahead model error prediction using time-delay neural networks combined with chaos theory." *Journal of Hydrology*, 395(1–2), 109–116. <https://doi.org/10.1016/j.jhydrol.2010.10.020>
13. Tiwari, S., Babbar, R., and Kaur, G. (2018). "Performance evaluation of two ANFIS models for predicting water quality index of River Satluj (India)." *Advances in Civil Engineering*, 2018, Article ID 8971079. <https://doi.org/10.1155/2018/8971079>
14. Kabolizadeh, M., Rangzan, K., Zareie, S., Rashidian, M., and Delfan, H. (2022). "Evaluating quality of surface water resources by ANN and ANFIS networks using Sentinel-2 satellite data." *Earth Science Informatics*, 15(1), 523–540. <https://doi.org/10.1007/s12145-021-00741-z>
15. Bhagowati, B., Talukdar, B., Narzary, B. K., and Ahamad, K. U. (2022). "Prediction of lake eutrophication using ANN and ANFIS by artificial simulation of lake ecosystem." *Modeling Earth Systems and Environment*, 8, 5289–5304. <https://doi.org/10.1007/s40808-022-01377-8>
16. Khayat, M., Noorossana, R., Soleimani, P., and Raissi, S. (2024). "Hybrid algorithm for early detection of water pollution impact on environmental indicators using wavelet techniques and

- RBF neural network learning.” *Pollution*, 10(4), 1074–1091. <https://doi.org/10.22059/poll.2024.372064.2246>
17. U.S. Environmental Protection Agency. (2013). *Water Quality Event Detection System Challenge: Methodology and Findings*. EPA 817-R-13-002, Office of Water, U.S. Environmental Protection Agency. https://www.epa.gov/sites/default/files/2015-07/documents/water_quality_event_detection_system_challenge_methodology_and_findings.pdf
 18. U.S. Geological Survey. (2017). *National Hydrography Dataset High-Resolution Flowline Data*. The National Map. Accessed 20 May 2017. <https://www.usgs.gov/national-hydrography/nhdplus-high-resolution>
 19. Byrand, K. (2010). “Nature and History in the Potomac Country: From Hunter-Gatherers to the Age of Jefferson.” *Journal of Historical Geography*, 36(2), 233–234. <https://doi.org/10.1016/j.jhg.2010.02.011>
 20. Zhang, Y. F., and Thorburn, P. J. (2022). “A deep surrogate model with spatio-temporal awareness for water quality sensor measurement.” *Expert Systems with Applications*, 200, 116914. <https://doi.org/10.1016/j.eswa.2022.116914>
 21. Kunz, J. V., Hensley, R., Brase, L., Borchardt, D., and Rode, M. (2017). “High frequency measurements of reach scale nitrogen uptake in a fourth order river with contrasting hydromorphology and variable water chemistry (Weiße Elster, Germany).” *Water Resources Research*, 53(1), 328–343. <https://doi.org/10.1002/2016WR019355>
 22. Shi, B., Wang, P., Jiang, J., and Liu, R. (2018). “Applying high-frequency surrogate measurements and a wavelet–ANN model to provide early warnings of rapid surface water quality anomalies.” *Science of the Total Environment*, 610–611, 1390–1399. <https://doi.org/10.1016/j.scitotenv.2017.08.232>
 23. Kuo, J. T., Wang, Y. Y., and Lung, W. S. (2006). “A hybrid neural-genetic algorithm for reservoir water quality management.” *Water Research*, 40(7), 1367–1376. <https://doi.org/10.1016/j.watres.2006.01.046>
 24. Christensen, V. G. (2001). *Characterization of Surface-Water Quality Based on Real-Time Monitoring and Regression Analysis*, Quivira National Wildlife Refuge, South-Central Kansas, December 1998 through June 2001. U.S. Geological Survey, Water-Resources Investigations Report 2001-4248. <https://doi.org/10.3133/wri014248>
 25. Erkyihun, S. T., Rajagopalan, B., Zagana, E., Lall, U., and Nowak, K. (2016). “Wavelet-based time series bootstrap model for multidecadal streamflow simulation using climate indicators.” *Water Resources Research*, 52(5), 4061–4077. <https://doi.org/10.1002/2016WR018696>
 26. Shupe, S. M. (2017). “High resolution stream water quality assessment in the Vancouver, British Columbia region: A citizen science study.” *Science of the Total Environment*, 603–604, 745–759. <https://doi.org/10.1016/j.scitotenv.2017.02.195>
 27. Kirchgässner, G., Wolters, J., and Hassler, U. (2013). *Introduction to Modern Time Series Analysis*. 2nd ed., Springer, Berlin and Heidelberg. <https://doi.org/10.1007/978-3-642-33436-8>

28. Chen, N., Tu, H., Duan, X., Hu, L., and Guo, C. (2023). "Semi-supervised anomaly detection of multivariate time series based on a variational autoencoder." *Applied Intelligence*, 53(5), 6074–6098. <https://doi.org/10.1007/s10489-022-03829-1>
29. Lundgren, A., and Jung, D. (2022). "Data-driven fault diagnosis analysis and open-set classification of time-series data." *Control Engineering Practice*, 121, 105006. <https://doi.org/10.1016/j.conengprac.2021.105006>
30. Karasu, S., and Altan, A. (2022). "Crude oil time series prediction model based on LSTM network with chaotic Henry gas solubility optimization." *Energy*, 242, 122964. <https://doi.org/10.1016/j.energy.2021.122964>
31. Jin, X. B., Gong, W. T., Kong, J. L., Bai, Y. T., and Su, T. L. (2022). "PFVAE: A planar flow-based variational auto-encoder prediction model for time series data." *Mathematics*, 10(4), 610. <https://doi.org/10.3390/math10040610>
32. Du, L., Gao, R., Suganthan, P. N., and Wang, D. Z. W. (2022). "Bayesian optimization based dynamic ensemble for time series forecasting." *Information Sciences*, 591, 155–175. <https://doi.org/10.1016/j.ins.2022.01.010>
33. Chen, H., Li, Q., and Zhu, F. (2022). "A new class of integer-valued GARCH models for time series of bounded counts with extra-binomial variation." *AStA Advances in Statistical Analysis*, 106(2), 243–270. <https://doi.org/10.1007/s10182-021-00414-8>
34. Tan, W. Y., Lai, S. H., Teo, F. Y., and El-Shafie, A. (2022). "State-of-the-art development of two-waves artificial intelligence modeling techniques for river streamflow forecasting." *Archives of Computational Methods in Engineering*, 29, 5185–5211. <https://doi.org/10.1007/s11831-022-09763-2>
35. Hajirahimi, Z., and Khashei, M. (2023). "Hybridization of hybrid structures for time series forecasting: A review." *Artificial Intelligence Review*, 56, 1201–1261. <https://doi.org/10.1007/s10462-022-10199-0>
36. Makridakis, S., and Hibon, M. (1997). "ARMA models and the Box–Jenkins methodology." *Journal of Forecasting*, 16(3), 147–163. <https://doi.org/10.1002/%28SICI%291099-131X%28199705%2916%3A3%3C147%3A%3AAID-FOR652%3E3.0.CO%3B2-X>
37. Engle, R. F. (1982). "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation." *Econometrica*, 50(4), 987–1007. <https://doi.org/10.2307/1912773>
38. Heberling, M. T., Price, J. I., Nietch, C. T., Elovitz, M., Smucker, N. J., Schupp, D. A., Safwat, A., and Neyer, T. (2022). "Linking water quality to drinking water treatment costs using time series analysis: Examining the effect of a treatment plant upgrade in Ohio." *Water Resources Research*, 58(5), e2021WR031257. <https://doi.org/10.1029/2021WR031257>
39. Song, C., and Yao, L. (2022). "Application of artificial intelligence based on synchrosqueezed wavelet transform and improved deep extreme learning machine in water quality prediction." *Environmental Science and Pollution Research*, 29(25), 38066–38082. <https://doi.org/10.1007/s11356-022-18757-3>

40. Du, K., Zhao, Y., and Lei, J. (2017). "The incorrect usage of singular spectral analysis and discrete wavelet transform in hybrid models to predict hydrological time series." *Journal of Hydrology*, 552, 44–51. <https://doi.org/10.1016/j.jhydrol.2017.06.019>