Journal of Urban Environmental Management

Journal homepage: https://jumee.kgut.ac.ir
Vol.4, No.2; 2026.pp: 1-25

Research paper

(Received 3 June 2026 Accepted 4 July 2026)

Monitoring Biochemical Oxygen Demand in Surface Layers of Lakes using Machine Learning Model

Parsa Toroghi[1], Mohammad Hossein Niksokhan[2*]

[1] *MSc., Faculty of Environment, University of Tehran, Tehran, Iran*
[2] *Professor, Faculty of Environment, University of Tehran, Tehran, Iran*

Abstract

This study addresses the challenge of estimating Biochemical Oxygen Demand over 5 days (BOD_5) in a remote polar lake, where direct measurement is expensive, time-consuming, and often constrained by limited monitoring data. Using more than four decades of historical water quality records from Lake Oulujärvi, Finland, the research developed predictive models based on observations collected from 34 stations across six sub-basins, with particular attention to stratified periods and depth-resolved measurements. The data were first preprocessed through normalization, detection of missing values, removal of invalid observations, and stratification analysis to improve consistency and model reliability. Three machine learning approaches, including Convolutional Neural Networks (CNN), Random Forest (RF), and XGBoost, were then trained to estimate BOD_5 from routinely monitored water quality variables. The selected predictors included dissolved oxygen (DO), pH, turbidity, electrical conductivity, Secchi depth, water temperature, and sampling time, all of which are readily measurable and physically linked to BOD_5 dynamics. Model performance was optimized using grid search and cross-validation. Among the tested models, the CNN achieved the best results, with validation R^2 values of 0.88–0.89, followed by RF, and XGBoost. Feature importance analysis highlighted DO, pH, turbidity, electrical conductivity, and sampling time as the most influential variables. The findings show that machine learning can provide accurate, cost-effective, and scalable alternatives for BOD_5 estimation in low-accessibility aquatic systems, supporting improved water quality assessment and management in fragile lake ecosystems under environmental change.

Keywords: water quality, convolutional neural network, random forest, polar lake, thermal stratification.

* Corresponding Author: Mohammad Hossein Niksokhan
Email: *niksokhan@ut.ac.ir*
Phone: 09122275086

Doi: 10.48306/juem.2026.584810.1160

نشریه علمی مدیریت محیط زیست شهری

آدرس نشریه: https://jumee.kgut.ac.ir

سال چهارم/شماره دوم/ تابستان ۱۴۰۵/صفحه ۱-۲۵

مقاله پژوهشی

تاریخ دریافت: ۱۴۰۵/۳/۱۳ تاریخ پذیرش: ۱۴۰۵/۴/۱۳

پایش اکسیژن خواهی بیوشیمیایی در لایه های سطحی دریاچه با بهره گیری از یادگیری ماشین

پارسا طرقی[۱]، محمد حسین نیک سخن[۲*]

[۱] کارشناسی ارشد، دانشکده محیط زیست، دانشگاه تهران، تهران، ایران

[۲] استاد، دانشکده محیط زیست، دانشگاه تهران، تهران، ایران

چکیده

با افزایش رشد فناوری و جمعیت، ورود آلاینده‌ها به منابع آبی رو به افزایش است و دریاچه‌های طبیعی در معرض آلودگی و تغذیه‌گرایی، از مهم‌ترین منابع آب کشورها به شمار می‌روند. در این پژوهش، از داده‌های کیفیت آب برای تخمین غلظت اکسیژن‌خواهی بیوشیمیایی پنج‌روزه (BOD_5) در شش زیرحوضه و ۳۴ ایستگاه نظارتی یک دریاچه قطبی با دسترسی سخت و محدود استفاده شد. هدف، کاهش هزینه‌های پایش، افزایش کارایی نظارت بر کیفیت آب و ارزیابی رویدادهای بحرانی بود. برای این منظور، عملکرد سه مدل یادگیری ماشین و یادگیری عمیق شامل جنگل تصادفی(Random Forest) ، XGBoost و شبکه عصبی پیچشی (CNN) با معماری متفاوت ارزیابی شد. متغیرهای ورودی شامل پارامترهای فیزیکی، شیمیایی و زمانی مانند pH، کدورت، اکسیژن محلول، عمق دیسک سکی، رسانایی، دمای آب و زمان نمونه‌برداری بودند. داده‌ها در بازه‌ای ۴۰ ساله از ایستگاه‌های مختلف گردآوری شد و شامل حدود ۶۵۰ مشاهده بودند. با تحلیل ماتریس همبستگی، ضرایب پیرسون بین BOD_5 و متغیرهای ورودی محاسبه شد. نتایج نشان داد هر سه الگوریتم در پیش‌بینی BOD_5 مؤثرند. در مدل جنگل تصادفی، مقدارضریب تعیین (R^2) در اعتبارسنجی بین ۰٫۸۵ تا ۰٫۸۶ بود و طبق تحلیل اهمیت متغیرها اکسیژن محلول، pH، کدورت و زمان نمونه‌برداری مهم‌ترین پارامترهای تخمین بودند. مدل XGBoost نیز R^2 برابر ۰٫۸۳ تا ۰٫۸۴ داشت و متغیرهای مهم شامل اکسیژن محلول، pH، کدورت و تاریخ نمونه‌برداری بودند. مدل CNN بهترین عملکرد را نشان داد و به R^2 برابر ۰٫۸۸ تا ۰٫۸۹ رسید. بر اساس تحلیل مدل برگزیده CNN، رسانایی، اکسیژن محلول، pH، کدورت و ساعت نمونه‌برداری مهم‌ترین پیش‌بینی‌کننده‌های BOD_5 بودند.

کلمات کلیدی: کیفیت آب، شبکه عصبی کانولوشن، جنگل تصادفی، دریاچه قطبی،لایه بندی حرارتی

* نویسنده مسئول: محمد حسین نیک سخن

ایمیل: *niksokha@ut.ac.ir*

شماره تماس: ۰۹۱۲۲۲۷۵۰۸۶

Doi: 10.48306/juem.2026.584810.1160

۱-مقدمه

با توجه به افزایش مقیاس فعالیت‌های انسانی در بخش‌هایی مانند کشاورزی، دامداری و صنایع مختلف که ناشی از رشد سریع جمعیت و پیشرفت فناوری است، دستیابی به زیرساخت‌های بهبود یافته و بهره‌وری بیشتر اجتناب‌ناپذیر شده است. با این حال، چنین پیشرفت‌هایی اغلب منجر به تولید فاضلاب حاوی انواع آلاینده‌ها، به ویژه از رواناب‌های کشاورزی، دامداری و فرآیندهای صنعتی می‌شود. این تخلیه‌ها در نهایت به منابع آبی طبیعی مانند رودخانه‌ها، دریاچه‌ها و دریاها راه پیدا می‌کنند و تهدیدات جدی را برای کمیت و کیفیت منابع آب شیرین مورد استفاده برای آشامیدن و آبیاری ایجاد می‌کنند. در مجموع، این فشارها، نظارت و مدیریت کیفیت آب را به یک چالش پیچیده و فزاینده تبدیل کرده است. این چالش نظارتی به ویژه در مناطق دورافتاده و بکر، مانند دریاچه‌های واقع در مناطق قطبی، که داده‌های کیفیت آب در مقایسه با دریاچه‌های در دسترس‌تر و مورد مطالعه گسترده‌تر در سراسر جهان کمیاب است، برجسته‌تر است. پایش کیفیت آب در دریاچه‌های قطبی یکی از ابزارهای کلیدی برای درک اکوسیستم‌های حساس آنها و تغییرات زیست‌محیطی است که متحمل می‌شوند. این مناطق، به دلیل شرایط سخت آب و هوایی و محدودیت‌های بیولوژیکی، در برابر تغییرات اقلیمی و آلودگی‌های انسانی بسیار آسیب‌پذیر هستند. پایش کیفیت آب شامل اندازه‌گیری پارامترهای فیزیکی، شیمیایی و بیولوژیکی است و این داده‌ها برای ارزیابی سلامت اکوسیستم و پیش‌بینی تغییرات زیست‌محیطی آینده ضروری هستند. یکی از شاخص‌های کلیدی کیفیت آب، اکسیژن خواهی بیوشیمیایی پنج روزه (BOD_5) است که میزان اکسیژن مصرف شده توسط میکروارگانیسم‌ها را در تجزیه مواد آلی اندازه‌گیری می‌کند. در دریاچه‌های قطبی (با فعالیت بیولوژیکی کم)، نوسانات BOD_5 می‌تواند نشان‌دهنده هجوم مواد آلی خارجی یا تغییرات در اکوسیستم باشد. افزایش BOD_5 معمولاً با کاهش اکسیژن محلول(DO) مطابقت دارد که ممکن است تهدیدی برای حیات آبزیان باشد. بنابراین، BOD_5 نه تنها به عنوان معیاری از وضعیت فعلی کیفیت آب، بلکه به عنوان یک شاخص اولیه آلودگی و تغییرات اکولوژیکی نیز عمل می‌کند. آلودگی و وضعیت تغذیه گرایی اکوسیستم‌های دریاچه‌ای تنوع وسیع را نشان می‌دهد و بسیاری از این سیستم‌ها عمدتاً به دلیل افزایش فعالیت‌های انسانی دچار زوال قابل توجهی شده‌اند. این امر به ویژه در دریاچه‌های قطبی، جایی که ورود سطوح بالای مواد آلی می‌تواند تعادل اکسیژن را در اکوسیستم مختل کند، اهمیت دارد. نظارت بر BOD_5، در کنار سایر شاخص‌های کیفیت آب، دانشمندان را قادر می‌سازد تا علائم اولیه تغذیه‌گرایی یا دگرگونی زیست محیطی را تشخیص دهند و از این طریق از اجرای استراتژی‌های مدیریتی مؤثر پشتیبانی کنند. همانطور که در بررسی‌های اخیر برجسته شده است، اهمیت BOD_5 به عنوان یک پارامتر اصلی در ارزیابی کیفیت آب در اکوسیستم‌های حساس مانند اکوسیستم‌های قطبی غیرقابل انکار است [۲]. به طور گسترده پذیرفته شده است که تخمین و اندازه‌گیری BOD_5 چالش برانگیزتر از پیش‌بینی بسیاری از پارامترهای رایج دیگر کیفیت آبهاست. در نتیجه، مطالعات متعددی به دنبال کاهش زمان مورد نیاز برای تعیین آن بوده‌اند [۶-۳]. در یکی از اولین پژوهش‌ها اما، اولیویرا-اسکویره و همکاران مدل‌های پیش‌بینی حالت پایدار و پویا را با استفاده از رگرسیون خطی چندگانه (MLR) و حداقل مربعات جزئی، برای امکان تخمین آنلاین پارامتر BOD پیشنهاد کردند [۷]. در زمینه تحقیقات، بسته به منابع آب موجود و اهداف مدیریتی خاص، هر دو رویکرد مدل‌سازی عددی و داده‌محور اغلب برای شبیه‌سازی پارامترهای کیفیت آب (WQP) استفاده می‌شوند. تکنیک‌های عددی مدت‌هاست که به عنوان ابزارهای مؤثر برای شبیه‌سازی WQP، با تأکید ویژه بر مدل‌سازی BOD_5، در نظر گرفته می‌شوند [۸]. شایان ذکر است که رویکردهای داده‌محور عموماً در مقایسه با مدل‌های عددی مرسوم، انعطاف‌پذیری بیشتری نسبت به چالش‌های مختلف ارائه می‌دهند. به عنوان مثال، چاوز و چانگ در پژوهشی از یک شبکه عصبی مصنوعی پیش‌خور (FFANN) برای مدل‌سازی پارامترهای کیفیت آب در مخزن شیهمن، تایوان استفاده کردند. مطالعه ایشان از ۴۷ روز رکورد کیفیت آب جمع‌آوری‌شده از پنج محل نمونه‌برداری در سه لایه عمقی استفاده کرد. متغیرهای آب و هواشناسی به عنوان ورودی‌های مدل و WQP به عنوان خروجی تعیین شدند. نتایج نشان داد که FFANN قابلیت پیش‌بینی قوی برای تخمین شاخص‌های کیفیت آب از خود نشان می‌دهد [۹]. در تحقیقی دیگر ویررو و همکاران با استفاده از الگوریتم جنگل تصادفی برای تخمین غلظت نیتروژن و فسفر در رودخانه های استونی، پتانسیل یادگیری ماشینی را برای پیش‌بینی کیفیت آب در مناطق کم‌داده نشان دادند. با استفاده از ۸۲ متغیر مربوط به کاربری زمین، خاک، هیدرولوژی و هواشناسی، این مدل به دقت بالایی دست یافت و عوامل کاربری زمین و هیدرولوژیکی به عنوان تأثیرگذارترین پیش‌بینی‌کننده‌ها شناسایی شدند. یافته‌های ایشان نشان می‌دهد که چگونه الگوریتم‌های پیشرفته می‌توانند بینش‌های ارزشمندی را از مجموعه داده‌های محدود استخراج کرده و از مدیریت کیفیت آب پشتیبانی کنند، و بهبودهای بیشتری از طریق ادغام با داده‌های ماهواره‌ای و نظارت طولانی‌مدت پیشنهاد شده است [۱۰]. در مجموع بررسی ها نشان

می دهد که اکثر مطالعات اخیر تخمین کیفیت آب نیز بر روی رودخانه‌ها و دریاچه‌های غیرقطبی متمرکز بوده‌اند و از روش های هوش مصنوعی رایج و بعضا سنجش از دور بهره برده اند [۱۶-۱۱]. نجف‌زاده و قائمی نیز برای نمونه از دو روش یادگیری ماشین یعنی حداقل مربعات – ماشین بردار پشتیبان (LS-SVM) و رگرسیون انطباقی چندمتغیره با اسپلاین (MARS) برای پیش‌بینی BOD_5و CODدر رودخانه کارون استفاده کردند. با ۲۰۰ نمونه میدانی و ۹ پارامتر ورودی مانند هدایت الکتریکی، کاتیون‌ها، نیترات، کدورت و pH مدل‌ها را توسعه دادند. عملکرد این روش‌ها با شبکه‌های عصبی، سیستم عصبی-فازی و رگرسیون سنتی مقایسه شد. نتایج نشان داد LS-SVM و MARS دقت بیشتری دارند و بر محدودیت‌های روابط تجربی غلبه می‌کنند. این پژوهش نشان می‌دهد که مدل‌های داده‌محور ابزاری کارآمد برای پایش و مدیریت آلودگی رودخانه‌ها هستند [۱۷].

هریسون و همکاران در بررسی‌های خود، پیش‌بینی مواد مغذی در جریان‌های رودخانه‌ای را با اعمال رگرسیون RF به داده‌های سنسور با فرکانس بالا، با غلظت نیتروژن و فسفر به عنوان متغیرهای هدف، بررسی کردند. این مدل پارامترهایی مانند دما، رسانایی الکتریکی، اکسیژن محلول و کدورت را در بر گرفت و امکان ثبت تغییرات کوتاه‌مدت در کیفیت آب را فراهم کرد. نتایج نشان‌دهنده دقت پیش‌بینی بالا، با مقادیر ضریب تعیین (R^2) حدود ۰/۹۰ بود و تجزیه و تحلیل حساسیت نشان داد که دما و کدورت تأثیرگذارترین پیش‌بینی‌کننده‌ها بودند. این مطالعه از اولین مطالعاتی بود که نظارت با وضوح و رزولوشن بالا را با یادگیری ماشین برای ارزیابی مواد مغذی در رودخانه‌ها ادغام کرد و پیامدهای امیدوارکننده‌ای برای کنترل تغذیه گرایی آب، ارائه داد [۱۸]. اوی و همکاران نیز یک تحقیق میدانی در چین انجام دادند که در آن نمونه‌های آب از دو دریاچه با استفاده از یک مجموعه داده کوچک و خصوصی جمع‌آوری‌شده مورد تجزیه و تحلیل قرار گرفتند. این مطالعه از چندین الگوریتم یادگیری ماشین از جمله RF، SVR و MLP برای پیش‌بینی غلظت BOD_5 بر اساس طیف وسیعی از پارامترهای کیفیت فیزیکی و شیمیایی آب استفاده کرد. مدل‌ها بر اساس ویژگی‌های ورودی آموزش داده شدند و با هدف ایجاد یک رویکرد کارآمدتر و قابل اعتمادتر برای تخمین BOD_5، با داده‌های آزمایش اعتبارسنجی شدند. یافته‌ها نشان داد که روش‌های یادگیری ماشین در مقایسه با تکنیک‌های آزمایشگاهی مرسوم، دقت پیش‌بینی بالاتر و پردازش سریع‌تری دارند [۱۹].

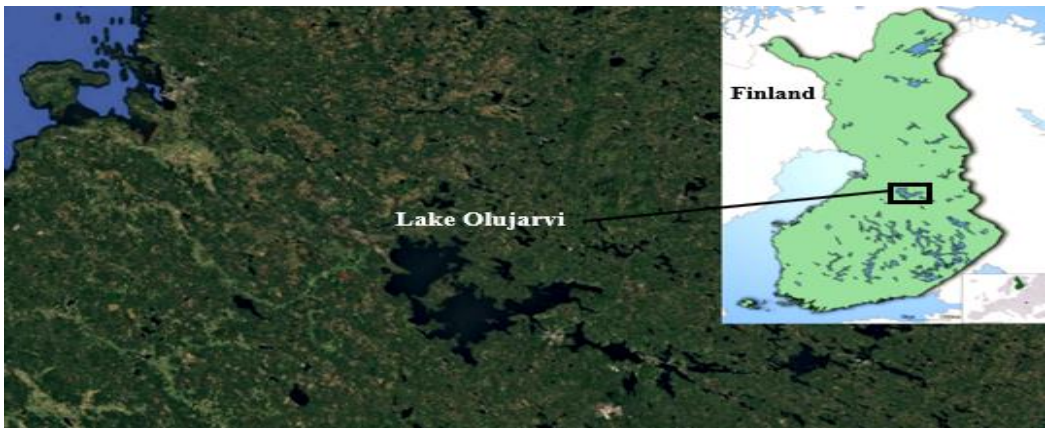
زیاد سامی و همکاران در نتیجه بررسی‌های خود با استفاده از رویکردهای یادگیری ماشین بر روی داده‌های نظارت بلندمدت، مطالعه‌ای در مورد پیش‌بینی DO در مخزن فی-تسوی ارائه دادند. آن ها با استفاده از مجموعه داده‌هایی شامل ۲۹ سال سابقه کیفیت آب، یک ANN بهینه شده از طریق تنظیم نورون‌ها در یک لایه پنهان، در کنار مدل‌های مقایسه‌ای مانند RF و SVR توسعه دادند. متغیرهای ورودی شامل دما، pH، رسانایی الکتریکی، جامدات معلق و سایر پارامترهای فیزیکی وشیمیایی بودند. مدل ANN، که با ارزیابی آماری و تحلیل حساسیت پشتیبانی می‌شد، قابلیت پیش‌بینی قوی را نشان داد و به ضریب تعیین R^2=۰/۹۸ دست یافت، و نتایج نشان داد که تنها چهار متغیر برای تخمین قابل اعتماد DO کافی هستند. با وجود R^2 بالای گزارش شده درمجموعه آزمایش، اتکای مطالعه به مشاهدات ماهانه طی ۲۹ سال (≈۳۴۸ رکورد) و وابستگی مدل نهایی به تنها چند پیش‌بینی‌کننده (به عنوان مثال، دما، اکسیژن خواهی بیوشیمیایی، آهن) توانایی آن را در ثبت دینامیک‌های کوتاه‌مدت محدود می‌کند. علاوه بر این، فقدان اعتبارسنجی مستقل، نمونه آزمایشی نسبتاً کوچک، محدودیت در همگن‌سازی داده‌های تاریخی، خطر بیش‌برازش را افزایش داده و اعتماد به تعمیم‌پذیری مدل به سایر مخازن یا سناریوهای نظارتی عملیاتی را تضعیف می‌کند [۲۰]. هیتالا و همکاران در پژوهشی شیوه‌های مدیریت آب‌های سطحی شهری در جنوب شرقی فنلاند واقع در سواحل دریاچه سایما، یک منبع حیاتی برای تأمین آب آشامیدنی را بررسی کردند. این مطالعه تأکید کرد که فشارهای شهرنشینی و تغییرات اقلیمی تهدیدهای اصلی برای کیفیت آب‌های سطحی در این منطقه هستند. برای رسیدگی به این چالش‌ها، شهرداری یک برنامه مدیریت یکپارچه را اجرا کرد که زیرساخت‌های مرسوم را با راه‌حل‌های مبتنی بر طبیعت ترکیب می‌کند. این رویکرد، اثربخشی ترکیب فناوری‌های پیشرفته با مداخلات زیست‌محیطی را برای افزایش حفاظت از آب‌های سطحی و تضمین مدیریت پایدار اکوسیستم دریاچه برجسته می‌کند [۲۱]. اکهولم و میتیکا نیز، ۲۰ دریاچه تحت تأثیر کشاورزی در فنلاند را بررسی کردند. نتایج آنها نشان داد که این دریاچه‌های کشاورزی مواد مغذی، کلروفیل-آ و سطح کدورت قابل توجهی بالاتری را نشان می‌دهند، که شش مورد به عنوان یوتروفیک و چهارده مورد به عنوان هیپرتروفیک طبقه‌بندی شده‌اند. تجزیه و تحلیل از سال ۱۹۷۶ تا ۲۰۰۲ نشان داد که اکثر دریاچه‌ها یا شرایط ثابت ماندن وضع موجود یا بدتر شدن تغذیه گرایی و کدورت را تجربه کرده‌اند، در حالی که کاهش مواد مغذی تنها در یک دریاچه اصلاح شده ثبت شد و هیچ کاهشی در کلروفیل-آ در کل مجموعه داده‌ها مشاهده نشد. این یافته‌ها تأکید می‌کنند که مواد مغذی کشاورزی همچنان عامل اصلی تخریب کیفیت آب

در فنلاند است [۲۲]. مدل‌های مبتنی بر یادگیری ماشین، انعطاف‌پذیری و سازگاری بالایی را نشان می‌دهند، به‌ویژه هنگامی که بر روی مجموعه داده‌های تاریخی شامل متغیرهای متعدد کیفیت آب و فاضلاب آموزش داده می‌شوند. این مدل‌ها پاسخگویی به تغییرات ناگهانی کیفیت را نیز بهبود می‌بخشند و از تصمیم‌گیری در زمان واقعی در مدیریت منابع آب پشتیبانی می‌کنند. این رویکرد یک راه‌حل جدید و عملی برای پرداختن به چالش‌های مدیریت پایدار کیفیت آب، به‌ویژه در سیستم‌های آبی حساس و دورافتاده، ارائه می‌دهد. برای پرداختن به این چالش‌ها، پیاده‌سازی فناوری‌های پیشرفته نظارت و تعمیق درک ما از فرآیندهای اکولوژیکی درمحیط‌های منحصربفرد بسیار مهم است. [۲۷-۲۳] با توجه به اینکه پارامترهای کیفیت آب در شرایط لایه‌بندی حرارتی معنا و رفتار متفاوتی پیدا می‌کنند، شناسایی و توصیف دقیق این بازه‌های لایه‌بندی، پیش‌نیاز اصلی هر تحلیل معتبر در این حوزه است. در مورد نوآوری و سهم علمی این پژوهش میتوان گفت که بررسی مطالعات پیشین، نشان می‌دهد که در این زمینه همچنان یک خلأ پژوهشی وجود دارد؛ زیرا تعیین دوره‌های وقوع لایه‌بندی حرارتی مستلزم تحلیل دقیق پروفایل‌های عمودی دمای ثبت‌شده در بازه‌های تاریخی است، در حالی‌که بسیاری از پژوهش‌ها این مرحله را به‌صورت نظام‌مند انجام نداده‌اند. افزون بر این، بخش قابل توجهی از تحقیقات پیشین تنها به یک یا دو مدل مرسوم و تکراری محدود شده‌اند و با وجود در دسترس بودن داده‌های گسترده و در عین حال محدودیت‌های ناشی از کمبود داده، هنوز چارچوبی جدید، دقیق و ساختاریافته از جمله مدل‌های مستعد وجدیدی مانند شبکه عصبی کانولوشنی (CNN) برای تحلیل کیفیت آب ارائه نکرده‌اند. از سوی دیگر، ویژگی‌های خاص دریاچه‌های قطبی و دور از دسترس با اکوسیستم حساس و الگوی لایه‌بندی منحصربه‌فرد آن‌ها نیز در اغلب مطالعات مورد توجه کافی قرار نگرفته است. بر این اساس، پژوهش حاضر با بهره‌گیری از سه روش قدرتمند یادگیری ماشین، یعنی CNN، RF و XGBoost، و با تمرکز بر دریاچه‌های قطبی و استفاده از یک مجموعه‌داده تاریخی غنی، تلاش می‌کند این شکاف را پوشش دهد. این مدل‌ها بر پایه بیش از ۴۰ سال داده تاریخی کیفیت آب، به‌ویژه در دوره‌هایی که لایه‌بندی حرارتی در دریاچه رخ داده است، ارزیابی شده‌اند و در نهایت می‌توانند به‌عنوان جایگزینی کارآمد برای پایش BOD_5 در چنین سیستم های آبی به‌کار گرفته شوند.

۲-مواد و روش‌ها

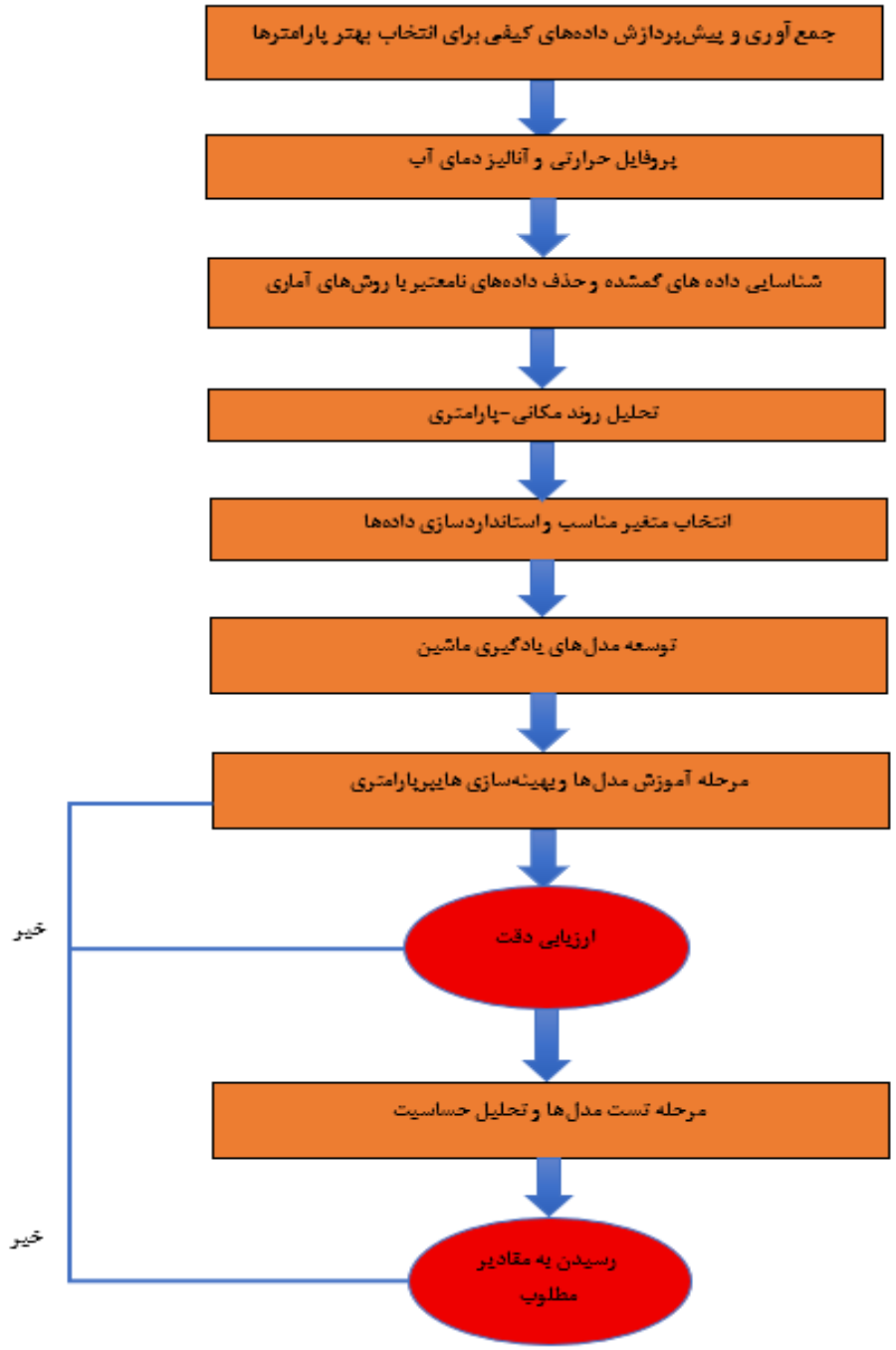
۲-۱-معرفی منطقه مورد مطالعه

دریاچه اولوجاروی، واقع در مرکز فنلاند با مساحت حوضه آبریز تقریباً ۲۲۹۲۵ کیلومتر مربع و وسعت سطحی بین ۷۷۸ کیلومتر مربع تا ۹۴۴ کیلومتر مربع، چهارمین دریاچه بزرگ این کشور شناخته می‌شود. تأثیرات انسانی مانند ساخت سد، توسعه نیروگاه‌های برق آبی و گسترش کشاورزی رژیم هیدرولوژیکی و تعادل اکولوژیکی آن را به طور قابل توجهی تغییر داده است [۲۸]. سطح آب تنظیم‌شده، پویایی جریان های طبیعی و زیستگاهی را مختل کرده است. این دریاچه معمولاً یک دوره بدون یخ حدود ۱۸۰ روز در سال را تجربه می‌کند. نظارت مداوم بر کیفیت آب از دهه ۱۹۶۰ در ۶ زیرحوضه و ۳۴ ایستگاه نمونه‌برداری انجام شده است که داده‌ها توسط موسسه محیط زیست فنلاند (SYKE) ارائه شده است.



شکل ۱- موقعیت دریاچه اولوجاروی و زیرحوضه‌های آن

طرح کلی روش تحقیق به شرح زیر است:



شکل ۲- فلوچارت کلی مطالعه

همانطور که در شکل ۲ نشان داده شده است، چارچوب کلی روش‌شناسی با جمع‌آوری داده‌های کیفیت آب آغاز می‌شود و پس از آن یک مرحله پیش‌پردازش شامل نرمال‌سازی داده‌ها برای افزایش پایداری مدل انجام می‌شود. سپس آنالیز لایه بندی آب انجام می شود. شناسایی داده های گمشده و حذف داده‌های نامعتبر با روش‌های آماری و تحلیل روند مکانی-پارامتری مراحل بعدی هستند. پس از این قسمت به انتخاب متغیر مناسب و استانداردسازی داده‌ها برای تخمین هر چه بهتر BOD_5 پرداخته شد. متعاقباً، چندین

مدل یادگیری ماشین با استفاده از الگوریتم‌های مختلف توسعه داده شده و از طریق تنظیم سیستماتیک هایپرپارامتر بهینه شده‌اند. در نهایت، عملکرد پیش‌بینی مدل پیشنهادی با مدل‌های موجود گزارش شده در مقالات مقایسه شد تا بر بهبودها و سهم روش‌شناسی این مطالعه تأکید شود. این فرآیند گامی مهم در استفاده از داده‌های تاریخی و تکنیک‌های یادگیری ماشین برای افزایش پیش‌بینی و مدیریت کیفیت آب در دریاچه‌های با سطح دسترسی پایین است. در مورد جمع آوری، پیش پردازش و شناسایی داده های گمشده ذکر این نکته ضروری است که با توجه به تغییرات زمانی، ممکن است چالش‌هایی مانند داده‌های پرت و بخش‌های گمشده داده‌ها ایجاد شود. بنابراین، پیش‌پردازش و تصحیح داده‌ها قبل از تجزیه و تحلیل ضروری بود. می‌توان از استراتژی‌های مختلفی مانند استفاده از کتابخانه پایتون برای جایگذاری ورودی‌های گمشده با شناسایی داده های مشابه و تخمین مقادیر غایب بر اساس متغیرهای مرتبط استفاده کرد. از طرف دیگر، می‌توان مقادیر گمشده را با استفاده از مشاهدات از نمونه‌های تقریبی زمانی یا مکانی جایگزین کرد و به طور موثر عدم قطعیت در مدل را کاهش داد. توصیه می‌شود این روش‌های جایگذاری زمانی اعمال شوند که داده‌های گمشده ۱ تا ۳ درصد از کل مجموعه داده‌ها را تشکیل دهند. فراتر از این آستانه، دقت مدل کاهش می‌یابد و خطای تخمین ممکن است افزایش یابد. توابع فعال‌سازی و انتقال مختلف به طور سیستماتیک بررسی و بهینه می‌شوند تا کارایی یادگیری مدل، پایداری همگرایی و عملکرد پیش‌بینی در شرایط مختلف کیفیت آب افزایش یابد. مدیریت داخلی مقادیر گمشده، آموزش روی مجموعه داده‌های محیطی عملی که اغلب شامل شکاف‌ها هستند را ساده می‌کند [۲۹].

توالی زیر، مدیریت بهینه داده‌های گمشده در مدل‌سازی آماری را تضمین می‌کند: ابتدا شناسایی ستون‌های حاوی مقادیر گمشده در مجموعه داده‌ها و سپس جایگذاری ورودی‌های گمشده با استفاده از جایگذاری Scikit-learn، تخمین مقادیر مناسب برای مشاهدات مفقود مهم است. و درنهایت ادغام متغیرهای جایگزین شده در مدل نهایی و بهینه‌سازی بعدی عملکرد پیش‌بینی انجام می‌گردد. پس از جایگذاری، مجموعه داده‌های کیفیت آب مرتب شده از اکسل به فرمت CSV [1]صادر می‌شود تا پیش نیازهای ذخیره‌سازی به حداقل برسد و بارگذاری داده‌ها در طول اجرای مدل تسریع شود. به طور مشابه، چندین الگوریتم آموزشی به طور سیستماتیک ارزیابی خواهند شد تا موثرترین رویکرد برای دستیابی به دقت پیش‌بینی بالا شناسایی شود. در ادامه تحلیل روند مکانی-پارامتری بررسی می شود. توسعه مدل پیش‌بینی برای تخمین BOD_5 بر اساس سایر پارامترهای قابل اندازه‌گیری آب سطحی، در درجه اول به دلیل هزینه و زمان بالای مورد نیاز برای تجزیه و تحلیل مستقیم BOD_5 است. در مقابل، پارامترهایی مانند دمای آب، اکسیژن محلول، کدورت، رسانایی الکتریکی، pH، عمق سکی و زمان نمونه‌برداری، به راحتی و به طور مکرر پایش می‌شوند. بنابراین، این متغیرها به دلیل در دسترس بودن و ارتباط قوی فیزیکی وشیمیایی آنها با دینامیک BOD_5 به عنوان پیش‌بینی‌کننده‌های ورودی انتخاب شدند. به طور خاص، سطوح بالاتر BOD_5 باعث تسریع در کاهش اکسیژن محلول در سیستم‌های آبی می‌شود و وابستگی متقابل بین این شاخص‌ها را برجسته کرده و از مناسب بودن آنها برای توسعه مدل پشتیبانی می‌کند.

دادز و همکاران به این نکته اشاره کردند که دما تأثیر مستقیمی بر فعالیت میکروبی دارد. با افزایش دما، متابولیسم میکروبی و تجزیه مواد آلی نیز افزایش می‌یابد و منجر به افزایش سطح BOD_5 می‌شود. در دریاچه‌های قطبی، دمای پایین فعالیت میکروبی را کاهش می‌دهد و بر این اساس بر BOD_5 تأثیر می‌گذارد. همچنین مشاهده کردند که تغییرات دمای فصلی مستقیماً بر BOD_5 تأثیر می‌گذارد، به طوری که مقادیر بالاتر در تابستان به دلیل افزایش فعالیت میکروبی و مقادیر پایین‌تر در زمستان است [۳۰]. وتزل دما را به عنوان یک عامل اصلی مؤثر بر میزان متابولیسم میکروبی در سیستم‌های آبی تأکید کرد، جایی که دمای بالا فعالیت میکروبی و مصرف اکسیژن را افزایش می‌دهد و در نتیجه BOD_5 را افزایش می‌دهد. برعکس، دمای پایین در مناطق قطبی این فرآیندها را کند می‌کند [۳۱]. پروس و همکاران تأیید کردند که دما عامل اصلی تجزیه مواد آلی در دریاچه‌های قطبی است، جایی که دمای پایین به طور قابل توجهی زمان تجزیه را طولانی‌تر می‌کند [۳۲]. روهلند و همکاران نیز تأیید کردند که حتی افزایش جزئی در دمای آب سطحی می‌تواند منجر به تغییرات قابل توجهی در سطح BOD_5 در دریاچه‌های قطبی شود. در مورد لایه‌بندی حرارتی نیز، در فصول گرم‌تر، دریاچه‌ها اغلب تحت لایه‌بندی حرارتی قرار می‌گیرند که تبادل اکسیژن بین لایه‌های سطحی و عمیق را محدود می‌کند. این می‌تواند منجر به کاهش اکسیژن در آب‌های عمیق‌تر و افزایش BOD_5 شود [۳۳]. پروس و همکاران در مورد چگونگی تشدید اثرات لایه‌بندی ناشی از آب و هوا در دریاچه‌های قطبی بر روی اثرات مرتبط با BOD_5 بحث کردند [۳۲]. گورهام و بویس نشان دادند که لایه‌بندی حرارتی در طول تابستان به دلیل محدودیت در دوباره پر کردن اکسیژن در لایه های آب، BOD_5 را در لایه‌های عمیق‌تر

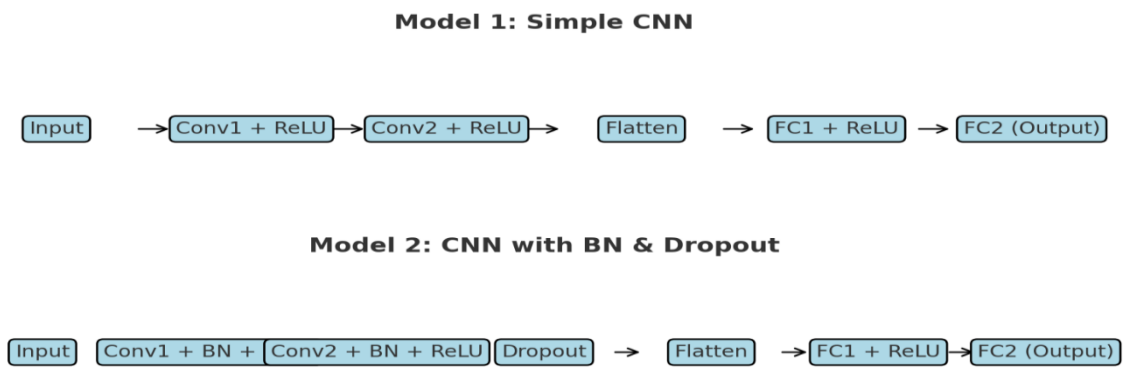
[1] Comma-Seperated Values

افزایش می‌دهد و منجر به تشدید مصرف اکسیژن می‌شود [۳۴]. به طور مشابه، pH بر تجزیه میکروبی مواد آلی تأثیر می‌گذارد. میکروارگانیسم‌ها در pH حدود ۶,۵ تا ۸,۵ بیشترین فعالیت را دارند. در خارج از این محدوده، فعالیت آنها کاهش می‌یابد و به طور بالقوه BOD_5 را کاهش می‌دهد. وتزل گزارش داد که شرایط اسیدی یا قلیایی شدید، تجزیه مواد آلی را به تأخیر می‌اندازد و در نتیجه فرآیند BOD_5 را کند می‌کند [۳۱]. کدورت نشان‌دهنده وجود ذرات معلق و مواد آلی قابل تجزیه است که به عنوان منبع غذایی برای میکروب‌ها عمل می‌کنند و در نتیجه BOD_5 را افزایش می‌دهند. کدورت بالاتر همچنین نفوذ نور را کاهش می‌دهد، فتوسنتز جلبک‌ها را محدود می‌کند و DO را کاهش می‌دهد که تقاضای اکسیژن را تشدید می‌کند. کالف در کتاب خود خاطرنشان کرد که جریان رسوب ناشی از فرسایش و کشاورزی در دریاچه‌های قطبی، کدورت و متعاقباً BOD_5 را افزایش می‌دهد. علاوه بر این، رسانایی نشان دهنده غلظت یون‌های محلول (مانند نیترات، فسفات، کلرید) است. رسانایی بالا اغلب با افزایش سطح مواد مغذی و مواد آلی مرتبط است که به دلیل افزایش تنزل میکروبی، BOD_5 را افزایش می‌دهد. مطالعات در دریاچه‌های قطبی، رسانایی بالا را به ورودی‌های انسانی (کشاورزی، صنعت، رواناب شهری) مرتبط کرده‌اند که منجر به تغذیه گرایی و BOD_5 بالاتر می‌شود [۳۵]. عمق سکی شفافیت آب را اندازه‌گیری می‌کند و با غلظت ذرات معلق رابطه معکوس دارد. عمق سکی پایین‌تر نشان دهنده کدورت و بار آلی بالاتر است که BOD_5 را افزایش می‌دهد. کاهش نفوذ نور، تولید اکسیژن فتوسنتزی را محدود می‌کند، DO را کاهش می‌دهد و سطح BOD_5 را تشدید می‌کند. دادز و همکاران مشاهده کردند که دریاچه‌هایی با عمق سکی کم، به دلیل فراوانی مواد معلق و آلی، عموماً BOD_5 بالاتری را نشان می‌دهند [۳۰]. بنابراین، پارامترهای انتخاب شده مستقیماً با فرآیندهای تقاضای اکسیژن مانند BOD_5 مرتبط هستند. این متغیرها به عنوان نماینده‌های مؤثر برای BOD_5 در مدل عمل می‌کنند. مطالعات متعددی نیز همبستگی قوی بین BOD/COD و غلظت مواد مغذی در اکوسیستم‌های آبی را گزارش کرده‌اند. مجیدزاده و همکاران (۲۰۱۷) نتیجه گرفتند که هر دو فرآیند BOD و COD نقش مهمی در انتقال مواد مغذی در محیط‌های آبی دارند [۳۶]. در نهایت، این پارامترها به عنوان نماینده‌هایی مؤثر برای BOD_5 در هر سه مدل عمل می‌کنند.

۲-۲-توسعه مدل ها

سه نوع مدل CNN و Random Forest و مدل XGBoost برای پیش‌بینی غلظت BOD_5 در دریاچه اولوجاروی پیشنهاد شد. در مدل‌سازی کیفیت آب، یادگیری ماشین (ML) با هدف توسعه سیستم‌هایی که قادر به یادگیری خودکار از داده‌های ورودی بدون برنامه‌نویسی صریح هستند، طراحی شده است. به جای کدگذاری دستی همه قوانین، داده‌ها به الگوریتم‌های عمومی وارد می‌شوند که داده‌ها را برای بهینه‌سازی مدل‌های پیش‌بینی تجزیه و تحلیل می‌کنند [۳۷]. با ظهور کلان‌داده، یادگیری ماشین به یک عامل کلیدی برای خودکارسازی فرآیندها، شناسایی الگوها و ایجاد بینش در حوزه‌های مختلف تبدیل شده است [۳۸]. از نظر تاریخی، تکنیک‌های داده‌کاوی با محدودیت‌های محاسباتی محدود می‌شدند. با این حال، با دسترسی به داده‌های کافی و ظهور الگوریتم‌های یادگیری عمیق، یادگیری ماشین به پیشرفت‌های چشمگیری در زمینه‌های مختلف دست یافته است [۳۹]. تکنیک‌های یادگیری ماشین به ماشین‌ها اجازه می‌دهند تا عملکرد خود را از طریق یادگیری به جای برنامه‌نویسی مبتنی بر قانون بهبود بخشند. معماری مدل پیشنهادی را می‌توان در شکل ۳ مشاهده کرد. مدل پیشنهادی شامل یک لایه ورودی است که پارامترهای ورودی مورد استفاده برای توسعه مدل را ارائه می‌دهد. در مقابل، لایه خروجی، خروجی مدل را که غلظت BOD_5 است، ارائه می‌دهد. وزن‌ها و بایاس‌ها، لایه‌های ورودی و خروجی را به لایه پنهان متصل می‌کنند. لایه پنهان از چندین نورون تشکیل شده است. در این مطالعه، معماری خاص و قدرتمندی از CNN یک بعدی با تفاوت‌های ساختاری متمایز نسبت به مدل پیش فرض و ساده بررسی خواهد شد. مدل های ساده دارای طراحی شبکه ابتدایی تری است که عمدتاً از لایه‌های کانولوشن و توابع فعال‌سازی تشکیل شده است. این ساختار امکان آموزش سریع‌تر را فراهم می‌کرد و زمانی که مجموعه داده‌ها بزرگ یا حاوی الگوهای واضح تعریف‌شده بود، به طور مؤثر کار می‌کند. با این حال، به دلیل عدم وجود مکانیسم‌های نرمال‌سازی و منظم‌سازی، این مدل های ابتدایی بیشتر مستعد بیش‌برازش داده‌های آموزشی می‌تواند باشد و در نتیجه توانایی تعمیم آن را هنگام اعمال به مجموعه داده‌های دیده نشده کاهش می‌دهد. معماری کلی مبتنی بر ارتباط متوالی لایه‌های کانولوشن و به دنبال آن لایه‌های کاملاً متصل بود که با هدف استخراج ویژگی‌ها و تولید تخمین و پیش‌بینی‌ها انجام می‌شد. از سوی دیگر، مدل CNN به کار رفته در این تحقیق مکانیسم‌های اضافی را برای افزایش استحکام و کاهش بیش‌برازش در خود جای داده است. در این معماری، لایه‌های نرمال‌سازی دسته‌ای (Batch Normalization) برای تثبیت فرآیند یادگیری

معرفی شدند، در حالی که لایه‌های حذف برای حذف تصادفی(dropout) بخشی از نورون‌ها در طول آموزش اعمال شدند. این تنظیمات نه تنها انعطاف‌پذیری مدل را در برابر تغییرپذیری داده‌ها و نویز بهبود بخشید، بلکه پایداری پیش‌بینی آن را در داده‌های دیده نشده نیز افزایش داد. اگرچه افزودن این پارامترها عموماً در مقایسه با مدل ساده‌تر به زمان آموزش طولانی‌تری نیاز دارد، اما نتایج سازگارتر و قابل اعتمادتری به همراه دارد. پس اگرچه هر دو مدل از چارچوب CNN های تک‌بعدی پیروی می‌کنند، اما چیدمان و ترکیب لایه‌ها منجر به ویژگی‌های عملکردی و تعمیم دهی متمایز و متفاوتی شد [۴۰]. شکل ۳ مقایسه شماتیکی از دو معماری را ارائه می‌دهد: اولی ساختار CNN ساده و پیش فرض را نشان می‌دهد و دومی طراحی توسعه‌یافته با نرمال‌سازی دسته‌ای و لایه‌های dropout را نشان می‌دهد که در این تحقیق از این مدل استفاده شده است که پایداری بالا را ارائه داد و از بیش برازش جلوگیری می‌کند.



شکل۳- تفاوت بین مدل CNN ساده و مدل CNN مورد استفاده در پژوهش

مدل جنگل تصادفی یکی از الگوریتم‌های قدرتمند یادگیری ماشین مبتنی بر روش‌های یادگیری جمعی یا گروهی (ensemble) است. این مدل از ترکیبی از درخت‌های تصمیم و مفهوم یادگیری گروهی برای ارائه نتایج دقیق‌تر و قابل اعتمادتر در مقایسه با مدل‌های منفرد مانند یک درخت تصمیم واحد، استفاده می‌کند. جنگل تصادفی به طور گسترده در هر دو چالش رگرسیون و طبقه‌بندی استفاده می‌شود. ساختار و عملکرد مدل جنگل تصادفی شامل مجموعه‌ای از درخت‌های تصمیم است که به صورت تصادفی آموزش داده می‌شوند. هر درخت با استفاده از یک زیرمجموعه تصادفی از داده‌های آموزشی و ویژگی‌ها ساخته می‌شود. این فرآیند که به عنوان تجمیع بوت‌استرپ (Bagging) شناخته می‌شود، به کاهش واریانس مدل کمک می‌کند. Bagging روشی مبتنی بر نمونه‌گیری مکرر با جایگزینی[1] از مجموعه داده‌های اصلی است و یکی از تکنیک‌های قدرتمند در یادگیری ماشین برای بهبود دقت مدل و کاهش واریانس است. این بدان معناست که چندین زیرمجموعه تصادفی با جایگزینی از داده‌های آموزشی تولید می‌شوند. هر زیرمجموعه ممکن است شامل نمونه‌های تکراری از داده‌های اصلی باشد. سپس، یک مدل جداگانه (مانند درخت تصمیم) برای هر زیرمجموعه ساخته می‌شود. پیش‌بینی نهایی با ترکیب خروجی‌های این مدل‌های منفرد به دست می‌آید. برای مسائل طبقه‌بندی، پیش‌بینی نهایی با استفاده از رأی‌گیری اکثریت انجام می‌شود، در حالی که برای مسائل رگرسیون، میانگین خروجی‌های مدل‌ها (درخت‌ها) محاسبه می‌شود. در این رویکرد، هر درخت به طور مستقل از یک بردار تصادفی نمونه‌برداری شده و طبق توزیع یکنواخت در تمام درختان جنگل ساخته می‌شود. با افزایش تعداد درختان، خطای کلی جنگل تصادفی تمایل به همگرایی به یک مقدار تقریباً ثابت دارد. برای ساخت هر درخت D_i، یک مجموعه داده S با استفاده از نمونه‌گیری بوت‌استرپ انتخاب می‌شود. در هر گره از درخت، به جای در نظر گرفتن همه ویژگی‌ها، فقط یک زیرمجموعه تصادفی از ویژگی‌ها انتخاب می‌شود. سپس بهترین تقسیم درخت بر اساس این ویژگی‌های انتخاب شده تعیین می‌شود. این رویکرد همبستگی بین درختان را کاهش می‌دهد و در نتیجه، دقت نهایی مدل را افزایش می‌دهد [۴۱].

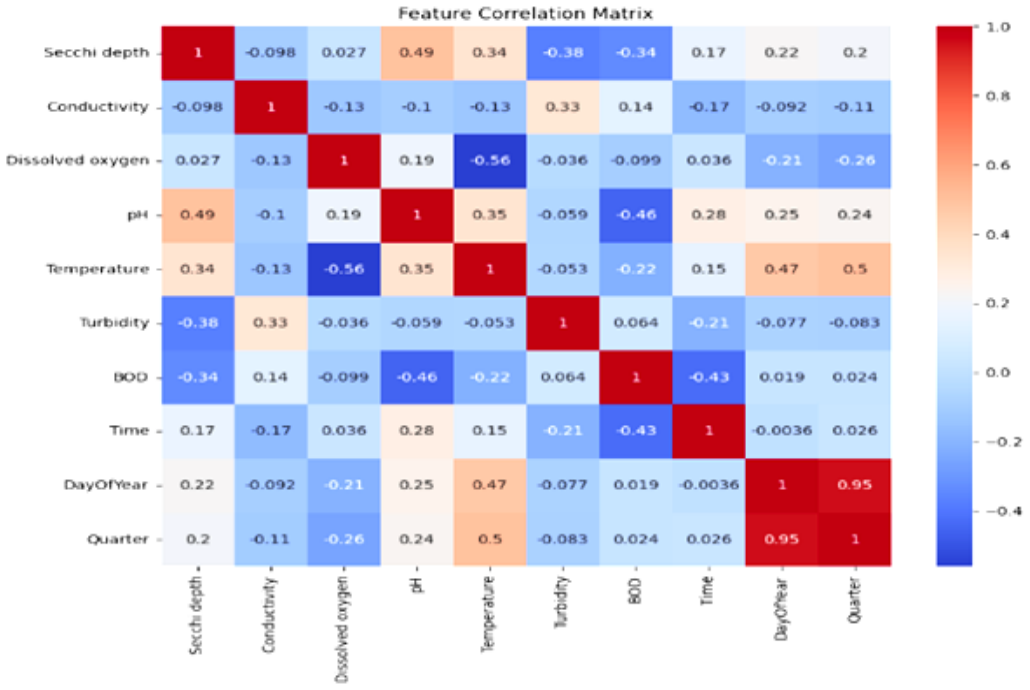
تقویت گرادیان شدید یا Extreme Gradient Boosting (XGBoost) یک الگوریتم گروهی با کارایی بالا است که تقویت گرادیان را روی درخت‌های تصمیم‌گیری پیاده‌سازی می‌کند. XGBoost به جای ساخت درخت‌های مستقل به صورت موازی، درخت‌ها

[1] Bootstrap Sampling

را به صورت متوالی اضافه می‌کند: هر درخت جدید برای کاهش خطاهای باقیمانده گروه ساخته شده آموزش داده می‌شود. XGBoost با ترکیب به‌روزرسانی‌های مبتنی بر گرادیان با روش‌های اکتشافی ساخت درخت و منظم‌سازی صریح، عملکرد پیش‌بینی قوی ارائه می‌دهد. این مدل با به حداقل رساندن یک تابع هدف منظم که یک عبارت ضرر و زیان را با جریمه‌های پیچیدگی صریح (مثلاً L۲/L۱) ترکیب می‌کند، درخت‌های متوالی می‌سازد. در هر تکرار، تقسیم‌های کاندید با استفاده از مشتقات مرتبه اول و دوم زیان (گرادیان و هسین) ارزیابی می‌شوند، که به الگوریتم اجازه می‌دهد به‌روزرسانی‌های آگاهانه‌تر و آگاه‌تر از انحنای درخت نسبت به نزول گرادیان ساده انجام دهد. XGBoost همچنین از نمونه‌برداری فرعی ردیف و ویژگی پشتیبانی می‌کند و می‌تواند به طور تصادفی زیرمجموعه‌هایی از داده‌ها یا متغیرها را برای هر درخت انتخاب کند، در نتیجه همبستگی بین درخت‌ها را کاهش داده و تعمیم را بهبود می‌بخشد [۲۹ و ۴۲]. برای اطمینان از تعمیم پذیری مدل و کاهش احتمال بیش‌برازش، از اعتبارسنجی متقابل K-fold با طرح‌های تقسیم‌بندی چندگانه استفاده خواهد شد. تمام فرآیندهای مدل‌سازی و محاسباتی با استفاده از زبان برنامه‌نویسی پایتون پیاده‌سازی می‌شوند که کتابخانه‌های قوی و انعطاف‌پذیری را برای توسعه یادگیری ماشین و بهینه‌سازی عملکرد فراهم می‌کند. تنظیم صحیح هایپرپارامترهای مدل برای شناسایی پیکربندی بهینه برای مدل‌های یادگیری ماشین ضروری است، زیرا هیچ تنظیم پیش‌فرض جهانی برای همه انواع مسائل وجود ندارد [۴۳]. پارامترهای مدل به مقادیر داخلی آموزش داده شده اشاره دارند و نقش محوری در آموزش و پیش‌بینی ایفا می‌کنند. در مقابل، هایپر پارامترهای مدل خارجی هستند و نمی‌توان آنها را مستقیماً از داده‌های ورودی استنباط کرد. آنها بر رفتار آموزشی مدل تأثیر می‌گذارند و باید با استفاده از روش‌هایی مانند جستجوی شبکه‌ای (Grid Search) یا جستجوی تصادفی (Random Search) به صورت دستی تنظیم یا بهینه شوند [۴۴]. این روش‌ها برای افزایش کارایی مدل‌ها استفاده شده‌اند و هایپرپارامترهای مدل گزارش خواهند شد. هدف اصلی این فرآیند تنظیم، به حداقل رساندن میانگین مربعات خطا (MSE) و به حداکثر رساندن ضریب تعیین (R^2) است که در نتیجه دقت پیش‌بینی و استحکام مدل را افزایش می‌دهد.

انتخاب مدل‌های CNN، RF و XGBoost در این پژوهش و به طور کلی پژوهش‌های کیفی آب بر پایه ماهیت پیچیده و غیرخطی روابط میان متغیرهای کیفی آب و شاخص BOD_5 انجام شده است. از آنجا که تغییرات BOD_5 تحت تأثیر برهم‌کنش هم‌زمان چندین پارامتر فیزیکی، شیمیایی و زیستی قرار دارد، استفاده از مدل‌هایی که توانایی استخراج الگوهای پنهان و روابط چندبعدی را داشته باشند، ضروری است.RF به دلیل پایداری مناسب در برابر نویز، توانایی کار با داده‌های پرابعاد و ارائه اهمیت نسبی متغیرها، گزینه‌ای قابل اعتماد برای مسائل محیطی محسوب می‌شود. XGBoost نیز با اتکا بر ساختار تقویتی خود، در شناسایی الگوهای پیچیده و افزایش دقت پیش‌بینی عملکرد مطلوبی دارد و در داده‌های ناهمگون معمولاً نتایج رقابتی ارائه می‌کند. در مقابل، CNNکه روشی جدیدتر هم هست با قابلیت یادگیری خودکار ویژگی‌ها، امکان استخراج وابستگی‌های نهفته در داده‌های تاریخی و ساختارمند را فراهم می‌سازد. بنابراین، به‌کارگیری این سه مدل در کنار یکدیگر، هم امکان مقایسه رویکردهای کلاسیک و پیشرفته یادگیری ماشین را فراهم می‌کند و هم چارچوبی مناسب برای تخمین دقیق BOD_5 در دریاچه‌های قطبی ارائه می‌دهد.

همچنین ماتریس همبستگی ویژگی(Feature Correlation Matrix)، که برای ارزیابی روابط خطی بین متغیرهای ورودی استفاده می‌شود، در مدل‌سازی مبتنی بر داده ضروری است. با آشکار کردن ویژگی‌های خطی یا بسیار مرتبط، به شناسایی متغیرهایی که ممکن است دقت و تفسیرپذیری مدل را به خطر بیندازند، کمک می‌کند. این امر امکان حذف یا تجمیع ورودی‌های اضافی را فراهم می‌کند و در نتیجه ساختار مدل کارآمدتر و قابل اعتمادتری ایجاد می‌کند. شکل ۴ ارزیابی آماری و مقادیر ماتریس همبستگی بین پارامترهای کیفیت آب انتخاب شده و غلظت BOD_5 را نشان می‌دهد که ارتباط مستقیم یا معکوس قوی ای برای این پارامتر به صورت پیش فرض وجود ندارد و بر اهمیت تحقیق می‌افزاید.



شکل ۴- ماتریس همبستگی ویژگی‌ها

برای ارزیابی عملکرد پیش‌بینی مدل‌های توسعه‌یافته، از چهار شاخص آماری - میانگین خطای مطلق (MAE)، میانگین مربعات خطا (MSE)، ضریب همبستگی (r) و ضریب تعیین یعنی R^2 استفاده شد. فرمول‌بندی‌ها و تفاسیر ریاضی این شاخص‌ها از آنچه که توسط نجاح و همکاران شرح داده شده است، پیروی می‌کنند [۴۵]. علاوه بر این، تحلیل‌های حساسیت و عدم قطعیت برای اطمینان از استحکام و قابلیت اطمینان مدل پیشنهادی انجام شد.

۳-نتایج و بحث‌ها

این فصل به جزئیات گام های تحقیق و روش‌های مدل‌سازی برای هر یک از مدل‌های اعمال‌شده می‌پردازد و خروجی‌های حاصل از آن‌ها را ارائه می‌دهد. این داده‌ها - که عمدتاً ماهانه در طول یک دوره ۴۰ ساله جمع‌آوری شده‌اند - در اعماق و مکان‌های مختلف نمونه‌برداری شده و سپس در یک پایگاه داده اکسل سازماندهی شده‌اند. تجزیه و تحلیل سری زمانی این سوابق نشان می‌دهد که نمونه‌برداری عمدتاً در فواصل ماهانه رخ داده است.

۳-۱-دوره‌های لایه‌بندی حرارتی در دریاچه اولوجاروی

به دلیل عرض جغرافیایی شمالی و عمق متوسط نسبتاً کم (تقریباً ۷/۶ متر)، دریاچه اولوجاروی، مانند بسیاری از دریاچه‌های فنلاند، در طول فصل گرم، لایه‌بندی حرارتی شدیدی را تجربه می‌کند. پس از دوره یخ‌زدایی در بهار، تابش خورشیدی لایه بالایی آب را گرم می‌کند و یک ترموکلاین تشکیل می‌دهد. با خنک شدن آب‌های سطحی در پاییز، اختلاط کامل دوباره در سراسر ستون آب رخ می‌دهد. این داده‌ها نشان می‌دهد که اوج گرمایش سطح در ماه ژوئیه رخ می‌دهد، دقیقاً زمانی که لایه‌بندی حرارتی بیشترین شدت را دارد.

- شروع لایه‌بندی: طبق داده‌های ثبت‌شده، ژوئن معمولاً پایان پوشش یخ و آغاز فصل آب‌های آزاد را نشان می‌دهد. داده‌های تاریخی نشان می‌دهند که شکستن یخ می‌تواند در هر زمانی بین اوایل ماه مه و اواخر ژوئن رخ دهد. بنابراین، از اواسط بهار به بعد، آب‌های آزاد در معرض گرمایش خورشیدی قرار می‌گیرند و لایه‌بندی حرارتی آغاز می‌شود.
- لایه‌بندی پایدار تابستانی: در طول تابستان، لایه بالایی آب به طور قابل توجهی گرم می‌شود و یک ترموکلاین پایدار تشکیل می‌دهد. مشاهدات نشان می‌دهد که دمای سطح در ماه ژوئیه تقریباً به ۱۸ درجه سانتیگراد می‌رسد که بسیار بالاتر از دمای لایه‌های عمیق‌تر است. این گرادیان دما تأیید می‌کند که دریاچه از ژوئن تا آگوست، به ویژه در جولای و اوایل آگوست، به شدت لایه‌بندی شده باقی می‌ماند.

• فروپاشی لایه‌بندی: با کاهش دمای هوا در اواخر آگوست و سپتامبر و افزایش تلاطم جوی، گرادیان‌های حرارتی ضعیف می‌شوند و اختلاط کامل ستون آب آغاز می‌شود. در این دریاچه ، همگنی دمای عمودی تا سپتامبر یا اکتبر بازمی‌گردد و با نزدیک شدن دریاچه به انجماد مجدد، دمای سطح کاهش می‌یابد. ملاحظات اقلیمی و داده‌ها نشان می دهند که گرچه پروفیل‌های عمق کامل و داده‌های روزانه برای کل دوره ۴۰ ساله به طور مداوم در دسترس نیستند، اما می‌توان با اطمینان فصل معمول آب‌های آزاد و لایه‌بندی در اولوجاروی را از ژوئن تا سپتامبر تخمین زد. الگو با رفتار کلی لیمنولوژیکی در دریاچه‌های فنلاند همسو است، جایی که لایه‌بندی قوی تابستانی قبل از تشکیل یخ رخ می‌دهد و اختلاط دوبار در سال (اختلاط بهار و پاییز) بر فصول انتقالی غالب است. در نتیجه، حداقل دوره لایه‌بندی از اواخر بهار تا اواخر تابستان ادامه دارد.

۳-۲-نتایج پیش‌پردازش داده‌ها

یک گام حیاتی در داده‌کاوی، اصلاح داده‌ها، به‌ویژه مدیریت مقادیر گمشده است. در این مطالعه، BOD_5 به عنوان متغیر هدف تعریف شده است. پارامترهای اندازه‌گیری شده کیفی آب شامل pH، دمای آب، رسانایی الکتریکی، کدورت، عمق دیسک سکی و اکسیژن محلول و زمان نمونه برداری به عنوان ورودی‌های مدل عمل می‌کنند. جدول ۱ گزیده‌ای از مجموعه داده‌های مورد استفاده در این تحقیق را نشان می‌دهد.

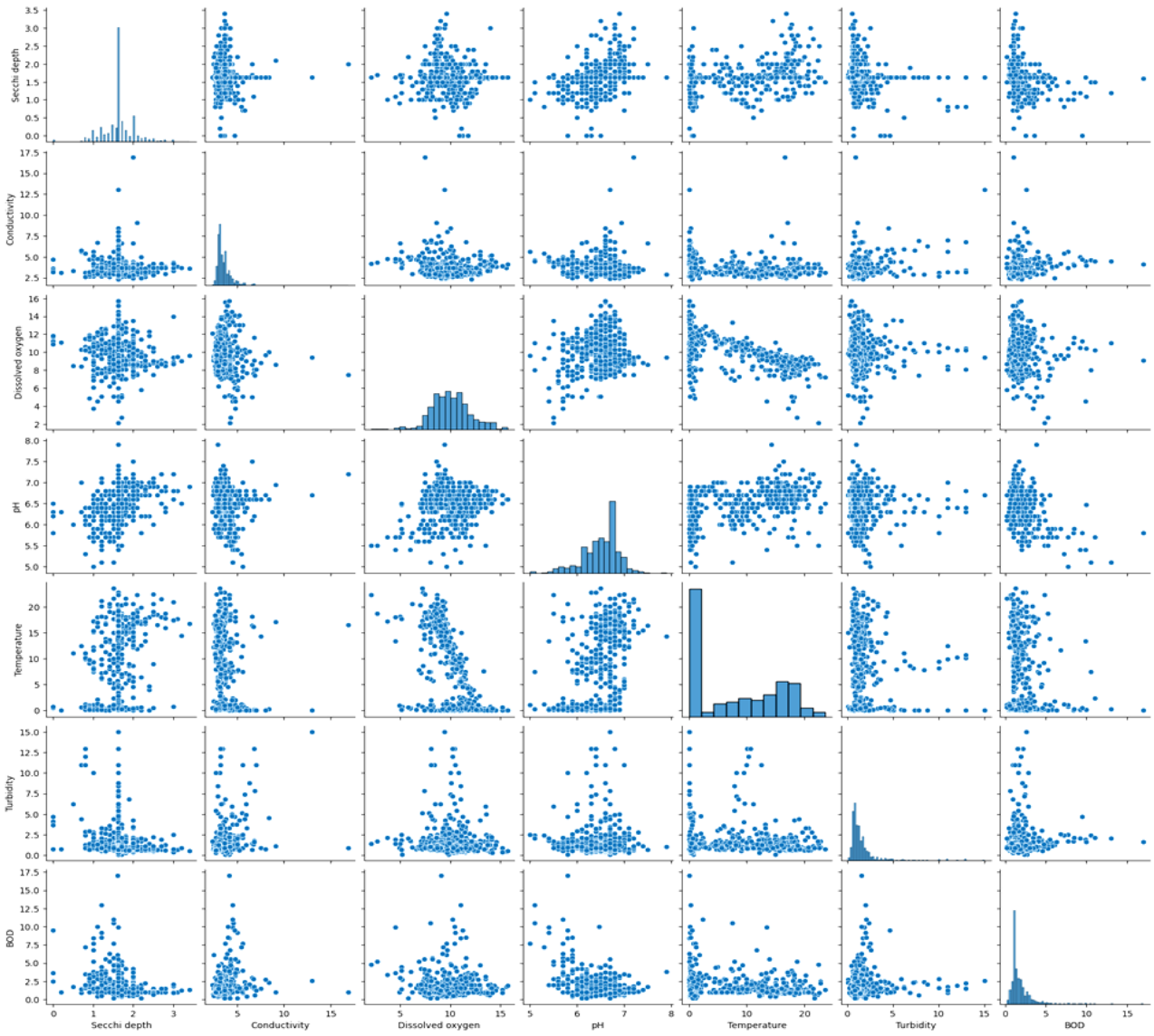
جدول ۱- بخشی از داده‌های خام ورودی به مدل‌ها

زمان نمونه گیری	BOD_5 (mg/l)	کدورت (FNU)	دما (°C)	pH (-)	اکسیژن محلول (mg/l)	رسانایی الکتریکی (mS/m)	عمق سکی (m)	شماره نمونه برداری
۱۹۷۸-۰۷-۳۱ ۰۰:۰۰	۱/۸	۰/۵۷	۲۱/۰	۷/۲	۹/۴	۳/۳	۳/۰	۱
۱۹۷۹-۰۳-۱۵ ۰۰:۰۰	۱/۰	۰/۳۸	۰/۱	۶/۳	۱۱/۵	۴/۰	۲/۰	۲
۱۹۷۹-۰۶-۲۷ ۱۴:۰۰	۱/۳	۰/۵۲	۱۶/۸	۶/۹	۹/۶	۳/۶	۳/۴	۳
۱۹۷۹-۰۷-۲۳ ۰۰:۰۰	۱/۰	۰/۶۵	۱۷/۹	۶/۸	۸/۷	۳/۸	۲/۸	۴
۱۹۷۹-۰۸-۱۵ ۱۱:۰۰	۱/۰	۰/۷۲	۱۷/۴	۷/۲	۸/۹	NaN	۳/۰	۵

در ابتدا، تمام کتابخانه‌های ضروری مانند NumPy، Pandas و Scikit-learn برای پردازش کارآمد داده‌ها وارد می‌شوند. مرحله اول شامل انجام یک تحلیل آماری اولیه از کل مجموعه داده‌ها برای تأیید کامل بودن آن و شناسایی هرگونه ورودی مفقود یا متناقض است. این تحلیل، که در جدول ۲ ارائه شده است، مروری بر توزیع داده‌ها و ویژگی‌های آماری ارائه می‌دهد.

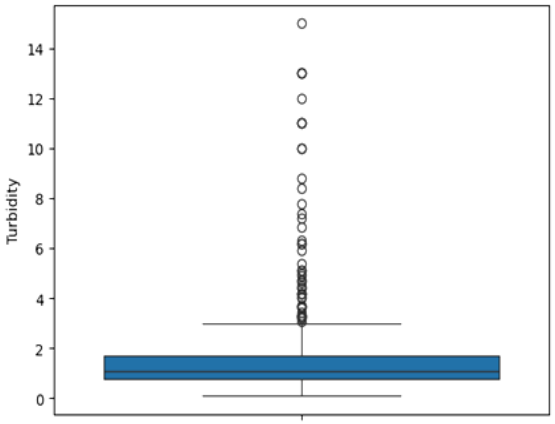
جدول ۲- مقادیر آماری مجموعه داده‌ها

نام پارامتر آماری	عمق سکی (m)	رسانایی الکتریکی (mS/m)	اکسیژن محلول (mg/l)	pH (-)	دما (°C)	کدورت (FNU)	BOD_5 (mg/l)	روز (day in year)
تعداد	۴۲۸	۶۲۶	۶۴۲	۶۳۹	۶۴۴	۳۶۳	۶۴۹	۶۴۹
میانگین	۱/۶۲۲۱	۳/۶۴۹۸	۹/۹۳۵۵	۶/۴۷۴	۷/۵۳۸	۱/۸۱۶۶	۲/۲۸۱۶	۱۶۸/۶۰۸
حداقل	۰	۲/۳	۲/۱	۵	۰	۰/۰۹	۰/۲	۲
۲۵٪	۱/۳	۳/۱	۸/۷	۶/۲۶۵	۰/۳۰	۰/۷۱	۱	۹۰
۵۰٪	۱/۶	۳/۳۵	۱۰	۶/۵	۵/۸	۱/۱	۱/۵	۱۶۵
۷۵٪	۲	۴	۱۱/۱	۶/۷	۱۵	۱/۹	۲/۵	۲۳۴
حداکثر	۳/۴	۱۶/۹	۱۵/۷	۷/۹	۲۳/۶	۱۵	۱۷	۳۶۵
انحراف معیار	۰/۵۱۷۲	۱/۱۰۱۸۸	۲/۰۱۵۷	۰/۳۸۵۳	۷/۴۲۵۷	۲/۱۵۴۲	۲/۲۳۹۲	۸۷/۸۱۴۵۹

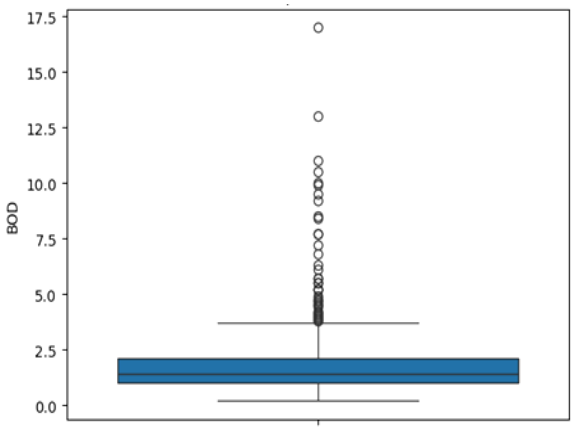


شکل ۵- ماتریس نمودار پراکندگی داده‌ها

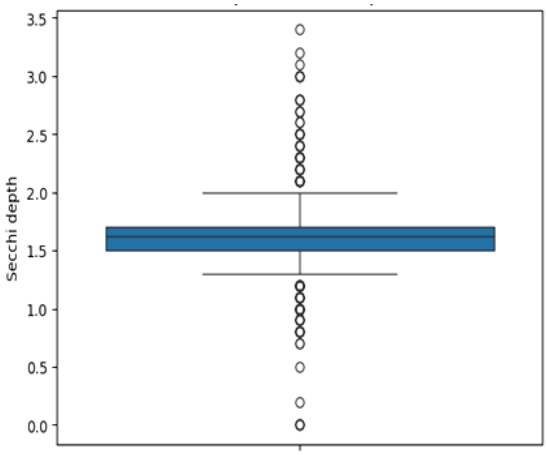
پس از غربالگری اولیه و آماده‌سازی مجموعه داده‌ها، داده‌ها به دو دسته تقسیم می‌شوند: متغیرهای ورودی و متغیر خروجی (هدف پیش‌بینی). در این مرحله، از کتابخانه NumPy برای تبدیل مجموعه داده‌ها به آرایه‌های عددی، شناسایی ردیف‌ها و ستون‌های مورد نیاز و ساختاردهی آنها برای پردازش مدل یادگیری ماشین استفاده می‌شود. این روش، کارایی عملکرد مدل را افزایش می‌دهد و امکان مدیریت داده‌های در مقیاس بزرگ را با اثربخشی محاسباتی بهبود یافته فراهم می‌کند.



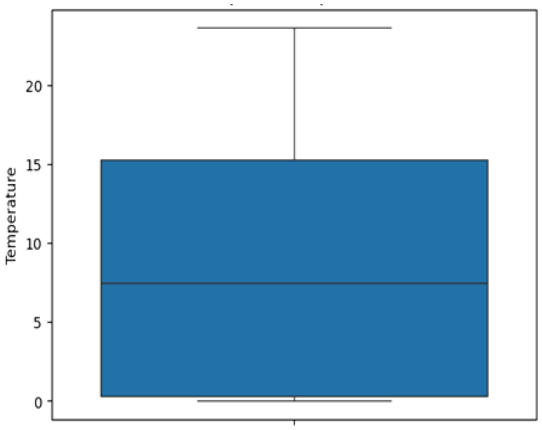
(الف)



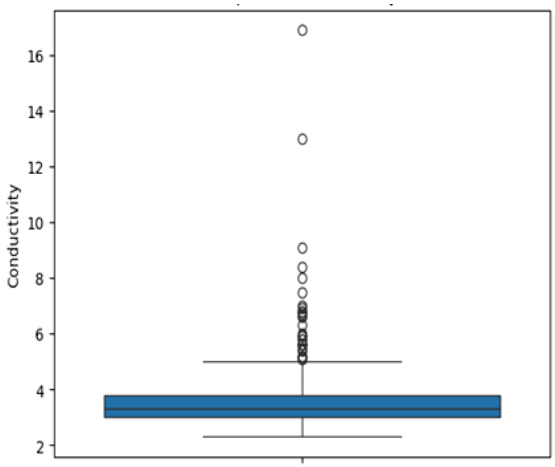
(ب)



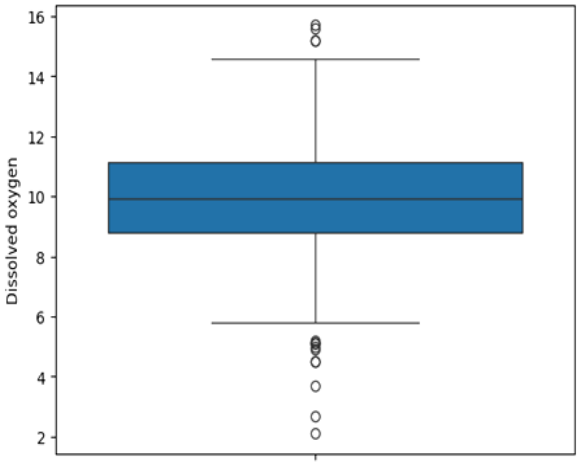
(ج)



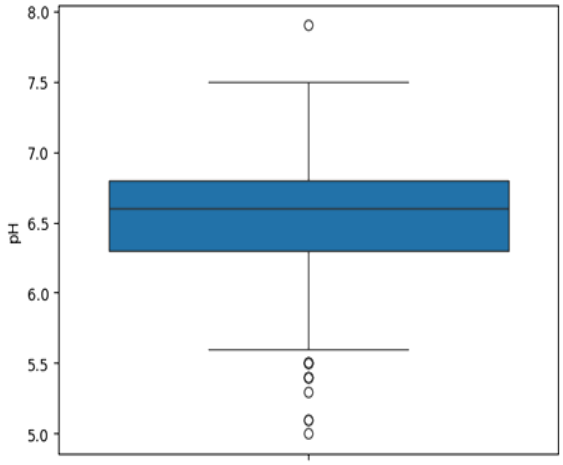
(د)



(ه)



(و)



(ز)

شکل ۶- نمودار جعبه‌ای داده‌ها پارامترهای مختلف: (الف) کدورت؛ (ب) BOD_5؛ (ج) عمق سکی؛ (د) دما؛ (ه) رسانایی؛ (و) DO؛ (ز) pH

غلظت‌های BOD_5 بیش از ۳ میلی‌گرم در لیتر در دریاچه‌های قطبی عموماً نشان‌دهنده آلودگی آلی یا تأثیر فعالیت‌های انسانی است که اهمیت این بررسی را در مطالعه موردی حاضر، کلیورد و همکاران، برجسته می‌کنند [۴۶]. یکی از مراحل اساسی در فرآیند مدل‌سازی، تقسیم مجموعه داده‌ها به زیرمجموعه‌های آموزشی و ارزیابی(تست) است. این روش با استفاده از کتابخانه Scikit-learn انجام شد تا اطمینان حاصل شود که مدل می‌تواند به طور مؤثر الگوهای اساسی را یاد بگیرد و متعاقباً تحت شرایط واقع‌بینانه ارزیابی شود. در این مرحله، ۸۰-۷۰٪ از داده‌ها برای آموزش و ۲۰-۳۰٪ باقی‌مانده برای ارزیابی اختصاص داده شد. علاوه بر این، از تابع شکل(Shape Function) برای تعیین تعداد ردیف‌ها و ستون‌ها در مجموعه داده‌ها استفاده شد و در نتیجه یک نمای کلی از ساختار آن ارائه شد. پس از اتمام مرحله آموزش، عملکرد مدل ارزیابی شد. در این مرحله مدل، پیش‌بینی‌هایی را برای داده‌های آزمایشی بدون دسترسی به مقادیر واقعی BOD_5 ایجاد کرد. در نهایت، خروجی‌های پیش‌بینی‌شده با مقادیر مشاهده‌شده در مجموعه داده‌های آزمایشی مقایسه شدند تا اعتبار مدل تأیید شود و دقت و میزان خطای آن کمی‌سازی شود. این ارزیابی، شناسایی نقاط ضعف بالقوه را ممکن ساخت و مبنایی برای بهینه‌سازی عملکرد مدل فراهم کرد.

۳-۲-۱-آستانه خطای قابل قبول [1]

برای اطمینان از ارزیابی معنادار عملکرد مدل، ابتدا یک آستانه خطای پایه (Baseline) تعیین شد. این خط پایه که در اینجا به عنوان میانگین روند BOD_5 تعریف می‌شود، خطای مرتبط با ساده‌ترین روش پیش‌بینی را بدون اعمال محاسبات پیشرفته نشان می‌دهد. این نقطه مرجع، ارزیابی این موضوع را امکان‌پذیر می‌کند که آیا مدل یادگیری ماشین بهبود قابل توجهی در دقت پیش‌بینی ارائه می‌دهد یا خیر. اگر خطاهای مدل از این خط پایه تجاوز کند، رویکرد انتخاب شده نامناسب تلقی می‌شود و استراتژی مدل‌سازی نیاز به بازنگری دارد.

جدول ۳- محاسبه خطای پایه مدل

پارامتر آماری	مقدار
Baseline_MAE	۱/۵۶۴۴
Baseline_MSE	۶/۰۳۰
Baseline_RMSE	۲/۴۵۵۶
Baseline_ R^2	-۰/۰۰۴۱

[1] Acceptable Error Threshold

۳-۳-آموزش مدل جنگل تصادفی

مدل رگرسیون جنگل تصادفی با استفاده از کتابخانه Scikit-learn پیاده‌سازی و با مجموعه داده‌های تعیین‌شده آموزش داده شد تا رابطه بین متغیرهای ورودی و BOD_5 ثبت شود. سپس عملکرد مدل با تولید تخمین هایی از متغیر هدف، صرفاً از متغیرهای ورودی و مقایسه آنها با اندازه‌گیری‌های واقعی BOD_5 ارزیابی شد. برای اطمینان از پایداری و تعمیم‌پذیری، اعتبارسنجی متقابل K-Fold نیز اعمال شد که تکرارپذیری و پایداری نتایج را بهبود بخشید و توانایی مدل را در یادگیری دقیق و بازتولید الگوهای موجود در داده‌ها تأیید کرد.دقت مدل رگرسیون جنگل تصادفی با استفاده از میانگین خطای مطلق و میانگین مربعات خطا ارزیابی شد. MAE میانگین انحراف پیش‌بینی‌ها از مقادیر واقعی را صرف نظر از جهت خطا کمّی می‌کند، در حالی که MSE، با مربع کردن انحرافات، تأکید بیشتری بر خطاهای بزرگتر دارد. هر دو معیار از طریق کتابخانه Scikit-learn محاسبه شدند و به عنوان شاخص‌های اصلی عملکرد مدل عمل کردند. بهینه‌سازی مدل جنگل تصادفی گامی برای اطمینان از دقت پیش‌بینی بالاتر و کاهش نرخ خطا است. از آنجایی که نسخه اولیه مدل به ندرت بهینه است، تنظیم هایپرپارامتر برای افزایش عملکرد آن اعمال می‌شود. هایپرپارامترها پارامترهایی خارجی هستند که به شدت بر کارایی مدل تأثیر می‌گذارند. فرآیند بهینه‌سازی شامل آزمایش ترکیبات مختلف این مقادیر و انتخاب مؤثرترین مجموعه بر اساس داده‌های آزمون است. از آنجایی که تنظیم دستی می‌تواند زمان‌بر و هم پیچیده باشد، از تکنیک‌های اعتبارسنجی متقابل (که از طریق کتابخانه Scikit-learn پیاده‌سازی می‌شوند) برای خودکارسازی و ساده‌سازی فرآیند استفاده می‌شود. روش جستجوی شبکه ای به طور سیستماتیک فضای هایپرپارامترها را بررسی می‌کند تا ترکیبی را شناسایی کند که MSE را به حداقل می‌رساند و در عین حال R^2 را به حداکثر می‌رساند. از طریق این رویکرد، مدل به پیش‌بینی‌های قابل اعتمادتر و دقیق‌تری دست می‌یابد.

۳-۳-۱-تعیین هایپرپارامتر

بهینه‌سازی هایپرپارامترهای مدل با استفاده از رویکردهای جستجوی شبکه‌ای و اعتبارسنجی متقابل انجام شد. در این فرآیند، تغییرات بر روی هایپرپارامترهای انتخاب‌شده اعمال شد تا موثرترین مقادیر شناسایی شود. با توجه به بار محاسباتی قابل توجه، نتایج مرحله نهایی (مجموعه بهینه هایپرپارامترهایی است که بهترین عملکرد مدل را به همراه داشته‌اند) گزارش شده است.

جدول ۴- هایپرپارامترهای مدل جنگل تصادفی

نام هایپرپارامتر	مقدار
N_estimators	۳۰۰
Test_size	۰/۲
Max_depth	۱۵
Random_state	۴۲

۳-۳-۲-نتایج نهایی مدل جنگل تصادفی

مدل با ترکیب‌های مختلفی از تمام پارامترهای موجود وابسته اجرا شد و در نهایت، بهینه‌ترین انتخاب پارامتر گزارش شد. متغیرهای انتخاب شده در مدل نهایی با بهترین دقت شامل DO، pH، کدورت آب و همچنین روز و زمان نمونه‌برداری بود. برای ارزیابی آماری عملکرد مدل، نتایج در جدول زیر ارائه شده است.

جدول ۵- نتایج حاصل از مدل جنگل تصادفی

نتایج آماری	مقدار
MAE_{Model}	۰/۵۹۹۷۵
MSE_{train}	۰/۳۳۵۶۷
R^2	۰/۸۵۳۲۰
$R^2_{adjusted}$	۰/۸۵۰۰۰

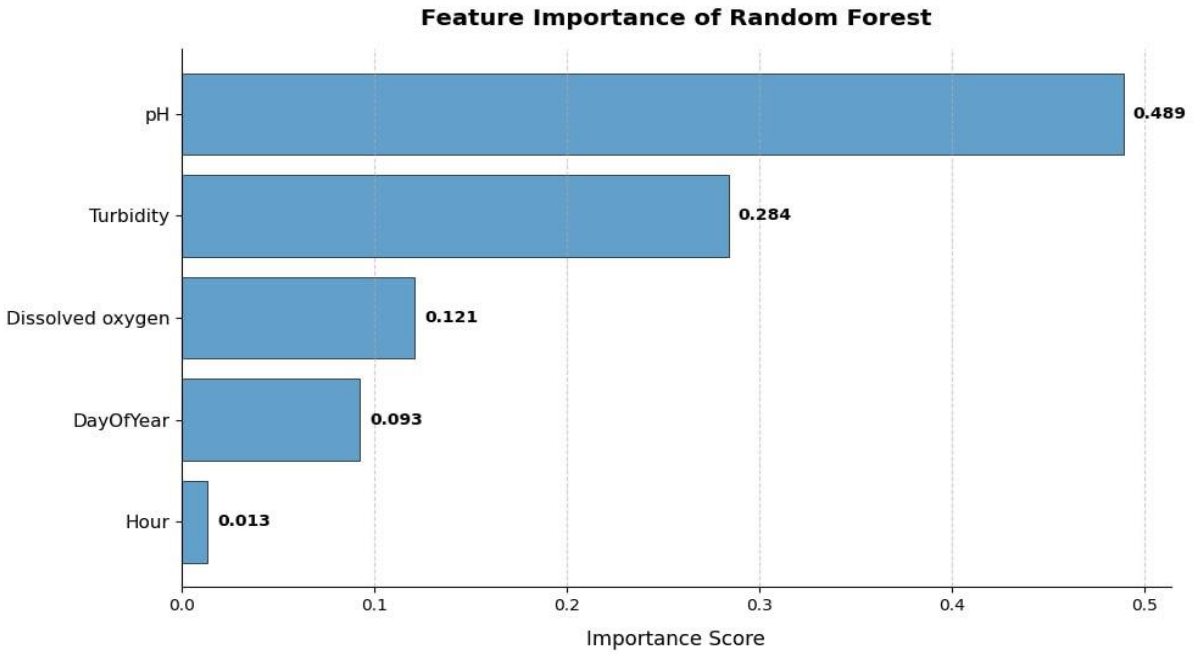
۳-۳-۳-ارزیابی اهمیت مدل جنگل تصادفی[1]

در این مرحله، سهم و تأثیر هر کدام از متغیرهای ورودی در مدل برتر پیش‌بینی BOD_5 بررسی شد. هدف از این تحلیل، ارزیابی میزان تأثیر هر متغیر بر تغییرات BOD_5 و تعیین اهمیت نسبی آنها در مدل بود. برای تعیین کمیت تأثیر پارامترهای ورودی در چارچوب جنگل تصادفی، مقادیر اهمیت متغیرها محاسبه شد و امکان ارزیابی سهم هر متغیر در فرآیند پیش‌بینی کلی فراهم شد. نتایج این ارزیابی، که میزان سهم هر پارامتر در پیش‌بینی‌های مدل را برجسته می‌کند، در جدول ۶ خلاصه شده است.

جدول ۶- اهمیت متغیرها در بهترین مدل جنگل تصادفی

پارامتر	امتیاز اهمیت هر متغیر
pH	۰/۴۸۹۰۷
کدورت	۰/۲۸۴۱۴
اکسیژن محلول	۰/۱۲۰۹۰
روز نمونه برداری	۰/۰۹۲۵۷
ساعت نمونه برداری	۰/۰۱۳۳۰

همانطور که در جدول نشان داده شده است، سه متغیر با بیشترین تأثیر بر پیش‌بینی BOD_5 به ترتیب pH، کدورت آب و اکسیژن محلول هستند. روز نمونه‌برداری و زمان نمونه‌برداری در جایگاه‌های بعدی قرار گرفتند. بر اساس این یافته‌ها، می‌توان پیشنهاد کرد که یک مدل ساده‌شده که صرفاً بر سه متغیر اول متکی باشد، نیز می‌تواند توسعه داده و اعمال شود. با این حال، باید توجه داشت که وقتی این مجموعه کاهش‌یافته از متغیرها در پایتون آزمایش شد، دقت مدل در مقایسه با مدل پنج متغیره کمتر بود. متعاقباً، با استفاده از کتابخانه Matplotlib، نموداری برای تجسم اهمیت پارامترهای ورودی مدل ایجاد شد. شکل زیر تأثیر هر پارامتر بر عملکرد مدل را نشان می‌دهد و دیدگاه روشنی برای ارزیابی نقش‌های مربوطه آنها ارائه می‌دهد.



شکل ۹- امتیاز اهمیت هر پارامتر برای ارزیابی مدل جنگل تصادفی

[1] Feature Importance of Random Forest

۳-۴-آموزش مدل XGBoost

مدل XGBRegressor با استفاده از کتابخانه xgboost پیاده‌سازی و متعاقباً بر روی مجموعه داده‌های آموزشی تعیین‌شده آموزش داده شد. هدف اصلی این فرآیند، قادر ساختن مدل به ثبت روابط اساسی بین متغیرهای ورودی و متغیر هدف و در نتیجه پیش‌بینی دقیق مقادیر BOD_5 بر اساس ورودی‌های داده‌شده بود. مشابه مدل جنگل تصادفی، ارزیابی و پیش‌بینی مدل با استفاده از مجموعه داده‌های آزمایشی انجام شد. برای ارزیابی دقت مدل، شاخص‌های آماری مانند MAE و MSE از طریق کتابخانه Scikit-learn و توابع مرتبط با آن محاسبه شدند. علاوه بر این، استفاده از اعتبارسنجی متقابل K-Fold به افزایش ثبات و قابلیت اطمینان به مدل کمک کرد.

۳-۴-۱-بهینه‌سازی هایپرپارامتری مدل XGBoost

بهینه‌سازی هایپرپارامترهای XGBoost با استفاده از روش‌هایی مانند جستجوی شبکه‌ای و اعتبارسنجی متقابل انجام شد. در این روش، تنظیمات تدریجی بر روی مقادیر ابرپارامتری‌ها اعمال شد تا موثرترین مقادیر تعیین شود. با توجه به حجم محاسبات این کار، جزئیات مراحل میانی حذف شده و تنها نتایج نهایی که نشان‌دهنده تنظیمات بهینه هایپرپارامترها هستند، در اینجا گزارش شده‌اند.

جدول ۷- هایپرپارامترهای مدل XGBoost

نام هایپرپارامتر	مقدار
N_estimators	۱۰۰
Test_Size	۰/۲
Max_depth	۱۵
Learning_rate	۰/۳

۳-۴-۲-نتایج نهایی مدل XGBoost

مدل XGBoost با استفاده از جایگشت و ترکیب متغیرهای موجود اجرا شد و در نهایت، ترکیبی با بهینه‌ترین مجموعه متغیرها شناسایی و گزارش شد. متغیرهای انتخاب شده در مدل نهایی شامل DO، pH، کدورت آب و روز نمونه‌برداری بود. برای ارزیابی آماری عملکرد مدل، نتایج مربوطه در جدول زیر خلاصه شده است.

جدول ۸- نتایج حاصل از مدل XGBoost

نتایج آماری	مقدار
MAE_{Model}	۰/۶۲۷۲۰
MSE_{train}	۰/۰۰۰۰۰۰۶
R^2	۰/۸۳۰۵
$R^2_{adjusted}$	۰/۸۱۰۳

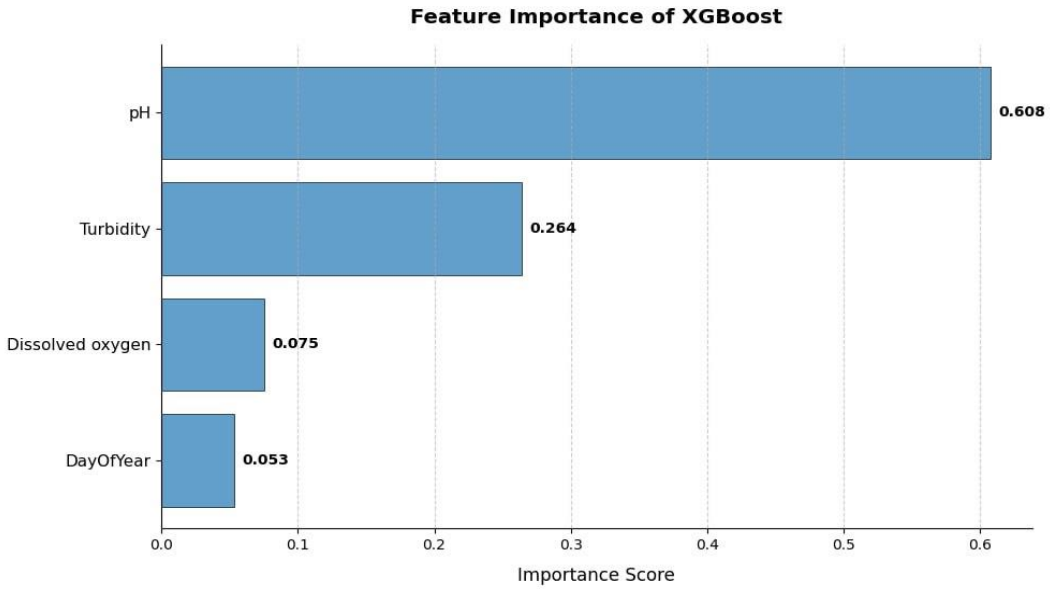
۳-۴-۳-ارزیابی اهمیت هر کدام از متغیرها در مدل XGBoost

در این مرحله از تحلیل، سهم هر متغیر ورودی در پیش‌بینی BOD_5 بررسی می‌شود. هدف، درک این موضوع است که هر پارامتر چقدر بر تغییرات مقادیر BOD_5 تأثیر می‌گذارد و اهمیت نسبی آنها در مدل تعیین شود. برای کمی‌سازی این اثرات، مقادیر اهمیت متغیرها محاسبه شد و امکان ارزیابی سهم هر یک از آنها در ساختار تخمین و پیش‌بینی کلی فراهم شد. نتایج این ارزیابی، که میزان تاثیرگذاری هر متغیر در فرآیند پیش‌بینی را برجسته می‌کند، در جدول ۹ خلاصه شده است.

جدول ۹- اهمیت ویژگی در مدل XGBoost

پارامتر	امتیاز اهمیت هر متغیر
pH	۰/۶۰۷۷۹
کدورت	۰/۲۶۳۹۷
اکسیژن محلول	۰/۱۲۰۹۰
روز نمونه برداری	۰/۰۵۳۲۴

سه متغیری که بیشترین تأثیر را بر پیش‌بینی BOD_5 دارند، به ترتیب pH، کدورت آب و اکسیژن محلول هستند. با کمک Matplotlib، نموداری برای نشان دادن اهمیت نسبی متغیرهای ورودی مدل ایجاد شد. در شکل زیر، میزان سهم هر متغیر در عملکرد مدل برجسته می‌شود. چنین نمایشی، دیدگاه برای ارزیابی نقش هر متغیر در شکل‌دهی به نتایج پیش‌بینی ارائه می‌دهد.



شکل ۱۰- امتیاز اهمیت هر پارامتر برای ارزیابی مدل XGBoost

۳-۵-آموزش مدل CNN

در این بخش، مدل CNN با استفاده از کتابخانه torch پیاده‌سازی و متعاقباً با مجموعه داده‌های آموزشی آماده، آموزش داده شد. هدف اصلی این فرآیند، قادر ساختن مدل به شناسایی رابطه اساسی بین متغیرهای ورودی و متغیر هدف، یعنی BOD_5، بود تا بتواند مقادیر BOD_5 را بر اساس ورودی‌های ارائه شده به طور قابل اعتمادی پیش‌بینی کند. مرحله ارزیابی و پیش‌بینی برای این مدل به همان شیوه دو مدل قبلی، با استفاده از مجموعه داده‌های آزمایشی انجام شد. برای کمّی‌سازی عملکرد مدل، معیارهای آماری مانند MAE و MSE محاسبه شدند. این شاخص‌ها با استفاده از کتابخانه Scikit-learn استخراج شدند و ثبات در فرآیند ارزیابی را تضمین کردند. در این تحقیق، CNN با ساختاری پیچیده و مهندسی شده (همانطور که در روش شناسی توضیح داده شد) طراحی و آزمایش شد که چارچوبی خوب برای برجسته‌تر کردن نقاط قوت و محدودیت‌های خود ارائه می‌دهد. پس توجه به این نکته ضروری است که از ساختار و معماری CNN با ویژگی های ساختاری پیچیده و متمایز به کار گرفته شد.

۳-۵-۱-بهینه‌سازی و تنظیم ابرپارامترهای مدل CNN

برای افزایش عملکرد پیش‌بینی مدل CNN، یک فرآیند تنظیم هایپرپارامتر سیستماتیک انجام شد. در این رویکرد، تنظیمات تدریجی روی مقادیر هایپرپارامترهای انتخاب‌شده انجام شد تا مؤثرترین ترکیب شناسایی شود. مراحل میانی در اینجا گزارش نشده‌اند و در واقع، فقط نتایج نهایی هایپرپارامترها که بهترین عملکرد مدل را به همراه داشته‌اند ارائه شده‌اند.

جدول ۱۲- هایپرپارامترهای مدل CNN

هایپرپارامتر	مقدار
Num_epochs	۸۸۱
Batch_size	۳۲
Learning_rate	۰/۰۱

۳-۵-۲-نتایج نهایی مدل CNN

در مدل CNN، ترکیبات مختلفی از تمام متغیرهای موجود آزمایش شدند و در نهایت، مدلی با مؤثرترین انتخاب پارامتر گزارش شد. متغیرهای کلیدی انتخاب شده در این مدل شامل عمق دیسک سکی، رسانایی الکتریکی، اکسیژن محلول، pH و کدورت آب است. برای ارزیابی آماری عملکرد مدل، نتایج مربوطه در جدول زیر ارائه شده است. ذکر این نکته موجه است که مدل CNN به صورت کلی بخاطر ویژگی‌ها و لایه‌های پنهانش امکان ارزیابی اهمیت هر پارامتر را ندارد ولی با توجه به مدل‌های دیگر و نتایجش میتوانیم از انتخاب بهترین پارامترها برای ساخت مدل بهتر بهره مند شد.

جدول ۱۳- نتایج حاصل از مدل CNN

نتایج آماری	مقدار
MAE_{Test}	۰/۵۵۸۹
MAE_{Train}	۰/۵۹۷۸
MSE_{Test}	۰/۷۱۵۷
MSE_{train}	۰/۹۹۱۶
R^2	۰/۸۸۰۸
$R^2_{adjusted}$	۰/۸۷۲۷

۳-۶- بیش برازش و مقایسه دقت مدل ها

برای ارزیابی احتمال بیش‌برازش، عملکرد مدل‌ها در داده‌های آموزش و آزمون به‌صورت هم‌زمان مقایسه شد. این بررسی بر این اصل استوار بود که یک مدل قابل‌اعتماد باید علاوه بر دستیابی به دقت مناسب در مرحله آموزش، در داده‌های ندیده نیز رفتار پایدار و قابل‌قبولی داشته باشد. در همین راستا، نزدیک بودن مقادیر R^2 در دو مرحله آموزش و اعتبارسنجی، در کنار نبود افت محسوس در عملکرد آزمون، به‌عنوان نشانه‌ای از تعمیم‌پذیری مناسب مدل در نظر گرفته شد. این رویکرد با رویه به‌کار رفته در مطالعات مشابه، از جمله مقاله نجف زاده و قایمی هم‌خوانی دارد که در آن نیز عملکرد مدل‌ها به‌طور جداگانه در مراحل آموزش و آزمون ارزیابی شده است. به بیان دیگر، اگر مدلی صرفاً در داده‌های آموزش عملکرد خوبی داشته باشد اما در مجموعه آزمون افت قابل توجهی نشان دهد، می‌توان آن را نشانه‌ای از Overfitting دانست. در این پژوهش اختلاف عملکرد میان train وtest به‌عنوان معیار اصلی کنترل بیش‌برازش مدنظر قرار گرفت. برای مثال در مدل جنگل تصادفی اختلاف ضریب تعیین فاز آموزش و تست هفت درصد است که باتوجه به ادبیات پژوهشی و تفاوت عدد ذکر شده تا آستانه ۱۰ الی ۱۵ درصد تفاوت عدم بیش برازش و تعمیم پذیری مناسب مدل را (در کنار روش هایی مثل K-Fold و Cross-Validation) را نشان میدهد و از طریق مقایسه شاخص‌های آماری، اطمینان حاصل شد که مدل‌ها تنها الگوهای خاص داده‌های آموزشی را حفظ نکرده‌اند، بلکه توانایی پیش‌بینی قابل اتکایی برای داده‌های جدید دارند [۱۷].

در مجموع، در میان سه مدل یادگیری ماشینی توسعه‌یافته در این مطالعه، مدل شبکه CNN بالاترین دقت و قابل اعتمادترین عملکرد را بر اساس ارزیابی های آماری به دست آورد و پس از آن مدل RF قرار گرفت. این مطالعه با موفقیت عوامل حیاتی مؤثر بر پیش‌بینی BOD_5 را شناسایی کرد و استراتژی‌هایی را برای بهبود ارزیابی کیفیت آب پیشنهاد داد. با توجه به پارامترهای تأثیرگذار مانند pH، کدورت، اکسیژن محلول و زمان نمونه‌برداری، نظارت و کنترل دقیق‌تر منابع آلودگی در حوضه دریاچه ضروری به نظر می‌رسد. جدول نهایی زیر خلاصه‌ای مقایسه‌ای از دقت پیش‌بینی و میانگین خطای مطلق به‌دست‌آمده توسط مدل‌های یادگیری ماشینی ارزیابی‌شده در این تحقیق را ارائه می‌دهد که از آستانه خطای قابل قبول هم فاصله قابل توجهی دارند.

جدول ۱۴- مقایسه آماری خطا و دقت مدل‌های یادگیری ماشین مورد استفاده

نام مدل تخمین گر	R^2	MAE_{Model}
Random Forest	۰/۸۵۳۲	۰/۵۹۹۷۵
XGBoost	۸۳۰۵۸/۰	۰/۶۲۷۲۰
CNN With BN& Dropout	۰/۸۸۰۸	۰/۵۵۸۹۰

۴-نتیجه‌گیری

به طور کلی، این مطالعه نشان می‌دهد که تقویت شبکه‌های پایش کیفیت آب از طریق ادغام فناوری‌های پیشرفته با روش‌های مرسوم می‌تواند به طور قابل توجهی قابلیت اطمینان به داده‌ها را افزایش داده و از مدیریت پیشگیرانه منابع آب در مواجهه با تغییرات اقلیمی پشتیبانی کند. چنین رویکردی نه تنها برای مدیریت پایدار دریاچه‌های قطبی، بلکه برای حفاظت از سایر اکوسیستم‌های آبی حساس نیز ارزشمند است. این تحقیق از مجموعه داده‌های کیفیت آب برای تخمین BOD_5 در شش زیرحوضه و ۳۴ ایستگاه در یک دریاچه قطبی استفاده کرد. اهداف این تحقیق کاهش هزینه‌های پایش، بهبود کارایی ارزیابی و بررسی رویدادهای تاریخی مهم بود. سه مدل یادگیری ماشین (RF، XGBoost و CNN) با استفاده از تقریباً ۶۵۰ رکورد جمع‌آوری‌شده طی ۴۰ سال با فواصل ماهانه تا فصلی توسعه داده شدند. مجموعه داده‌ها به زیرمجموعه‌های آموزش (۸۰-۷۰٪) و اعتبارسنجی (۲۰-۳۰٪) تقسیم شدند. مدل RF، با عملکرد مطلوب در آموزش و R^2 ۸۵ تا ۸۶ درصد را در اعتبارسنجی به دست آورد. تجزیه و تحلیل اهمیت وتاثیرگذاری متغیرها نشان داد که DO، pH، کدورت و زمان نمونه‌برداری به عنوان تأثیرگذارترین پیش‌بینی‌کننده‌ها هستند. مدل XGBoost با پیش‌بینی‌کننده‌های کلیدی مشابه به مقادیرایده آل در آموزش و ضریب تعیین ۸۳ تا ۸۴ درصد در اعتبارسنجی رسید. مدل CNN اما عملکرد بی نظیری داشت، با ضریب تعیین بالا در فاز آموزش و R^2 ۸۸ تا ۸۹ درصد در اعتبارسنجی، که در آن عمق سکی، رسانایی، DO، pH و کدورت به عنوان متغیرهای حیاتی مدل ظاهر شدند. هایپرپارامترهای بهینه شامل ۳۰۰ درخت برای جنگل تصادفی، ۱۰۰ درخت برای XGBoost (هر دو با حداکثر عمق درخت ۱۵) و برای CNN، نرخ یادگیری ۰/۰۱ و ۸۸۱ دوره آموزشی و اندازه بچ به مقدار ۳۲ بود. با وجود محدودیت‌های اندازه مجموعه داده‌ها و عدم قطعیت‌های پیش‌پردازش، همه مدل‌ها توانایی پیش‌بینی قوی را نشان دادند و تخمین‌های قابل اعتمادی از BOD_5 (یک شاخص نشان دهنده تغذیه‌گرایی) در دریاچه‌های قطبی ارائه دادند.

سهم اصلی این پژوهش در قیاس با پژوهش های دیگر در ارائه یک چارچوب مقایسه‌ای برای تخمین BOD_5 با استفاده از سه الگوریتم متفاوت یادگیری ماشین، بر پایه بیش از چهار دهه داده تاریخی از یک دریاچه قطبی، قابل تعریف است. در واقع تمرکز هم‌زمان بر دوره‌های لایه‌بندی حرارتی و شرایط خاص این اکوسیستم، دامنه مطالعه را از یک مدل‌سازی صرف فراتر برده و امکان ارزیابی دقیق‌تر رفتار BOD_5 و متغیرهای اثرگذار بر آن را فراهم کرده است. در این میان، مقایسه یک مدل یادگیری عمیق با دو روش یادگیری گروهی درخت‌محور نشان داد که رویکردهای داده‌محور می‌توانند در پایش شاخص‌های کیفی آب، به‌ویژه در محیط‌های کم‌داده و حساس، کارایی بالایی داشته باشند. از منظر کاربردی، نتایج این مطالعه می‌تواند مبنایی برای توسعه سامانه‌های پایش سریع‌تر و کم‌هزینه‌تر با صرف انرژی کمتر برای کیفیت آب باشد. با توجه به زمان‌بر و پرهزینه بودن اندازه‌گیری مستقیم BOD_5، مدل‌های پیشنهادی این امکان را فراهم می‌کنند که این شاخص با تکیه بر پارامترهای متداول کیفیت آب، به‌صورت قابل اعتماد برآورد شود. از طرفی ایده کار بر روی داده های کیفی لایه های عمیق آب (که شاخصی از وضعیت بلند مدت دریاچه است و استفاده از سنسورهای اندازه‌گیری و یا بهره مندی از داده های ماهواره ای در عمق بسیار مشکل و هزینه بر است) در پژوهش های آتی در دریاچه

های قطبی با الهام از تخمین ها و مدل سازی های این پژوهش میتواند جهت بررسی تاریخچه دریاچه های قطبی پیشنهاد گردد. این رویکرد به‌ویژه برای دریاچه‌های قطبی و یا سایر اکوسیستم‌های آبی با شرایط محیطی حساس و دورافتاده که با محدودیت دسترسی و پایش مستمر مواجه‌اند، ارزش عملی بالایی دارد و می‌تواند در ارزیابی بلندمدت، هشدار زودهنگام و تصمیم‌گیری مدیریتی مورد استفاده قرار گیرد.

۵-منابع

1. Adrian, R., C. M. O'Reilly, H. Zagarese, S. B. Baines, D. O. Hessen, W. Keller, D. M. Livingstone, R. Sommaruga, D. Straile, E. Van Donk, G. A. Weyhenmeyer, and M. Winder, *Lakes as sentinels of climate change*. Limnology and Oceanography, 2009. 54(6 Part 2): p. 2283-2297. https://doi.org/10.4319/lo.2009.54.6_part_2.2283
2. Banita, A., R. L. C. Pârvulescu, and M. Moldovan, *Integrating Remote Sensing Methods for Monitoring Lake Water Quality: A Comprehensive Review*. Remote Sensing, 2024. 16(7): p. 1192. https://doi.org/10.3390/rs16071192
3. Fouchal, A., Y. Tikhamarine, M. A. Benbouras, D. Souag-Gamane, and S. Heddam, *Biological oxygen demand prediction using artificial neural network and random forest models enhanced by the neural architecture search algorithm.* Modeling Earth Systems and Environment, 2025. 11(1): p. 9. https://doi.org/10.1007/s40808-024-02178-x
4. Jain, S. K., A. Kumar, S. Kumar, A. Kumar, and A. Choudhary, *Data-Driven Prediction of Effluent BOD5 from an Institutional Wastewater Treatment Plant*. In Advances in Data-driven Computing and Intelligent Systems: Selected Papers from ADCIS 2022, Volume 2. Singapore: Springer Nature Singapore, 2023: p. 217-224. https://doi.org/10.1007/978-981-99-0981-0_17
5. Toming, K., H. Liu, T. Soomets, E. Uuemaa, T. Nõges, and T. Kutser, *Estimation of the biogeochemical and physical properties of lakes based on remote sensing and artificial intelligence applications.* Remote Sensing, 2024. 16(3): p. 464. https://doi.org/10.3390/rs16030464
6. Mekaoussi, H., S. Heddam, N. Bouslimanni, S. Kim, and M. Zounemat-Kermani, *Predicting biochemical oxygen demand in wastewater treatment plant using advance extreme learning machine optimized by Bat algorithm.* Heliyon, 2023. 9(11). https://doi.org/10.1016/j.heliyon.2023.e21351
7. Oliveira-Esquerre, K. P., D. E. Seborg, R. E. Bruns, and M. Mori, *Application of steady-state and dynamic modeling for the prediction of the BOD of an aerated lagoon at a pulp and paper mill: Part I. Linear approaches.* Chemical Engineering Journal, 2004. 104(1-3): p. 73-81. https://doi.org/10.1016/j.cej.2004.05.009
8. Ning, S. K., N. B. Chang, L. Yang, H. W. Chen, and H. Y. Hsu, *Assessing pollution prevention program by QUAL2E simulation analysis for the Kao-Ping River Basin, Taiwan.* Journal of Environmental Management, 2001. 61(1): p. 61-76. https://doi.org/10.1006/jema.2000.0397
9. Chavez, P., and F. J. Chang, *Simulation of multiple water quality parameters using artificial neural networks*. In 7th International Conference on Hydroinformatics, Nice, France, 2006. https://doi.org/10.13140/RG.2.1.2580.0408

10. Virro, H., A. Kmoch, M. Vainu, and E. Uuemaa, *Random forest-based modeling of stream nutrients at national level in a data-scarce region.* Science of the Total Environment, 2022. 840: p. 156613. https://doi.org/10.1016/j.scitotenv.2022.156613
11. Deng, F., W. Liu, M. Sun, Y. Xu, B. Wang, W. Liu, Y. Yuan, and L. Cui, *Fine estimation of water quality in the Yangtze River Basin based on a geographically weighted random forest regression model.* Remote Sensing, 2025. 17(4): p. 731. https://doi.org/10.3390/rs17040731
12. Bai, Y., M. Peng, and M. Wang, *A river water quality prediction method based on dual signal decomposition and deep learning.* Water, 2024. 16(21): p. 3099. https://doi.org/10.3390/w16213099
13. Peng, C., Z. Xie, and X. Jin, *Using ensemble learning for remote sensing inversion of water quality parameters in Poyang Lake.* Sustainability, 2024. 16(8): p. 3355. https://doi.org/10.3390/su16083355
14. Guo, H., X. Zhu, J. J. Huang, Z. Zhang, S. Tian, and Y. Chen, *An enhanced deep learning approach to assessing inland lake water quality and its response to climate and anthropogenic factors.* Journal of Hydrology, 2023. 620: p. 129466. https://doi.org/10.1016/j.jhydrol.2023.129466
15. Leung, J. H., Y. M. Tsao, R. Karmakar, A. Mukundan, S. C. Lu, S. Y. Huang, and H. C. Wang, *Water pollution classification and detection by hyperspectral imaging.* Optics Express, 2024. 32(14): p. 23956-23965. https://doi.org/10.1364/OE.523456
16. Saeed, A., A. Alsini, and D. Amin, *Water quality multivariate forecasting using deep learning in a West Australian estuary.* Environmental Modelling & Software, 2024. 171: p. 105884. https://doi.org/10.1016/j.envsoft.2023.105884
17. Najafzadeh, M., and A. Ghaemi, *Prediction of the five-day biochemical oxygen demand and chemical oxygen demand in natural streams using machine learning methods.* Environmental Monitoring and Assessment, 2019. 191(6): p. 380. https://doi.org/10.1007/s10661-019-7446-8
18. Harrison, J. W., M. A. Lucius, J. L. Farrell, L. W. Eichler, and R. A. Relyea, *Prediction of stream nitrogen and phosphorus concentrations from high-frequency sensors using Random Forests Regression.* Science of the Total Environment, 2021. 763: p. 143005. https://doi.org/10.1016/j.scitotenv.2020.143005
19. Ooi, K. S., Z. Chen, P. E. Poh, and J. Cui, *BOD5 prediction using machine learning methods.* Water Supply, 2022. 22(1): p. 1168-1183. https://doi.org/10.2166/ws.2021.202
20. Ziyad Sami, B. Fadi, S. D. Latif, A. N. Ahmed, M. F. Chow, M. A. Murti, A. Suhendi, B. H. Ziyad Sami, J. K. Wong, A. H. Birima, and A. El-Shafie, *Machine learning algorithm as a sustainable tool for dissolved oxygen prediction: a case study of Feitsui Reservoir, Taiwan.* Scientific Reports, 2022. 12(1): p. 3649. https://doi.org/10.1038/s41598-022-06969-z
21. Hietala, R., H. Virkkunen, J. Salminen, P. Ekholm, J. Riihimäki, P. Laine, and T. Kirkkala, *Assessment of agricultural water protection strategies at a catchment scale: case of Finland.* Regional Environmental Change, 2024. 24(1): p. 2. https://doi.org/10.1007/s10113-023-02164-8

22. Ekholm, P., and S. Mitikka, *Agricultural lakes in Finland: current water quality and trends.* Environmental Monitoring and Assessment, 2006. 116(1): p. 111-135. https://doi.org/10.1007/s10661-006-7231-3
23. ElGharbawi, T., M. R. Kaloop, J. W. Hu, et al., *Hyperspectral reflectance-driven estimation of water quality parameters using hybrid random forest and support vector regression techniques.* Environmental Earth Sciences, 2025. 84: p. 376. https://doi.org/10.1007/s12665-025-12387-x
24. Kim, H. I., D. Kim, M. M. Salamattalab, M. Mahdian, S. M. Bateni, and R. Noori, *Machine learning-based modeling of surface water temperature dynamics in arctic lakes.* Environmental Science and Pollution Research, 2024. 31(49): p. 59642-59655. https://doi.org/10.1007/s11356-024-35173-x
25. Uddin, M. G., A. Bamal, M. T. M. Diganta, A. M. Sajib, A. Rahman, M. Abioui, and A. I. Olbert, *The role of optimizers in developing data-driven model for predicting lake water quality incorporating advanced water quality model.* Alexandria Engineering Journal, 2025. 122: p. 411-435. https://doi.org/10.1016/j.aej.2025.03.022
26. Chen, Y., H. Zhang, Y. You, J. Zhang, and L. Tang, *A hybrid deep learning model based on signal decomposition and dynamic feature selection for forecasting the influent parameters of wastewater treatment plants.* Environmental Research, 2025. 266: p. 120615. https://doi.org/10.1016/j.envres.2024.120615
27. Hamada, M. S., H. A. Zaqoot, and W. A. Sethar, *Using a supervised machine learning approach to predict water quality at the Gaza wastewater treatment plant.* Environmental Science: Advances, 2024. 3: p. 132-144. https://doi.org/10.1039/D3VA00170A
28. Vainikka, A., E. Jakubavičiūtė, and P. Hyvärinen, *Synchronous decline of three morphologically distinct whitefish (Coregonus lavaretus) stocks in Lake Oulujärvi with concurrent changes in the fish community.* Fisheries Research, 2017. 196: p. 34-46. https://doi.org/10.1016/j.fishres.2017.07.024
29. Ching, P. M. L., X. Zou, D. Wu, R. H. Y. So, and G. H. Chen, *Development of a wide-range soft sensor for predicting wastewater BOD5 using an eXtreme gradient boosting (XGBoost) machine.* Environmental Research, 2022. 210: p. 112953. https://doi.org/10.1016/j.envres.2022.112953
30. Dodds, W. K., and M. R. Whiles, Freshwater Ecology: *Concepts and Environmental Applications of Limnology.* 2nd ed. San Diego, CA: Academic Press, 2010.
31. Wetzel, R. G., *Limnology: Lake and River Ecosystems*. 3rd ed. San Diego, CA: Academic Press, 2001.
32. Prowse, T., K. Alfredsen, S. Beltaos, B. R. Bonsal, W. B. Bowden, C. R. Duguay, A. Korhola, et al., *Effects of changes in arctic lake and river ice.* Ambio, 2011. 40(Suppl 1): p. 63-74. https://doi.org/10.1007/s13280-011-0217-6
33. Rühland, K. M., A. M. Paterson, and J. P. Smol, *Lake diatom responses to warming: reviewing the evidence.* Journal of Paleolimnology, 2015. 54(1): p. 1-35. https://doi.org/10.1007/s10933-015-9837-3

34. Gorham, E., and F. M. Boyce, *Influence of lake surface area and depth upon thermal stratification and the depth of the summer thermocline.* Journal of Great Lakes Research, 1989. 15(2): p. 233-245. https://doi.org/10.1016/S0380-1330(89)71479-9
35. Kalff, J., *Limnology: Inland Water Ecosystems*. New Jersey: Prentice Hall, 2002, 592 p.
36. Majidzadeh, H., H. Uzun, A. Ruecker, D. Miller, J. Vernon, H. Zhang, S. Bao, M. T. K. Tsui, T. Karanfil, and A. T. Chow, *Extreme flooding mobilized dissolved organic matter from coastal forested wetlands.* Biogeochemistry, 2017. 136(3): p. 293-309. https://doi.org/10.1007/s10533-017-0394-x
37. Mitchell, T. M., *Does machine learning really work?* AI Magazine, 1997. 18(3): p. 11-20. https://doi.org/10.1609/aimag.v18i3.1303
38. Jordan, M. I., and T. M. Mitchell, *Machine learning: Trends, perspectives, and prospects.* Science, 2015. 349(6245): p. 255-260. https://doi.org/10.1126/science.aaa8415
39. Zhou, Z. H., *Machine Learning*. Singapore: Springer Nature, 2021. https://doi.org/10.1007/978-981-15-1967-3
40. Kiranyaz, S., O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, *1D convolutional neural networks and applications: A survey.* Mechanical Systems and Signal Processing, 2021. 151: p. 107398. https://doi.org/10.1016/j.ymssp.2020.107398
41. Breiman, L., Random forests. Machine Learning, 2001. 45(1): p. 5-32. https://doi.org/10.1023/A:1010933404324
42. Chen, T., and C. Guestrin, XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: p. 785-794. https://doi.org/10.1145/2939672.2939785
43. Kumar, K. V., P. Kumari, A. Chatterjee, and D. P. Mohapatra, *Software fault prediction using random forests.* In: Intelligent and Cloud Computing: Proceedings of ICICC 2019, Volume 1, 2020: p. 95-103. Singapore: Springer Singapore. https://doi.org/10.1007/978-981-15-5971-6_10
44. Han, S., and H. Kim, *Optimal feature set size in random forest regression.* Applied Sciences, 2021. 11(8): p. 3428. https://doi.org/10.3390/app11083428
45. Najah, A., A. El-Shafie, and O. Karim, *Prediction of Water Quality Parameters Using Artificial Intelligence: Case Study-Johor River Basin*. LAP Lambert Academic Publishing, 2011. https://doi.org/10.5555/2208226
46. Clilverd, H., D. White, and M. Lilly, *Chemical and Physical Controls on the Oxygen Regime of Ice-Covered Arctic Lakes and Reservoirs.* JAWRA Journal of the American Water Resources Association, 2009. 45(2): p. 500-511. https://doi.org/10.1111/j.1752-1688.2009.00305.x