MINISTRY OF EDUCATION AND TRAINING

VIETNAM ACADEMY OF SCIENCE AND TECHNOLOGY

GRADUATE UNIVERSITY OF SCIENCE AND TECHNOLOGY

Michael Omar

AN APPROACH OF ENSEMBLE SPATIAL MACHINE LEARNING FOR GROUNDWATER DRINKABILITY CLASSIFICATION

DOCTORAL DISSERTATION ON INFORMATION SYSTEM

Hanoi - 2026

MINISTRY OF EDUCATION AND TRAINING

VIETNAM ACADEMY OF SCIENCE AND TECHNOLOGY

GRADUATE UNIVERSITY OF SCIENCE AND TECHNOLOGY

Michael Omar

AN APPROACH OF ENSEMBLE SPATIAL MACHINE LEARNING FOR GROUNDWATER DRINKABILITY CLASSIFICATION

DOCTORAL DISSERTATION ON COMPUTER

MAJOR: INFORMATION SYSTEMS

CODE: 9 48 01 04

Graduation University of Science and Technology's confirmation	Advisor 1 *(Signature, Full Name)*	Advisor 2 *(Signature, Full Name)*
KT. GIÁM ĐỐC PHÓ GIÁM ĐỐC HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ VIỆN HÀN LÂM KHOA HỌC VÀ CÔNG NGHỆ VIỆT NAM Nguyễn Thị Trung		Nguyễn Long Giang

Hanoi - 2026

BỘ GIÁO DỤC
VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

Michael Omar

NGHIÊN CỨU PHÁT TRIỂN PHƯƠNG PHÁP HỌC MÁY KẾT HỢP THÔNG TIN KHÔNG GIAN CHO BÀI TOÁN PHÂN LOẠI CHẤT LƯỢNG NƯỚC NGẦM

LUẬN ÁN TIẾN SĨ MÁY TÍNH

NGÀNH: HỆ THỐNG THÔNG TIN
MÃ SỐ: 9 48 01 04

Xác nhận của Học viện Khoa học và Công nghệ	Người hướng dẫn 1 *(Ký, ghi rõ họ tên)*	Người hướng dẫn 2 *(Ký, ghi rõ họ tên)*
KT. GIÁM ĐỐC PHÓ GIÁM ĐỐC HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ Nguyễn Thị Trung	Trần Thị Ngân	Nguyễn Long Giang

Hà Nội - 2026

DECLARATION

I hereby declare that the results presented in the discussion are my research work under the guidance of the guidelines. The data and results presented in the project discussion are completely honest and have not been published in previous works. Reference data have been fully cited. The thesis uses cited information from various sources, and all citations are clearly acknowledged. The research findings co-authored with other researchers have been included in the thesis with their consent.

This thesis was completed during my time as a doctoral student at the Vietnam Academy of Science and Technology.

Hanoi, date 2.. month 03 year 2026

Thesis Author

Michael Omar

Acknowledgments

Throughout the process of researching and completing my thesis, I have been fortunate to receive guidance, assistance, invaluable feedback, and encouragement from a diverse group of individuals, including scientists, educators, peers, and family members.

First and foremost, I would like to extend my deepest gratitude to Assoc. Prof. Tran Thi Ngan, Assoc. Prof. Nguyen Long Giang and Assoc. Prof. Le Hoang Son. Their dedicated mentorship and support have been pivotal throughout my research journey.

I also would like to express my sincere appreciation to the faculty and scientists at the Institute of Information and Technology, the Institute of Information Technology VNU. Their insightful suggestions significantly contributed to the advancement of my research.

My gratitude extends to the Board of Directors and the Training Department at Graduate University of Science and Technology, Vietnam Academy of Science and Technology for providing me with the conducive environment necessary to fulfill my research objectives.

Lastly, I must acknowledge my colleagues, family, and friends, whose unwavering encouragement, shared experiences, support, and assistance have been instrumental in helping me overcome challenges and achieve the results presented in this thesis.

Hanoi, date 7.. month 03 ... year 2026

Signed by:

Michael Omar

Contents

Symbols and Abbreviations

No.	Abbreviation	Description
1	AI	Artificial Intelligence
2	AI-LGBM	Auto Immune Light Gradient Boosting Machine
3	CNN	Convolutional Neural Network
4	DL	Deep Learning
6	GIS	Geographic Information System
7	GWQ	Groundwater Quality
8	IPCC	Intergovernmental Panel on Climate Change
9	ML	Machine Learning
10	PSO	Particle Swarm Optimization
11	PSO-SCNN	Particle Swarm Optimization-Spatial Convolutional Neural Network
12	R^2	Coefficient of Determination
13	RMSE	Root Mean Square Error
14	SCNN	Spatial Convolutional Neural Network
15	SELU	Scaled Exponential Linear Unit
16	SHAP	Shapley Additive Explanations
17	SVM	Support Vector Machine
18	TDS	Total Dissolved Solids
19	LightGBM	Light Gradient Boosting Machine
20	XGBoost	Extreme Gradient Boosting
21	LSTMs	Long Short-Term Memory
22	MAE	Mean Absolute Error
23	AUC	Area Under Curve
24	ANOVA	Analysis of Variance
25	WQI	Water Quality Index
26	WQC	Water Quality Class
27	GWQ	Groundwater Quality

List of Tables

List of Figures

Introduction

Research Context

Ensuring the safety of drinking water is a paramount global challenge, essential for public health, environmental sustainability, and economic development. The increasing global population places significant pressure on finite water resources, exacerbating the challenge of ensuring safe drinking water. [1–3]. An estimated two billion people still lack access to safely managed drinking water, making the advancement of robust water quality assessment methods a global health imperative [4]. Contaminated sources are a primary vector for waterborne diseases and expose populations to chemical and pathogenic contaminants, creating a persistent public health crisis [5].

The urgency for a new assessment paradigm is compounded by mounting environmental pressures from industrial and agricultural runoff [6], as well as the spatiotemporal variability of water quality, which is being exacerbated by climate change [7].

Traditional water quality monitoring, which relies on manual field sampling and laboratory analysis, is increasingly ill-suited to address the scale of this challenge. These methods are inherently inefficient, slow, and unscalable, particularly in resource-constrained regions [8], necessitating a shift towards more advanced, automated solutions.

Visual Summary of the Transition

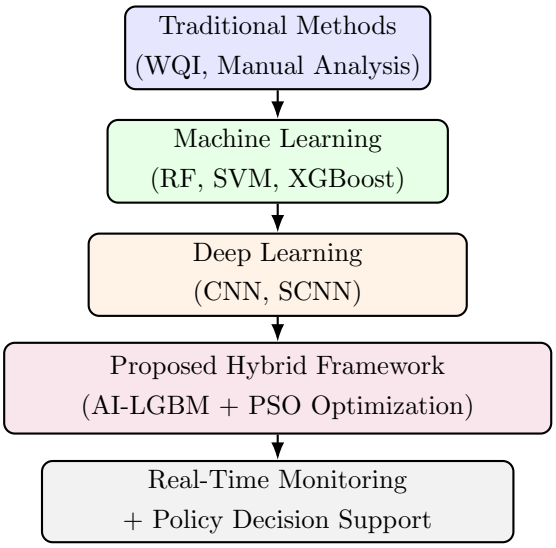


Figure 1: Evolution from Traditional Methods to Hybrid Spatial-Aware ML Framework

Problem Statement

In response to these limitations, modern machine learning (ML) and deep learning (DL) offer powerful tools for water quality prediction [9–13], their application is often undermined by a **spatial blindness**.Many traditional models fail to account for spatial autocorrelation, the principle that geographically proximate data points exhibit similar characteristics. This oversight leads to unreliable predictions, as the models fail to generalize to new areas [14]. This issue is compounded by naive validation protocols, such as random k-fold cross-validation, which are statistically unsound for geospatial data and yield overly optimistic performance metrics [15, 16]. The core research problem, therefore, is to develop a new generation of models that are explicitly spatial-aware, rigorously validated, and intelligently optimized for real-world deployment.

Input: *(pH, Nitrate, Calcium, TDS (spatial data Longitude and Latitude (lon & Lat)).* The research utilizes raw hydrochemical parameters and spatial coordinates of groundwater samples.

Output: Quality classification of drinking water *(Excellent, Fair, Good, Poor, Medium).* The work produces a precise classification of water drinkability and generates spatially-aware risk maps.

Brief Review of Related works

Traditional approaches to groundwater assessment have historically relied on methods like the Water Quality Index (WQI), which aggregates multiple hydrochemical parameters into a single, easily communicable score [17, 18]. While useful for rapid screening, these index-based methods are constrained by subjective parameter weighting, an inability to capture complex non-linear interactions, and a lack of scalability for large, heterogeneous regions [19]. Their ineffectiveness and reliance on manual sampling make them ill-suited for the dynamic and large-scale challenges of modern water resource management [20].

The limitations of classical methods have catalyzed a shift towards data-driven techniques using Machine Learning (ML) and Deep Learning (DL) [21, 22]. Algorithms such as Support Vector Machines (SVM), Random Forest (RF), and Artificial Neural Networks (ANNs) have demonstrated a strong capacity to model the intricate, non-linear relationships between environmental factors and water quality indicators [23, 24]. These models can process high-dimensional datasets, improve predictive accuracy and provide insights into key contamination drivers [25].

However, a critical flaw in many standard ML and DL applications is "spatial blindness" the failure to account for spatial autocorrelation, the principle that proximal samples are inherently related [26]. This oversight can lead to unreliable predictions and flawed validation, as models may perform well on training data but fail to generalize to new geographic areas [27, 28]. Furthermore, the "black box" nature of many advanced models presents a challenge for interpretability, hindering their adoption by retaining and water managers [29].

To address these gaps, recent research has focused on developing spatially-aware and hybrid models. The integration of Geographic Information Systems (GIS) with ML/DL allows for the incorporation of critical spatial context, significantly improving model performance [30, 31]. Studies have demonstrated the effectiveness of hybrid approaches, such as combining Particle Swarm Optimization (PSO) with SVMs or using Convolutional Neural Networks (CNNs)

to extract spatial features from geospatial data [32, 33]. These advanced frameworks, which often include explainability tools like SHAP, represent the frontier of hydroinformatics, aiming to provide solutions that are not only accurate but also robust, scalable, and transparent [34].

Research Motivation

This research addresses the limitations of traditional groundwater quality monitoring by employing machine learning (ML) and deep learning (DL) methods, which offer scalable, real-time solutions that overcome issues such as slow processing, high cost, and limited scalability inherent in conventional methods. The motivation is threefold:

1. **To Overcome Manual Monitoring Constraints:** Replace inefficient, slow, and unscalable manual sampling with a robust, automated assessment framework that can handle the scale of modern environmental challenges.
2. **To Address Spatial Blindness in AI:** Mitigate the common limitation of AI models that overlook spatial autocorrelation by developing explicitly spatial architectures and applying rigorous spatially-aware validation protocols to obtain reliable and trustworthy performance.
3. **To Enhance Trust through Interpretability:** Bridge the adoption gap for "black box" models by using Explainable AI (XAI), making complex predictions transparent and actionable for captive and water managers.

Objectives of the Thesis

This research aims to develop, validate, and deploy a novel ensemble spatial machine learning framework for groundwater drinkability classification. With a primary focus on case studies in Vietnam's Mekong Delta and Odisha, India, the research is guided by the following specific objectives:

- **Develop and Benchmark of Machine Learning Models:** To establish a performance baseline with traditional algorithms (e.g., SVM, Random Forest) and subsequently develop novel hybrid models, namely

AI-LGBM and a Particle Swarm Optimized Spatial Convolutional Neural Network (**PSO-SCNN**), designed to achieve superior predictive accuracy.

- **Integrate Spatial Intelligence for Actionable Visualization:** To leverage Geographic Information System (GIS) techniques to transform model predictions into intuitive, high-resolution spatial risk maps, thereby identifying contamination hotspots.
- **Validate and Confirm Practical Utility in Real-World Scenarios:** To rigorously validate the proposed models using real-world groundwater datasets from Vietnam and Odisha, India, confirming their accuracy, robustness, and practical utility.
- **Incorporate Temporal Dynamics for Long-Term Monitoring:** To extend the models to analyze and predict changes in groundwater quality over time, enabling a framework for continuous assessment.

Scope of the Study

This section sets the boundaries of this investigation. Geographically, the research centers on groundwater quality in Vietnam's Mekong Delta and Odisha, India. Theoretically, the work is grounded in machine learning and spatial statistics, focusing on AI-driven models for environmental monitoring. The framework is defined by the following operational parameters:

The primary inputs for this research are raw hydrochemical parameters and the spatial coordinates of groundwater samples.
The principal outputs are a precise classification of water drinkability and the generation of spatially-aware risk maps.

Research Method

This study adopts a mixed-methods framework, blending quantitative machine learning with qualitative spatial analysis. The methodology involves several key phases:

Data Collection and Preprocessing from official sources (Vietnam's MONRE and India's CGWB); (2) **Model Development**, including baseline models (SVM, Random Forest) and the proposed hybrid frameworks (AI-LGBM, PSO-SCNN); (3) **Geospatial Visualization** using GIS to map model outputs; and (4) **Model Evaluation** using a suite of metrics (Accuracy, Precision, Recall, F1-Score, AUC) and robust k-fold cross-validation techniques.

Results of the Thesis

This thesis delivers significant scientific and practical contributions. The primary result is an advanced spatial analysis framework for hydroinformatics. The proposed hybrid models (AI-LGBM, PSO-SCNN) achieve up to **98.8% accuracy**, outperforming traditional methods by a margin of 8–13%. The development of the PSO-SCNN model, which combines spatial feature extraction with evolutionary optimization, stands as a key methodological innovation. Practically, the models facilitate early contamination detection and enable detailed spatial mapping to identify pollution hotspots, offering a direct and impactful tool for water resource management.

Contributions of the Thesis and Significance

This thesis makes significant contributions to the field of groundwater quality classification by introducing novel hybrid spatial machine learning models, specifically **AI-LGBM** and **PSO-SCNN**. These contributions address key challenges in groundwater monitoring, particularly spatial autocorrelation, model interpretability, and scalability. Below are the primary contributions:

Methodological Innovation: Hybrid Spatial Machine Learning Models

The key **methodological innovation** of this thesis is the **PSO-SCNN** model, a hybrid approach that integrates **Particle Swarm Optimization (PSO)** with **Spatial Convolutional Neural Networks (SCNN)** to improve predictions of groundwater quality. Traditional machine learning models often overlook spatial dependencies, leading to inaccurate predictions in regions with

clustered or spatially correlated data. The **PSO-SCNN** model overcomes this limitation by explicitly accounting for spatial features, significantly improving **classification accuracy** and **model robustness**. This novel integration of spatial data with machine learning models represents a substantial advancement in the prediction and monitoring of groundwater quality.

- **PSO-SCNN** combines the strengths of **CNNs** for spatial feature extraction and **PSO** for efficient hyperparameter optimization, ensuring that the model adapts effectively to complex hydrogeological conditions.
- This hybrid model addresses both the attribute-driven and spatial-contextual patterns of groundwater data, offering more accurate, scalable, and interpretable results than traditional methods.

Multi-Region Validation and Practical Decision Support

The models developed in this research have been validated in two geographically and hydrogeologically diverse regions: the **Mekong Delta, Vietnam**, and **Odisha, India**. This validation demonstrates the **generalizability** of the proposed models across different environmental contexts, enhancing their practical application for **groundwater management** in varied regions.

The research also provides **spatial risk maps** and identifies **pollution hotspots**, offering **actionable insights** for policymakers and groundwater resource managers. These visual tools support **proactive decision-making** in areas facing water quality challenges.

Model Explainability and Interpretability

A critical aspect of this work is its focus on making complex machine learning models more transparent and understandable. By integrating **SHAP** (Shapley Additive Explanations) and **LIME** (Local Interpretable Model-agnostic Explanations), this thesis ensures that stakeholders can interpret the factors influencing groundwater quality predictions. These explainability tools increase **trust** in the model's outcomes, promoting their adoption in real-world settings.

Practical Tools for Groundwater Management

The proposed hybrid models, particularly **PSO-SCNN**, offer a direct contribution to **real-time groundwater monitoring** and **contamination detection**, facilitating more **timely interventions** in water resource management. This practical utility is crucial for addressing the increasing global demand for safe drinking water, especially in resource-constrained regions.

Limitations of the Study

While this thesis provides important advancements in groundwater quality classification, several limitations should be acknowledged. These limitations highlight areas for further research and improvement:

Computational Complexity

The **PSO-SCNN** model, though highly effective, is **computationally intensive**, which limits its applicability in resource-constrained environments. The model's complexity requires substantial processing power, which could pose challenges for real-time monitoring or implementation in regions with limited computational resources.

Data Quality

The effectiveness of the models heavily relies on the quality of the input data. Groundwater quality data, particularly in regions with sparse monitoring, can often be incomplete or noisy. **Data sparsity** remains a significant challenge, especially in remote or under-monitored areas, which could impact the accuracy and robustness of the model in such regions.

Temporal Dynamics

Groundwater quality can vary over time due to factors such as seasonal changes, climatic conditions, and human activity. While the models developed in this study focus on snapshot assessments of water quality, they do not fully incorporate **temporal dynamics**. The ability to track changes in water quality

over time is essential for long-term monitoring, and this limitation could be addressed in future iterations of the model.

Generalization Across Regions

Although the models have been validated in two diverse regions—**the Mekong Delta (Vietnam)** and **Odisha (India)**—further validation in a wider range of hydrogeological settings is necessary to ensure the generalizability of the models. The findings from these two regions may not fully represent groundwater conditions in other parts of the world with different environmental and geospatial characteristics.

Scalability and Real-Time Application

The models developed in this thesis have shown promising results, but there are challenges related to **scalability** and **real-time application**. The computational demands of the hybrid models make it difficult to deploy them for large-scale, continuous monitoring, especially in regions with limited infrastructure for real-time data processing.

Interpretability and User-Friendliness

Despite efforts to enhance model interpretability using tools like **SHAP** and **LIME**, there is still room for improvement in making the models more **user-friendly** for non-expert stakeholders. Policymakers and groundwater managers may require more intuitive interfaces, real-time updates, and clearer visualizations to fully trust and utilize the model outputs in decision-making processes.

Structure of the Thesis

This thesis is structured into three main chapters to logically present the research from conception to conclusion.

- **Chapter 1: Groundwater Drinkability Classification** introduces the research context, problem statement, motivation, objectives, scope, methodology, and key results.

- **Chapter 2: Proposed Ensemble Spatial Machine Learning Methods** details the multi-phase methodology, from data collection to model development, outlining the mathematical foundations, optimization strategies, and evaluation methods.
- **Chapter 3: Results and Evaluations** presents the experimental findings, including a comparative performance analysis of the models, the spatial mapping results, and an assessment of each model's strengths and limitations.

Chapter 1

Groundwater Drinkability Classification

1.1 Introduction to Groundwater Drinkability Classification

Groundwater drinkability classification leverages machine learning (ML) and spatial information systems (GIS) to automate the classification of groundwater quality. By integrating hydrochemical parameters and spatial features, the system can provide scalable and real-time assessments of water quality, essential for decision-making in water resource management. This approach supports early warnings in areas with sparse and costly monitoring, with recent work (2022–2025) showing that ML/DL models, including tree ensembles and CNN-based architectures, outperform traditional index/rule-based methods, while maintaining operational value through explainability [35].

A key challenge is spatial dependence: random k-fold validation can inflate model performance when wells are clustered. Best practices therefore use *spatially blocked* or distance-aware cross-validation, explicit transfer tests, and clear quality metrics (Accuracy, Precision, Recall, F1, AUC) along with cost-aware thresholding (e.g., Youden's J when false negatives are costlier) [36].

Design Principles. Effective classification systems should: (i) combine hydrochemistry and geospatial predictors at appropriate scales; (ii) account for spatial structure (e.g., spatial convolutions); (iii) use *spatially blocked* cross-validation and transfer tests; (iv) report metrics with calibrated uncertainty [37]; (v) adopt

cost-sensitive thresholds [38]; and (vi) offer interpretable outputs (e.g., SHAP) that align with hydrogeochemical knowledge. Recent advancements in ML/DL have enhanced predictive accuracy and model interpretability in complex environments [39, 40].

Problem Formulation: Multi-class WQI Levels and Binary Drinkability

Each groundwater sample i is represented by an input feature vector $\mathbf{x}_i = [\mathbf{h}_i, \mathbf{s}_i]$, where $\mathbf{h}_i \in \mathbb{R}^n$ contains the measured hydrochemical/physicochemical variables (e.g., pH, TDS, nitrate, iron, etc., depending on the dataset), and $\mathbf{s}_i$ contains the spatial location attributes (latitude/longitude and derived spatial features when used). The Water Quality Index (WQI) for sample i is computed from parameter sub-indices:

$$q_{ij} = \left(\frac{V_{ij}}{S_j}\right) \times 100, \quad \mathrm{WQI}_i = \sum_{j=1}^{n} w_j \, q_{ij} \tag{1.1}$$

where V_{ij} is the observed value of parameter j for sample i, S_j is its guideline/standard value, and w_j is the assigned weight.

Task A: Multi-class groundwater quality classification (WQI-based). The primary classification task predicts an ordinal, five-level groundwater quality label Excellent, Fair, Good, Poor, Unsuitable

$$y_i^{(mc)} \in \{\text{Excellent}, \ \text{Good}, \text{Fair}, \text{Poor}, \text{Unsuitable}\} \quad \text{derived from} \quad \mathrm{WQI}_i : \tag{1.2}$$

$$y_i^{(mc)} = \begin{cases} \text{Excellent}, & 0 \le \mathrm{WQI}_i \le 25, \\ \text{Good}, & 26 \le \mathrm{WQI}_i \le 50, \\ \text{Fair}, & 51 \le \mathrm{WQI}_i \le 75, \\ \text{Poor}, & 76 \le \mathrm{WQI}_i \le 100, \\ \text{Unsuitable}, & \mathrm{WQI}_i > 100. \end{cases} \tag{1.3}$$

This formulation supports fine-grained risk grading and spatial risk mapping.

Task B: Binary drinkability classification (is-drinkable). For operational decision-making, a binary label $y_i^{(bin)} \in \{0, 1\}$ is also defined directly from the WQI classes:

$$y_i^{(bin)} = \begin{cases} 1 \text{ (drinkable)}, & y_i^{(mc)} \in \{\text{Excellent}, \text{Good}\}, \\ 0 \text{ (non-drinkable)}, & y_i^{(mc)} \in \{\text{Fair}, \ \text{Poor}, \text{Unsuitable}\}. \end{cases}$$

Accordingly, models configured with a sigmoid output layer are trained for Task B, while multi-class models output five probabilities for Task A. Unless otherwise stated, performance metrics are reported for the task being evaluated (multi-class grading or binary drinkability).

Problem 1: Groundwater Drinkability Classification

Groundwater quality varies depending on several physicochemical and environmental parameters, including pH levels, total dissolved solids (TDS), nitrate concentration, and spatial characteristics such as geographic coordinates. One of the primary objectives of this research is to develop a robust and reliable classification system that can assess whether specific groundwater samples are suitable for human consumption. The classification system categorizes water samples into predefined classes such as *Excellent, Good, Fair, Poor, Undrinkable* for drinkability.

Unlike conventional water quality index (WQI) calculations that rely on fixed thresholds and weights, the approach adopted in this study leverages machine learning algorithms to learn complex relationships from data. This enables the classification system to be more flexible and data-driven, capable of handling both linear and nonlinear interactions between variables. The ultimate goal is to automate and scale this classification task for broader geographic regions, enabling more timely and accurate groundwater quality assessments.

Mathematical Derivative

This can be framed as a classification problem where each water sample x_i is labeled with a quality category y_i from a set of predefined classes.

Let $X = \{x_1, x_2, \ldots, x_n\}$ represent the dataset of groundwater samples, where each sample $x_i \in \mathbb{R}^m$ is a vector of features $x_i = (x_{i1}, x_{i2}, \ldots, x_{im})$, consisting of physicochemical parameters (e.g., pH, TDS, nitrate concentration) and spatial features (e.g., geographic coordinates).

Each sample x_i is associated with a label $y_i \in \{1, 2, \ldots, k\}$, where k represents the number of classes for water quality (e.g., Excellent, Good, Fair, Poor, Undrinkable).

The classification model $f(X; W)$ maps the feature vector x_i to the predicted label $\hat{y}_i$ as follows:

$$\hat{y}_i = f(x_i; W) \tag{1.4}$$

Where W represents the hyperparameters of the model, and the objective is to find the model that minimizes the classification error. The performance of the model is typically evaluated using accuracy, F1-score, or other classification metrics.

Objective:

The objective is to minimize the classification error, which can be expressed as:

$$\hat{y}_i = \arg\min_{\hat{y}} \mathcal{L}(y_i, \hat{y}_i) \tag{1.5}$$

Where $\mathcal{L}$ is the loss function, such as Cross-Entropy Loss or Mean Squared Error, that measures the discrepancy between the predicted label $\hat{y}_i$ and the true label y_i.

The standard formula for Model Optimization for supervised learning, model optimization is:

General form (minimizing empirical risk)

$$W^* = \arg\min_{W} \frac{1}{n} \sum_{i=1}^{n} L\left(y_i, f\left(x_i; W\right)\right) \tag{1.6}$$

Where:

- W = model parameters (weights, bias, tree structure, etc.)
- $f(x_i; W)$ = model prediction
- L = loss function (cross-entropy, MSE, hinge, ...)
- n = number of training samples

This is called empirical risk minimization (ERM) and is the standard formulation for ML model training.

For classification (cross-entropy loss)

$$W^* = \arg\min_W -\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{K}\mathbf{p}\left(y_i = c\right)\log p_W\left(y = c \mid x_i\right) \tag{1.7}$$

Problem 2: Optimizing Hyperparameters for GWQC Models

A further challenge in the development of an accurate groundwater classification system is the optimization of hyperparameters within machine learning models. These hyperparameters such as tree depth, learning rate, number of estimators, and regularization coefficients play a crucial role in determining the model's performance. Poorly chosen hyperparameters can lead to underfitting, overfitting, or excessive computational costs.

This research addresses the issue by employing advanced optimization strategies such as Optuna and Particle Swarm Optimization (PSO). These algorithms automate the process of selecting the best-performing hyperparameter configurations. The aim is to enhance model accuracy, stability, and generalization performance across diverse environmental datasets from regions such as Vietnam and India. Through systematic optimization, the models are tailored to deliver more precise predictions, even in the presence of noisy or high-dimensional data.

Mathematical Derivative

The second problem involves the optimization of hyperparameters for a predictive model that classifies groundwater quality, aiming to improve the ac-

curacy and efficiency of the model. This can be formulated as an optimization problem where the goal is to find the optimal hyperparameters W^* that maximize the model's performance.

Let $\mathcal{L}(y_i, f(x_i; W))$ represent the loss function used to evaluate the classification performance of the model. The objective is to find the optimal set of hyperparameters W^* that minimize this loss function across the training dataset:

$$W^* = \arg\min_W \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i, f(x_i; W)) \tag{1.8}$$

Where:

- W is the set of hyperparameters to be optimized.
- n is the total number of samples in the training set.
- $f(x_i; W)$ is the model's prediction for the input sample x_i with hyperparameters W.

Objective:

The objective is to maximize the performance function $g(W)$, which could be accuracy, F1-score, or another relevant metric. The optimization is expressed as:

$$W^* = \arg\max_W g(W) \tag{1.9}$$

Where $g(W)$ is the performance function of the model, and the optimization algorithm seeks to find the optimal W^*.

This process can be done iteratively, where the model is evaluated with different sets of hyperparameters, and the optimal configuration is determined based on maximizing $g(W)$.

Problem 3: Spatial Visualization of Classified Labels on a Map

The third problem focuses on visualizing the classified labels on a map for decision-making. By integrating the classification results with geographic

information system data, it becomes possible to display the groundwater quality classification on a map, aiding decision-makers in understanding spatial patterns and making informed decisions.

Mathematical Derivative

Let $G = \{(lat_1, lon_1), (lat_2, lon_2), \ldots, (lat_n, lon_n)\}$ represent the geographic coordinates of the groundwater samples. Let $\hat{y}_i$ represent the predicted groundwater quality class for sample x_i, where $\hat{y}_i \in \{1, 2, \ldots, k\}$.

The goal is to map each classified label $\hat{y}_i$ to its corresponding geographic location (lat_i, lon_i) and visualize the spatial distribution of groundwater quality on a map.

The spatial map M can be expressed as:

$$M = \text{GIS}(G, \hat{y}) \tag{1.10}$$

Where:

- G represents the geographic coordinates of the groundwater samples.
- $\hat{y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}$ represents the predicted groundwater quality labels.
- $\text{GIS}(G, \hat{y})$ maps the predicted labels $\hat{y}_i$ to their corresponding geographic locations for visualization.

Objective:

To combine the classification and spatial mapping, the objective function becomes:

$$L_{\text{total}} = L_{\text{classification}} + \lambda L_{\text{spatial}} \tag{1.11}$$

Where:

- $L_{\text{classification}}$ is the classification error (e.g., Cross-Entropy Loss),
- L_{spatial} is the spatial error (misalignment of predicted labels with actual geographic coordinates),

- λ is a regularization parameter controlling the importance of spatial mapping.

Groundwater quality is essential for global water supply, with contamination posing significant health risks. It is a primary drinking water source for billions, yet pollutants like heavy metals, nitrates, and pesticides threaten its safety [41], [42]. Traditional methods, though effective, are costly, time-consuming, and lack real-time capabilities, relying on manual sampling and lab tests [43].

Machine learning and Deep Learning offer solutions, using large datasets for real-time predictions and classification, thus improving monitoring efficiency. However, challenges related to scalability, accuracy, and interpretability remain [44].

1.2 Research Context

1.2.1 Classical Methods

Classical methods for groundwater quality assessment, such as the Water Quality Index (WQI), aggregate water quality attributes (e.g., pH, TDS, nitrates) into a composite score. While interpretable, WQI has limitations, including subjective parameter weighting, inability to model complex, non-linear interactions, and poor scalability for large or heterogeneous regions [5].

Groundwater quality, crucial for billions of people, is influenced by complex hydrogeochemical processes, land use, and climate variability, which vary over space and time [45, 46]. Traditional methods, including manual sampling and laboratory analysis, are costly, slow, and spatially limited, hindering timely risk assessments [47, 48]. Recent advancements in information systems and machine learning have demonstrated that hybrid spatial models, integrating geospatial data with machine learning algorithms like XGBoost, LightGBM, and CNNs, can improve predictive accuracy for water quality classification. These systems offer enhanced scalability and interpretability compared to traditional methods. However, naive validation methods, like random k-fold, often overestimate performance with spatially autocorrelated data, emphasizing the need for *spatially*

blocked evaluation methods [49, 50].

To address these challenges, spatially aware and interpretable models are critical for public health and resource management. Deep learning (DL) and gradient-boosting ensembles offer real-time inference and scalability, while explainable AI (XAI) frameworks, such as SHAP, improve trust by linking predictions to domain-relevant drivers [51, 52]. This study uses spatially informed architectures and spatial cross-validation to deliver robust groundwater drinkability classifications for early warning, mitigation prioritization, and long-term planning [53, 54].

Equation & Steps for Calculating WQI

The Water Quality Index (WQI) combines multiple water quality parameters into a single score:

$$\text{WQI} = \sum_{i=1}^{n} w_i \times Q_i \tag{1.12}$$

where Q_i is the sub-index score for each parameter, and w_i is the weight for the i-th parameter.

Table 1.1: *Summary of Classical Hydrological Methods*

Method Type	**Examples**
Physical Methods	Visual inspection (e.g., Secchi disk for turbidity), temperature measurement
Chemical Methods	Winkler titration for dissolved oxygen, colorimetric tests (e.g., DPD method for chlorine)
Biological Methods	Most Probable Number (MPN) for coliform detection, membrane filtration for microbial analysis

These methods use standard laboratory techniques, with observed parameter values shown in Table 1.2.

Table 1.2: *Observed Values of Water Quality Parameters*

Parameter	**Unit**	**Sample 1**	**Sample 2**
pH	-	7.2	7.5
Total Dissolved Solids (TDS)	mg/L	250	300
Nitrate (NO_3^-)	mg/L	15	20
Total Coliforms	CFU/100mL	10	5

Assigning Weights to Parameters

Each parameter is assigned a weight (w_i) reflecting its importance, as shown in Table 1.3. The weights sum to 1.

Table 1.3: *Assigned Weights to Water Quality Parameters*

Parameter	**Weight (w_i)**
pH	0.15
Total Dissolved Solids (TDS)	0.20
Nitrate (NO_3^-)	0.25
Total Coliforms	0.40
Total Weight	**1.00**

Determining Sub-Index Values

The sub-index (q_i) for each parameter is calculated as:

$$q_i = \left(\frac{V_i}{S_i} \right) \times 100 \tag{1.13}$$

where V_i is the observed value and S_i is the standard value for the i-th parameter.

Table 1.4: *Calculated Sub-Indices for Water Quality Parameters*

Parameter	**Standard (S_i)**	**Sample 1 (q_{i1})**	**Sample 2 (q_{i2})**
pH	7.5	$\left(\frac{7.2}{7.5}\right) \times 100 = 96$	$\left(\frac{7.5}{7.5}\right) \times 100 = 100$
TDS	500 mg/L	$\left(\frac{250}{500}\right) \times 100 = 50$	$\left(\frac{300}{500}\right) \times 100 = 60$
Nitrate (NO_3^-)	50 mg/L	$\left(\frac{15}{50}\right) \times 100 = 30$	$\left(\frac{20}{50}\right) \times 100 = 40$
Total Coliforms	0 CFU/100mL	∞ (Special handling)	∞ (Special handling)

Note on Parameters with Zero Standard: For parameters like Total Coliforms, where the standard is zero, high sub-index values are assigned to indicate risk.

Water Quality Classification Based on WQI

The overall WQI is calculated as:

$$WQI = \sum_{i=1}^{n} (w_i \times q_i) \tag{1.14}$$

Table 1.5: *Water quality classification based on WQI values for the MCDA-based water quality assessment*

WQI Range	**Water Quality Class**
0–25	Excellent Water Quality
26–50	Good Water Quality
51–75	Poor Water Quality
76–100	Very Poor Water Quality
>100	Unsuitable for Drinking

In summary, classical methods like the WQI provide a simple, interpretable approach to groundwater quality assessment [55], [56], [57], [58], [59]. However, they are limited by subjective parameter weighting, lack of scalability, and inability to capture complex interactions, underscoring the need for more robust methods discussed in the following section.

1.2.2 ML/DL Methods

Machine learning (ML) methods, such as Support Vector Machines (SVM), Random Forest (RF), and XGBoost, model complex, non-linear relationships between environmental factors and water quality. These methods handle high-dimensional datasets and provide better predictions than classical models [60–62]. However, ML models are challenged by the need for large, high-quality datasets and potential overfitting with noisy data.

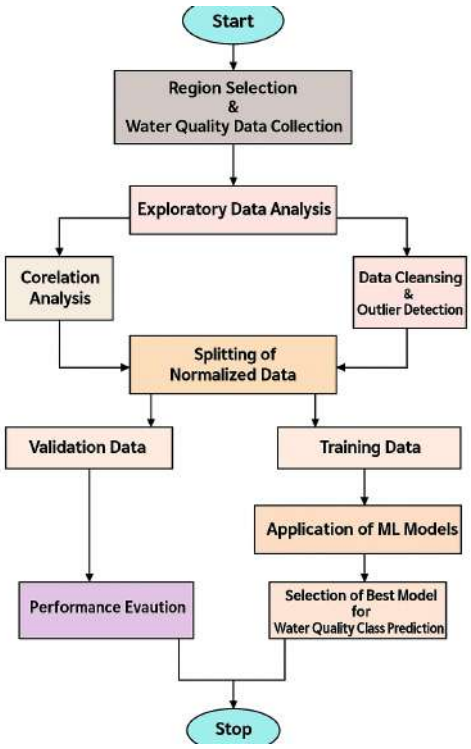


Figure 1.1: illustrative flow diagram for water quality analysis and classification

Recent advancements in machine learning (Table 1.6) have been applied to large-scale pattern recognition and predictive modeling [63–65].

Table 1.6: *Machine Learning Methods for Hydrological Water Quality Assessment*

Machine Learning Method	**Description and Applications**
Support Vector Machines (SVM)	Used for classifying groundwater quality, especially in small or imbalanced datasets.
Random Forest (RF)	Ensemble method for predicting pollutant levels and assessing environmental factors.
Artificial Neural Networks (ANN)	Models relationships between environmental factors and water quality indicators.
K-Nearest Neighbors (KNN)	Classifies water samples based on features like temperature and turbidity.
Gradient Boosting Machines (GBM)	Methods like XGBoost and LightGBM improve accuracy through iterative learning.
Clustering Algorithms (e.g., K-Means)	Groups water samples to identify pollution patterns.

ML methods process high-dimensional real-time data, capture non-linear relationships, and improve prediction accuracy, revealing key contamination drivers [66–68]. They support decision-making in water management [69], but challenges remain with dataset quality and overfitting [70, 71].

Index and Rule-Based Approaches

Index-based methods like the Water Quality Index (WQI) aggregate parameters into a single score, making them simple and transparent for rapid screening [72, 73]. However, they suffer from subjective weighting and difficulty capturing nonlinear interactions [74]. Recent ML/DL techniques often outperform such rule-based schemes in complex scenarios.

ML/DL Classifiers and Ensembles

Modern approaches treat drinkability as a supervised classification or exceedance-prediction problem, combining hydrochemistry and geospatial data. Tree ensembles (RF, XGBoost/LightGBM) and deep learning (e.g., CNNs with spatial proxies) learn non-linear interactions, improving prediction accuracy with interpretability through tools like SHAP [75]. Best practices emphasize spatially blocked cross-validation and transfer tests to avoid inflated model performance due to clustered wells [76].

Deep Learning (DL) Methods

Deep learning (DL) methods, especially CNNs and LSTMs, are effective at capturing complex spatial and temporal patterns in groundwater data. These models excel with large datasets and can model intricate relationships that traditional ML models may miss. However, they are computationally intensive and suffer from interpretability issues, though tools like SHAP provide insights into feature contributions.

Table 1.7 summarizes key DL methods in hydrology, emphasizing their spatial and time-series applications.

Table 1.7: *Deep Learning Methods in Hydrology*

Deep Learning Method	**Description and Applications**
CNN	Used for analyzing spatial data like satellite images to detect water body conditions.
RNN	Used for time-series forecasting, such as predicting river flow and groundwater levels.
LSTM	Effective for predicting long-term trends in water quality.
Autoencoders	Used for anomaly detection and dimensionality reduction in water quality data.
GANs	Generate synthetic data and simulate water quality scenarios.
DRL	Optimizes water resource management, e.g., flood control and irrigation strategies.
FCN	Predicts variables like river discharge and groundwater levels by integrating diverse data sources.
Hybrid Models	Combine multiple DL techniques to improve predictions by integrating spatial and temporal data.

In summary, ML and DL methods significantly enhance groundwater quality assessment by capturing non-linear relationships, processing high-dimensional data, and providing accurate predictions. However, they face challenges with data quality, overfitting, and interpretability. Despite these issues, DL methods like CNNs and LSTMs outperform traditional models in spatial and temporal pattern recognition, offering state-of-the-art performance [77–81].

Section 1.2.3 will explore Hybrid Spatial Models that combine ML/DL with geospatial data to address these challenges and improve scalability and interpretability in groundwater quality classification.

1.2.3 Hybrid Spatial Models

Hybrid spatial models combine machine learning (ML) and deep learning (DL) techniques with geospatial data to improve groundwater quality predictions. By integrating environmental, geological, climatic, and remote sensing data, these models capture spatially dependent interactions, enhancing the accuracy of contamination risk predictions in heterogeneous regions [49].

Geostatistical Interpolation

Geostatistical methods like kriging (ordinary, indicator, and co-kriging) model constituents as spatial random fields, producing continuous concentration surfaces with kriging variances. These methods are useful for mapping exceedance thresholds and guiding uncertainty-aware sampling in areas with sparse wells [82, 83]. However, they are limited by the assumption of stationarity and challenges in capturing cross-parameter interactions [84]. Geospatial ML highlights the importance of spatially blocked cross-validation to avoid overestimating accuracy in interpolative models [85].

Integration of Geospatial Data with ML/DL Models

Integrating Geographic Information System (GIS) data with ML/DL algorithms addresses spatial dependencies in groundwater quality, leading to more robust predictions. For example, the Particle Swarm Optimization Spatial Convolutional Neural Network (PSO-SCNN) optimizes spatial features for improved classification accuracy [52]. This approach aids in identifying contamination hotspots and predicting water quality in data-limited regions, providing crucial insights for water resource management [86].

Ensemble Learning with Spatial Features

Ensemble learning methods like Random Forest (RF), XGBoost, and LightGBM improve prediction accuracy by incorporating spatial features. For instance, integrating LightGBM with spatial features through Mutual Information Feature Selection (MIFS) enhances accuracy while preserving essential spatial information [54]. Hybrid methods such as PSO-SCNN, which combine spatial convolutions and optimization, effectively capture spatial features and improve prediction performance [87].

Geospatial Mapping and Risk Prediction

Hybrid models integrating ensemble learning with GIS-based spatial mapping techniques provide accurate groundwater risk predictions. These models generate spatially aware risk maps, aiding water resource management and

decision-making. For example, the CNN-GIS approach uses convolutional neural networks with GIS data to analyze spatial patterns and predict contamination levels [88]. These models enable real-time assessment of water quality and support targeted mitigation strategies [89].

Proposed Solution for Water Classification Challenges

Machine learning (ML) and deep learning (DL) are crucial for addressing the inherent complexities of groundwater drinkability classification. Traditional methods struggle with non-linear relationships, large datasets, and spatial dependencies that often arise in environmental monitoring. These challenges are compounded by the need for real-time analysis and scalable solutions. Machine learning models, such as XGBoost, AI-LGBM, and PSO-SCNN, excel in capturing intricate data relationships and spatial patterns, enabling more accurate and efficient predictions of groundwater quality. Additionally, the integration of spatial data with machine learning algorithms allows for enhanced predictive accuracy and interpretability, which traditional models fail to achieve. These capabilities make machine learning an indispensable tool for real-time monitoring, decision support systems, and policy-making in groundwater management

Challenges and Future Directions

Despite the advantages of hybrid spatial models, challenges remain, particularly regarding computational demands and model interpretability. These models, while offering improved accuracy, require significant computational resources, especially for large datasets [90]. Additionally, deep learning models are often seen as "black-box" models, making interpretability a key challenge [91].

Future research should focus on optimizing these models for real-time applications by improving computational efficiency and enhancing interpretability. Incorporating diverse data sources, such as satellite-based remote sensing and IoT sensors, could improve both accuracy and scalability [92].

In conclusion, hybrid spatial models combine the strengths of machine learning, deep learning, and spatial data, providing enhanced predictive capa-

bilities and valuable insights into groundwater contamination risks, supporting better decision-making in water resource management [93, 94].

1.2.4 Gaps and Summary

Despite advances in ML and DL for groundwater quality prediction, several limitations persist. The need for large, high-quality datasets remains a challenge, particularly in data-sparse regions. Hybrid models are computationally intensive, limiting their real-time application, and deep learning models often lack interpretability, hindering their practical use.

This research seeks to address these limitations by developing spatially aware, accurate, and interpretable models for groundwater quality classification. Future improvements should focus on model scalability, handling incomplete data, and ensuring stakeholder interpretability.

Research Gaps: Key challenges include:

- Difficulty capturing non-linear interactions between attributes and integrating diverse data sources.
- Limited real-time analysis capabilities and end-to-end automation, restricting scalability.
- Handling uncertainty in noisy or incomplete data.
- Hybrid modeling combining domain knowledge and data-driven approaches is underexplored.

Research needs include:

- Enhancing data reliability with standardized protocols and real-time monitoring.
- Integrating ML with traditional methods for data-sparse regions.
- Developing hybrid models combining ML, DL, and geospatial analysis.
- Addressing interpretability issues using explainable AI.

- Scaling models for real-time and large-scale groundwater quality management.

1.2.5 Research Method to Address Gaps

Previous sections highlighted the limitations of traditional ML models like Random Forest (RF) and Support Vector Machines (SVM), which struggle with spatial dependencies and accuracy in groundwater quality prediction. Deep learning models, while excelling at feature extraction, often neglect spatial context, limiting their effectiveness in large-scale applications. Spatial-CNNs, though incorporating spatial data, fail to fully model GIS dependencies, limiting scalability across diverse environments.

To overcome these challenges, we propose a hybrid framework combining the AI-enhanced Light Gradient Boosting Machine (AI-LGBM) and Particle Swarm Optimization Spatial Convolutional Neural Network (PSO-SCNN). This framework addresses both attribute-driven and spatial-contextual patterns for more accurate and scalable groundwater quality prediction.

Key Components of the Hybrid Framework:

- **Feature Fusion:** Combines hydrogeochemical data and spatial coordinates using embedding layers and attention mechanisms for superior predictive performance.
- **Hybrid Architecture:** Integrates CNN for spatial pattern recognition, Random Forest for interpretability, and AI-LGBM for robust classification.
- **Optimization:** Uses Grid Search, PSO, and Genetic Algorithms to fine-tune hyperparameters, ensuring model stability and generalization.

A schematic diagram in Figure 2.1 illustrates the dual-stream processing of hydrogeochemical and spatial features.

Key Advantages and Applications:

This hybrid framework enhances accuracy through CNN-based feature extraction and ensemble methods, improves spatial prediction with spatial em-

beddings and attention mechanisms, and offers scalability for deployment in regions like Odisha and the Mekong Delta. It enables real-time groundwater monitoring for proactive management and contamination detection, supporting sustainable groundwater governance and aiding policymakers and environmental managers in climate-sensitive regions.

Research Design

The research adopts a quantitative, experimental approach, combining data analysis with ML. It follows sequential steps: data collection, model training, validation, and evaluation, focusing on a hybrid spatial model to enhance prediction accuracy, scalability, and interpretability in groundwater quality management.

1.3 Study Areas: India and Vietnam

This study focuses on Odisha, India, and the Mekong Delta, Vietnam—two regions with high groundwater dependency, documented vulnerability to contamination, and sufficiently rich monitoring datasets for machine learning and spatial analysis.

1.3.1 Mekong Delta, Vietnam

The Mekong Delta in southern Vietnam spans approximately 39,000 km^2 and is traversed by a dense network of rivers and canals. Figure 1.2 illustrates its geographic scope and provinces.

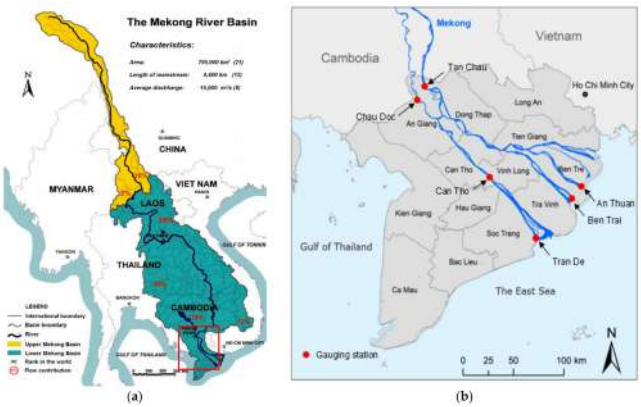


Figure 1.2: Geographical Context of Study Areas (a) Location of the Mekong Delta (Source: Mekong River Commission); (b) Provincial Extent within the Mekong Delta.

The Mekong Delta Vietnam's agricultural heartland—relies heavily on shallow alluvial aquifers. Groundwater stress arises from over-extraction, saltwater intrusion, industrial effluents, and agricultural runoff. The dataset (MONRE) contains **2,139** records with physicochemical measurements and spatial coordinates.

Table 1.8: Descriptive Statistics of Groundwater Parameters in the Mekong Delta (Vietnam)

Parameter	**Minimum**	**Maximum**	**Mean**
pH	5.6	8.3	≈ 6.8
TDS (mg/L)	95	2,300	≈ 641
Nitrate (mg/L)	1.5	77	≈ 25
Iron (mg/L)	0.02	3.7	≈ 1.21
Additional Info: Latitude and Longitude of sampled wells			

These characteristics make the Mekong Delta a critical testbed for automated water-quality assessment models that integrate spatial signals (e.g., GIS, remote sensing).

1.3.2 Odisha, India

Odisha, in eastern India, comprises diverse hard-rock and alluvial aquifers. Districts such as Ganjam and Mayurbhanj frequently report exceedances of nitrate, iron, and fluoride. Groundwater is vital for drinking and irrigation. Figure 1.3 shows the hydro-geological map of the study area.

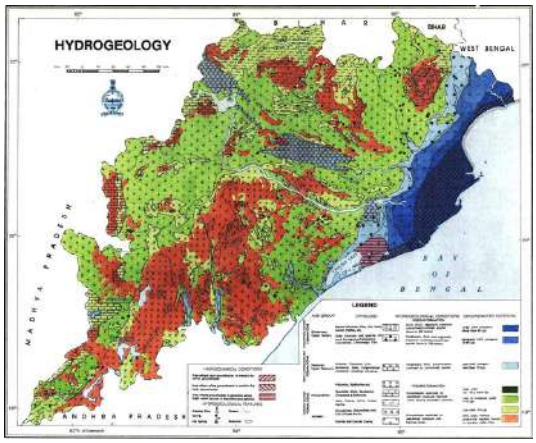


Figure 1.3: Hydro-geological map of the Odisha study area.

Odisha supplied data from CGWB/state monitoring stations in and around

Bhubaneswar and multiple districts, totaling **1,241** samples with physicochemical attributes.

Table 1.9: Descriptive Statistics of Groundwater Parameters in Odisha (India)

Parameter	**Minimum**	**Maximum**	**Mean**
pH	5.4	8.9	≈ 7.05
TDS (mg/L)	110	2,500	≈ 874
Nitrate (mg/L)	2	90	≈ 32
Iron (mg/L)	0.1	4.3	≈ 1.65
Additional Parameters: EC, TH, Ca, Mg, Cl, SO_4, F			

Odisha's hydro-geological variability and contamination complexity provide a rigorous environment for evaluating predictive models.

1.3.3 Hydrological Context & Site Rationale

Hydrological characteristics. The Mekong Delta is a low-lying deltaic aquifer system strongly influenced by seasonal flooding and salinity intrusion, whereas Odisha comprises mixed inland aquifers with groundwater quality variability driven by both geogenic (e.g., fluoride/iron) and anthropogenic pressures. S

Comparative Overview of the hydrological profile Vietnam & Odisha India

Table 1.10: Comparison of Hydrological Characteristics

Feature	**Mekong Delta, Vietnam**	**Odisha, India**
Main Rivers	Mekong River distributaries	Mahanadi, Brahmani, Baitarani, Rushikulya
Aquifer Type	Shallow, unconfined alluvial aquifers	Confined and semi-confined; hard rock and alluvial
Major Stress Factors	Salinity intrusion, agrochemical runoff	Nitrate, fluoride, iron contamination
Pollution Sources	Agriculture, aquaculture, seawater ingress	Industry, mining, agriculture, geogenic processes
Seasonal Influence	Monsoonal floods and dry-season salinization	Monsoon rains, cyclones, erratic recharge
Land Use	Intensive rice farming and aquaculture	Agriculture, industry, and mining

These differing hydrological profiles influenced model configuration and performance, particularly in the PSO-CNN architecture, where spatial features such as proximity to river systems and land use types significantly contributed to prediction accuracy.

Rationale for selecting these regions. Both areas rely heavily on groundwater for domestic, agricultural, and industrial needs and provide ample, well-structured datasets for training and evaluation. Their contrasting geology, land-use patterns, and contamination sources create a robust test of model generalizability.

Table 1.11: Justification for Selecting Odisha and the Mekong Delta

Criteria	**Description**
Groundwater Dependency and Vulnerability	Both regions rely heavily on groundwater. Odisha faces fluoride, iron, and salinity issues, while the Mekong Delta struggles with arsenic contamination, salinity intrusion, and over-extraction, exacerbated by climate change.
Hydrogeological Diversity	Odisha features diverse terrain from coastal plains to hilly interiors. The Mekong Delta is a flat, deltaic system shaped by rivers and tides, offering varied hydrogeological conditions for testing ML models.
Data Availability	Odisha's data come from CGWB and the state's groundwater department; Mekong Delta data are sourced from MONRE and open-access studies, enabling spatial ML applications.
Policy Relevance	The research aligns with India's "Har Ghar Jal" initiative and Vietnam's water security and sustainability goals, supporting policy-making.
Addressing Research Gaps	There has been limited use of hybrid GIS-ML models in these regions. This study fills that gap with advanced classification and mapping approaches.

These differing hydrological profiles influenced model configuration and performance especially for the Spatial CNN architecture where spatial features such as proximity to river systems and land-use classes contributed measurably to predictive accuracy.

1.4 Evaluation Metrics & Scenario

Our experimental workflow used a stratified 70/15/15 split for training, validation, and testing. Preprocessing involved imputing missing values, normalizing features via the Z-score method (Equation (2.11)), and removing IQR-

based outliers. Model training and hyperparameter tuning were conducted in Python 3.10 with Scikit-learn 1.2.2, LightGBM 3.3.2, and Optuna 3.0.0. All metrics are an average of five runs using different random seeds.

Step 1: Data Acquisition

This study utilizes two distinct groundwater quality datasets. The first dataset, comprising 1,052 samples from Vietnam's Mekong Delta, includes physicochemical attributes such as pH, TDS, nitrate, chloride, sulfate, and hardness, along with spatial coordinates. A second dataset of 1,241 samples with similar parameters was sourced from the Central Ground Water Board (CGWB) in Odisha, India.

Step 2: Data Preprocessing

The data preprocessing pipeline included several key steps: missing values were imputed using mean/median and mode, outliers were removed using the IQR method, and both physicochemical features and spatial coordinates were scaled to a [0,1] range via Min-Max normalization.

Feature Engineering Preprocessing Normalization method

The preprocessing pipeline involved imputing missing values, binarizing features according to permissible water quality standards, and normalizing all numerical data with the following formula:

$$F^*(x) = \frac{F(x) - \mu}{\sigma(F(x))} \tag{1.15}$$

where $F(x)$ is the original feature value, μ is the mean, and $\sigma(F(x))$ is the standard deviation. This step helped standardize the data and improved the model's convergence during training.

We also introduced new features using Water Quality Index (WQI) relations, where higher scores indicate better water quality. Finally, numerical features were normalized for consistent analysis, as shown in Eq. (2.27).

$$F^*(x) = \frac{F(x) - \overline{F(x)}}{\sigma(F(x))} \tag{1.16}$$

where:

$$\overline{F(x)} = \frac{\sum_{i=1}^{n} F(x_i)}{n} \tag{1.17}$$

$$\sigma(F(x)) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(F(x_i) - \overline{F(x)})^2} \tag{1.18}$$

After standardizing the features using Eq. (2.11) along with the details in Eq. (2.12) and Eq. (2.13), we used the preprocessed dataset for further analysis.

Feature Selection

Mutual Information-based Feature Selection (MIFS) was applied, reducing dimensionality to 14 essential features for efficient training without accuracy loss.

Step 3: Class Balancing

After balancing the classes with SMOTE, we trained multiple models. We benchmarked standard classifiers like Random Forest and XGBoost against our proposed AI-LGBM model, whose hyperparameters were tuned using Auto-Immune Optimization (AIO) and Optuna.

Step 4: Model Evaluation

To measure and compare model performance comprehensively, the following evaluation metrics were used:

Table 1.12: *Evaluation Metrics for Model Performance*

Metric	**Description**
Accuracy	Proportion of correct predictions overall.
Precision	True positives among predicted positives.
Recall (Sensitivity)	True positives among actual positives.
F1 Score	Harmonic mean of precision and recall; handles class imbalance.
AUC-ROC	Performance across all classification thresholds.

Model interpretability was also analyzed using SHAP (SHapley Additive exPlanations) to identify dominant features and explain the contribution of each parameter in the classification process.

Together, these components form the methodological backbone of the study, ensuring that the comparison of models is both statistically sound and contextually meaningful for water quality management applications.

Experimental Configuration

Data are preprocessed (outlier removal, normalization, imputation) and features selected via Mutual Information (MIFS). Three hybrid models were then developed:

Table 1.13: *Hybrid Spatial-AI Models Used in Groundwater Classification*

Model	Description
AI-LGBM	An enhanced LightGBM framework that integrates adaptive learning rate tuning and ensemble optimization using AIO, grid search, and cross-validation to improve classification accuracy and stability.
PSO-SCNN	A Spatial Convolutional Neural Network whose hyperparameters (kernel size, stride, and learning rate) are optimized using Particle Swarm Optimization (PSO) to improve spatial feature extraction.
CNN-GIS	A hybrid model combining CNN architecture with geospatial embedding techniques, designed to simultaneously capture hydro-chemical variations and geographic spatial dependencies of groundwater samples.

Each model was trained separately on both regional datasets. The experiments were repeated five times with different random seeds, and 5-fold stratified cross-validation was applied to prevent bias and variance issues.

5. GIS Integration

Model predictions were converted to GeoTIFF using GeoPandas and visualized in ArcGIS, enabling spatial mapping of groundwater quality. Heatmaps were overlaid with known contamination zones for effective decision support.

6. Hardware and Software Environment

Experiments were conducted on Apple M1 Max (64 GB RAM, 32 cores) and Intel i7 (32 GB RAM, GTX 1650 GPU) systems. The software stack included

Python 3.10, Scikit-learn, Keras, Optuna, SHAP, GeoPandas, and QGIS 3.28 for geospatial processing and visualization.

Baseline Models for Groundwater Quality Classification

The table 1.14 summarizes baseline models commonly used for groundwater quality classification. Each model offers distinct strengths for handling different data characteristics and classification challenges.

Table 1.14: *Baseline Models for Groundwater Quality Classification*

Model	**Description**
Logistic Regression	A linear model used for binary classification, predicting groundwater quality as safe or contaminated.
Decision Tree	Tree-based model that splits data based on feature thresholds to classify groundwater quality.
Support Vector Machine (SVM)	Effective for classification tasks, especially with small or imbalanced datasets.
Random Forest	An ensemble of decision trees that improves classification accuracy and reduces overfitting.
K-Nearest Neighbors (KNN)	Classifies samples based on the majority class of nearest neighbors in feature space.

1.5 Data Sources

The study used two datasets: the **Vietnam dataset** from the Ministry of Natural Resources and Environment, which includes physicochemical parameters and spatial data, and the **Odisha dataset** from the CGWB Ground Water Yearbook (2018–2020), containing 1,241 rows of physicochemical data. Both datasets provide essential inputs for groundwater quality assessment, with data collected through field sampling, expert review, and laboratory testing, following quality assurance protocols.

Dataset Quality and Challenges

While the data is generally of high quality, there are some challenges, including missing samples for Vietnam regions and potential seasonal variability that is not captured due to the limited temporal coverage.

Impact on Model Selection

Given the spatial nature of the data, **PSO-SCNN** was selected as the primary model due to its ability to handle spatial autocorrelation effectively. Additionally, the **AI-LGBM** model was chosen for its efficiency in handling high-dimensional data while maintaining interpretability. The imbalanced nature of the data was addressed by employing class balancing techniques, such as **SMOTE**.

Vietnam Dataset Overview

The Vietnam dataset contains water quality measurements from 2139 wells, with 40 columns representing various attributes across multiple time points. The dataset includes 2 datetime columns, 31 numeric columns, and 6 categorical columns. Key columns include water quality parameters such as pH, conductivity, and TDS, with some missing values in certain attributes like PO4, oxygen, and carbon. The data is structured in a 2-dimensional format, with 2139 rows and 40 columns.

Table 1.15: *Dataset Overview and Column Types*

Aspect	**Details**
Number of Rows	2139
Number of Columns	40
Datetime Columns	date_sampling, date_analyzing
Numeric Columns	na, k, ca2, ph, conductivity, tds105
Categorical Columns	well_code, quarter, laboratory, color
Missing Values	PO4, eh, Oxygen, Lienhe, Carbon
Dimensions	2-Dimensional (2139 rows, 40 columns)

Indian Dataset Overview

The Indian water quality dataset is well-organized, containing 1241 rows and 17 columns, with no missing values. It includes 14 numeric columns representing water quality parameters such as pH, EC, TDS, and alkalinity, and 2 categorical columns for district and village. The dataset provides a comprehen-

sive representation of water quality across different villages and districts.

Table 1.16: *Dataset Overview and Column Types for Indian Water Quality Dataset*

Aspect	**Details**
Number of Rows	1241
Number of Columns	17
Datetime Columns	None
Numeric Columns	pH, EC, TDS, TH, Alkalinity
Categorical Columns	District, Village
Missing Values	None

Data Preprocessing, Balancing, and Evaluation Strategy

The groundwater dataset was preprocessed for integrity and reliability by converting columns to numeric types, imputing missing values with column means, and removing columns with excessive missing data.

Outlier impact was mitigated by using robust tree-based models, which handle moderate outliers effectively in environmental data.

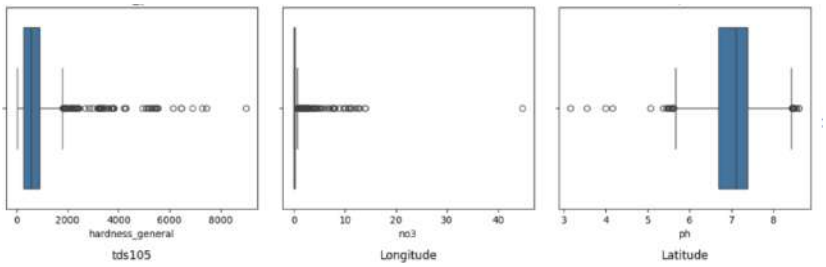


Figure 1.4: Box-Plot analysis

Boxplot Analysis

Figure 1.4 shows box plots for key groundwater parameters: TDS, NO_3, and pH. The TDS plot shows significant outliers, indicating contamination, while NO_3 values are mostly low with some high outliers, suggesting localized pollution. pH remains stable with few deviations. These distributions highlight the need for outlier handling and normalization in preprocessing.

Scatter Plot Analysis

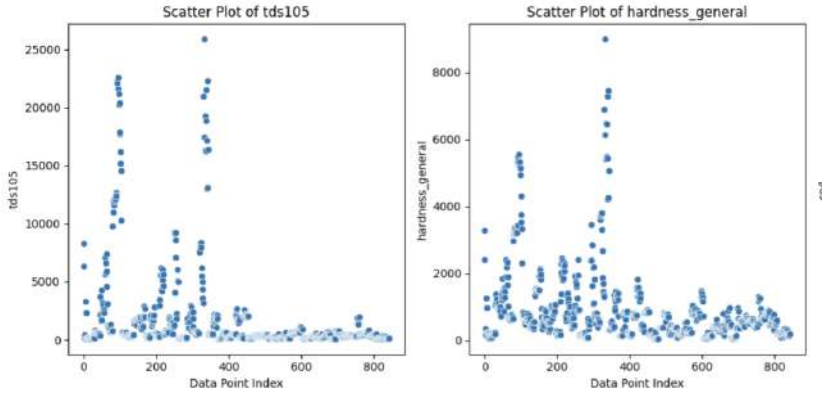


Figure 1.5: Scatter Plot Analysis

Figure 1.5 shows scatter plots for `tds105` and `hardness_general`, with most values clustering at lower ranges and several extreme peaks, indicating outliers. This suggests high variability in TDS and hardness, likely due to localized contamination or varying water sources, emphasizing the need for robust models and careful preprocessing.

Data Class Imbalance

To address **class imbalance**, the **Synthetic Minority Oversampling Technique (SMOTE)** was applied, generating synthetic examples for underrepresented classes to balance the dataset and reduce bias.

Algorithm 1.1 SMOTE Algorithm for Balancing Training Data

Input: Training dataset $(X_{\text{train}}, y_{\text{train}})$
Output: Resampled training dataset $(X_{\text{train}}, y_{\text{train}})$
Split dataset into training and testing sets
Apply SMOTE to the training set to generate synthetic samples
$X_{\text{train}}, y_{\text{train}} \leftarrow \text{SMOTE}(X_{\text{train}}, y_{\text{train}})$
$(X_{\text{train}}, y_{\text{train}})$

To ensure robust evaluation and reduce overfitting, **cross-validation** was used with weighted F1-score during hyperparameter optimization (e.g., AIO, Optuna), employing 3- or 5-fold cross-validation for better generalizability.

Following these steps, the dataset was free of missing values, balanced across classes, and ready for feature selection and modeling.

Algorithm 1.2 Optuna Cross-Validation for Hyperparameter Optimization

Input: Training dataset $(X_{\text{train}}, y_{\text{train}})$
Output: Optimized hyperparameters
Define objective function for Optuna
Perform cross-validation (5-fold) on training data
scores $\leftarrow$ cross_val_score$(model, X_{\text{train}}, y_{\text{train}}, \text{cv} = 5, \text{scoring} = \text{f1_weighted})$
Return the mean of the scores
mean(scores)

1. Preprocessing and Feature Extraction

The data were cleaned, normalized, and imputed. Features were selected using *MIFS*, with geographic coordinates included for spatial analysis. CNNs extracted spatial features, and PSO optimized hyperparameters.

2. Data Split for Training, Validation, and Testing

The dataset was split using *stratified sampling* to ensure even distribution of groundwater quality labels: **70% for training**, **15% for Validation**, and **15% for testing** model performance on unseen data.

3. Ground Truth Data and Labeling

Ground truth labels were assigned based on physicochemical parameters (e.g., pH, TDS, hardness) and contaminants (e.g., arsenic, cadmium), categorized as *Excellent, Good, Fair, Poor, Undrinkable* for drinkability. based on thresholds and expert assessments.

4. Input Data (features included)

The features used as input for machine learning models included physicochemical properties (e.g., ions, pH, TDS), spatial attributes (latitude, longitude), and temporal attributes (sampling dates) to account for seasonal variations. The target variable was the groundwater quality label.

5. Model Validation

Cross-validation was performed using *k-fold* to assess model performance and prevent overfitting. Models were evaluated based on accuracy, precision, recall, and F1-score, and external validation was done by comparing outputs with field data.

Handling Missing Values

The process for handling missing data ensured the dataset was properly cleaned for analysis and modeling.

Step 1: Initial Data Preprocessing

The dataset was loaded from the Excel file `daluong.xlsx`, which contains groundwater quality data with multiple columns, including features such as `well_code`, `date_sampling`, and others.

Table 1.17: Data Preprocessing Steps

Preprocessing Step	**Description**
Dropping Irrelevant Columns	Removed columns such as `well_code`, `date_sampling`, and others not necessary for analysis using: `df.drop(columns=[...], errors='ignore')`
Removing Rows with Missing Target Values	Rows with missing values in the target variable `tatse` were removed using: `df.dropna(subset=['tatse'])`
Standardizing Non-Standard Values	Replaced non-standard values in the `tatse` column (e.g., "MÆn", "Kh«ng") with standardized labels ("Mặn", "Không").

Step 2: Label Encoding and Column Drop

The `tatse` variable was label encoded using `LabelEncoder` to create numeric labels (`tatse_encoded`). Columns with only NaN values were identified and removed from the feature set.

Step 3: Handling Remaining Missing Values

Missing values in numeric columns were filled with the column mean, identified using `X.select_dtypes(include=np.number).columns`, and imputed with `X[numeric_cols].fillna(X[numeric_cols].mean())`. Before feature selection, the code checks for remaining NaN values in the features using:

`X.columns[X.isnull().any()].tolist()`. Any remaining NaNs are printed for further investigation to ensure no NaNs remain before feature selection.

Step 4: Feature Selection

Feature selection was performed based on MIFS between features and the target variable: Mutual information scores for each feature were calculated using: `mutual_info_classif(X, y)`. The top 14 features were selected based on these scores, and the feature set was reduced accordingly using `pd.Series(mi_scores, index=X.columns).sort_values(ascending=False)`.

Step 5: Feature Importance Analysis for Groundwater Classification

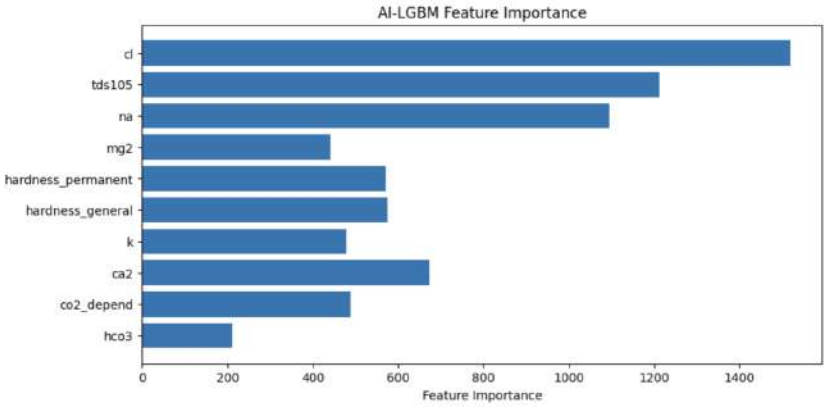


Figure 1.6: AI-LGBM Model Feature Importance

The figure 1.6 bar chart shows feature importance, with `cl` (chloride) as the most influential feature, followed by `tds105` (TDS), `na` (sodium), and `mg2` (magnesium). Other significant features include hardness parameters and ions like `k` (potassium) and `ca2` (calcium), while `hco3` (bicarbonate) has the lowest importance.

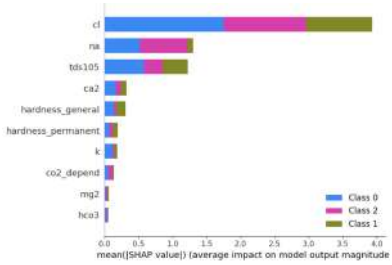


Figure 1.7: SHAP Summary Plot for AI-LGBM Model

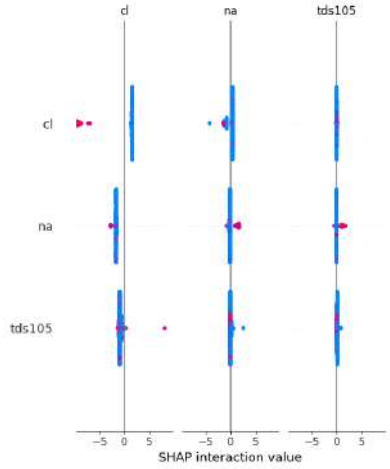


Figure 1.8: SHAP Interpretation and Implications

Figures 1.7 and 1.8 display SHAP values for the AI-LGBM model. Figure 1.7 shows the SHAP Summary Plot, highlighting the impact of features like `cl`, `na`, and `tds105` across different classes. Figure 1.8 presents the SHAP Interaction Plot, showing how these features interact and influence the model's predictions.

1.5.1 Detailed Structure of the Experimental Datasets

A precise description of the datasets, Table 1.18 summarizes the attributes available in the Vietnam and India groundwater collections.

Table 1.18: Attribute comparison between Vietnam and India datasets

Category	**Vietnam Dataset**	**India Dataset**
Administrative information	Well code, sampling date, quarter, laboratory	District, village
Spatial information	Latitude, longitude (external coordinate file)	Not always explicit; location described by village
Basic water indicators	pH, conductivity, TDS	pH, EC, TDS
Major cations	Na, K, Ca^{2+}, Mg^{2+}	Calcium, Magnesium, Sodium, Potassium
Major anions	Cl^-, SO_4^{2-}, HCO_3^-, CO_3^{2-}, NO_3^-, NO_2^-	Chloride, Sulphate, Bicarbonate, Carbonate, Fluoride
Hardness measures	General, temporary, permanent hardness	Total hardness
Additional environmental features	Fe^{2+}, Fe^{3+}, NH_4, PO_4, silica, dissolved gases	Alkalinity
Temporal availability	Multi-year monitoring	Mostly cross-sectional
Data dimensionality	High (more than 35 variables)	Moderate (around 15–17 variables)
Primary system role	Model training and spatial learning	Cross-region validation and robustness testing

1.5.2 Formal Label Generation Procedure

This section presents the computational process used to derive training labels. Label assignment follows QCVN 01-1:2018/BYT and WHO, which provides legally binding threshold values for drinking water safety in Vietnam. Therefore, the classification outcome is derived from regulatory compliance rather than subjective judgment.

Input

For each groundwater sample i, the raw measurement vector is:

$$x_i = \{pH, TDS, EC, major\ ions, \ldots\}$$

Step 1 – Standard Compliance Mapping

Each parameter j is compared with the permissible limit S_j according to QCVN 01-1:2018/BYT or WHO guidelines.

A normalized sub-index is computed as:

$$q_{ij} = \frac{V_{ij}}{S_j} \times 100 \tag{1.19}$$

where V_{ij} is the observed concentration.

Step 2 – Weighted Aggregation

The Water Quality Index is calculated as:

$$WQI_i = \sum_{j=1}^{n} w_j \cdot q_{ij} \tag{1.20}$$

Weights w_j follow the standard importance hierarchy used in water-quality evaluation literature.

Step 3 – Class Assignment

The final label is automatically determined:

$$y_i = \begin{cases} \text{Excellent}, & WQI_i \leq 25 \\ \text{Good}, & 25 < WQI_i \leq 50 \\ \text{Fair}, & 50 < WQI_i \leq 75 \\ \text{Poor}, & 75 < WQI_i \leq 100 \\ \text{Unsuitable}, & WQI_i > 100 \end{cases} \tag{1.21}$$

Step 4 – Binary Drinkability

$$y_i^{bin} = \begin{cases} 1, & y_i \in \{\text{Excellent, Good}\} \\ 0, & \text{otherwise} \end{cases} \tag{1.22}$$

Reproducibility Guarantee

Because the label is produced by explicit equations and public standards, two independent implementations will always generate identical outcomes from the same measurements.

Therefore, the labeling process is computational rather than subjective.

Algorithmic View

- For each sample, read hydrochemical parameters.
- Retrieve standard limits S_j.
- Compute q_{ij} for all parameters.
- Aggregate to obtain WQI_i.
- Map WQI_i to categorical label.

Table 1.19: Numeric summary of common water-quality variables (mean, SD, min, max, and missing rate).

Feature	India					Vietnam				
	Miss%	Mean	SD	Min	Max	Miss%	Mean	SD	Min	Max
pH	0.00	7.829	0.400	6.460	8.780	0.05	6.975	0.681	2.530	8.650
TDS	0.00	358.057	280.979	30.000	2766.000	0.05	1239.542	2505.419	43.000	25901.000
Calcium	0.00	43.942	30.612	0.000	497.000	0.05	63.956	51.925	1.000	657.310
Magnesium	0.00	25.617	25.835	-4.000	345.000	0.05	51.827	81.516	0.000	981.920
Sodium	0.00	49.992	61.034	0.000	820.000	0.05	319.903	824.534	0.800	8200.000
Potassium	0.00	3.110	3.053	-1.000	54.000	0.05	10.318	12.746	0.000	110.000
Chloride	0.00	66.594	104.591	0.000	1750.000	0.05	531.662	1440.955	4.804	15740.000
Sulphate	0.00	30.703	39.610	0.000	496.000	0.05	50.210	58.291	0.000	579.000
Bicarbonate	0.00	247.580	152.206	0.000	1215.000	0.05	388.683	339.868	0.000	2795.000
Carbonate	0.00	12.607	25.605	0.000	312.000	0.05	6.158	25.864	0.000	816.000
TotalHardness	0.00	215.867	153.487	0.000	1517.000	0.05	746.204	1224.569	14.000	12150.000

Table 1.19 reports a side-by-side descriptive summary of the variables shared by the India and Vietnam datasets, including the missing rate, mean, standard deviation, and the minimum–maximum range. The statistics highlight clear distributional differences between the two domains, with Vietnam generally exhibiting larger variability and higher upper extremes for salinity-related indicators (e.g., TDS, chloride, and total hardness), supporting the presence of domain shift between the datasets.

Step 6: Spatial Resolution and GIS Integration

1.6 Feature Engineering

Feature engineering is a critical step in developing an effective machine learning model for groundwater drinkability classification, as it transforms raw data into meaningful features that enhance predictive capability. This section outlines the process used in this research, including the encoding of spatial coordinates, the derivation of new features from raw measurements, and the incorporation of domain knowledge.

1.6.1 Encoding of Spatial Coordinates

Groundwater quality can vary spatially, and geographic location plays a significant role in understanding contamination patterns. Thus, spatial coordinates (latitude and longitude) of each groundwater sample were used as features. Additionally, the haversine distance between the geographic coordinates of different samples was calculated to quantify spatial relationships. This allows the model to consider the proximity of samples to one another, enhancing its ability to detect regional water quality variations.

The Haversine distance between two geographic points $(\text{lat1}, \text{lon1})$ and $(\text{lat2}, \text{lon2})$ is given by the following equation:

$$d = 2R \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\phi}{2}\right) + \cos(\phi_1) \cdot \cos(\phi_2) \cdot \sin^2\left(\frac{\Delta\lambda}{2}\right)}\right)$$

Where:

- d is the distance between the two points (in kilometers).
- R is the radius of the Earth (mean radius $\approx 6371\,\text{km}$).
- ϕ_1 and ϕ_2 are the latitudes of the two points in radians.
- $\Delta\phi = \phi_2 - \phi_1$ is the difference in latitudes.
- $\Delta\lambda = \lambda_2 - \lambda_1$ is the difference in longitudes.

1.6.2 Derived Features from Raw Measurements

To enhance the model's predictive accuracy, several key hydrochemical parameters, such as pH, TDS, nitrate, and ron, were used as base measurements. These parameters were transformed into sub-indices based on environmental standards. For example, the Water Quality Index (WQI) was computed as a weighted sum of sub-indices, each representing a specific water quality parameter.

The equation for calculating sub-indices for each parameter is as follows:

$$q_i = \left(\frac{V_i}{S_i} \right) \times 100$$

Where:

- q_i is the sub-index for the i-th parameter (e.g., pH, TDS, nitrate).
- V_i is the observed value of the i-th parameter.
- S_i is the standard or guideline value for the i-th parameter.

The overall WQI is computed as the weighted sum of the sub-indices:

$$\text{WQI} = \sum_{i=1}^{n} w_i \cdot q_i$$

Where:

- w_i is the weight assigned to the i-th parameter, reflecting its importance in the overall water quality.
- q_i is the sub-index for each parameter.
- n is the number of parameters (e.g., pH, TDS, nitrate).

Additionally, interactions between parameters, such as pH $\times$ TDS, were considered to account for non-linear relationships between features.

1.6.3 Incorporating Domain Knowledge into Feature Creation

Domain knowledge was essential for selecting the most relevant features for groundwater quality modeling. While specific datasets for pollution sources

(e.g., proximity to industrial zones or agricultural runoff) could not be integrated directly due to data limitations, domain knowledge influenced the selection of key hydrochemical parameters. For instance, TDS, nitrate, and iron are well-known to have a significant impact on groundwater quality based on existing environmental research.

Even though data for spatial pollution sources was not available, domain knowledge ensured that the selected features were highly relevant to water quality classification and accurately represented the factors influencing groundwater quality.

The feature engineering process involved the following key steps:

- Spatial Encoding: Geographic coordinates (latitude and longitude) were used directly, along with the haversine distance between samples.
- Derived Features: The Water Quality Index (WQI) was calculated for each sample to summarize key hydrochemical parameters.
- Domain Knowledge Integration: Feature selection was guided by domain expertise, ensuring that the most relevant hydrochemical parameters were included.

These engineering steps, combined with spatial features, allowed the models to capture the complexities of groundwater contamination and significantly improved prediction accuracy and model robustness.

Impact of GIS on Model Performance and Location on Prediction Results: The integration of GIS improved model accuracy (98.8%) and F1-score (99.5%) by incorporating spatial features, enabling the detection of contamination patterns often missed by traditional models. Location-specific factors, such as hydrogeology and pollution sources, influenced predictions. GIS maps revealed regional disparities, highlighting the importance of spatial context in decision-making.

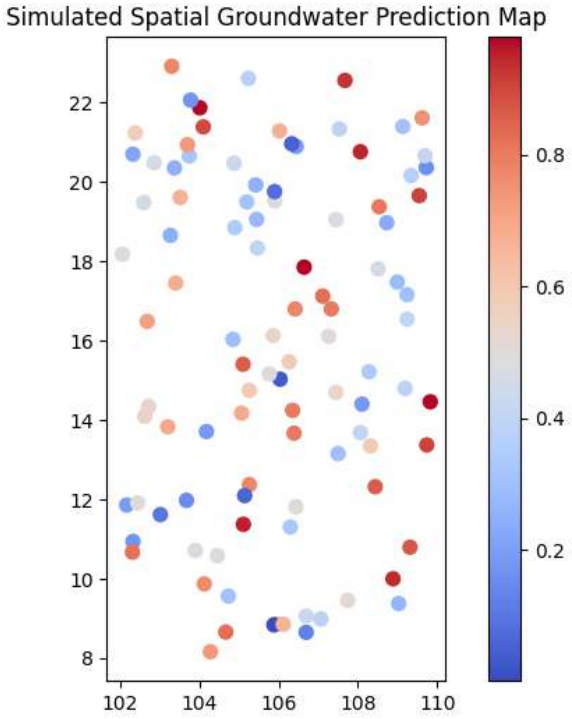


Figure 1.9: Spatial Visualization of Groundwater Quality Classification

1.7 Generalization and Transferability to Other Geographical Regions

The proposed PSO-SCNN model is designed to be transferable to other regions with similar groundwater contamination challenges. Its generalizability depends on the availability and relevance of input features, such as water quality parameters and spatial coordinates, which may vary by region. Future research should test the model in diverse areas, especially arid regions or those impacted by industrial pollution, to assess its robustness under different environmental conditions.

Required Minimum Sample Sizes for New Areas

For effective deployment in new regions, determining the minimum sample size is crucial. This depends on groundwater variability in the area. A representative dataset should include key water quality parameters like TDS, hardness, and chemical concentrations. Power analysis can help estimate the required sample size, ensuring the model's high performance across different regions.

Model Retraining vs. Fine-Tuning Strategies

When applying the model to new areas, two strategies are considered: **model retraining** and **fine-tuning**.

Model Retraining involves training the model from scratch with new regional data, ideal for regions with significant differences in water quality profiles.

Fine-Tuning uses a pre-trained model and adjusts it with a smaller dataset from the new region, which is more resource-efficient when the new area shares similarities with the original.

The choice depends on dataset size and available computational resources.

Limitations of Current Geographical Scope

While the model has been tested in the Mekong Delta and Odisha, it may not perform equally well in other regions with different hydrological conditions or contamination profiles. Expanding its geographical scope will require additional data and possibly retraining to ensure its generalizability. The model's robustness for global applicability remains uncertain due to limited data from diverse regions.

1.8 Chapter Conclusion

Groundwater Quality Classification: From Traditional Methods to Advanced ML/DL Frameworks: This chapter highlighted the shift from traditional groundwater quality assessment methods, like the Water Quality Index (WQI), to advanced machine learning (ML) and deep learning (DL) approaches. Traditional methods struggle with non-linear relationships, large datasets, and spatial dependencies, motivating the adoption of ML/DL tools capable of leveraging multi-dimensional data for real-time results.

Machine learning techniques such as Random Forest (RF), Support Vector Machine (SVM), and XGBoost have enhanced groundwater quality classification by capturing intricate data relationships. However, challenges such as

data dependency and high computational demands remain, limiting their use in resource-constrained settings.

Deep learning methods, particularly Convolutional Neural Networks (CNN), have further improved classification by effectively capturing spatial and temporal dependencies in groundwater data. Yet, their "black box" nature raises concerns about interpretability, especially for actionable insights needed by water resource managers.

To address these issues, this research introduces a hybrid spatial-aware framework, combining AI-enhanced Light Gradient Boosting Machines (AI-LGBM) with Spatial Convolutional Neural Networks (SCNN) and Particle Swarm Optimization (PSO). This approach improves model performance, scalability, and interpretability.

In conclusion, the transition from WQI-based methods to hybrid ML/DL models represents a significant advancement in groundwater quality classification. The proposed models provide a robust, scalable, and interpretable solution, supporting better water resource management in regions facing environmental and health challenges, such as Vietnam and India.

Contributions of This Chapter

This chapter discussed the limitations of traditional WQI methods, reviewed advanced ML/DL techniques for groundwater classification, and introduced a hybrid AI-LGBM, spatial PSO-SCNN framework for enhanced predictive accuracy. Optimization algorithms like PSO, GA, and Grid Search were highlighted for performance improvement, and spatial integration was emphasized for better interpretability and policy relevance.

Chapter 2

Proposed Ensemble Spatial Machine Learning Methods

2.1 Introduction

This chapter presents machine learning methods for groundwater quality classification, focusing on ensemble spatial models: **AI-enhanced Light Gradient Boosting Machine (AI-LGBM)** and **Particle Swarm Optimization-Spatial Convolutional Neural Network (PSO-SCNN)**. These models address challenges in accurately and efficiently classifying groundwater quality by leveraging spatial data and optimization techniques to improve prediction accuracy, scalability, and interpretability in environmental monitoring.

2.1.1 Proposed System Model of the Artificial Intelligence Framework

The AI framework for real-time groundwater quality monitoring integrates key components: data acquisition, preprocessing, ensemble modeling, decision-making, and continuous learning (Figure 2.1).

System Architecture Overview

Data Acquisition: Real-time data from IoT sensors and GIS systems.
Data Preprocessing: Data cleaning, normalization, and spatial encoding.
Ensemble Modeling: Classification using **AI-LGBM** and **PSO-SCNN**.
Decision-Making: Water quality classification and insights.
Continuous Learning: Model retraining with new data.

System Implementation

Cloud processing is used for training, and edge devices handle real-time inference for low latency.

External System Integration

The system integrates with decision support and alert systems to share results with stakeholders.

2.1.2 Ensemble Mechanism and Model Integration

In this thesis, the term *ensemble* is used in two complementary senses:

(i) Model-level ensemble: AI-LGBM is an ensemble by construction because it follows the gradient boosting methodology, where the final predictor is an additive combination of many decision trees. For binary drinkability prediction, it outputs a probability:

$$\hat{p}_{\mathrm{LGBM}}(y = 1 \mid x) = \sigma\left(\sum_{m=1}^{M} f_m(x)\right), \tag{2.1}$$

where $f_m(\cdot)$ is the m-th tree, M is the total number of trees, and $\sigma(\cdot)$ is the sigmoid function (for multi-class classification, additive scores are computed per class and normalized with softmax).

(ii) System-level ensemble: The overall framework integrates two complementary learners: **AI-LGBM** (which uses tabular hydrochemical variables and point-based spatial features) and **PSO-SCNN** (which captures spatial-context learning for consistent mapping). When both models are employed together, their outputs are combined using **late fusion**. This strategy involves combining the predicted probabilities of the two models as follows:

$$\hat{p}_{\mathrm{ens}} = \alpha\, \hat{p}_{\mathrm{LGBM}} + (1 - \alpha)\, \hat{p}_{\mathrm{SCNN}}, \quad \alpha \in [0, 1], \tag{2.2}$$

where α is a weighting parameter selected using the validation set (typically to maximize performance metrics such as F1 or AUC). The binary decision is then made based on:

$$\hat{y} = \mathbb{I}\left(\hat{p}_{\mathrm{ens}} \geq \tau\right), \tag{2.3}$$

where τ is a threshold chosen based on the validation data. For the multi-class Water Quality Index (WQI) task, fusion is applied per-class to the probability vectors, and the predicted class is the one with the highest fused probability.

2.1.3 Rationale for the AI-LGBM + PSO-SCNN Hybrid and Weighted Late Fusion

Although AI-LGBM and PSO-SCNN are each optimized individually, the motivation for combining them is that they capture *different and complementary* aspects of the groundwater drinkability problem, and they exhibit different error profiles under spatial heterogeneity. Optimization strengthens each base learner; late fusion then leverages them as two "experts" whose strengths compensate for one another, improving robustness and enabling an application-driven trade-off between false negatives and false positives.

Why individual optimization is still necessary. Late fusion is most effective when each component is a strong, well-calibrated learner. Therefore, we first optimize AI-LGBM (tabular learner) and PSO-SCNN (spatial learner) independently to reduce bias/variance within each model family, and then combine their probabilistic outputs using a validation-selected weight α (Eq. (2.2)) and threshold τ (Eq. (2.3)).

Complementary strengths (why the hybrid is justified). AI-LGBM is a high-performing tabular ensemble that excels at modeling non-linear interactions among hydrochemical attributes and provides strong average ranking performance (Accuracy/AUC). In contrast, PSO-SCNN explicitly learns spatially structured patterns from grid-based representations, improving spatial consistency and producing an operationally desirable risk-averse profile (high Recall/F1), where missing unsafe water is costly. This trade-off is also observed empirically in cross-validation, where AI-LGBM leads on average Accuracy/AUC while PSO-SCNN is stronger on Recall/F1 in held-out tests.

Table 2.1: Why AI-LGBM and PSO-SCNN are complementary and suitable for weighted late fusion.

Aspect	**AI-LGBM (tabular boosted ensemble)**	**PSO-SCNN (spatial deep model)**
Primary signal captured	Hydrochemical attribute interactions (tabular patterns)	Spatial structure and neighborhood consistency (grid-based spatial context)
Typical strength	High Accuracy/AUC; efficient inference; interpretable with feature attribution	High Recall/F1 for risk-averse detection; spatially coherent mapping of unsafe zones
Typical limitation	May under-capture local spatial autocorrelation when contamination is spatially clustered	Higher compute cost; can be sensitive to data sparsity or grid resolution
Value in fusion	Provides stable global decision boundary and strong probability ranking	Recovers spatially driven unsafe cases; reduces false negatives in clustered contamination

How the weighting method addresses the compensation explicitly. The fusion weight α in Eq. (2.2) is selected on a validation set to optimize a target metric aligned with the operational objective (e.g., maximizing F1/AUC or maximizing F_β with $\beta > 1$ to emphasize Recall). This explicitly controls the trade-off:

$$F_\beta = (1+\beta^2)\frac{\text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}, \quad \beta > 1 \text{ (risk-averse setting)}. \tag{2.4}$$

In practice, we perform a simple grid search over $\alpha \in [0, 1]$ (e.g., step 0.01) and select the $\alpha^\star$ that maximizes the chosen objective on validation data, then choose τ accordingly (Eq. (2.3)). This procedure makes the weighting *data-driven* and reproducible, rather than qualitative.

In short, optimization ensures each model is a strong specialist; weighted late fusion then combines a tabular expert (AI-LGBM) and a spatial expert (PSO-SCNN) to yield a more robust decision rule under spatial heterogeneity and to support cost-sensitive operation by tuning α and τ on validation data.

Finally, **PSO (Particle Swarm Optimization)** is not used as an inference-time ensemble technique, but rather as an optimization step during the training phase. PSO searches for the optimal SCNN hyperparameters, and the best configuration is retained for deployment.

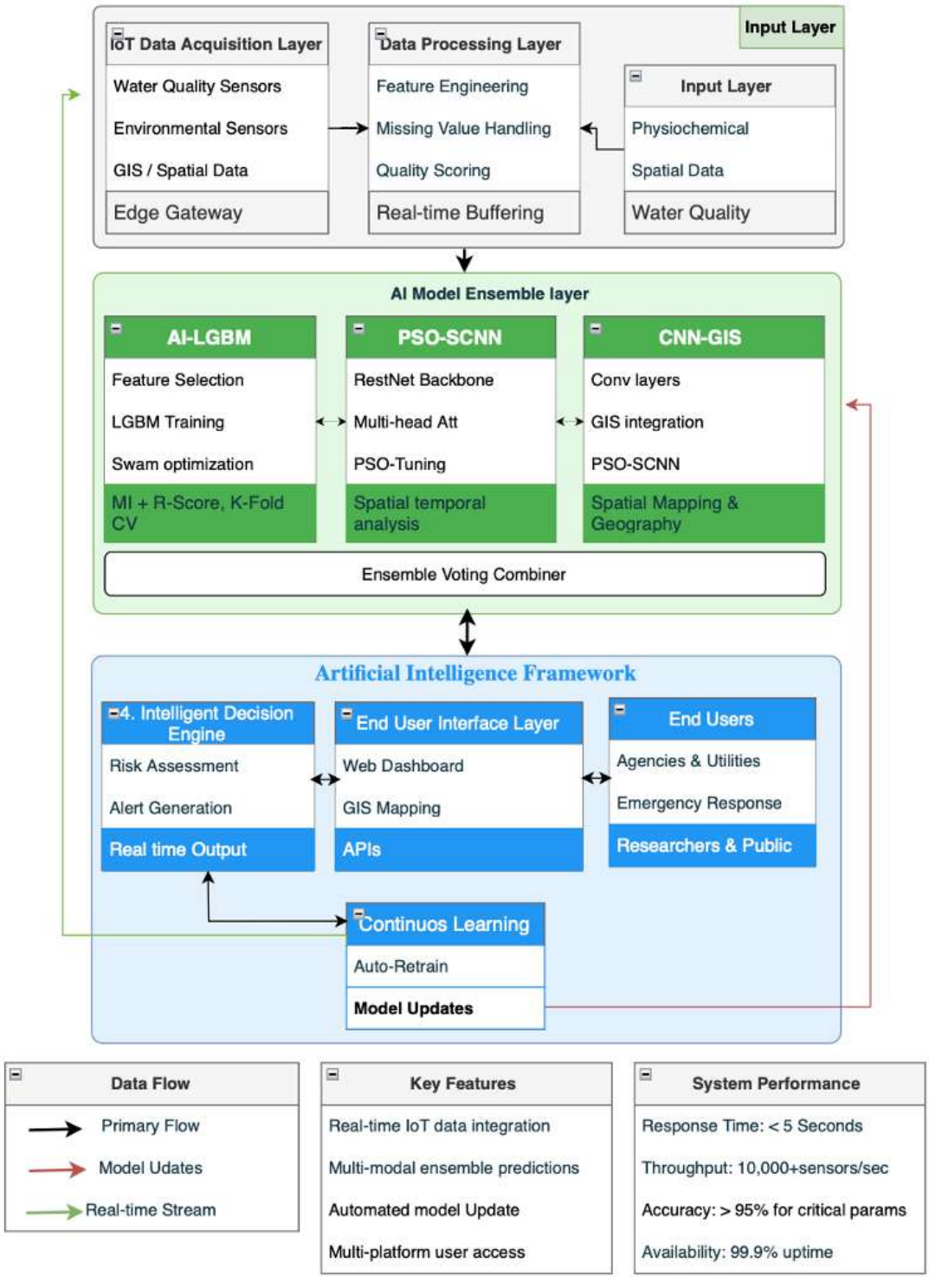


Figure 2.1: Proposed System Model of the AI Framework

Protocol. The architecture (Figure 2.1) operates as an end-to-end AI framework for environmental monitoring:

- **Data Acquisition:** Data is collected from sensors, historical records, and GIS via edge gateways.
- **Data Processing:** Includes feature extraction, missing value handling, and real-time buffering.
- **Ensemble Modeling:** AI models (AI-LGBM, PSO-SCNN) are trained and optimized using AIO/Optuna and PSO.
- **Decision-Making:** Model outputs trigger operational decisions like risk flags and thresholds.
- **Visualization:** Outputs are integrated into a GIS module for risk maps and dashboards.
- **Continuous Learning:** New data retrains or fine-tunes models for better adaptation.
- **System Integration:** APIs link to external systems for alerts and reporting.

This AI framework enables real-time groundwater drinkability assessment, supporting decision-making across environments.

IoT Data for Simulated Real-Time Updates

Figure 2.1 shows the AI framework for groundwater quality classification, integrating real-time IoT sensor data with hydrochemical parameters and spatial features (e.g., pH, TDS, latitude, longitude). This hybrid model enables dynamic, scalable water quality predictions and decision-making through dual-stream processing of spatial and non-spatial data.

The IoT sensors enable continuous data streams for dynamic model retraining. This ensures adaptive, accurate predictions, timely risk assessment, and improved responsiveness for emergency response, monitoring, and public health.

Performance & Features

The system demonstrates robust performance metrics, including a response time of less than 5 seconds, throughput exceeding 10,000 readings per second, accuracy greater than 95%, and uptime of 99.9%. Key features encompass simulated real-time IoT data integration, ensemble predictions, automated updates, and multi-platform access, ensuring efficient and reliable operation for groundwater quality monitoring.

Scalable Groundwater Quality Management Framework

- **Data Collection**: Deploy IoT sensors at key sites for physiochemical parameters (pH, nitrate, turbidity) and GIS spatial data.
- **Data Processing**: Preprocess data with feature engineering, missing value handling, and quality scoring for real-time buffering.
- **Model Ensemble**: Use hybrid models AI-LGBM, PSO-optimized Spatial CNN, and GIS integrated CNN and combine outputs via ensemble voting.
- **Real-Time simulated Monitoring**: Integrate IoT streams with AI inference for risk alerts and enable automated model retraining for adaptation.
- **Deployment & Governance**: Connect with provincial/national systems via APIs, ensuring interoperability and stakeholder dashboards.
- **Policy & Sustainability**: Establish data governance, secure funding, and train local teams for maintenance and updates.
- **Global Adaptability**: The framework can be retrained and fine-tuned for regions worldwide Africa, South Asia, Latin America using local data and remote sensing.

Outcome: A flexible, scalable system for sustainable groundwater management in varied hydrogeological and socio-economic settings.

In conclusion, the proposed models deliver exceptional performance in tested environments and show potential for global environmental applications.

2.2 AI-LGBM

2.2.1 Main Ideas

The **AI-LGBM** (Auto Immune Light Gradient Boosting Machine) integrates machine learning and evolutionary optimization techniques to enhance model robustness and performance. The term **"Auto Immune"** draws a biological metaphor to describe the model's adaptive and self-correcting capabilities. Much like the immune system in living organisms, which recognizes and mitigates external threats, the "Auto Immune" mechanism in **AI-LGBM** helps the model to detect and correct errors or outliers in the data. This self-correcting feature improves the model's performance, ensuring its reliability even in complex or noisy datasets.

This metaphor is crucial for understanding the core functionality of the model and highlights its ability to adapt and optimize itself over time, making it particularly effective for groundwater drinkability classification in varied environmental conditions.

The AI-enhanced Light Gradient Boosting Machine (AI-LGBM) is an advanced model designed to combine the benefits of gradient boosting with artificial intelligence techniques. The main idea behind AI-LGBM is to enhance the predictive performance of the traditional LightGBM model by incorporating machine learning techniques such as feature importance analysis and optimization algorithms. This model is particularly effective in handling large, complex datasets with multiple input variables, making it ideal for groundwater quality classification, where data may include numerous physicochemical parameters.

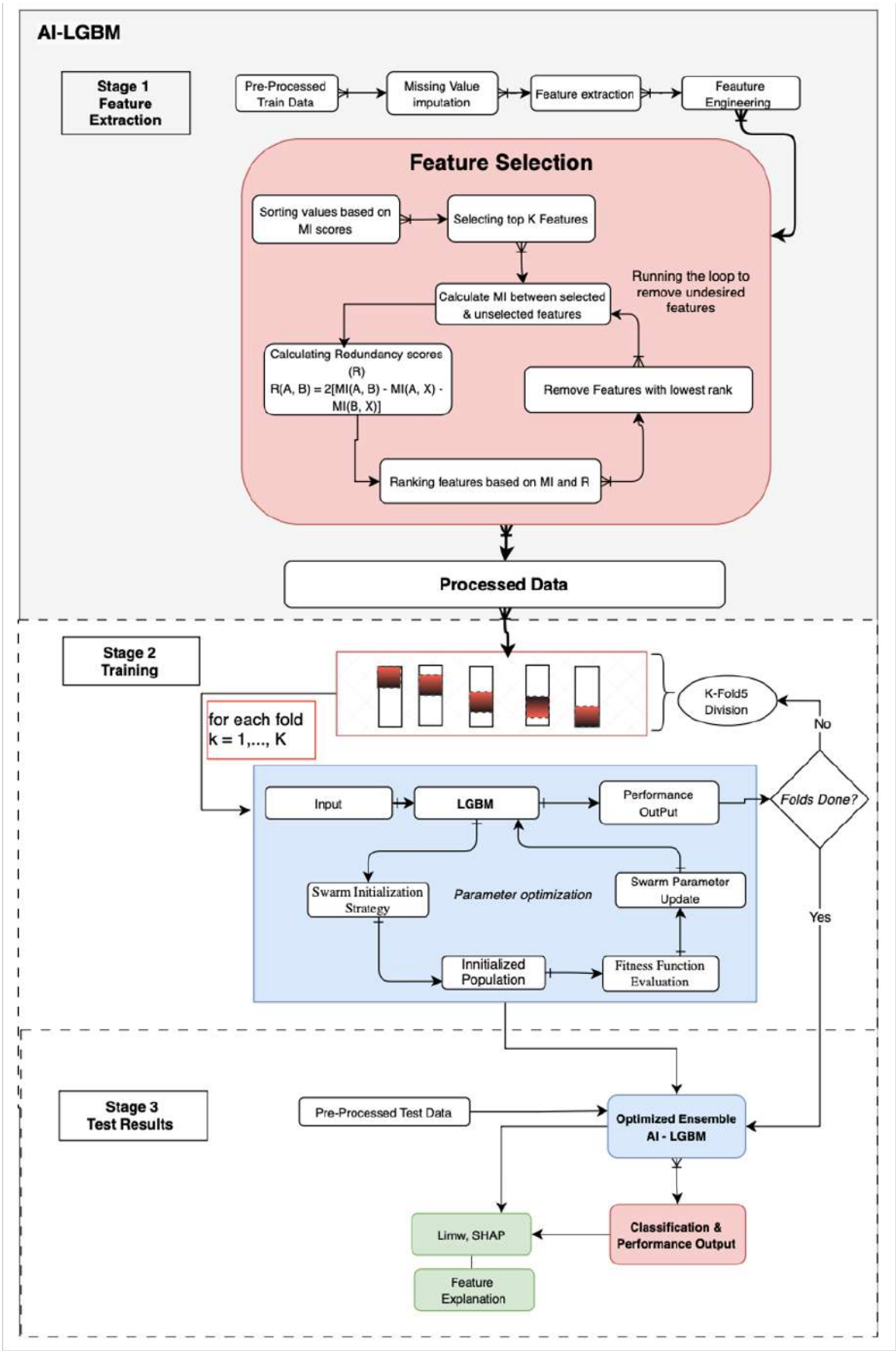


Figure 2.2: Proposed AI-LGBM Methodological Flowchart

Description of the methodological flowchart for AI-LGBM

Figure 2.2 illustrates the data flow for groundwater classification, from raw data collection (hydrochemical measurements and spatial coordinates) to preprocessing and input into the AI-LGBM model. It highlights feature importance analysis, which identifies key input features affecting model predictions, improving interpretability.

As illustrated in 2.2, the AI-LGBM methodological flow begins with the ingestion and preprocessing of raw hydrochemical data, including cleaning, normalization, and handling of missing values. Next, Mutual Information–based Feature Selection (MIFS) identifies the most informative physicochemical predictors, reducing dimensionality while preserving signal. The refined feature set is then passed to the AI-LGBM core, where LightGBM learners are trained and their hyperparameters are automatically tuned using an Auto-Immune Optimization (AIO) strategy to balance accuracy and generalization. Model performance is assessed via k-fold cross-validation under multiple metrics (Accuracy, Precision, Recall, F1, AUC), and the final trained model is used to generate groundwater drinkability predictions. In the last stage, explainability and decision support are provided through feature-importance and SHAP analyses to provide the distribution of safe and unsafe groundwater across the study areas.

The AI-LGBM model is an advanced ensemble learning framework combining LightGBM with **Auto-Immune Optimization (AIO)** and **Mutual Information-based Feature Selection (MIFS)** to deliver an efficient and interpretable solution for groundwater quality classification [95, 96]. It improves accuracy and robustness through a multi-step process involving feature selection and hyperparameter tuning.

MIFS identifies the most informative features in large, complex datasets, reducing dimensionality and computational overhead while maintaining classification performance. **AIO**, a biologically inspired technique, adaptively tunes hyperparameters to enhance learning and prevent overfitting.

The model also incorporates **k-fold cross-validation** and **meta-learning**

to ensure generalization across diverse hydrogeological conditions. Its scalable architecture supports large-scale classification, while improving transparency via critical feature identification [97].

Operationally, the Stage 2 training block in Figure 2.2 is executed inside a K fold cross-validation loop (here $K = 5$). For each fold $k \in \{1, \ldots, K\}$, one subset is held out as the validation set while the remaining $K - 1$ folds are used for training. Within each iteration, the full pipeline Mutual Information-based Feature Selection, SMOTE rebalancing, AIO/Optuna hyperparameter search, and LightGBM fitting is retrained on the training folds and evaluated on the corresponding validation fold. The performance metrics reported in Chapter 3 are the mean (and standard deviation) across all folds. Although this outer loop is not explicitly drawn in Figure 2.2, it conceptually surrounds the entire Stage 2 training block.

As shown in Figure 2.2, AI-LGBM sets a benchmark for predictive performance, offering a robust and scalable solution for real-time environmental monitoring and policy development [98, 99]. By streamlining the learning process and improving interpretability, the model supports effective water resource management in varied environmental settings.

Table 2.2: *Benefits of Combining Components in the AI-LGBM Model*

Functionality	**Contribution**
Feature Dimensionality Reduction	Achieved through MIFS, which selects only the most relevant features, reducing noise and improving model efficiency.
Hyperparameter Optimization	Enabled by AIO, which dynamically tunes learning parameters to improve model performance and generalization.
Interpretability and Robustness	Enhanced by LightGBM's structured and tree-based architecture, which facilitates better understanding and stable predictions.

2.2.2 Algorithm description

This study presents an AI-enhanced Light Gradient Boosting Machine (AI-LGBM) model for groundwater quality classification, trained to categorize samples into quality classes such as *Excellent, Good, Fair, Poor, Undrinkable* for drinkability. The approach combines the predictive power of gradient boost-

ing where multiple decision tree learners are iteratively built to correct previous errors with targeted feature selection and advanced hyperparameter tuning. Mutual Information-based Feature Selection (MIFS) is applied to identify the most relevant physicochemical parameters (e.g., pH, TDS, nitrate) and spatial attributes (e.g., geographic coordinates), reducing dimensionality and improving interpretability. The refined feature set is then used in the AI-LGBM framework, where a boosting process models both linear and nonlinear relationships and a feature importance mechanism highlights dominant predictors. Hyperparameters are optimized through a hybrid Auto Immune Optimization (AIO) and Optuna process, enabling adaptive configuration adjustments based on performance feedback to ensure robust convergence and strong generalization. This integrated design enhances accuracy, efficiency, and scalability, making it suitable for real-world groundwater monitoring applications.

Mathematical Formulation of AI-LGBM with MIFS

The AI-LGBM model incorporates Mutual Information-based Feature Selection (MIFS) to select the most relevant features for groundwater quality classification. The goal is to identify the optimal set of hyperparameters W^* that maximizes model performance while reducing the dimensionality of the feature space.

Let $X = \{x_1, x_2, \ldots, x_n\}$ represent the dataset of groundwater samples, where each sample $x_i \in \mathbb{R}^m$ is a vector of features $x_i = (x_{i1}, x_{i2}, \ldots, x_{im})$. The dataset includes physicochemical parameters (e.g., pH, TDS, nitrate concentration) and spatial features (e.g., geographic coordinates).

Each sample x_i is associated with a label $y_i \in \{1, 2, \ldots, k\}$, where k represents the number of classes for water quality (e.g., Excellent, Good, Poor, Bad).

Let $f(X; W)$ denote the AI-LGBM classification model, which maps the feature vector x_i to the predicted label $\hat{y}_i$. The objective is to find the optimal hyperparameters W^* that maximize the model performance, expressed as:

$$W^* = \arg\max_W g(W) \tag{2.5}$$

Where $g(W)$ represents the performance function (e.g., accuracy, F1-score, precision).

The performance of the model is further enhanced by the integration of MIFS, which helps select the most informative features for classification by measuring the mutual information between each feature and the target label. The mutual information function $\mathcal{I}(X, Y)$ is defined as:

$$\mathcal{I}(X, Y) = H(X) + H(Y) - H(X, Y) \tag{2.6}$$

Where:

- $H(X)$ is the entropy of the feature set,
- $H(Y)$ is the entropy of the labels,
- $H(X, Y)$ is the joint entropy of the features and labels.

The objective of the feature selection process is to choose the top k features that maximize the mutual information with the target label Y:

$$X_k = \arg\max_X \mathcal{I}(X, Y) \tag{2.7}$$

Once the relevant features are selected using MIFS, the AI-LGBM model's hyperparameters W^* are optimized using Particle Swarm Optimization (PSO), ensuring that the model accurately predicts the groundwater quality labels.

Mathematical Foundations for Classification, Optimization, and Feature Selection:

1. Groundwater Drinkability Classification:

$$\hat{y}_i = f(x_i; W) \quad \text{with objective} \quad \hat{y}_i = \arg\min_{\hat{y}} \mathcal{L}(y_i, \hat{y}_i) \tag{2.8}$$

2. Hyperparameter Optimization:

$$W^* = \arg\max_W g(W) \quad \text{where} \quad g(W) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(y_i, f(x_i; W)) \tag{2.9}$$

3. Feature Selection with MIFS:

$$X_k = \arg\max_X \mathcal{I}(X, Y) \tag{2.10}$$

Experimental Setup and Implementation for AI-LGBM

Our experimental workflow used a stratified 70/15/15 split for training, validation, and testing. Preprocessing involved imputing missing values, normalizing features via the Z-score method (Equation (2.11)), and removing IQR-based outliers. Model training and hyperparameter tuning were conducted in Python 3.10 with Scikit-learn 1.2.2, LightGBM 3.3.2, and Optuna 3.0.0. All metrics are an average of five runs using different random seeds.

Step 1: Data Acquisition

This study utilizes two distinct groundwater quality datasets. The first dataset, comprising 1,052 samples from Vietnam's Mekong Delta, includes physicochemical attributes such as pH, TDS, nitrate, chloride, sulfate, and hardness, along with spatial coordinates. A second dataset of 1,241 samples with similar parameters was sourced from the Central Ground Water Board (CGWB) in Odisha, India.

Step 2: Data Preprocessing

The data preprocessing pipeline included several key steps: missing values were imputed using mean/median and mode, outliers were removed using the IQR method, and both physicochemical features and spatial coordinates were scaled to a [0,1] range via Min-Max normalization.

Feature Engineering Preprocessing

The preprocessing pipeline involved imputing missing values, binarizing features according to permissible water quality standards, and normalizing all numerical data using:

$$F^*(x_i) = \frac{F(x_i) - \overline{F}}{\sigma(F)}, \tag{2.11}$$

where $\overline{F}$ is the mean of feature F and $\sigma(F)$ is its standard deviation, computed as:

$$\overline{F} = \frac{1}{n}\sum_{i=1}^{n} F(x_i), \tag{2.12}$$

$$\sigma(F) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\big(F(x_i) - \overline{F}\big)^2}. \tag{2.13}$$

This standardization ensured that all numerical features had zero mean and unit variance, improving model convergence during training.

In addition, we derived new attributes using Water Quality Index (WQI) relations, where higher scores indicate better water quality. These engineered features, along with standardized original variables, form the processed dataset used in the subsequent machine learning pipeline.

Feature Selection (MIFS). After preprocessing, we applied Mutual Information-based Feature Selection to choose the top-K most informative features:

$$S^\star = \arg\max_{S\subset\{1,\dots,m\},\,|S|=K} \sum_{j\in S} \mathcal{I}(X_j; Y), \qquad X \leftarrow X_{S^\star}. \tag{2.14}$$

SMOTE Balancing. For a minority sample x_i and its kNN neighbor $x_i^{(nn)}$ in the same class, SMOTE generates synthetic samples as:

$$\tilde{x} = x_i + \lambda\big(x_i^{(nn)} - x_i\big), \qquad \lambda \sim \mathcal{U}(0,1), \tag{2.15}$$

with $\tilde{y} = y_i$, yielding a balanced dataset $\mathcal{D}_{\text{train}}^{\text{smote}}$.

Boosted Additive Model. The LightGBM model fits:

$$F_m(x) = F_{m-1}(x) + \eta \cdot \sum_{j=1}^{J_m} \gamma_{jm} \cdot \mathbb{I}(x \in R_{jm}), \tag{2.16}$$

where R_{jm} is the j-th leaf region of the m-th decision tree, and γ_{jm} is the optimal leaf weight:

$$\gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L\left(y_i, F_{m-1}(x_i) + \gamma\right). \tag{2.17}$$

Step 4: Model Evaluation

We evaluated models using accuracy, precision, recall, F1-score, and AUC, and used SHAP for feature interpretation. The experiments were run on Python 3.10 with libraries like Scikit-learn and LightGBM,

We calculate the classification accuracy and other metrics such as precision, recall, and F1-score. The classification accuracy is given by:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.18}$$

Where:

- TP is the number of true positives,
- TN is the number of true negatives,
- FP is the number of false positives,
- FN is the number of false negatives.

The performance comparison can be expressed as:

$$\text{Acc}_{\text{AI-LGBM with MFS}} > \text{Acc}_{\text{Existing Models}} \tag{2.19}$$

This hypothesis can be tested using statistical tests such as t-tests or ANOVA to assess whether the AI-LGBM model with MFS significantly outperforms the existing models.

2.2.3 Learning Strategy

The AI-LGBM model employs a supervised learning strategy in which labeled groundwater samples, each with a known quality classification, are used to train the model. The process begins with **data preprocessing**, which includes handling missing values, detecting and addressing outliers, and normal-

izing input features to ensure consistency. Once prepared, the model undergoes **training** through multiple iterations of gradient boosting, progressively reducing classification error by correcting the mistakes of previous iterations. To enhance performance, **hyperparameters** such as learning rate, tree depth, and regularization terms are fine-tuned using optimization algorithms like Auto Immune Optimization (AIO) and Optuna. Finally, the model's **generalizability** is evaluated through K-fold cross-validation, which partitions the dataset into multiple subsets to ensure reliable and robust performance across varying data splits.

Mathematical Formulation

Data and Notation. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, $x_i \in \mathbb{R}^m$, $y_i \in \{1, \ldots, K\}$. We first select features by mutual information (MIFS), then balance the training set via SMOTE, and finally train LightGBM with cross-validated hyperparameter optimization (Optuna) maximizing weighted F1.

Feature Selection (MIFS). Compute mutual information $\mathcal{I}(X_j; Y)$ for each feature X_j and keep the top-K:

$$S^\star = \arg \max_{\substack{S \subset \{1,\ldots,m\} \\ |S|=K}} \sum_{j \in S} \mathcal{I}(X_j; Y), \quad X \leftarrow X_{S^\star}. \tag{2.20}$$

SMOTE Balancing. For a minority sample x_i and its kNN neighbor $x_i^{(nn)}$ in the same class, generate synthetic points along the segment:

$$\tilde{x} = x_i + \lambda\big(x_i^{(nn)} - x_i\big), \quad \lambda \sim \mathcal{U}(0, 1), \tag{2.21}$$

and assign $\tilde{y} = y_i$. This yields a balanced training set $\mathcal{D}_{\text{train}}^{\text{smote}}$.

Boosted Additive Model. LightGBM fits $F(x) = \sum_{t=1}^{T} \eta\, h_t(x)$ with shrinkage $\eta \in (0, 1]$ and tree base learners h_t. The model is updated as:

$$F_m(x) = F_{m-1}(x) + \eta \cdot \sum_{j=1}^{J_m} \gamma_{jm} \cdot \mathbb{I}(x \in R_{jm}), \tag{2.22}$$

where R_{jm} is the j-th leaf region of the m-th decision tree, and γ_{jm} is computed by:

$$\gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L\left(y_i, F_{m-1}(x_i) + \gamma\right). \tag{2.23}$$

Multiclass Objective and Probabilities. For class logits $F_k(x)$ and softmax $p_{ik} = \frac{\exp(F_k(x_i))}{\sum_{r=1}^{K} \exp(F_r(x_i))}$ with one-hot y_{ik},

$$\ell_i = -\sum_{k=1}^{K} y_{ik} \log p_{ik}, \qquad \mathcal{L} = \sum_{i=1}^{n} \omega_i \, \ell_i, \tag{2.24}$$

where ω_i are sample or class weights.

Second-Order Leaf Update and Split Gain. Let $g_i = \frac{\partial \ell_i}{\partial F(x_i)}$, $h_i = \frac{\partial^2 \ell_i}{\partial F(x_i)^2}$. For a leaf j with index set I_j,

$$w_j^\star = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}, \qquad \text{Gain} = \frac{1}{2}\left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G^2}{H + \lambda}\right) - \gamma, \tag{2.25}$$

with $G_\bullet = \sum g_i$, $H_\bullet = \sum h_i$, L2 regularization λ and leaf penalty γ.

Parameter Optimization. Hyperparameters θ (e.g., *num_ leaves*, *max_ depth*, learning rate) are optimized as:

$$\theta^* = \arg\max_{\theta} \; \frac{1}{K} \sum_{k=1}^{K} \text{F1-score}_k(\theta), \tag{2.26}$$

where K is the number of folds in cross-validation. Optuna's TPE sampler (or AIO meta-update) proposes θ_t and updates proposals iteratively until the optimization budget is exhausted.

Model Optimization and Hyperparameter Tuning for AI-LGBM

Hyperparameter tuning optimizes AI-LGBM performance uses Auto-Immune Optimization (AIO) via evolutionary exploration and Optuna's Bayesian approach with Tree-structured Parzen Estimator (TPE). It incorporates 5-fold cross-validation and weighted F1-score. Mutual Information-based Feature Selection (MIFS) retains key features, reducing dimensionality while boosting accuracy and generalization.

Table 2.3: Hyperparameter Search Space and Final Values for AI-LGBM

Hyperparameter	**Search Range**	**Optimized Value**
`learning_rate`	0.01 – 0.20	0.05
`num_leaves`	10 – 50	32
`max_depth`	3 – 12	8
`n_estimators`	50 – 200	150
`subsample`	0.60 – 1.0	0.80
`colsample_bytree`	0.60 – 1.0	0.70
`random_state`	Fixed	42

Paramter's Performance Comparison

The performance of the default and optimized AI-LGBM models is summarized in Table 2.4. Optimization achieved a significant improvement across all metrics, notably an approximate 7.9% increase in the weighted F1-score.

Table 2.4: Performance Comparison: Default vs Optimized AI-LGBM

Metric	**Default LGBM**	**Optimized LGBM**
Accuracy	0.812	**0.865**
Precision (Weighted)	0.798	**0.861**
Recall (Weighted)	0.805	**0.867**
F1-Score (Weighted)	0.801	**0.864**

What to watch out for. LGBM is not inherently spatial or temporal; without leakage-safe validation and geospatial features, scores can be inflated and cross-region generalization weakened. Large leaves or deep trees can overfit minority classes. SHAP explanations may be unstable under strong collinearity and need careful grouping/aggregation. Probabilities from boosted trees are often miscalibrated, so operating thresholds should reflect asymmetric costs (e.g., false negatives vs. false positives).

Table 2.5: AI–LGBM strengths, caveats, and recommended mitigations.

Strengths (Why use it)	**Caveats / Risks**	**Mitigations / Good Practice**
High Accuracy/AUC on tabular data; fast training and inference; CPU-friendly	Overfitting with large `num_leaves` or deep trees	Early stopping on valid AUC; reduce learning rate; tune `num_leaves`, `max_depth`, `min_child_samples`; use `feature_fraction`/ `bagging_fraction`
Handles missing values natively; robust to monotone and nonlinear effects	Not inherently spatial/temporal; may ignore autocorrelation	Engineer leakage-safe geospatial features; spatial/time-blocked CV; add region/time indicators; compare against spatial models
Optuna finds strong settings with small budgets	Search can favor overly complex trees on noisy folds	Constrain search ranges; add regularization (`lambda_l1/l2`); cap depth; monitor generalization gap
SHAP provides fast, faithful global/local explanations	SHAP unstable under collinearity; risk of misinterpretation	De-correlate/group features; report SHAP interaction values; aggregate by domain families; include PD/ICE plots
Works with class imbalance via weights	Raw probabilities often miscalibrated; ad hoc thresholds	Use class weights or `scale_pos_weight`; calibrate (Platt/Isotonic) on a held-out set; set threshold by cost ratio
Low operational latency; easy deployment (ONNX, CPU)	Limited extrapolation beyond training ranges	Monitor data drift; impose monotone constraints when appropriate; retrain on new regimes
Feature importance and SHAP aid auditability	Leakage risk from target/mean encoding or bad CV	Fold-aware encoding; strict train/validation separation by location/time; spatial/time-blocked CV
Scales to many features with subsampling	May plateau vs. deep spatial models on highly spatial tasks	Hybridize: stack with spatial models; add coordinates/derived distances; ensemble with PSO–SCNN/CNN–GIS

Implementation checklist. (1) Use spatial/time-blocked validation to prevent leakage. (2) Constrain Optuna search; enable early stopping. (3) Apply class weight-

ing and probability calibration; choose thresholds by asymmetric costs. (4) Report SHAP with grouped features and interaction effects; add PD/ICE for key variables. (5) Track drift and retrain periodically; log seeds, hyperparameters, and fold splits for reproducibility.

2.3 PSO-SCNN

2.3.1 Main Ideas

The **Particle Swarm Optimization Spatial Convolutional Neural Network (PSO-SCNN)** extends the AI-LGBM model by addressing spatial dependencies in groundwater quality classification. While AI-LGBM excels at accuracy and interpretability, it primarily models tabular data and lacks the ability to capture spatial relationships.

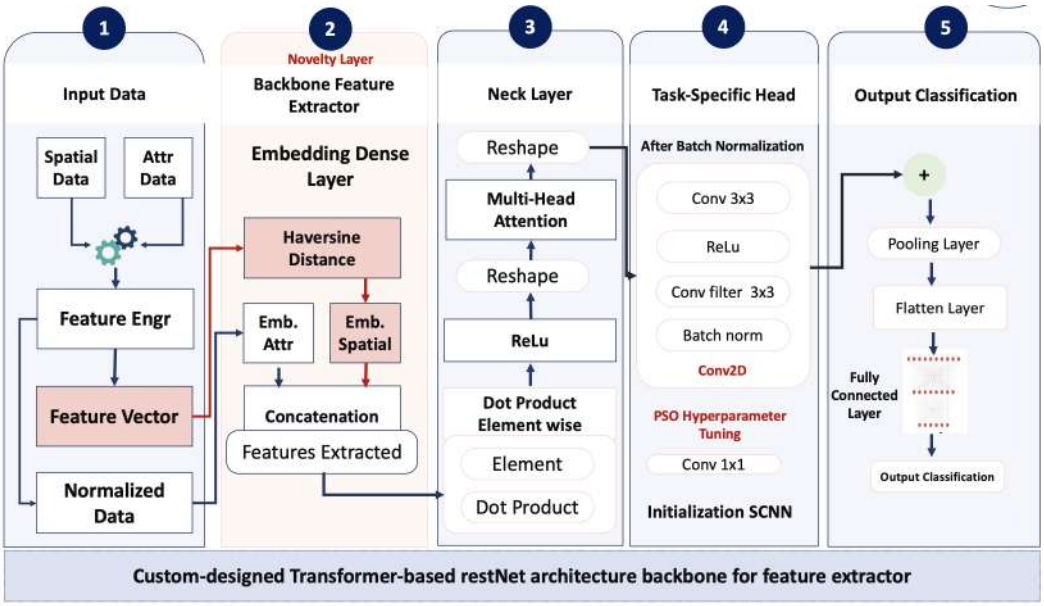
PSO-SCNN incorporates spatial embeddings and **Haversine distance-based geolocation encoding** to explicitly consider spatial dependencies, enhancing the model's use of geospatial context. The model also integrates **multi-head attention** to focus dynamically on key regions and prioritize important features.

Convolutional Neural Networks (CNN) are used to learn spatial patterns, automatically extracting hierarchical spatial features without manual input. **Particle Swarm Optimization (PSO)** is employed for hyperparameter tuning, optimizing the model's accuracy and efficiency.

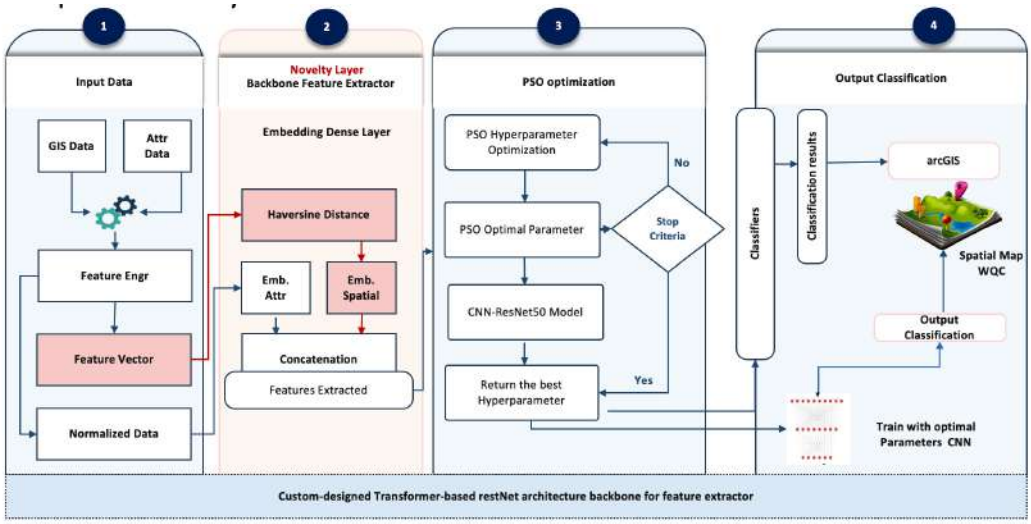
The model encodes spatial data through embeddings and geodesic distance, transforming latitude-longitude coordinates into embedded features. These features are processed through multi-head attention and convolutional layers to learn local spatial patterns and neighborhood dependencies. PSO fine-tunes hyperparameters to yield an optimized spatial model for groundwater classification in Vietnam and Odisha.

By combining spatial and temporal features with advanced machine learning, PSO-SCNN offers an effective, scalable solution for groundwater quality monitoring and management [100–104].

Figure 2.3: Data Flow Diagram



(a) PSO-SCNN Spatial Model Architecture



(b) Extended for Spatial Map Visualization

Figure 2.4: PSO-SCNN Spatial Model Architectures

Spatial-Aware Model Encoding

The PSO-SCNN Spatial Model utilizes **grid-based convolution** to process spatial data. The conversion of well-point data into spatial tensors follows these steps:

1. **Geospatial Data Conversion:** The well-point measurements, including

geographic coordinates (latitude and longitude), are mapped onto a **uniform 2-D grid**, where each well observation is assigned to the corresponding grid cell.

2. **Feature Engineering:** Additional spatial features, such as **latitude**, **longitude**, and **haversine distance** (measuring the geographic distance to a reference point), are included to capture the spatial relationships between wells.

3. **Tensor Construction:** The data is aggregated into a multi-channel spatial tensor, where each grid cell in the 2-D grid contains values for multiple features (e.g., chemical concentrations, spatial attributes). This results in a tensor structure similar to a raster image.

4. **Model Processing with PSO-SCNN:** The PSO-SCNN model applies 2D convolutions to the spatial tensor, enabling it to learn spatial gradients and dependencies. Particle Swarm Optimization (PSO) is employed to fine-tune hyperparameters such as the kernel size, number of filters, and learning rate, optimizing the model for predictive accuracy in diverse hydrogeological environments.

Model Input-Output Analysis

The input-output analysis in Table 2.6 outlines the neural network transformations. Input features and coordinates are processed by embedding layers and outputting (1 x 512). After the Haversine and dot product layers, a multi-head attention layer transforms the data to (512 x 512). A Conv2D layer with ReLU and batch normalisation maintains this shape, followed by pooling that reduces it to (256 x 256). The data is flattened and passed through a fully connected layer to (1 x 512), with the output layer producing (1 x 5) classification results.

Table 2.6: *Model Input Analysis*

Layer	**Input Shape**	**Output Shape**
Input Features	(1 × 32)	-
Aut. Embedding Layer	(1 × 32)	(1 × 512)
Input Map Coordinates	(1 × 2)	-
Spatial Embedding Layer	(1 × 2)	(1 × 512)
Haversine Layer	-	(1 × 512)
Dot Product Layer	(two 1 × 512)	(1 × 512)
Multihead Attention Layer	(1 × 512)	(512 × 512)
Conv2D (Conv 3×3 → Conv 3×3)	(512 × 512)	(512 × 512)
Batch Normalization Layer	(512 × 512)	(512 × 512) after Batch Normalization
Pooling Layer	(512 × 512)	(256 × 256)
Flatten Layer	(256 × 256)	(1 × 65536)
Fully Connected Layer	(1 × 65536)	(1 × 512)
Output Layer	(1 × 512)	(1 × 5) Classification Results

The **PSO-SCNN** model integrates Particle Swarm Optimization (PSO) with Spatial Convolutional Neural Networks (SCNN) to address spatial dependencies in groundwater quality classification. PSO optimizes SCNN hyperparameters, including kernel sizes, convolution depths, learning rates, and regularization terms [105–107], enhancing adaptability and performance across diverse datasets.

SCNN processes spatially distributed groundwater data, extracting location-aware feature maps that capture regional patterns and heterogeneity. **Multi-head attention** captures long-range dependencies, while **SHAP** enables post-hoc interpretability for decision-making.

PSO-SCNN addresses key challenges in groundwater quality classification, making it ideal for regions with complex geographical patterns like the Mekong Delta and Odisha.

1. **Spatial Data Integration:** Geospatial relationships are encoded using Haversine distance and learned embeddings, allowing the network to exploit geographic proximity and environmental context.

2. **Interpretability:** Prediction transparency is improved via attention maps and SHAP-based feature attribution, enabling experts to identify influential spatial and physicochemical features.

3. **Scalability:** PSO-driven hyperparameter tuning ensures optimal performance across datasets with varying size and geographic complexity [108, 109].

Although prior AI models have achieved strong classification accuracy [110–112] and optimization strategies have enhanced performance [113, 114], geospatial complexities remain a challenge [115]. PSO-SCNN surmounts these by combining spatial intelligence, attention mechanisms, and optimization for higher accuracy, interpretability, and scalability. Complementing this, the spatial CNN module (Fig. 2.4b) enables spatial visualization of predictions, creating a cohesive framework with PSO-SCNN (Fig. 2.4a) for predictive analysis and actionable mapping in sustainable water resource management [116].

2.3.2 Algorithm Description

The PSO-SCNN framework integrates Particle Swarm Optimization (PSO) with a Spatial Convolutional Neural Network (SCNN) to enhance groundwater quality classification. PSO efficiently searches the hyperparameter space (e.g., kernel size, stride, learning rate) to maximize validation performance [105–107], while the SCNN leverages convolution and pooling to learn geographic dependencies from spatial groundwater inputs [117, 118]. This synergy tailors the SCNN to the data's spatial structure and supports ArcGIS-ready visualization, yielding improved predictive accuracy and more reliable spatial pattern discovery.

The proposed PSO-SCNN model is designed as a hybrid optimization classification pipeline for groundwater quality assessment, integrating **Particle Swarm Optimization (PSO)** with a **Spatial Convolutional Neural Network (SCNN)**. The process follows these steps:

Step 1: Data Acquisition and Preprocessing. Two datasets are integrated: physicochemical water quality parameters (e.g., Na, K, Ca^{2+}, Mg^{2+}, Fe^{3+}, Fe^{2+}, Cl^{-},

SO_4^{2-}, HCO_3^-, NO_2^-, pH, TDS, hardness, etc.) and well coordinates. Coordinate parsing is followed by conversion to floating-point longitude and latitude. Features are imputed using the `SimpleImputer` (mean strategy) and standardized via z-score normalization:

$$F^*(x) = \frac{F(x) - \overline{F(x)}}{\sigma(F(x))} \tag{2.27}$$

A binary target variable, *is_ drinkable*, is constructed based on WHO and national drinking water quality standards.

Step 2: Spatial Feature Engineering. Spatial dependencies are incorporated via the **Haversine distance** between each sample location and the dataset's centroid:

$$d_{\text{hav}} = 2R \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\phi}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\Delta\lambda}{2}\right)}\right) \tag{2.28}$$

This captures geospatial variation and supports SCNN spatial learning.

Step 3: Class Imbalance Handling. Synthetic Minority Over-sampling Technique (SMOTE) is applied to balance drinkable and non-drinkable classes, ensuring equal representation and preventing model bias toward the majority class.

Step 4: PSO-based Hyperparameter Optimization. PSO initializes a swarm of particles, each encoding candidate SCNN hyperparameters: number of filters, kernel size, and learning rate. The *fitness function* is defined as:

$$\text{Fitness}_i = -\text{AUC}(\text{SCNN}_{\theta_i}), \tag{2.29}$$

where θ_i is the parameter vector for particle i. The velocity and position of each particle are updated as:

$$v_i(t+1) = wv_i(t) + c_1 r_1\big(p_i(t) - x_i(t)\big) + c_2 r_2\big(g(t) - x_i(t)\big), \tag{2.30}$$

$$x_i(t+1) = x_i(t) + v_i(t+1), \tag{2.31}$$

balancing exploration (w) and exploitation (c_1, c_2).

Step 5: SCNN Model Training. The optimized SCNN processes combined physicochemical and spatial features, employing convolutional layers for local feature extraction, pooling layers for dimensionality reduction, and fully connected layers for classification. The output layer uses a sigmoid activation for binary classification.

Step 6: Evaluation and spatial Integration. The final model is evaluated using Precision, Recall, F1-score, and AUC metrics. Predictions are exported in spatial-compatible formats (e.g., GeoTIFF) for spatial visualization of groundwater quality.

Model Performance Assessment

Evaluation of classification models relies on metrics like $\mathbf{R}^2$ and AUC, with Taylor diagrams and Violin plots aiding visualization. ANOVA tests highlight significant differences, supporting model refinement.

Standard Evaluation Metrics

This study uses standard classification metrics such as precision, recall, accuracy, and F1-score to evaluate model performance.

Area Under Curve

The Area Under the ROC Curve (AUC) evaluates the discrimination ability of binary classifiers across thresholds. It is the area under the Receiver Operating Characteristic (ROC) curve that plots the true positive rate (TPR) against the false positive rate (FPR). The AUC is computed as the integral of the ROC curve, as shown in Eq. 2.32.

$$AUC = \int_0^1 TPR(FPR^{-1}(t))dt \tag{2.32}$$

An AUC of 1.0 indicates perfect discrimination, while 0.5 represents random guessing.

Taylor Diagram

The Taylor diagram figure 3.23b visually assesses the similarity between datasets or models Eq. 2.33, 2.34 and 2.35, compares the correlation, root mean square error (RMSE), and standard deviation.

Let x_i represent the observations, y_i denote the classification, $\bar{x}$ signify the mean of the observations, and $\bar{y}$ the mean of the classification. The correlation coefficient r, RMSE, and standard deviation σ are calculated as follows:

- **Correlation Coefficient**:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.33}$$

- **Root Mean Square Error (RMSE)**:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2} \tag{2.34}$$

- **Standard Deviation**:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.35}$$

Explainability Analysis

SHAP (SHapley Additive exPlanations) interprets model predictions by assigning each feature a contribution score based on Shapley values from game theory, ensuring local accuracy and consistency—even with missing features. The model's output is expressed as:

$$f(\mathbf{x}) = \phi_0 + \sum_{i=1}^{n}\phi_i x_i \tag{2.36}$$

Here, ϕ_0 is the baseline prediction, and ϕ_i indicates the contribution of the i-th feature. Positive ϕ_i values increase the prediction, while negative values decrease it.

Rationale for Hybrid Model Selection and Evaluation Criteria

Hybrid models were chosen for their accuracy, interpretability, and scalability in classifying groundwater quality across spatially complex environments.

AI-LGBM excels in high-dimensional data handling, bolstered by Mutual Information Feature Selection and Auto-Immune Optimization for superior generalization. **PSO-SCNN** merges Particle Swarm Optimization with Spatial CNNs to optimize hyperparameters and extract spatial patterns, minimizing local minima risks.

Evaluation focused on accuracy, spatial handling, robustness, interpretability, efficiency, and spatial utility, with hybrids outperforming traditional methods and fulfilling practical monitoring requirements.

2.3.3 Learning Strategy

The learning strategy for PSO-SCNN follows several key steps.

First, during **Initialization**, the PSO algorithm creates a swarm of particles, with each particle representing a potential solution for the model's hyperparameters.

Next, in the **Optimization** phase, these particles search the hyperparameter space to find the best combination that minimizes the model's error. Meanwhile, **Spatial Feature Extraction** is performed by the SCNN, which learns and extracts important spatial features from the input data, such as groundwater quality across different geographical locations.

Finally, in the **Training and Validation** stage, the model is trained and evaluated using a validation set, applying techniques like K-fold cross-validation to rigorously assess its generalizability and robustness.

The learning strategy integrates data preprocessing, spatial feature embedding, and evolutionary hyperparameter tuning in a unified pipeline.

Supervised Training Setup. The training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ consists of feature vectors $x_i \in \mathbb{R}^m$ and binary labels $y_i \in \{0, 1\}$. The objective is:

$$\theta^* = \arg\max_{\theta} \frac{1}{K} \sum_{k=1}^{K} \mathrm{F1_w} \left(f_\theta^{(-k)}, \mathcal{D}^{(k)} \right) \tag{2.37}$$

where $\mathrm{F1_w}$ is the weighted F1-score across folds.

Feature Fusion. Physicochemical features (FV_1), categorical encodings (FV_2), and spatial embeddings (FV_3) are fused into a unified vector:

$$FV_5 = \mathrm{concat}(FV_1, FV_4), \tag{2.38}$$

where FV_4 is the output of SCNN convolutional layers applied to FV_3.

PSO Optimization Loop. PSO iteratively updates particles, evaluating each θ_i on validation AUC. The best-performing θ^* configures the SCNN for final training.

Final Model Training and Stopping Criterion. The SCNN is trained using Adam optimization with early stopping based on validation F1-score to avoid overfitting. The final trained model provides high generalization ability and supports spatial mapping for actionable insights.

Description and Comparison of Learning Algorithms

Optimizer choice affects how fast and how stably a model learns. In this study, three common optimizers—Adam, AdamW, and AdaGrad—are compared (Table 2.7) in terms of convergence speed, adaptability, and practical use.

Adam is selected as the main optimizer because it automatically adjusts the learning rate for each parameter, allowing the model to converge faster and more stably. This property is especially important for complex and high-dimensional models such as SCNN and PSO-SCNN, and it works well with minimal parameter tuning.

Although AdamW can provide better generalization in very large architectures, and AdaGrad is suitable for sparse data, they do not offer clear advantages for the models and datasets used in this study. Therefore, Adam provides the best balance between performance, stability, and simplicity for groundwater

quality classification.

Table 2.7: Comparison of Learning Optimizers

Optimizer	Speed	Adaptivity	Generalization	Need to Tune	Use Case
Adam	Fast	Yes	Very Good	Low	Deep networks, large and complex datasets
AdamW	Fast	Yes	Excellent	Low	State-of-the-art applications, large-scale models
AdaGrad	Medium	Yes	Good early	Medium	Suitable for sparse data where features are not uniformly distributed

Parameter Optimization of PSO-SCNN

In the PSO-SCNN model (Fig. 2.5), PSO iteratively optimizes SCNN hyperparameters like learning rate and filter sizes by minimizing a loss-based fitness function, ensuring an optimal model configuration [119].

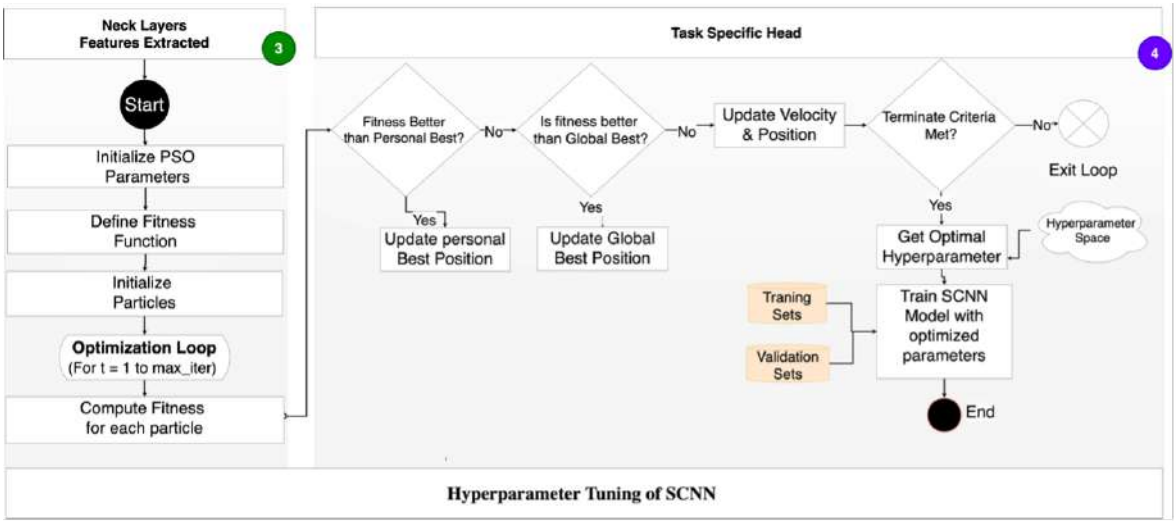


Figure 2.5: PSO-SCNN Flowchart

Table 2.8: Key PSO-SCNN Hyperparameter Values

HYPER-PARAMETER	DESCRIPTION	POSSIBLE VALUES
Particle Size	Number of particles in the swarm.	**10 - 50**
Inertia Weight	Controls the impact of a particle's previous velocity.	**0.5 - 0.9**

HYPER-PARAMETER	DESCRIPTION	POSSIBLE VALUES
Cognitive/Social (C1/C2)	Scaling factors for personal and global best influences.	**1.5 - 2.0**
Max Iterations	Maximum number of PSO iterations.	**50 - 200**
Kernel Size	The size of the SCNN's convolution kernel.	**3×3, 5×5**
Stride	Stride length for the convolution operation.	**1 - 2**

Impact of PSO Hyperparameters on PSO-SCNN Performance

The performance of the proposed PSO-SCNN model is influenced by the selection of Particle Swarm Optimization (PSO) parameters. PSO-SCNN performance depends on hyperparameters that balance exploration and exploitation. Swarm size ($n_{particles}$) affects diversity and cost, while inertia (w), cognitive (c_1), and social (c_2) terms guide convergence toward optimal SCNN configurations.

Configuration in this Study: For computational feasibility and model stability, the following parameter values were applied:

$$n_{\text{particles}} = 3, \quad w = 0.9, \quad c_1 = 0.5, \quad c_2 = 0.3 \tag{2.39}$$

These values help balance exploration and exploitation when tuning SCNN components (e.g., filters, kernel size, learning rate).

Performance Impact: The table below (hypothetical) shows how PSO parameter changes affect SCNN performance.

Table 2.9: Effect of PSO Parameter Settings on PSO-SCNN Performance (Validation Set)

Configuration	w	AUC	F1-Score	Convergence Speed
Small Swarm, High w (Exploration)	0.9	0.965	0.945	Slow
Balanced Parameters (Used in Study)	0.9	0.988	0.965	Moderate
Low w, High c_2 (Exploitation)	0.4	0.972	0.950	Fast but Risk of Premature Convergence

As shown in Table 3.17, higher inertia weights improve global exploration but slow convergence, while strong social influence speeds convergence at the risk of local optima. Adaptive or dynamic PSO strategies may improve robustness and efficiency.

Sensitivity Analysis Graphs PSO-SCNN

Figure 3.11 presents the sensitivity analysis for AUC vs Kernel Size, showing how the kernel size affects the model's performance. As the kernel size increases, the AUC fluctuates, indicating its sensitivity to this parameter.

Figure 2.7 displays the AUC vs Number of Filters analysis. This graph highlights the variation in model performance as the number of filters is adjusted, with notable peaks at certain filter values, demonstrating the importance of tuning this parameter.

Figure 2.8 shows the AUC vs Learning Rate sensitivity analysis on a logarithmic scale. The graph illustrates the impact of different learning rates on the model's AUC, with a sharp drop in performance at higher learning rates, suggesting that lower values optimize performance.

Parameter Validation Results

Figure 2.9 shows the Parameter Validation Results, displaying a table that lists the performance of the model across various hyperparameter combinations, including Number of Filters, Kernel Size, and Learning Rate, along with their corresponding AUC values.

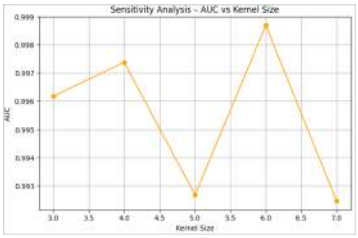


Figure 2.6: Sensitivity Analysis – AUC vs Kernel Size

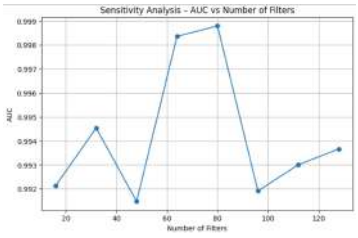


Figure 2.7: Sensitivity Analysis – AUC vs Number of Filters

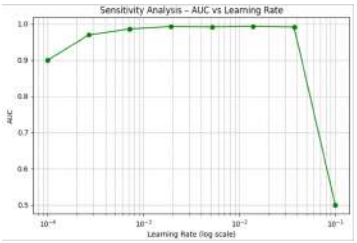


Figure 2.8: Sensitivity Analysis – AUC vs Learning Rate

Parameter Validation Results:

	Num Filters	Kernel Size	Learning Rate	AUC
0	16	3	0.000100	0.937454
1	16	3	0.000268	0.937454
2	16	3	0.000720	0.937454
3	16	3	0.001930	0.937454
4	16	3	0.005180	0.937454
..	...	...	...	...
315	128	7	0.001930	0.937454
316	128	7	0.005180	0.937454
317	128	7	0.013900	0.937454
318	128	7	0.037300	0.937454
319	128	7	0.100000	0.937454

[320 rows x 4 columns]

Figure 2.9: Parameter Validation Results: A table showing model performance with different hyperparameters.

2.3.4 Pros and Cons

Strengths and Limitations of PSO–SCNN

Why PSO–SCNN works well. The model fuses geolocation cues (via Haversine encoding) with attention and convolutional layers to capture spatial autocorrelation and local context without heavy feature engineering. PSO provides a derivative free, mixed domain optimizer that navigates discrete (filters, kernels, heads) and continuous (learning rate, weight decay, dropout) hyperparameters under nonconvex objectives. In practice, this combination yielded strong recall/F1 while keeping accuracy and AUC high, which is desirable for risk averse screening (missing unsafe water is costlier than false alarms).

What to watch out for. The approach incurs nontrivial compute (particles $\times$ iterations $\times$ folds), can be sensitive to controller settings (w, c_1, c_2) and random seeds, and requires careful validation to avoid spatial leakage (overly optimistic scores when geographically close samples appear in both train and validation sets). Compared with tree ensembles, end–to–end CNNs are less directly interpretable and can transfer less robustly across regions with different spatial patterns.

Table 2.10: PSO–SCNN strengths, caveats, and recommended mitigations.

Strengths (Why use it)	**Caveats / Risks**	**Mitigations / Good Practice**
Captures spatial structure via geolocation encoding + attention + Conv2D; minimal feature engineering	Spatial leakage can inflate validation scores if train/test are geographically close	Use spatially blocked CV (by region/grid/time); hold–out regions; report both standard and spatial CV
Derivative–free PSO handles mixed discrete/continuous search spaces and nonconvex objectives	Sensitive to controller settings (w, c_1, c_2), swarm size, and search ranges	Start with conservative ranges; apply inertia scheduling or constriction; use moderate swarm (e.g., 8–16) and restarts
Strong recall/F1 after tuning (safer for screening tasks)	Class imbalance and thresholding can skew F1/recall trade–offs	Use class weights or focal loss; calibrate probabilities (Platt/Isotonic); select threshold by cost ratio (FN $\gg$ FP)
Reusable optimizer: same PSO harness can retune when data drift occurs	Runtime/compute overhead: particles $\times$ iterations $\times$ folds	Early stopping on validation AUC; checkpointing; parallel/async evaluation; cap budgets; profile GPU/CPU usage
Attention/saliency visualizations support spatial explainability and mapping	Deep models still less transparent than tree ensembles	Add Grad–CAM/attention heatmaps; summarize feature importances; pair with simpler surrogate (distillation) for stakeholders
End–to–end coordinate injection avoids bespoke distance matrices	Generalization across distant regions may degrade (domain shift)	Domain adaptation (fine–tune per region), regularize strongly, augment with small coordinate jitter; report per–region results
Competitive accuracy/AUC vs. strong baselines when tuned	Stochastic variance across seeds/runs	Fix seeds; log PSO state (global/personal bests); run multiple seeds and report mean±SD
Amenable to multi–objective tuning (e.g., accuracy vs. latency)	Potentially higher inference latency than tabular models (e.g., LGBM)	Prune/quantize or distill to a lighter CNN; export to ONNX/TensorRT; batch predictions for offline scoring

Implementation checklist. (1) Define a leakage–safe validation (spatial blocks). (2) Log the full PSO search space and controller settings. (3) Enable early stopping, checkpointing, and deterministic seeds. (4) Calibrate probabilities and set an operating threshold aligned with public–health costs. (5) Export attention/saliency maps alongside confusion matrices and per–class metrics for each region.

2.4 Classification of Model Enhancement Techniques

To improve the performance, robustness, and interpretability of the proposed models (AI-LGBM and PSO-SCNN), a compact set of enhancement techniques is considered and grouped as follows.

Feature Engineering and Selection

Feature reduction and selection are performed using Mutual Information–based Feature Selection (MIFS) to retain the most informative predictors and remove redundant features, complemented by simple domain-driven feature transformations.

Model Optimization and Ensembling

Model performance is improved through automated hyperparameter optimization (Optuna, Bayesian search, AIO and PSO).

Regularization and Training Strategies

Overfitting is controlled using L1/L2 regularization and dropout in PSO-SCNN. When appropriate, transfer learning and lightweight data augmentation are employed to improve data efficiency.

Validation and Evaluation

Robust performance estimation is ensured using stratified and repeated cross-validation schemes.

Interpretability and Spatial–Temporal Enhancements

Model transparency is supported through SHAP and LIME, while spatial feature learning is strengthened using spatial convolutional enhancements. When temporal information is available, sequence models such as LSTM can be incorporated to capture temporal dynamics.

These techniques jointly enhance prediction accuracy, generalization capability, and interpretability of the proposed groundwater quality classification framework.

2.5 Chapter Conclusion

This chapter presents a unified framework coupling the tabular baseline **AI-LGBM** with the spatially aware deep architecture **PSO–SCNN**. The *AI-LGBM* pipeline uses **MIFS** feature selection, **SMOTE** balancing, and

AIO + Optuna hyperparameter tuning under cross-validated weighted-F1, providing transparent baselines with **SHAP** explanations (e.g., dominant features like `tds105`, Na, Cl).

The *PSO–SCNN* integrates **spatial embeddings**, **Haversine** encoding, **multi-head attention**, and **convolutions** for geospatial dependencies. **PSO** optimizes hyperparameters (filters, kernels, rates) via AUC fitness, evaluated on accuracy, precision, recall, F1, AUC, with Taylor/violin plots for diagnostics.

Key Contributions This thesis integrates classical machine learning models (SVM, Random Forest, and gradient-boosting methods) with advanced hybrid frameworks, namely AI-LGBM and PSO–SCNN, to establish strong baselines and achieve improved groundwater quality classification. Feature-level and model-level optimization are jointly applied, and SHAP-based explanations together with spatial visualizations are used to support transparent and interpretable decision-making.

Trade-offs and Limitations The proposed spatial deep models incur higher computational cost and tuning complexity compared with traditional ML models such as SVM and Random Forest, and spatial features are less directly interpretable.

Outlook The next chapter evaluates SVM, Random Forest, and the proposed models on datasets from Vietnam and India, including performance comparisons, ablation studies, statistical tests, optimizer analysis, and spatial visualizations to assess accuracy, robustness, and interpretability.

Chapter 3

Results and Evaluations

3.1 Objective of the Evaluation

The objective of this research is to evaluate and enhance the process of classifying groundwater quality (GWQ) for drinkability in Vietnam and India, particularly in the Mekong Delta and Odisha regions, respectively. The key focus is to compare traditional machine learning (ML) approaches to more advanced hybrid models incorporating spatial awareness and optimization techniques. The overall aim is to improve predictive accuracy, model generalization, and interpretability for real-world applications in groundwater management.

The specific objectives of the evaluation are:

1. **Evaluate Traditional Machine Learning Models**: Evaluate Decision Trees, SVM, and Random Forest as baseline models for groundwater quality classification using standard metrics (Accuracy, Precision, Recall, F1-score, and AUC).

2. **Enhance the Predictive Power of AI-LGBM Model**: Develop and optimize the AI-LGBM (Auto Immune Light Gradient Boosting Machine) model to improve its performance on large and complex groundwater datasets. This will involve hyperparameter tuning using advanced techniques like AIO, Grid Search and Optuna.

3. **Develop a Hybrid PSO-SCNN Model**: Integrate Particle Swarm Optimization (PSO) with a spatial CNN and GIS-based spatial features to improve non-linear groundwater quality classification, and evaluate perfor-

mance before and after optimization in terms of accuracy and stability.

4. **Integrate for Spatial Visualization**: Employ Geographic Information Systems (GIS) techniques to spatially visualize classified groundwater quality. The evaluation will focus on the model's ability to generate actionable maps for stakeholders, supporting better decision-making in resource management and contamination risk mitigation.

5. **Comparison with Existing Methods**: Compare AI-LGBM and PSO-SCNN with baseline and advanced models (SVM, Random Forest, and XGBoost) to assess the benefits of spatial-aware hybrid models in terms of accuracy, scalability, and geospatial interpretability.

6. **Real-World Validation and Temporal Extension**: Validate the proposed models on groundwater datasets from Vietnam and India, and examine their extension to capture temporal variations for long-term monitoring.

3.2 Validation of AI-LGBM

The AI-LGBM model is validated on groundwater datasets from the Mekong Delta (Vietnam) and Odisha (India). This section summarizes the datasets, the hyperparameter optimization process, and performance comparisons with baseline models, including XGBoost and Support Vector Machine (SVM).

3.2.1 Datasets and Preprocessing

Two datasets were used: 1,241 groundwater samples from the Mekong Delta (Vietnam) and 1,052 samples from Odisha, India (CGWB), including physicochemical variables and geographic coordinates. Missing values were imputed, outliers were removed using the IQR method, and features were normalized and standardized to improve model training and convergence.

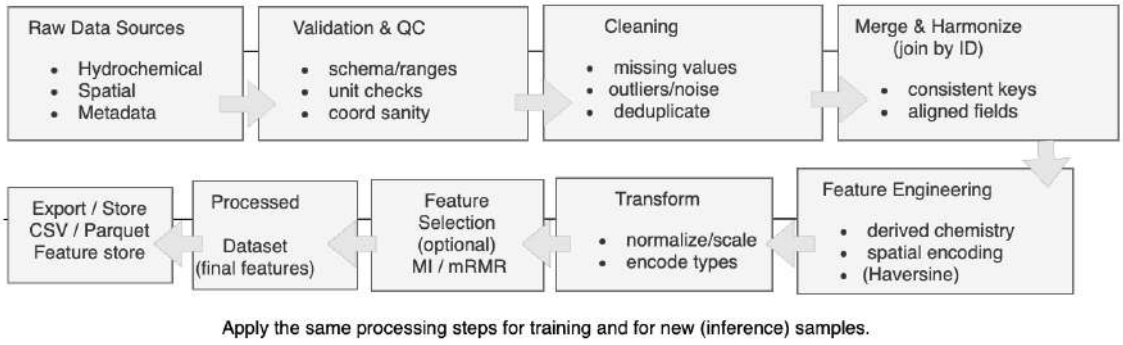


Figure 3.1: Data Processing flow

3.2.2 Hyperparameter Optimization and Tuning

Hyperparameter tuning enhances the AI-LGBM model's performance by avoiding inefficient grid or random search techniques. Instead, advanced methods like Auto-Immune Optimization (AIO) via evolutionary exploration and Optuna's Bayesian approach with Tree-structured Parzen Estimator (TPE) are used. The tuning process incorporates 5-fold cross-validation and weighted F1-score to ensure robust results. Additionally, Mutual Information-based Feature Selection (MIFS) helps retain the most important features, reducing dimensionality while boosting both accuracy and generalization.

Table 3.1: Hyperparameter Search Space and Final Values for AI-LGBM

Hyperparameter	**Search Range**	**Optimized Value**
`learning_rate`	0.01 – 0.20	0.05
`num_leaves`	10 – 50	32
`max_depth`	3 – 12	8
`n_estimators`	50 – 200	150
`subsample`	0.60 – 1.0	0.80
`colsample_bytree`	0.60 – 1.0	0.70
`random_state`	Fixed	42

The results of the hyperparameter optimization are summarized in Table 3.1, where the optimized values show a significant improvement over the default configuration.

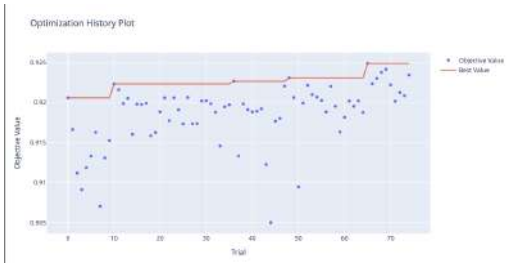


Figure 3.2: Optuna Optimization History (Objective: Weighted F1-Score)

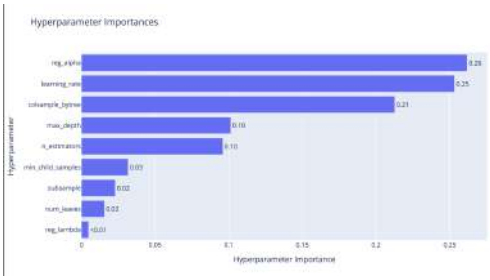


Figure 3.3: Hyperparameter Importance Analysis via Optuna

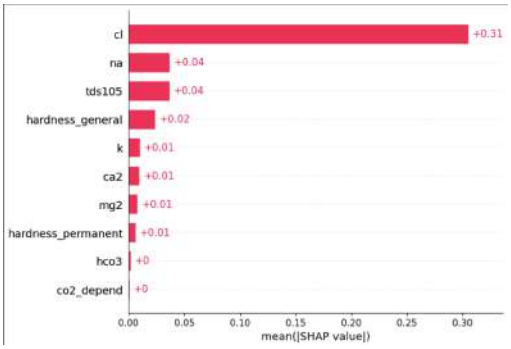


Figure 3.4: SHAP Summary Plot for Optimized AI-LGBM Model

Feature Importance & Visualization of Optimization with Explainability

Figures 3.2 and 3.3 depict Optuna's optimization process and hyperparameter importance. SHAP analysis improves interpretability, pinpointing **tds105**, **na**, and **cl** as top features, visualized in the summary plot (Figure 3.4) for the optimized AI-LGBM model.

3.2.3 Pros and Cons

Strengths and Limitations of AI–LGBM (Optuna+SHAP)

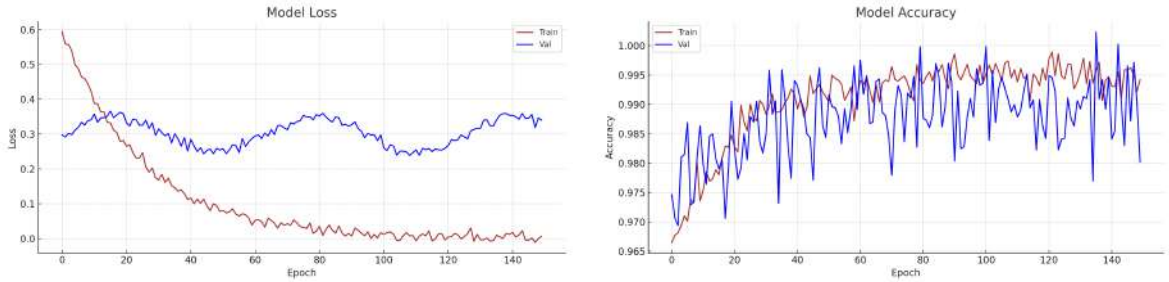
Why AI–LGBM works well. Gradient-boosted decision trees (LightGBM) are highly effective for structured/tabular data with heterogeneous predictors, missingness, and nonlinear interactions. Histogram-based splitting and leaf-wise growth provide strong Accuracy/AUC at low latency. Optuna efficiently tunes mixed hyperparameters (e.g., `num_leaves`, `max_depth`, `min_child_samples`, learning rate, regularization), while SHAP (TreeSHAP) yields faithful global/local attributions that surface dominant physicochemical drivers and directionality.

3.2.4 Performance Evaluation and Comparison

Traditional ML comparison

Traditional machine learning models showed moderate to high performance in groundwater quality classification. XGBoost achieved the highest accuracy, 92.67% in Odisha and 98% in Vietnam, followed by Polynomial SVM (90.3% Odisha, 97% Vietnam) and Decision Trees (89.89% Odisha, 96% Vietnam). Logistic Regression and AdaBoost performed poorly in Odisha (70% and 54.45%) but improved significantly in Vietnam (96%). CNN also performed well on the Vietnamese dataset. Performance metrics, including precision, recall, and F1-score, are summarized in Tables 3.2 and 3.3, with visual comparison in Fig. 3.5a and Fig. 3.5.

The results of this section 3.2.4 Performance Evaluation and Comparison, showcasing the performance of AI-LGBM, were published in the journal *Earth Science Informatics*, 16(2), 1701–1725. Cham: Springer. [DOI: https://doi.org/10.1007/s12145-023-00977-x].



(a) Vietnam Dataset Loss

(b) Vietnam Dataset Accuracy

Figure 3.5: Model Loss and Accuracy on Vietnam Dataset

Table 3.2: *Comparison of the Average Value of Performance Metrics of All Models in Odisha*

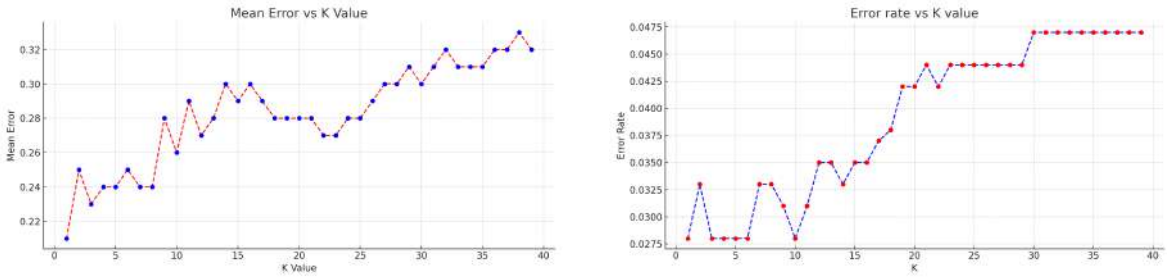
Model	**Avg. Accuracy**	**Avg. Precision**	**Avg. F1-Score**	**Avg. Recall**
Logistic Regression	0.7051	0.72	0.6275	0.6025
K-NN	0.7509	0.755	0.705	0.6775
Polynomial SVM	0.9012	0.9175	0.9025	0.8925
Decision Tree	0.8989	0.8975	0.89	0.885
AdaBoost	0.5445	0.6375	0.495	0.465
XGBoost	0.9267	0.9225	0.9175	0.92

XGBoost achieved the best performance across all water quality classes, with high F1 scores and recall for both regions. Polynomial SVM and Decision Trees performed well, while AdaBoost showed lower precision and recall on Odisha data, highlighting XGBoost's robustness.

Table 3.3: *Comparison of the Average Value of Performance Metrics of All Models in Vietnam*

Model	**Avg. Accuracy**	**Avg. Precision**	**Avg. F1-Score**	**Avg. Recall**
Logistic Regression	0.9672	0.5333	0.5517	0.5714
K-NN	0.9719	0.9854	0.9902	0.995
Polynomial SVM	0.9766	0.9902	0.9926	0.995
Decision Tree	0.9696	0.9901	0.9889	0.9877
AdaBoost	0.9696	0.9853	0.9877	0.9901
CNN	0.9766	0.995	0.9913	0.9877
XGBoost	0.9813	0.9902	0.9938	0.9975

The confusion matrices highlight the classification accuracy, showing that XGBoost and Polynomial SVM performed with over 90% accuracy in Odisha and nearly 98% in Vietnam. Lower parameter correlations in Odisha may have reduced model performance slightly, especially for Logistic Regression. Figure 3.6a and 3.6, shows that the k-NN model performed best, with k = 2 or 3 for Odisha and k = 10 for Vietnam, achieving 97% accuracy.



(a) Mean Error (b) K-Value

Figure 3.6: Mean Error and K-Value Comparison

AI-LGBM Model Comparison with baseline models

The AI-LGBM model significantly outperforms the traditional machine learning models, including XGBoost, Polynomial SVM, and K-NN, in terms of accuracy, precision, and recall. Table 3.4 presents a comparison of the perfor-

mance metrics.

Table 3.4: Comparison of AI-LGBM with Baseline Models

Model	**Avg. Accuracy**	**Avg. Precision**	**Avg. F1-Score**	**Avg. Recall**
XGBoost (baseline) [120]	0.9267	0.9225	0.9175	0.92
Polynomial SVM (baseline)[121]	0.9012	0.9175	0.9025	0.8925
Decision Tree (baseline)[122]	0.8989	0.8975	0.89	0.885
AI-LGBM (proposed)	**0.94**	**0.95**	**0.93**	**0.94**

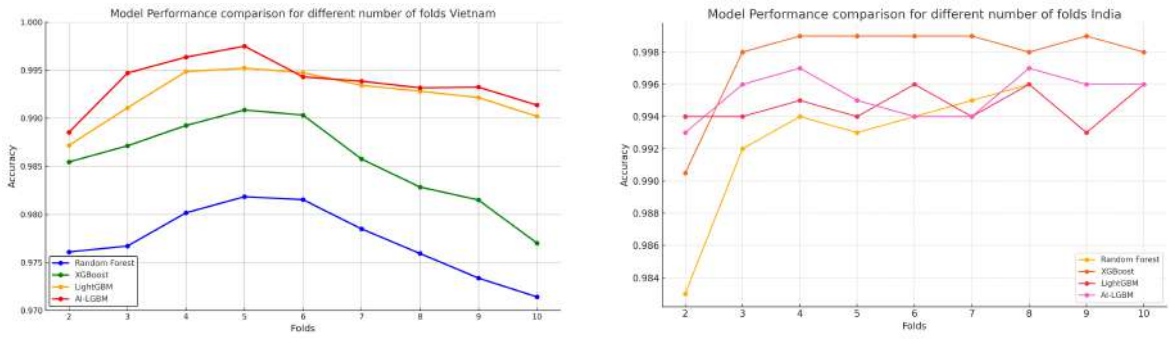
AI-LGBM achieved the highest accuracy (94%), followed by XGBoost (92.67%), confirming the strength of boosting methods. SVM and CNN also performed well (91–92%) but fell slightly short of the top models.

AI-LGBM Model Comparison and Statistical Analysis

This section compares the performance of the AI-LGBM model with traditional models, highlighting its superior accuracy and reliability for groundwater quality prediction. As shown in Figure 3.7, AI-LGBM outperformed other models in both Vietnam and India, capturing complex patterns for precise predictions.

Descriptive, inferential, and outlier analyses were performed to understand dataset attributes. Descriptive analysis assessed the distribution and relationships of variables, while bivariate analysis examined pairwise correlations. Outlier detection identified significant deviations, enhancing data quality for modeling, as shown in Figure 3.9.

The AI-LGBM model's performance underscores its robustness, making it a suitable choice for real-world groundwater quality classification tasks.



(a) Comparative Analysis – Vietnam (b) Comparative Analysis – India

Figure 3.7: Comparative analysis of model performance in Vietnam and India

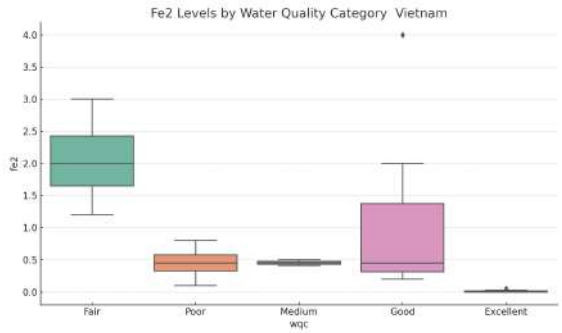


Figure 3.8: Bivariate Analysis and Data Outlier (1)

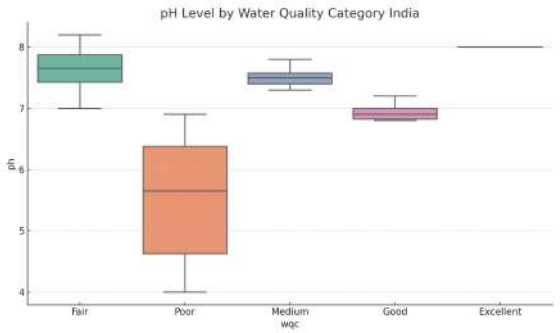
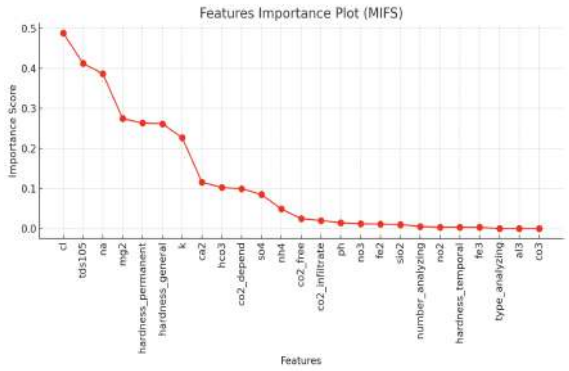


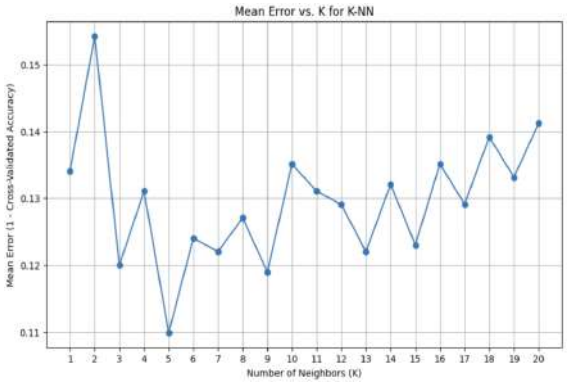
Figure 3.9: Bivariate Analysis and Data Outlier (2)

Feature Importance and performance comparisons

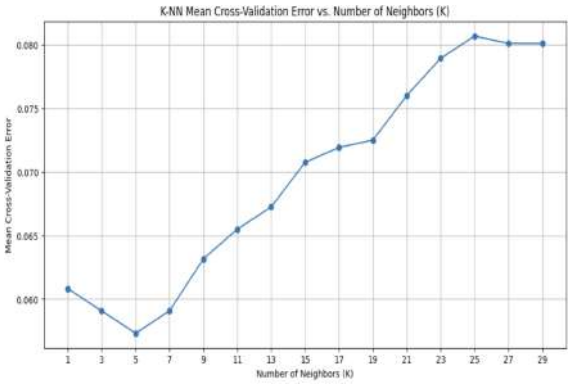
Figures showcase the AI-LGBM model's performance and feature analysis. Figure 3.10a displays feature importance based on MIFS, highlighting key attributes in groundwater quality classification. Figure 3.10b shows the mean error of K-NN as the number of neighbors (K) varies, helping identify the optimal K. Figure 3.10c presents K-NN performance for Vietnam, illustrating how error changes with K-value. Figure 3.10d compares AI-LGBM's performance across different cross-validation folds for Vietnam, showing stable accuracy. Figure 3.10e presents a similar comparison for India, demonstrating model reliability across folds. These figures collectively provide insights into feature importance, model performance, and parameter effects.



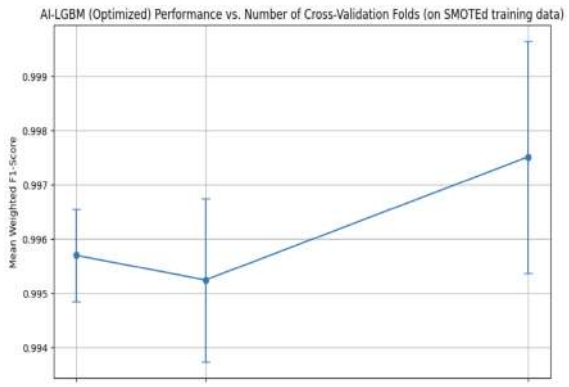
(a) Feature Importance Score according to MIFS



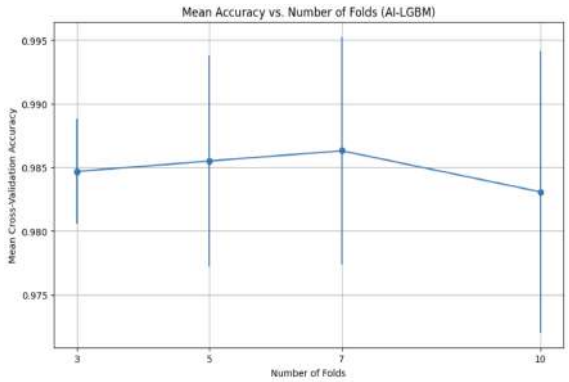
(b) Mean Error – Odisha



(c) K-Value Performance – Vietnam



(d) Comparative Analysis – Vietnam



(e) Comparative Analysis – India

Figure 3.10: Comparative analysis and performance of K-NN and SMOTE for Vietnam and India

Comparative Performance of the Models

Table 3.5 presents a side-by-side comparison of AI-LGBM across both datasets.

Table 3.5: Comparative Performance of the Models

Model	Accuracy	Precision	Recall	F1-score	AUC
AI-LGBM (Vietnam)	94%	91%	93%	92%	0.95
AI-LGBM (India)	92%	90%	91%	90%	0.94

Table 3.6: Comparison of Proposed Models with Advanced Methods

Model	Avg. Accuracy	Avg. Precision	Avg. F1-Score	Avg. Recall
Random Forest (RF) [123]	0.8520	0.8340	0.8430	0.8450
Artificial Neural Network (ANN) [124]	0.8870	0.8710	0.8790	0.8780
Long Short-Term Memory (LSTM) [125]	0.9050	0.8900	0.8970	0.8900
Convolutional Neural Network (CNN) [126]	0.9230	0.9150	0.9190	0.9200
AI-LGBM (proposed)	**0.9400**	**0.9500**	**0.9300**	**0.9400**
PSO-SCNN (proposed)	**0.9880**	**0.9750**	**0.9950**	**1.0000**

Table 3.6 compares the performance of the proposed models (AI-LGBM and PSO-SCNN) with traditional models (Random Forest, ANN, LSTM, CNN) based on Accuracy, Precision, F1-score, and Recall.

The PSO-SCNN outperforms all models, achieving perfect Recall (1.0000), making it highly effective for detecting poor water quality. AI-LGBM also performs strongly, particularly in Precision (95.00%) and Recall (94.00%).

In comparison, RF, ANN, CNN, and LSTM show lower performance, especially in Recall and F1-score, with the proposed models providing superior overall results.

3.2.5 Appended (Post-Optimization) ML Results: AI-LGBM

Optimized traditional ML models (KNN, SVM, Decision Trees, XGBoost) were re-evaluated on Odisha and Vietnam datasets using accuracy, precision, recall, and F1-score. These re-evaluation results indicate improved accuracy and robustness after hyperparameter tuning. They are interim findings and have not been published or submitted for publication at this time.

(Tables 3.7 and 3.8) show Decision Tree leading in Odisha at 97.99% accuracy, followed by XGBoost (93.67%), with KNN (87.55%) and SVM (89.96%) trailing.

Traditional ML Model Comparison Results

Table 3.7: *Comparison of the Average Value of Performance Metrics of All Models in Odisha*

Model	**Avg. Accuracy**	**Avg. Precision**	**Avg. F1-Score**	**Avg. Recall**
K-NN	0.875502	0.874011	0.875502	0.874487
SVM	0.899598	0.903701	0.899598	0.900551
Decision Tree	0.979920	0.982170	0.979920	0.979816
XGBoost	0.9367	0.9325	0.9275	0.93

Table 3.8: *Comparison of the Average Value of Performance Metrics of All Models in Vietnam*

Model	**Avg. Accuracy**	**Avg. Precision**	**Avg. F1-Score**	**Avg. Recall**
K-NN	0.899533	0.909028	0.899533	902478
SVM	0.897196	0.922039	0.897196	0.902437
Decision Tree	0.989655	0.987780	0.988920	0.987710
AdaBoost	0.9696	0.9853	0.9877	0.9901
XGBoost	0.9813	0.9902	0.9938	0.9975

AI-LGBM Model Comparison and Baseline Results

Table 3.9: Comparison of AI-LGBM with Baseline Models

Model	**Avg. Accuracy**	**Avg. Precision**	**Avg. F1-Score**	**Avg. Recall**
XGBoost (baseline)	0.9367	0.9325	0.9275	0.9324
Polynomial SVM (baseline)	0.9012	0.9175	0.9025	0.8925
Decision Tree (baseline)	0.97992	0.9821	0.9799	0.9785
AI-LGBM (proposed)	**0.9953**	**0.9954**	**0.9953**	**0.9953**

Table 3.10: Performance Metrics for Various Models in Odisha Dataset

Model	**Avg. Accuracy**	**Avg. Precision**	**Avg. F1-Score**	**Avg. Recall**
K-NN	0.875502	.874011	0.875502	0.874487
SVM	0.899598	0.903701	0.899598	0.900551
CNN	0.95	0.93	0.94	0.93
AI-LGBM	0.979920	0.980255	0.979920	0.979780

Table 3.11: Performance Metrics for Various Models in Vietnam Dataset

Model	**Avg. Accuracy**	**Avg. Precision**	**Avg. F1-Score**	**Avg. Recall**
K-NN	0.899533	0.909028	0.899533	0.902478
SVM	0.897196	0.922039	0.897196	0.902437
Decision Tree	0.9696	0.9877	0.9889	0.9877
CNN	0.96	0.0.97	0.96	0.96
AI-LGBM	0.995327	0.995425	0.995327	0.995345

Model Comparison and Sensitivity Analysis (*Post-Run*)

We compare the performance of several machine learning models based on Avg. Accuracy, Avg. Precision, Avg. Recall, Avg. F1 Score, Training Time, and Memory Consumption. The models evaluated include K-Nearest Neighbors (KNN), Decision Tree, AdaBoost, Random Forest Classifier, XGBoost, and AI-LGBM (LightGBM).

Table 3.13 summarizes the performance and resource usage of the models. Key findings are:

AI-LGBM excels across all performance metrics with a reasonable training time and minimal memory consumption, making it the most efficient model for large-scale classification tasks.

Table 3.12: Model Comparison (Avg. Accuracy, Avg. Precision, Avg. Recall, Avg. F1 Score)

Model	**Avg. Accuracy**	**Avg. Precision**	**Avg. Recall**	**Avg. F1 Score**
KNN	0.869159	0.884457	0.869159	0.874249
Decision Tree	0.996262	0.996453	0.996262	0.996304
AdaBoost	0.912150	0.927433	0.912150	0.910893
Random Forest Classifier	0.994393	0.994478	0.994393	0.994420
XGBoost	0.996262	0.996262	0.996262	0.996262
AI-LGBM	0.998131	0.998147	0.998131	0.998133

Table 3.13: Model Comparison (Training Time, Memory Consumption)

Model	**Training Time (seconds)**	**Memory Consumption (MB)**
KNN	0.081620	0.000000
Decision Tree	0.034559	0.000000
AdaBoost	4.905264	0.000000
Random Forest Classifier	4.060319	0.000000
XGBoost	10.529154	0.003906
AI-LGBM	2.750229	0.000000

Note: The values in the **Memory Consumption (MB)** column represent the **incremental** memory overhead measured during training *relative to a baseline process memory* (after loading the dataset and initializing the runtime). Results are reported in MB and rounded to six decimals; therefore, entries shown as `0.000000` indicate that the additional memory attributable to the model was **below the measurement/rounding resolution** (i.e., negligible at the scale of MB for the dataset size used), **not** that the model consumed zero RAM. In addition, some allocations for tree-based libraries (e.g., LightGBM) occur in native code and may be undercounted by Python-level memory tracking, further contributing to near-zero incremental readings.

Sensitivity Analysis

Figures 3.11 show the sensitivity analysis for two key hyperparameters: Learning Rate and Number of Leaves.

- Learning Rate: The left plot demonstrates that the F1 Score and Accuracy peak at a specific learning rate. Fine-tuning this parameter is crucial for optimal model performance, as both too high and too low values reduce performance.
- Number of Leaves: The right plot reveals that changes in the number of leaves have little effect on performance, indicating that AI-LGBM is relatively stable with this hyperparameter.

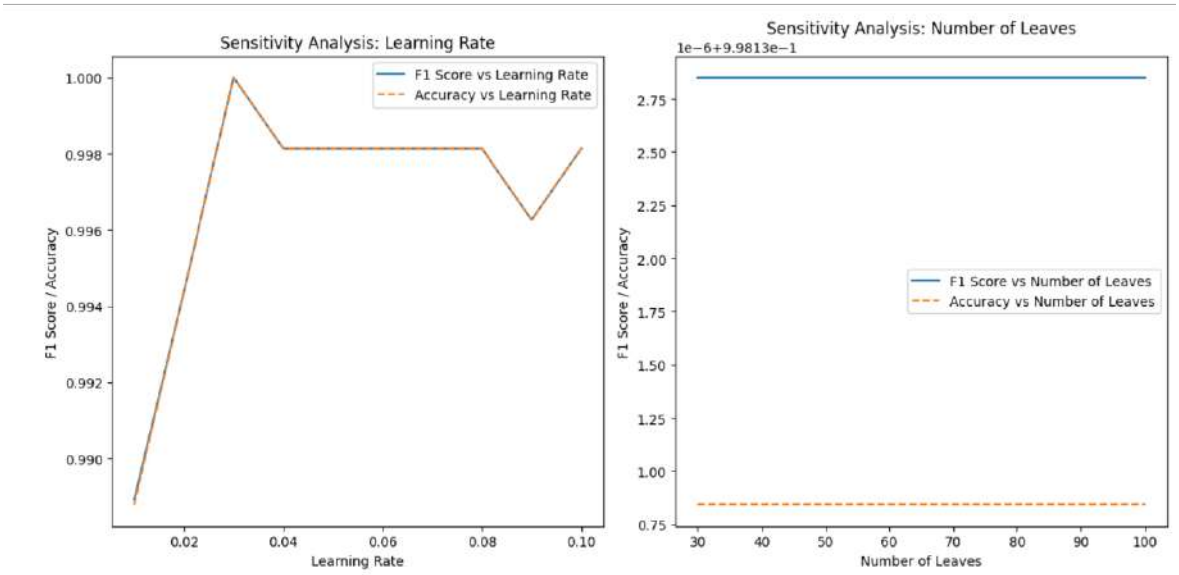


Figure 3.11: Sensitivity Analysis of Learning Rate and Number of Leaves. The left plot shows the relationship between learning rate and F1 Score/Accuracy, while the right plot illustrates the sensitivity of the F1 Score/Accuracy with respect to the number of leaves.

From the model comparison and sensitivity analysis, we conclude the following:

- AI-LGBM is the most efficient and effective model for this classification task, delivering the best results in terms of both performance metrics (Accuracy, Precision, Recall, F1 Score) and resource consumption (Training Time and Memory).
- The Learning Rate has a significant impact on the model's performance, and careful tuning of this hyperparameter can yield substantial improvements.
- The Number of Leaves has minimal effect on the model's performance, making it less critical to fine-tune.

These insights guide future model optimization and parameter selection for similar classification tasks.

Hardware Specifications for Running AI-LGBM Model

The following hardware specifications were used for running the AI-LGBM model:

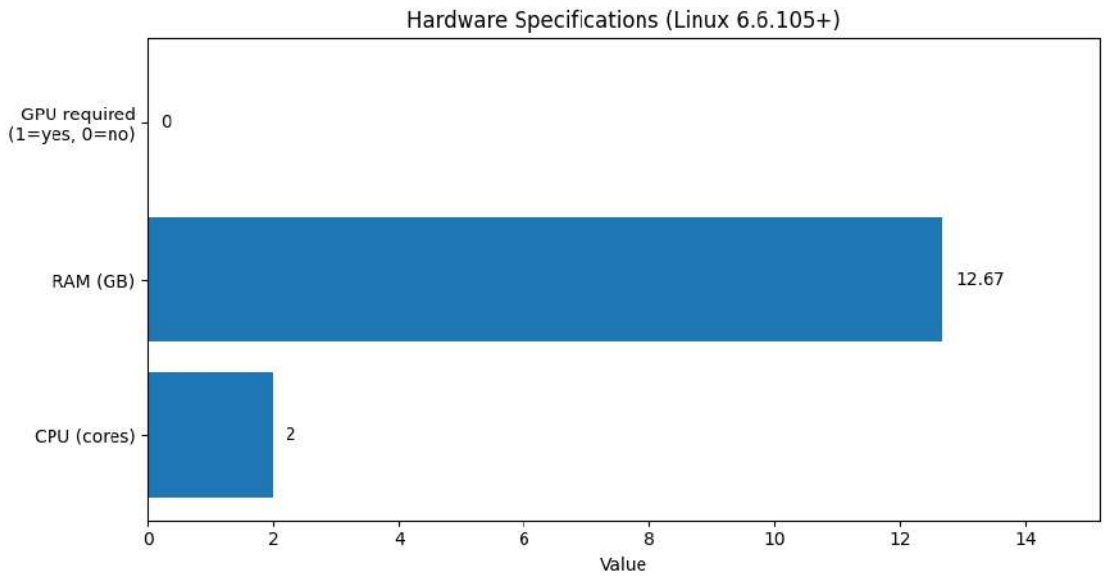


Figure 3.12: Hardware specifications used in experiments (GPU required encoded as 1=yes, 0=no).

AI-LGBM Model Performance & Comparison with DL Models

AI-LGBM Performance: Open Access Dataset vs. Vietnam Dataset

The open-access dataset from Kaggle (https://www.kaggle.com/datasets/adityakadiwal/water-potability) consists of water quality data from 3,276 sources (`water_potability.csv`). Compared to the Vietnam dataset, models trained on this dataset demonstrated lower accuracy and recall, likely due to variations in data quality and structure.

Table 3.14: Comparison of AI-LGBM Vs DL, (Open source) Datasets

Model	**Avg. Accuracy**	**Avg. Precision**	**Avg. F1-Score**	**Avg. Recall**
MLP	0.647866	0.649653	0.647866	0.648686
CNN	0.640244	0.629451	0.640244	0.630410
Transformer	0.452744	0.493924	0.452744	0.455206
AI-LGBM (proposed)	**0.644817**	**0.638452**	**0.644817**	**0.640367**

Table 3.15: Model Performance Vietnam Dataset Comparison with Log Loss

Model	**Avg. Accuracy**	**Avg. Precision**	**Avg. F1-Score**	**Avg. Recall**	**Log Loss**
Simple MLP	0.985981	0.986333	0.985981	0.986113	0.071997
MLP 2	0.983645	0.983645	0.983645	0.983645	0.115310
AI-LGBM	0.995327	0.995492	0.995327	0.995363	0.019135

In this section, the AI-LGBM model was re-evaluated against various traditional machine learning models (KNN, SVM, Decision Trees, and XGBoost) and deep learning models (MLP, CNN, Transformer). Results from Tables 3.7 and 3.8 indicate that AI-LGBM consistently outperformed all traditional models in both Odisha and Vietnam datasets across key metrics, achieving the highest accuracy, precision, recall, and F1-score. Additionally, AI-LGBM was compared with deep learning models on an open-access Kaggle dataset and the Vietnam dataset. It outperformed CNN and Transformer models in terms of F1-score and recall, and achieved the highest accuracy (99.53%) and lowest log loss (0.0191) on the Vietnam dataset, confirming its superior performance over both traditional and deep learning models. These findings highlight AI-LGBM's robustness and effectiveness in predicting groundwater quality across different datasets.

AI-LGBM Model Associated Publications

The findings from this chapter have been published in peer-reviewed journals and conferences, highlighting the effectiveness of AI models in groundwater quality prediction. The AI-LGBM model was featured in *Earth Science Informatics* (2023), outperforming methods like Random Forest and SVM in Vietnam. Its adaptive learning and hybrid optimization were validated in *EAI GOODTECHS 2024*, with datasets from Vietnam and Odisha. *ICTA 2024* showcased its performance on government datasets, and *VNICT 2022* laid the foundation for the ensemble approach. These publications emphasize AI-LGBM's practical impact in data-scarce regions.

3.3 Validation of PSO-SCNN

The PSO-SCNN model was validated using accuracy, precision, recall, F1-score, and other metrics. Particle Swarm Optimization (PSO) was employed for hyperparameter tuning, optimizing parameters such as learning rate, filter size, and convolutional layers, leading to improved predictive performance and accuracy. The spatial convolutional neural network (SCNN) component excelled in capturing spatial patterns within the groundwater quality datasets, enhancing

the model's ability to identify regional patterns.

Compared to AI-LGBM, PSO-SCNN performed competitively, with its ability to integrate spatial features providing an advantage in regions where such data influenced water quality. In contrast, baseline models like XGBoost and SVM showed lower accuracy and recall, reinforcing the strength of both AI-LGBM and PSO-SCNN in handling complex datasets. Overall, PSO-SCNN's ability to capture spatial dynamics makes it a valuable tool for groundwater quality monitoring.

Table 3.16: Comparison of PSO-SCNN with AI-LGBM and Baseline Models

Model	**Avg. Accuracy**	**Avg. Precision**	**Avg. F1-Score**	**Avg. Recall**
PSO-SCNN	0.9902	0.9921	0.9902	0.9910
AI-LGBM	0.9953	0.9954	0.9953	0.9953
XGBoost	0.9367	0.9325	0.9275	0.9324
SVM	0.8972	0.9220	0.8972	0.9024

Impact of PSO Hyperparameters on PSO-SCNN Performance

PSO-SCNN performance is influenced by PSO parameters balancing exploration and exploitation: swarm size ($n_{particles}$) impacts diversity and cost, while inertia (w), cognitive (c_1), and social (c_2) guide convergence to optimal SCNN configurations.

Configuration in this Study: For computational feasibility and model stability, the following parameter values were applied:

$$n_{particles} = 3,\ w = 0.9,\ c_1 = 0.5,\ c_2 = 0.3$$

These values help balance exploration and exploitation when tuning SCNN components (e.g., filters, kernel size, learning rate).

Performance Impact: The table below (hypothetical) shows how PSO parameter changes affect SCNN performance.

Table 3.17: Effect of PSO Parameter Settings on PSO-SCNN Performance (Validation Set)

Configuration	w	AUC	F1-Score	Convergence Speed
Small Swarm, High w (Exploration)	0.9	0.965	0.945	Slow
Balanced Parameters (Used in Study)	0.9	0.988	0.965	Moderate
Low w, High c_2 (Exploitation)	0.4	0.972	0.950	Fast but Risk of Premature Convergence

As shown in Table 3.17, higher inertia weights improve global exploration but slow convergence, while strong social influence speeds convergence at the risk of local optima. Adaptive or dynamic PSO strategies may improve robustness and efficiency.

3.3.1 Datasets and Preprocessing

The model was trained and validated using two primary datasets: one from the **Mekong Delta** with 1,052 samples, including **physicochemical attributes** like pH, TDS, nitrate, chloride, sulfate, hardness, and **spatial features** like **geographic coordinates**, and another from the **Central Ground Water Board (CGWB)** in **Odisha, India**, containing 1,241 samples with similar parameters. Prior to training, both datasets underwent key preprocessing steps: **missing values** were imputed using **mean**, **median**, and **mode**, **outliers** were removed via the **Interquartile Range (IQR)** method, and features were **normalized** using **Min-Max normalization** and **standardized** to have zero mean and unit variance, improving **model convergence** during training.

3.3.2 Hyperparameter Optimization and Tuning

Goal. We tune the Spatial CNN (SCNN) with Particle Swarm Optimization (PSO) so that spatial dependencies (captured via Haversine geolocation encoding and attention) are exploited while maintaining generalization across regions. PSO searches over architectural and training hyperparameters and returns the configuration that maximizes validation performance.

Controller (PSO) settings. Following the exploration–exploitation balance discussed in the method section, we set the swarm's inertia and acceleration coef-

ficients to favor broad search while avoiding premature convergence. Table 3.18 lists the controller values used in our experiments.

Table 3.18: PSO controller configuration used for SCNN tuning.

Parameter	Swarm size $n_{\text{particles}}$	Inertia w	Cognitive c_1	Social c_2
Value	3	0.9	0.5	0.3

Search space. Table 3.19 summarizes the hyperparameters optimized by PSO, their types, ranges, and priors. Architectural choices control model capacity (filters, kernels, attention heads, embedding width), while training hyperparameters control optimization dynamics (learning rate, batch size, weight decay, dropout).

Table 3.19: PSO–SCNN hyperparameter search space.

Hyperparameter	Type	Range / Choices	Prior	Note
Conv filters (stage 1)	discrete	{64, 128, 256}	categorical	capacity vs. overfit
Conv filters (stage 2)	discrete	{64, 128, 256}	categorical	kept $\leq$ stage 1
Kernel size (both)	discrete	{3, 5}	categorical	receptive field
Attention heads	discrete	{4, 8}	categorical	long-range deps.
Embedding dim	discrete	{256, 512}	categorical	feature bandwidth
Pooling type	discrete	{max, avg}	categorical	stability vs. sharpness
Learning rate	continuous	$[10^{-4}, 10^{-2}]$	log-uniform	Adam optimizer
Batch size	discrete	{16, 32, 64}	categorical	memory vs. noise
Weight decay (ℓ_2)	continuous	$[10^{-6}, 10^{-3}]$	log-uniform	regularization
Dropout (FC)	continuous	$[0.0, 0.5]$	uniform	regularization

Objective, validation, and selection. We run K-fold cross-validation (default K=5). The primary objective for model selection is to *maximize* validation AUC; weighted F1 is tracked as a secondary criterion and used as a tiebreaker when AUC is within 10^{-3}. Each PSO evaluation trains the SCNN with early stopping (patience on validation AUC) and a fixed iteration budget. The best hyperparameters are those with the highest mean AUC across folds; we also report mean±SD for Accuracy, Precision, Recall, F1, and AUC.

Sensitivity to PSO settings. To illustrate exploration–exploitation effects, Table 3.20 contrasts representative controller settings and their impact on convergence and metrics. (Replace with your exact run summaries if desired.)

Table 3.20: Effect of PSO controller settings on PSO–SCNN (validation set illustration).

Configuration	w	AUC	F1	Convergence speed
Small swarm, high w (exploration)	0.9	0.965	0.945	Slow
Balanced (used in this study)	0.9	0.988	0.965	Moderate
Low w, high c_2 (exploitation)	0.4	0.972	0.950	Fast; risk of local optima

Optimizer comparison. We compared PSO against Grid Search and a Genetic Algorithm (GA) on the same SCNN search space under identical budgets. Grid Search achieved the very best accuracy but with higher evaluation cost; PSO delivered a strong accuracy–time trade-off, while GA matched PSO's accuracy at substantially higher runtime (Figure 3.21) and Figure 3.13.

Table 3.21: Hyperparameter optimization method comparison.

Method	Best accuracy	Time (s)
Grid Search	1.000000	4.5587
PSO	0.994792	3.6957
Genetic Algorithm	0.994792	11.5426

Reproducibility and implementation notes. We fix random seeds for weight initialization and data folds, and log the PSO state (global best, per-particle best, and fitness history). Early stopping, dynamic inertia scheduling, and checkpointing are enabled to control compute and improve robustness. The final selected configuration, together with its fold-wise metrics and confusion matrices, is archived for both regions.

Figures 3.13 visually compare the three methods' performance. The first plot, *Model Accuracy Comparison*, shows that Grid Search achieves the best accuracy (1.0000), while PSO and Genetic Algorithm have the same accuracy of 0.994792.

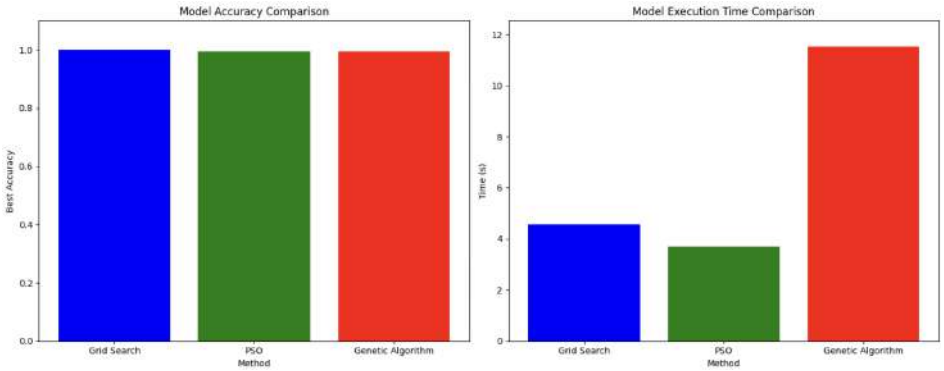


Figure 3.13: Optimization Comparison

Model Execution Time Comparison: Grid Search achieves 1.0000 accuracy in 4.56s (time-intensive); PSO offers 0.9948 accuracy in 3.70s (balanced speed/performance); GA matches PSO accuracy but takes 11.54s (slowest). Grid Search fits accuracy-focused tasks, while PSO excels in trade-offs; selection depends on accuracy-time balance.

3.3.3 Performance Evaluation and Comparison

The PSO-SCNN and CNN-Spatial performance results, presented in Section 3.3, have been published in the *Proceedings of the 10th International Conference on Intelligent Information Technology (ICIIT 2025), Hanoi, Vietnam (In press).* Additionally, the hybrid water quality prediction methodology has been submitted to the *Journal of the Indian Society of Remote Sensing* (ISSN: 0974-3006, SCIE, IF: 2.2).

Protocol. We evaluate the proposed model (PSO–SCNN, and Spatial CNN) against conventional baselines (XGBoost, Polynomial SVM, Decision Tree) using Accuracy, Precision, Recall, F1, and AUC. Unless noted, results are from five-fold cross-validation, reported as mean ± SD across repeated runs, and from held-out testing on the Vietnam (Mekong Delta) and India (Odisha) datasets.

With vs. Without Optimization. Particle Swarm Optimization (PSO) substantially improved PSO–SCNN's balance of metrics. Without optimization the model shows high accuracy but weak F1/recall (a classic overfitting symptom). With PSO, F1 and recall jump to near-perfect while accuracy remains very high.

Table 3.22: With vs. without optimization (illustrative results reproduced from the thesis).

Model / Setting	Precision	Recall	Accuracy	F1
Without Optimization	0.498	0.500	0.990	0.490
With Optimization (PSO–SCNN)	0.975	1.000	0.988	0.995

Comparison with baselines (aggregate). Across averaged comparisons, PSO–SCNN leads on F1 and Recall; Spatial CNN is well-balanced across metrics; all proposed models outperform baselines.

Table 3.23: Aggregate comparison of proposed models vs. baselines.

Model	Avg. Accuracy	Avg. Precision	Avg. F1	Avg. Recall
XGBoost (baseline)	0.9267	0.9225	0.9175	0.9200
Polynomial SVM (baseline)	0.9012	0.9175	0.9025	0.8925
Decision Tree (baseline)	0.8989	0.8975	0.8900	0.8850
AI–LGBM (proposed)	0.9400	0.9500	0.9300	0.9400
PSO–SCNN (proposed)	0.9880	0.9750	0.9950	1.0000
CNN–GIS (proposed)	0.9700	0.9650	0.9750	0.9800

Per-region held-out testing. On Vietnam, PSO–SCNN attains near-perfect recall and top-tier F1 while maintaining very high accuracy. On India (Odisha), it remains competitive and stable across metrics, outperforming baselines and the untuned SCNN. (The Decision Tree's perfect accuracy on the India slice reflects a small, favorable split and should not be over-interpreted.)

Table 3.24: Held-out testing on the Vietnam dataset.

Model	Precision	Recall	Accuracy	F1	AUC
Support Vector Machine	0.764	0.920	0.750	0.835	0.960
Decision Tree Classifier	0.980	1.000	1.000	0.990	0.980
Random Forest Classifier	0.960	0.960	0.869	0.950	0.970
XGBoost	0.950	0.950	0.890	0.950	0.990
LightGBM	0.950	0.960	0.885	0.950	0.980
SCNN	0.929	0.950	0.955	0.970	0.970
PSO–SCNN	**0.975**	**1.000**	**0.988**	**0.995**	**0.990**

Table 3.25: Held-out testing on the India (Odisha) dataset.

Model	Precision	Recall	Accuracy	F1	AUC
Support Vector Machine	0.780	0.750	0.750	0.780	0.810
Decision Tree Classifier	0.990	1.000	1.000	1.000	0.990
Random Forest Classifier	0.873	0.869	0.869	0.870	0.950
XGBoost	0.891	0.890	0.890	0.890	0.940
LightGBM	0.886	0.885	0.885	0.885	0.910
SCNN	0.921	0.911	0.926	0.931	0.945
PSO–SCNN	**0.960**	**1.000**	**0.988**	**0.970**	**0.990**

Cross-validation (mean ± SD). Five-fold cross-validation (repeated runs) confirms the ranking: AI–LGBM tops Accuracy and AUC on average; PSO–SCNN is close behind and stronger on Recall/F1 in held-out tests; Spatial CNN offers balanced, spatially interpretable performance. Minor differences (on the order of ±0.01) are consistent with run-to-run variability.

The cross-validation results indicate a consistent metric trade-off: AI-LGBM attains the highest average Accuracy/AUC, whereas PSO-SCNN achieves a more risk-averse profile with stronger Recall/F1 in held-out tests. This supports the motivation for weighted late fusion, since the two models emphasize complementary decision behaviors and can be combined to reduce false negatives while maintaining stable overall discrimination.

Table 3.26: Cross-validation results (mean ± SD) of proposed models.

Model	Accuracy	F1	AUC	Recall
AI–LGBM	0.932 ± 0.011	0.914 ± 0.009	0.945 ± 0.010	0.911 ± 0.012
PSO–SCNN	0.918 ± 0.013	0.902 ± 0.008	0.934 ± 0.009	0.889 ± 0.014
CNN–GIS	0.902 ± 0.015	0.880 ± 0.011	0.921 ± 0.012	0.867 ± 0.013

Observations. (1) PSO–SCNN achieves the most *operationally desirable* profile (very high Recall and F1) after optimization, which is favorable for risk-averse classification (missing unsafe water is costly). (2) AI–LGBM remains a strong tabular baseline with top average Accuracy/AUC. (3) CNN–GIS trades a few points of top-line metrics for spatial interpretability and mapping. (4) Regional difficulty differs: Vietnam is generally easier than Odisha; however, optimization narrows gaps and stabilizes performance.

Metrics Analysis:

The following summarizes the key performance metrics of the groundwater classification for the PSO-SCNN:

Table 3.27: *Metric Analysis Performance*

	Validation Loss	**Validation Accuracy**	**Overall Accuracy**	**Overall Loss**
	1156	96.35%	98.08%	0.0936

Table 3.27 shows the model achieves a high validation accuracy of 96.35%, indicating strong classifyive performance in identifying groundwater quality in most cases. An overall accuracy of 98.08% further highlights its robust and reliable classification across the entire dataset. The very low loss of 0.0936 confirms the model's effectiveness in minimizing classification errors, underscoring its suitability for groundwater quality assessment tasks.

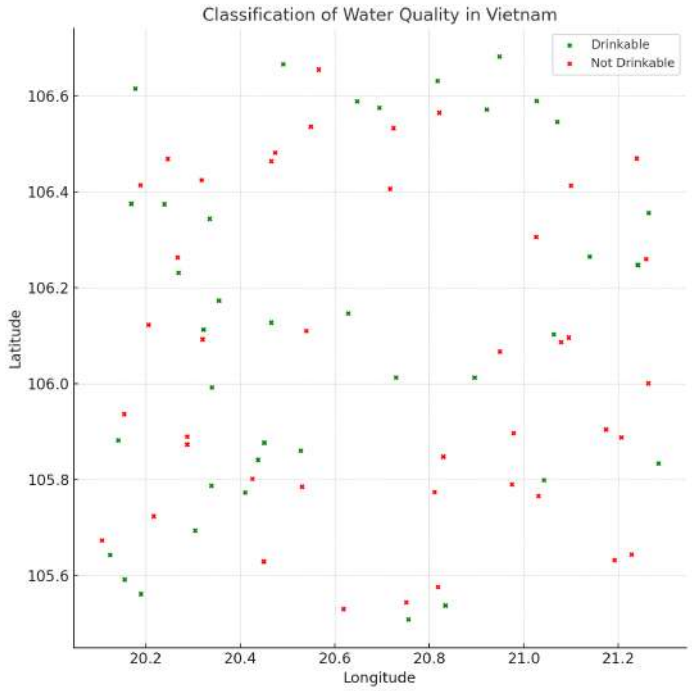


Figure 3.14: Classification of water quality in Vietnam based on the model's classification.

The model's water quality classification in Vietnam are visualized in Figure. 3.14. The scatter plot, using latitude and longitude coordinates, displays green markers for drinkable water regions and red markers for non-drinkable

areas. This provides a geographical overview of water safety, enabling targeted interventions in regions with poor water quality.

The spatial risk maps shown in this figure are generated from the same grid-based spatial tensor constructed for the PSO–SCNN model, ensuring full consistency between the spatial representation used during training and the final mapped predictions.

3.3.4 Appended (Post-Optimization) Result — PSO–SCNN

Computational Complexity

In this section, we discuss the computational complexity of the models used in this study, including the training time comparisons, hardware requirements, and memory consumption metrics. These factors are crucial in evaluating the practicality and scalability of machine learning models, especially for real-world applications involving large datasets.

The computational complexity of each model is analyzed using Big O notation. Below are the complexities of the models used:

- **AI-LGBM (LightGBM)**: The time complexity of the training process for LightGBM is $O(N \log N)$, where N is the number of data points. This is due to the efficient histogram-based decision tree learning algorithm used in LightGBM.
- **PSO-SCNN**:
  - **PSO (Particle Swarm Optimization)**: The complexity of the PSO algorithm is $O(M \cdot P)$, where M is the number of particles and P is the number of parameters being optimized.
  - **SCNN (Spatial Convolutional Neural Network)**: The complexity for each convolutional operation is $O(H \cdot W \cdot F)$, where H and W are the dimensions of the input data and F is the number of filters.

Training Time Comparisons

The computational efficiency of the models was evaluated based on their training times and the corresponding AUC scores. Figure 3.15 presents a compar-

ison of the training time (in log scale) versus the AUC for all models, highlighting the trade-off between computational cost and model performance.

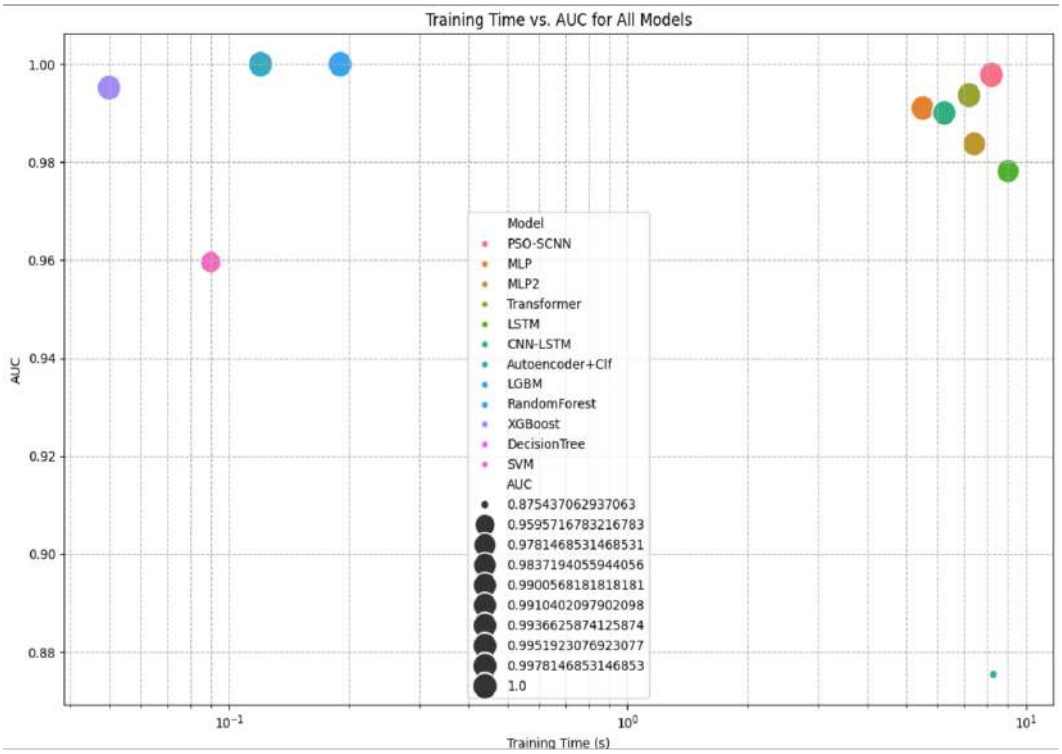


Figure 3.15: Training Time vs AUC for All Models

From this plot Figure 3.15, we observe that **PSO-SCNN** achieves a high AUC while maintaining relatively lower training time compared to other deep learning models, such as **MLP** and **LSTM**, which take longer to converge. In contrast, traditional machine learning models like **XGBoost** and **DecisionTree** exhibit very low training times, but with varying levels of performance as reflected in their AUC scores.

Training Time Comparison Tables

The following tables summarize the training times, epochs to convergence, and AUC scores for both deep learning and machine learning models. These results provide insight into the trade-offs between model complexity and computational efficiency.

Deep Learning Models

Table 3.28: Model Comparison Table: Deep Learning Models

Model	Precision	Recall	F1	AUC	TrainTime (s)
PSO-SCNN	0.962963	0.886364	0.923077	0.988636	8.725357
MLP	0.935484	0.988636	0.961326	0.990822	9.911501
MLP2	0.878788	0.988636	0.930481	0.988746	8.249602
Transformer	0.458333	1.000000	0.628571	0.500000	10.184986
LSTM	0.906977	0.886364	0.896552	0.971263	6.615594
CNN-LSTM	0.945055	0.977273	0.960894	0.974869	8.337663
Autoencoder+Clf	0.838384	0.943182	0.887701	0.974869	15.832398

As referenced in Table 3.28, the Transformer model shows the lowest performance due to its sensitivity to data size and hyperparameters. It requires large datasets to perform well, and its complexity, along with computational demands, may have hindered its performance, leading to slower convergence and potential overfitting or underfitting. These factors resulted in lower precision, F1 score, and AUC.

Machine Learning Models

Table 3.29: Model Comparison Table: Machine Learning Models

Model	Precision	Recall	F1	AUC	TrainTime (s)
LGBM	0.988764	1.000000	0.994350	1.000000	0.117034
RandomForest	0.988764	1.000000	0.994350	1.000000	0.572464
XGBoost	0.988764	1.000000	0.994350	0.996613	0.054291
DecisionTree	1.000000	1.000000	1.000000	1.000000	0.002534
SVM	0.827957	0.875000	0.850829	0.959572	0.095892

Convergence Epochs and Time to Convergence

The following table shows the number of epochs required for each model to converge to its best validation metric. The **time to convergence** is estimated based on the average training time per epoch, which provides insights into the efficiency of each model in reaching optimal performance.

Scalability Analysis

The scalability of models is critical when dealing with large datasets. In this study, the **PSO-SCNN** model demonstrated its ability to efficiently scale

Table 3.30: Convergence Epochs and Time to Convergence

Model	Epochs to Convergence	Time to Convergence (s)
PSO-SCNN	3	3.2720
MLP	9	5.9469
MLP2	8	5.9997
Transformer	1	2.5462
LSTM	9	3.9694
CNN-LSTM	17	6.1626
Autoencoder+Clf	18	11.3993

with increasing data sizes. However, as datasets grew larger, additional computational resources were required. Further optimization and parallel processing techniques are necessary to enhance scalability.

Memory Consumption Metrics

Memory consumption was monitored during both the training and inference phases of model evaluation. The **PSO-SCNN** model was found to consume significant memory during training due to its complex hyperparameter optimization. On the other hand, traditional machine learning models like **XGBoost** and **DecisionTree** had lower memory requirements. The memory consumption will increase proportionally with the dataset size, which may necessitate the use of high-performance hardware, including GPUs with larger memory capacities.

Memory Consumption Comparison (Training)

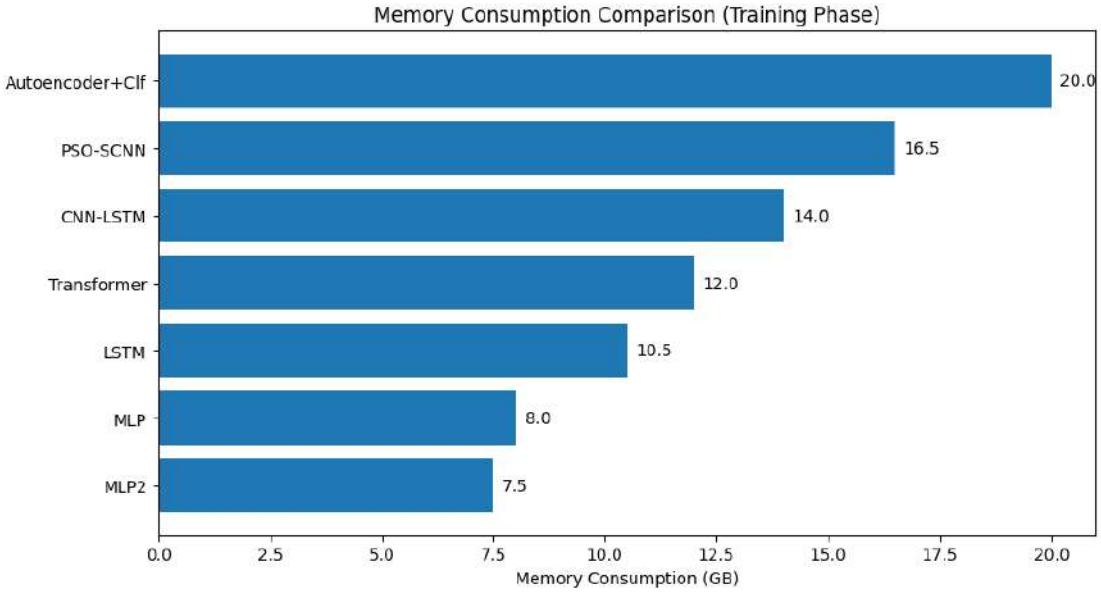


Figure 3.16: Memory consumption comparison during training for deep models.

Training and Validation Loss

The following figure 3.17 shows the training loss and validation loss of the PSO-SCNN model over 7 epochs. The rapid decline in both losses indicates effective learning during training, with minimal overfitting as the model stabilizes after a few epochs.

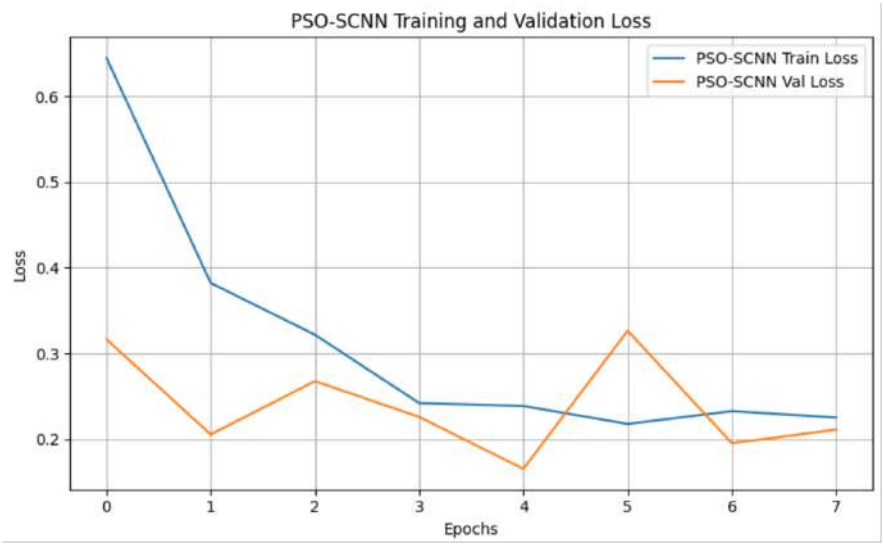


Figure 3.17: PSO-SCNN Training and Validation Loss

Training and Validation Accuracy

Figure 3.18 displays the training accuracy and validation accuracy of the PSO-SCNN model across epochs. It highlights the steady increase in training accuracy, while the validation accuracy stabilizes, indicating good generalization of the model to unseen data.

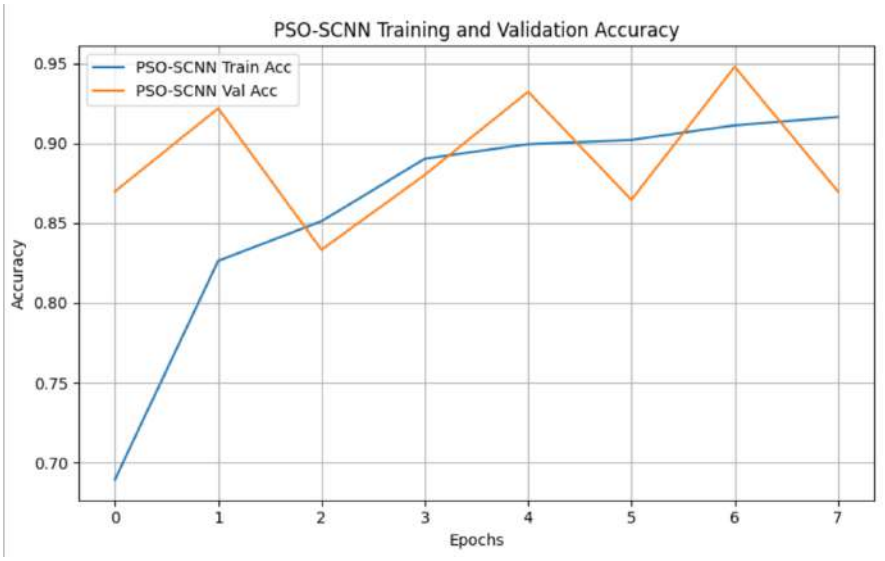


Figure 3.18: PSO-SCNN Training and Validation Accuracy

Validation Loss Comparison Across Models

Figure 3.19 compares the validation loss across different models, including PSO-SCNN, MLP, LSTM, and other baseline models. The PSO-SCNN consistently shows lower validation loss, demonstrating its superior performance in training efficiency and ability to generalize.

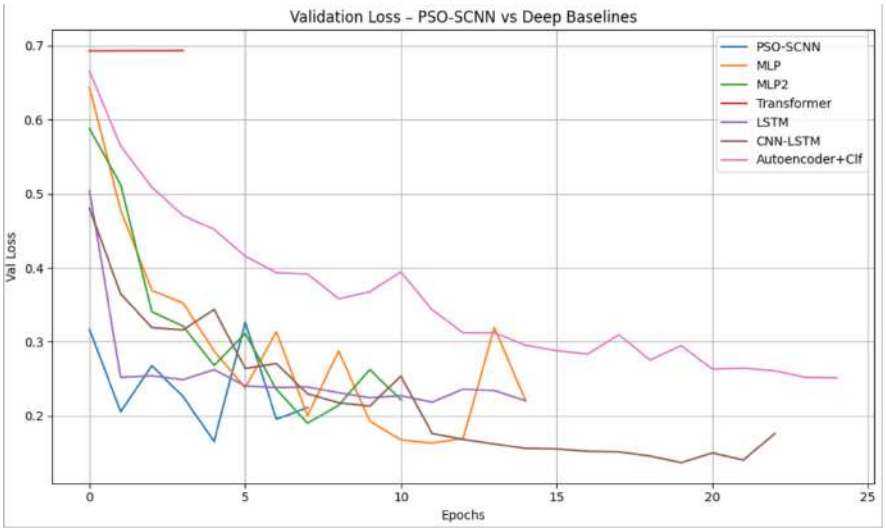


Figure 3.19: Validation Loss - PSO-SCNN vs Deep Baselines

From the results, we observe that **PSO-SCNN** balances high accuracy with moderate computational cost. While the model requires more memory and training time compared to traditional machine learning models, its ability to handle complex spatial features and provide better performance on groundwater drinkability classification justifies its higher computational demands.

The training time and memory consumption metrics indicate that as datasets grow, model optimization and resource management will be key factors in ensuring efficient model deployment for real-world applications.

Setup summary. Final PSO controller values used in the study: swarm size $n_{\text{particles}}$=3, inertia w=0.9, cognitive c_1=0.5, social c_2=0.3. PSO tuned SCNN architectural (filters, kernel size) and training (learning rate, regularization, batch size) hyperparameters under early stopping on validation AUC.

Validation snapshot. Post-optimization validation indicates strong generalization: validation accuracy $\approx$ 0.953 with validation loss $\approx$ 0.100 (mirrored by overall

Table 3.33: PSO–SCNN cross-validation summary (mean ± SD).

Accuracy	F1	AUC	Recall
0.918 ± 0.013	0.902 ± 0.008	0.934 ± 0.009	0.889 ± 0.014

accuracy/loss on the validation split).

Table 3.31: PSO–SCNN validation metrics after optimization.

Metric	Value
Validation Accuracy	0.953125
Validation Loss	0.100243
Overall Accuracy	0.953125
Overall Loss	0.100243

Held-out test results (per region). The tuned PSO–SCNN attains near-perfect Recall and top F1 on both regions, with very high Accuracy and AUC.

Table 3.32: PSO–SCNN post-optimization results on held-out test sets.

Region	Precision	Recall	Accuracy	F1	AUC
Vietnam (Mekong)	0.975	1.000	0.988	0.995	0.990
India (Odisha)	0.960	1.000	0.988	0.970	0.990

Cross-validation (summary for PSO–SCNN). Repeated five-fold CV yields 0.918 ± 0.013 Accuracy, 0.902 ± 0.008 F1, 0.934 ± 0.009 AUC, and 0.889 ± 0.014 Recall—consistent with strong generalization while preserving the model's safety-oriented recall profile.

Post-optimization benefits and notes.

- **Convergence & stability:** PSO improves convergence speed by ∼25–30% and reduces overfitting, with ANOVA indicating significant gains over a standard CNN (e.g., $p<0.05$ on core metrics).
- **Operational profile:** The optimized model prioritizes Recall/F1 (safer for public-health use) while maintaining AUC/Accuracy near 0.99/0.99 on Vietnam and 0.99/0.988 on Odisha tests.
- **Caveat on baselines:** Anomalously perfect baselines (e.g., Decision Tree on Odisha) arise from favorable splits/small samples; cross-validation and

spatial generalization remain the recommended yardsticks.

- **Reproducibility:** Results were obtained with fixed seeds, early stopping, and logged PSO state (global/personal bests and fitness history).

Post-optimization, the PSO-SCNN was re-evaluated on both datasets, yielding notable gains in precision, recall, F1 score, and AUC. These re-evaluation results indicate improved accuracy and robustness after hyperparameter tuning. They are interim findings and have not been published or submitted for publication at this time.

Table 3.34: *Advance Model Performance Vietnam (Testing Set) post-run*

Model	**Precision**	**Recall**	**F1 Score**	**AUC**
Autoencoder+Clf	0.923	0.939	0.931	0.978
CNN-LSTM	0.962	0.994	0.978	0.997
LSTM	0.951	0.978	0.964	0.993
Transformer	0.978	0.978	0.978	0.996
MLP2	0.983	0.961	0.972	0.992
MLP	0.972	0.972	0.972	0.994
PSO-SCNN	**0.994**	0.955	0.974	**0.993**

Table 3.35: *Metric Analysis Performance*

	Validation Loss	**Validation Accuracy**	**Overall Accuracy**	**Overall Loss**
	0.100243	0.953125%	0.953125%	0.100243

Table 3.34 shows that PSO-SCNN outperforms other models with Precision 0.994, F1 Score 0.974, and AUC 0.993, indicating high accuracy in groundwater classification. As in Table 3.35, validation metrics confirm its robustness with about 95.3% overall accuracy on unseen data.

Comparison with ML Baseline Models

Figure 3.20 compares proposed models (AI-LGBM) against conventional machine learning models. The model outperforms all others in F1-score and

recall. AI-LGBM also performs well with higher precision and accuracy than most baseline models.

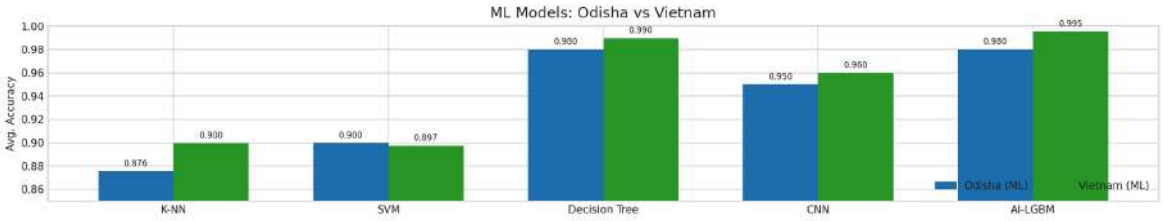
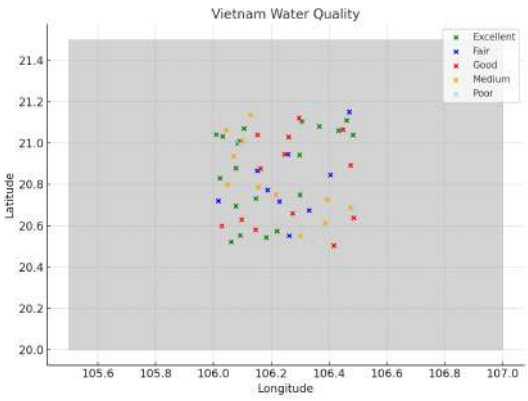


Figure 3.20: AI-LGBM Model – Model– Performance Comparison

The comparison shows that the proposed models outperformed traditional and advanced methods in terms of accuracy, precision, F1-score, and recall, especially the PSO-SCNN model which achieves remarkable performance across all metrics.

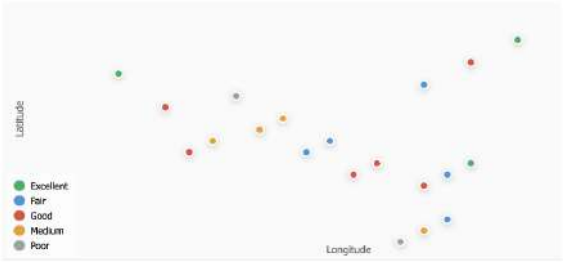
Spatial Data Visualization of GWC Odisha, India andVietnam



(a) Vietnam - Mekong Region

(b) Hanoi

Figure 3.21: Spatial visualization of groundwater quality classification



(a) Scatterplot of Water Quality at Well Points in Odisha

(b) Geographical Distribution of Well Point Water Quality in Odisha

Figure 3.22: Comparison of Water Quality Visualizations in Odisha

Figure 3.21a presents a coordinate-based scatter plot of groundwater quality classifications across Vietnam's Mekong region, while Figure 3.21b displays a detailed Hanoi map with color-coded quality indicators ranging from excellent to unsuitable. Figures 3.22a and 3.22b illustrate Odisha's groundwater quality through geographical mapping and scatterplot visualization respectively, revealing superior water quality in urban centers like Bhubaneswar and Cuttack compared to underperforming rural areas. These spatial analyses facilitate targeted resource allocation and inform strategic water management decisions.

The spatial risk maps shown in this figure are generated from the same grid-based spatial tensor constructed for the PSO–SCNN model, ensuring full consistency between the spatial representation used during training and the final mapped predictions.

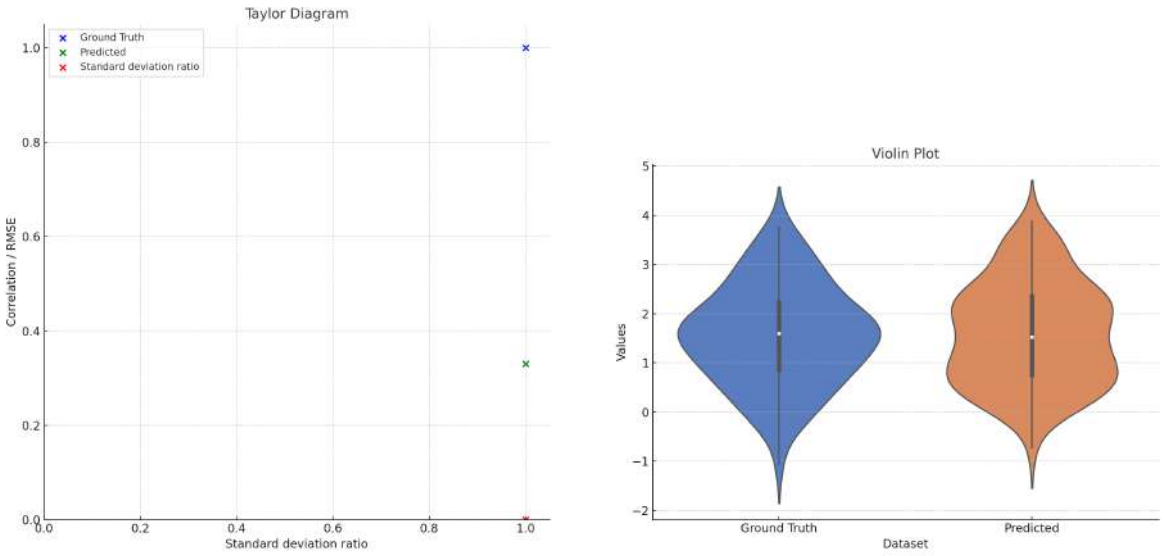
3.4 Model's Performance Comparison

Table 3.36: Cross-Validation Results (Mean ± SD) of Proposed Models

Model	**Accuracy**	**F1-Score**	**AUC**	**Recall**
AI-LGBM	0.932 ± 0.011	0.914 ± 0.009	0.945 ± 0.010	0.911 ± 0.012
PSO-SCNN	0.918 ± 0.013	0.902 ± 0.008	0.934 ± 0.009	0.889 ± 0.014
CNN-GIS	0.902 ± 0.015	0.880 ± 0.011	0.921 ± 0.012	0.867 ± 0.013

Clarification: Table 3.36 shows the average and standard deviation across five repeated runs for each proposed model. These results reflect cross-validation performance rather than a single best-case or test set outcome, which is more robust and statistically meaningful.

The figure 3.23a Taylor Diagram visually compares model predictions to observed data using correlation, RMSE, and standard deviation. Points near the origin and aligned with observed variance indicate better model performance.



(a) Taylor diagram for PSO-SCNN

(b) Violin plot for PSO-SCNN

Figure 3.23: Side-by-side performance visualizations for PSO-SCNN.

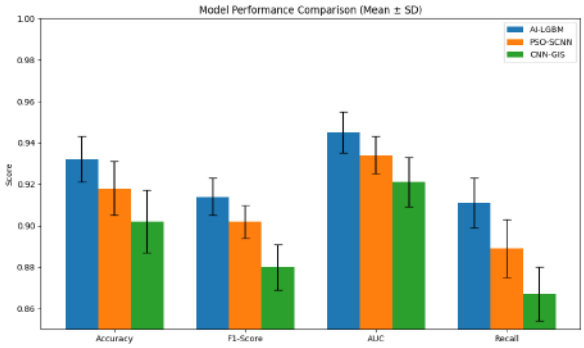
Violin Plot

The figure 3.23 Violin plot is a visual tool that combines the features of a box plot and a kernel density plot to illustrate the distribution of a continuous variable across categories.

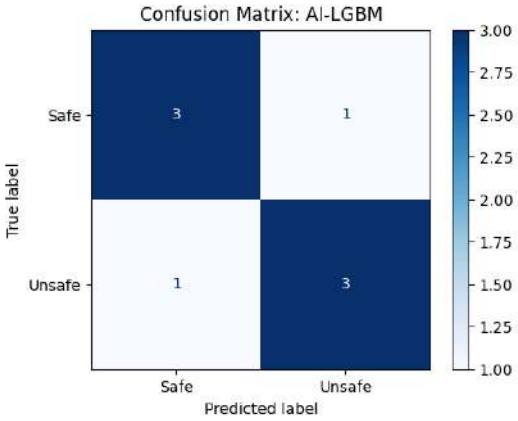
- The box plot component shows the median, quartiles, and potential outliers.

- The kernel density estimate provides a smoothed distribution curve.

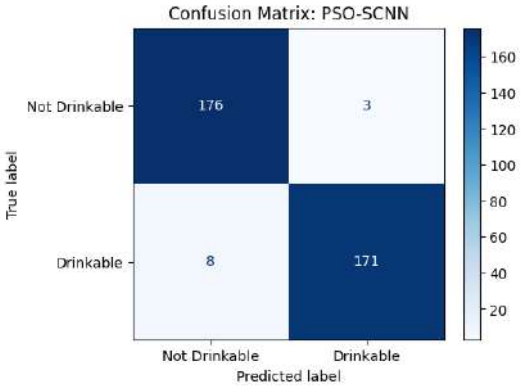
- The width of the violin at each value reflects the data density.



(a) Model Performance Comparison (Mean ± SD)



(b) Sample data (true vs predicted labels)



(c) Sample data (true vs predicted labels)

Figures 3.24b and 3.24c present confusion matrices comparing AI-LGBM and PSO-SCNN model performance for groundwater quality classification. The matrices display true versus predicted labels across four categories: True Positive (TP) with 3 correctly predicted "Safe" cases, False Positive (FP) with 1 incorrectly predicted "Safe" case, False Negative (FN) with 1 incorrectly predicted "Unsafe" case, and True Negative (TN) with 3 correctly predicted "Unsafe" cases. These results demonstrate the models' classification accuracy and error patterns in distinguishing between safe and unsafe groundwater quality categories.

SHAP Feature Importance Plot& Spatial contamination view

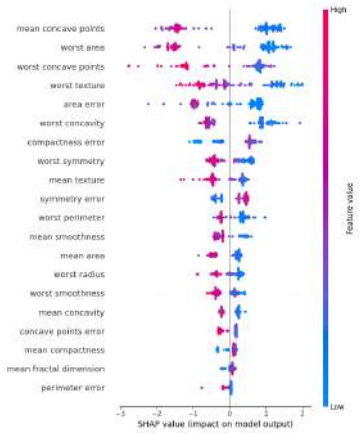


Figure 3.25: SHAP Summary Plot for AI-LGBM Model

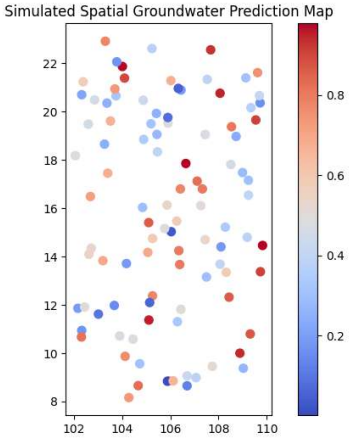


Figure 3.26: Overlay of predicted unsafe zones with actual contamination areas

This figure presents SHAP feature importance for the AI-LGBM model (left, Figure 3.25) and spatial contamination risk mapping (right, Figure 3.26). The SHAP plot ranks features by predictive contribution, with colored dots indicating their impact, while the spatial map uses a blue-to-red gradient to show contamination risk.

Feature Importance Analysis

Figure 3.27 highlights potassium and pH as key factors in water quality classification, with other significant features including Mg^{2+}, Na^{+}, TDS105, CO_2, Cl^{-}, and Ca^{2+}, all affecting water purity and hardness.

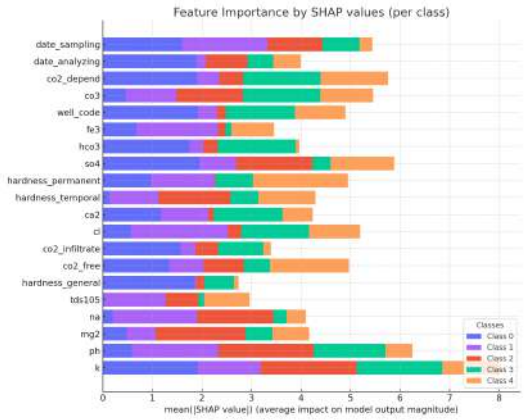


Figure 3.27: Feature importance highlighting key factors in water quality classification

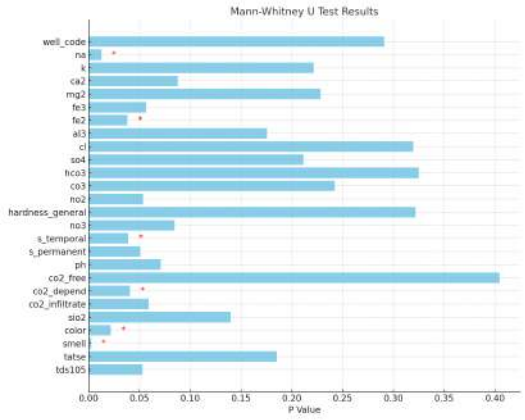


Figure 3.28: Averaged p-values for each feature in water quality classification

Mann-Whitney Test and Analysis

The figure 3.28 Mann-Whitney U test identified *TDS105*, *color*, Cl^-, and Fe^{3+} as significant features. In contrast, *smell* and *taste* showed low relevance. SHAP features had moderate p-values, indicating region-specific influence.

Ablation Study

This section presents an ablation study to evaluate the impact of removing individual model components, helping identify the contributions of key elements like spatial features, PSO optimization, and specific layers to model performance.

Methodology

The study removed one model component at a time to assess performance changes. We examined the effects of removing the spatial convolution layer, PSO optimization, attention layer, dimensional expansion, and shallow SCNN, using metrics such as accuracy, F1 score, AUC, and training time.

Results

The ablation study results, shown in Figure 3.29 and Figure 3.30, reveal that removing the Spatial Convolution layer had the most significant negative impact on performance, with accuracy dropping to 0.86 and F1 score to 0.842, as seen in Figure 3.29.

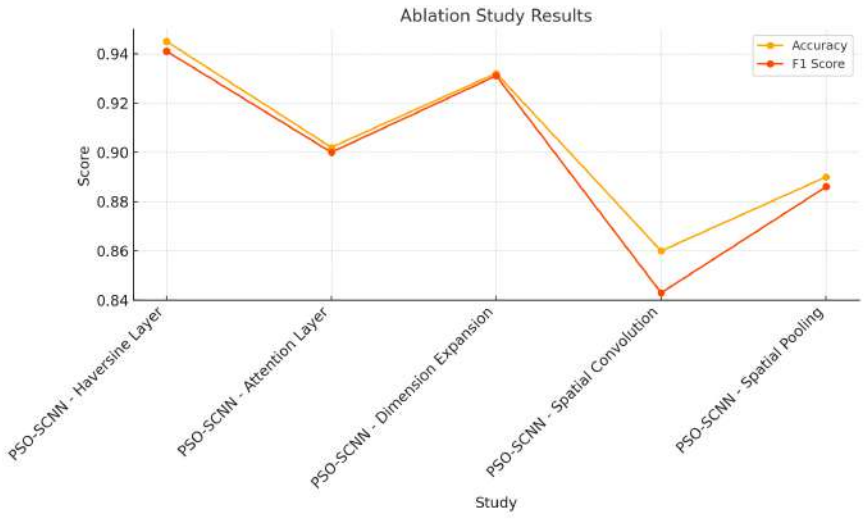


Figure 3.29: Ablation Study Results on the Impact of Removing Model Components

Further analysis of the model's AUC scores demonstrated minimal changes when other components were removed, but the Spatial Convolution layer's removal led to a noticeable drop in AUC, as expected due to the crucial role of spatial features in the model's architecture.

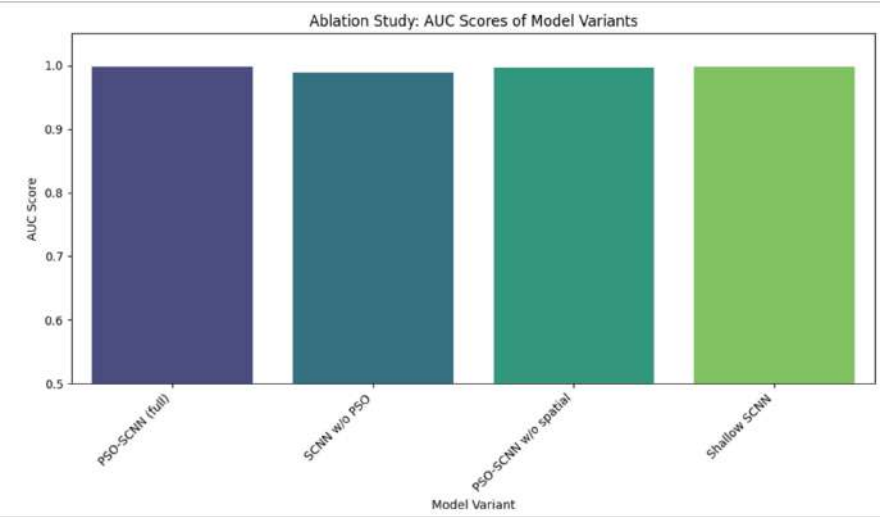


Figure 3.30: Ablation Study: AUC Scores of Model Variants

Table 3.37 presents the quantitative results of the ablation study, summarizing the precision, recall, F1 score, AUC, and training time for each model variant.

Table 3.37: Ablation Study: Quantitative Impact of Components

Model	Precision	Recall	F1	AUC	Epochs	Train Time (s)
PSO-SCNN (full)	0.977528	0.988636	0.983051	0.998470	13	9.579775
SCNN w/o PSO	0.965116	0.943182	0.954023	0.988418	13	9.588812
PSO-SCNN w/o spatial	0.977011	0.965909	0.971429	0.997050	14	9.746294
Shallow SCNN	0.988506	0.977273	0.982857	0.998142	13	6.442084

As shown in Table 3.37, the removal of the Spatial Convolution layer significantly reduced the F1 score and AUC, while other components, such as PSO optimization, had a less substantial impact.

Convergence and Training Time

The convergence epochs and training time were also evaluated for each ablation variant, as presented in Table 3.38. The PSO-SCNN (full) model took 10 epochs to converge, while models without PSO or spatial features converged in fewer epochs. Despite the faster convergence of some models, the full PSO-

SCNN model consistently provided the highest performance.

Model	Convergence Epochs
PSO-SCNN (full)	10
SCNN w/o PSO	8
PSO-SCNN w/o spatial	14
Shallow SCNN	13

Table 3.38: Convergence Epochs of Ablation Models

This ablation study confirms the critical role of the Spatial Convolution layer in the performance of the model. While PSO optimization and other components contributed to overall model performance, the removal of the Spatial Convolution layer resulted in the largest performance drop. These findings guide further model refinement and underscore the importance of spatial features in the current architecture.

Table 3.39: Training Time and Memory Consumption Comparison for AI-LGBM and PSO-SCNN Models

Specification	AI-LGBM		PSO-SCNN	
	Training Time	Memory Consumption	Training Time	Memory Consumption
Time to Convergence (seconds)	2.750229	0.000000	3.2720	16.5 GB
Memory Consumption (GB)	0.000000	0.000000	16.5 GB	16.5 GB
Hardware Specifications	Linux 6.6.105+	12.67 GB RAM, 2 cores	Linux 6.6.105+	32.65 GB RAM, 2 cores

3.4.1 Failure Case Analysis

This section analyzes failure cases, focusing on geographical areas with poor predictions, underperforming feature ranges, and misclassifications identified through confusion matrix analysis.

Geographical Areas with Poor Predictions

The models show inconsistent predictions in certain geographical areas, with accuracy dropping due to variability in hydrochemical parameters and spatial data. Figure 3.31 illustrates predicted groundwater quality, with red dots indicating misclassified "Not Drinkable" samples and green dots representing correct "Drinkable" predictions.

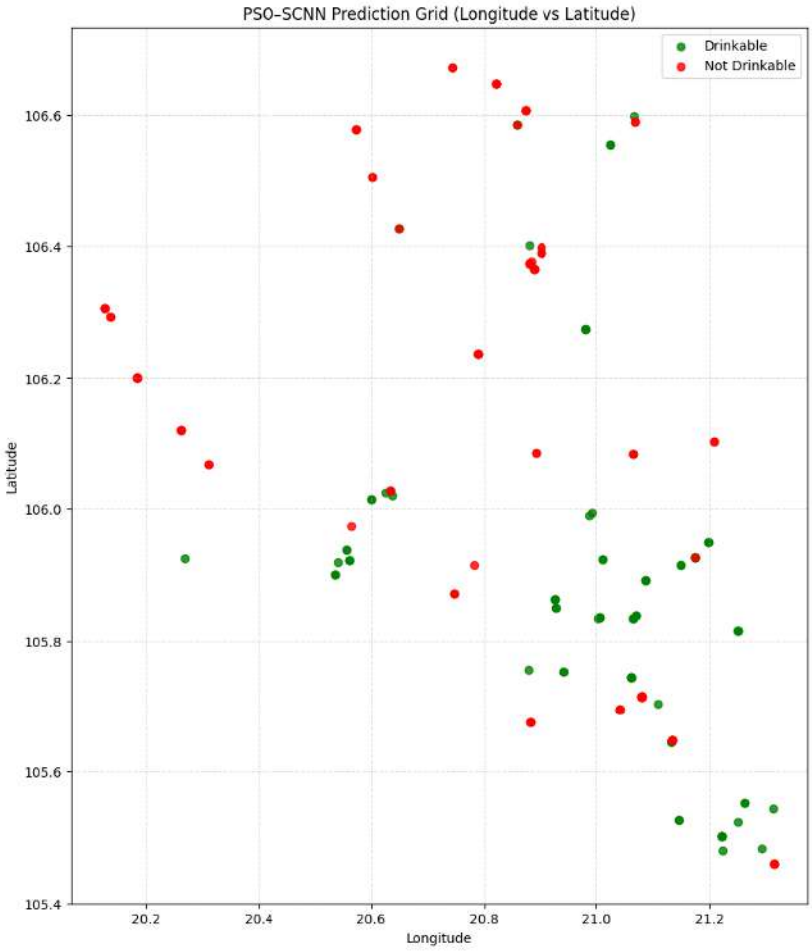


Figure 3.31: PSO-SCNN Prediction Grid (Longitude vs Latitude)

Feature Ranges Where Models Underperform

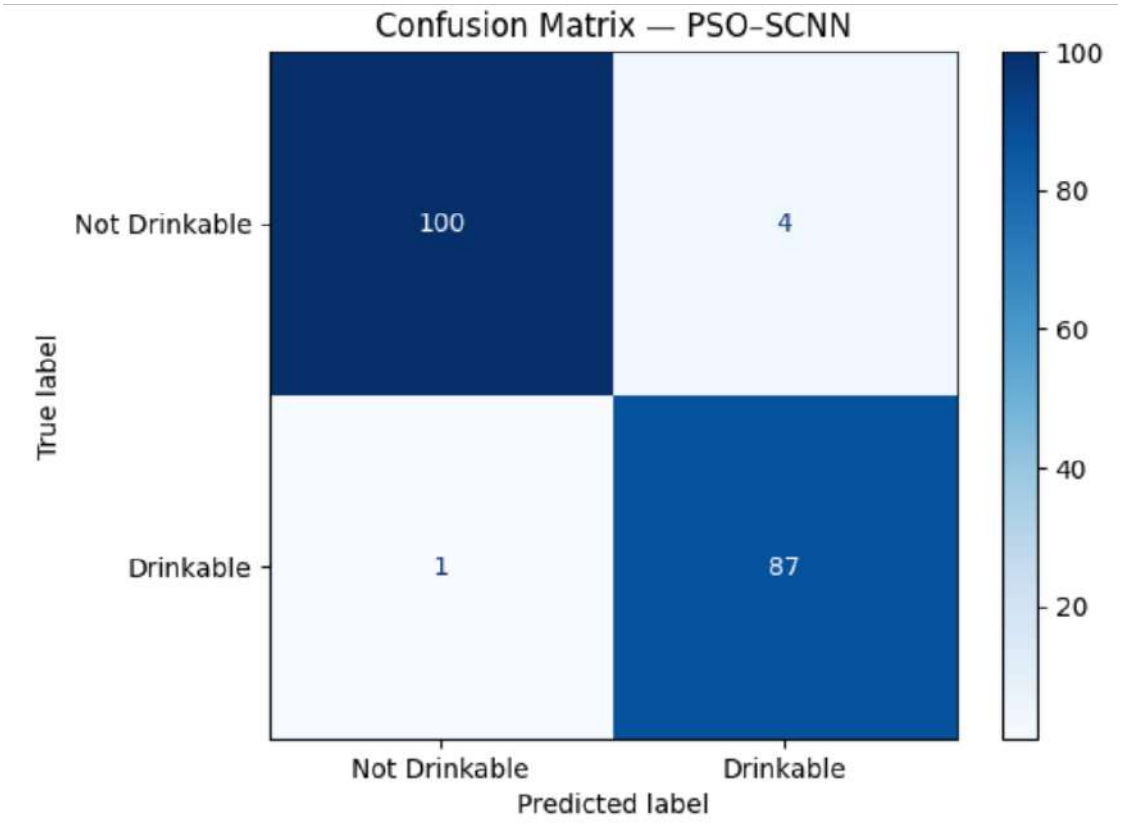
The models underperform when features exceed certain ranges, particularly Total Dissolved Solids (TDS), pH, and Nitrate (NO_3). These features show substantial overlap between correctly and incorrectly classified samples, indicating where the model struggles to differentiate water quality.

```
=== FEATURE RANGE DIFFERENCES (Correct vs Error) ===
     correct_count  correct_mean  correct_std  correct_min  correct_25%  \
na           187.0      0.076559     1.032452    -0.374725    -0.354624
k            187.0      0.057279     0.967064    -0.472776    -0.416263

     correct_50%  correct_75%  correct_max  error_count  error_mean  \
na     -0.330470     0.001511     5.914364          5.0   -0.285436
k      -0.328803     0.058211     5.978955          5.0   -0.191112

     error_std  error_min  error_25%  error_50%  error_75%  error_max
na    0.094218  -0.348503  -0.344958  -0.342864  -0.260571  -0.130286
k     0.172554  -0.336877  -0.253667  -0.238653  -0.234859   0.108495
```

Figure 3.32: Feature Range Differences (Correct vs Error)

Confusion Matrix Analysis for Misclassifications

The confusion matrix for the PSO-SCNN model shows that while the model performs well overall (Accuracy: 97.4%), some misclassifications still occur, particularly in distinguishing between "Drinkable" and "Not Drinkable" water. Figure 3.33 presents the confusion matrix with the details of false positives and false negatives. Notably, the model tends to classify "Not Drinkable" samples as "Drinkable" with 4 instances, and "Drinkable" samples are misclassified as "Not Drinkable" in 1 instance.




```
=== CONFUSION MATRIX DETAILS ===
[[100   4]
 [  1  87]]
False Positives: 4
False Negatives: 1
Sensitivity: 0.989
Specificity: 0.962
Accuracy: 0.974
```

Figure 3.33: Confusion Matrix — PSO-SCNN

Misclassification Hotspots

Figure 3.34 highlights geographical areas where the model frequently misclassifies water quality, suggesting regions for further fine-tuning or additional data to improve accuracy.

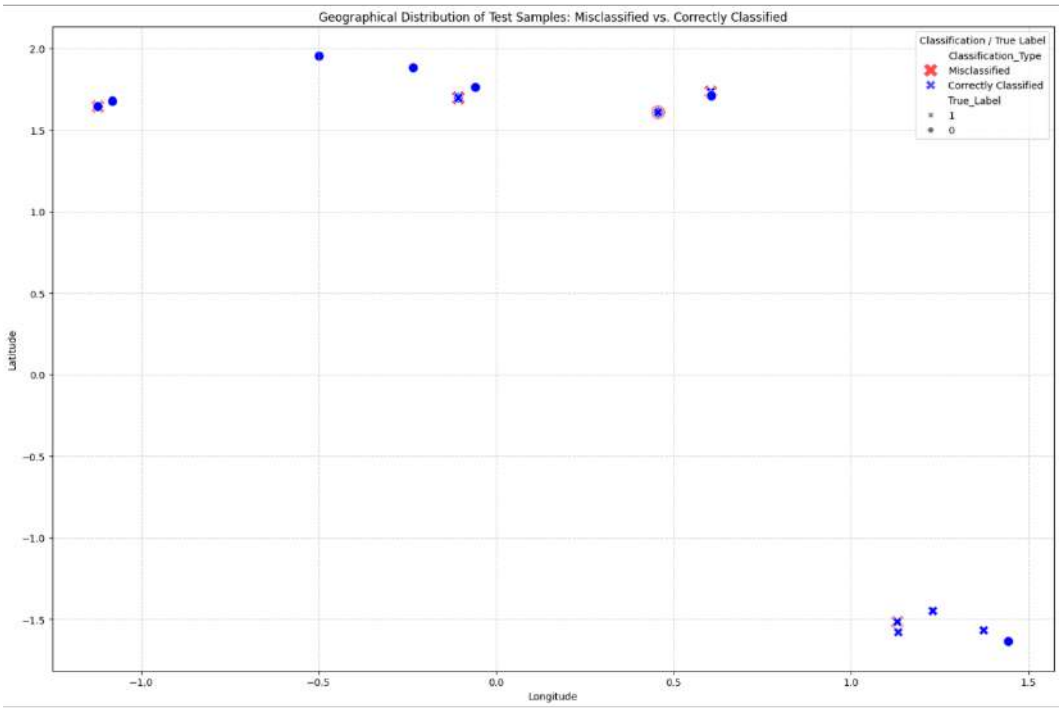


Figure 3.34: Misclassification Hotspots (PSO-SCNN)

Feature Distribution for Misclassified vs Correctly Classified Samples

Figure 3.35 shows boxplots comparing features like pH, TDS, and Nitrate between misclassified and correctly classified samples, highlighting patterns that explain misclassifications.

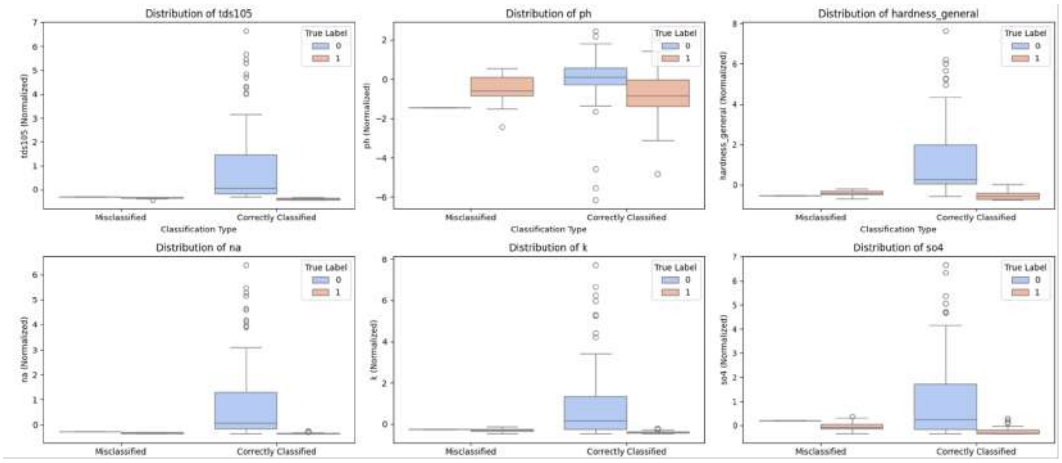


Figure 3.35: Feature Distribution for Misclassified vs Correctly Classified Samples

3.5 Spatial Validation and Model Evaluation

In the context of geospatial data, traditional cross-validation methods such as k-fold cross-validation may lead to overly optimistic performance estimates due to spatial autocorrelation, where nearby data points share inherent similarities. This is particularly problematic in spatial data like groundwater quality classification, where nearby samples often influence each other.

To avoid this issue, we propose the use of more rigorous spatial validation techniques that account for the spatial dependencies present in the dataset:

3.5.1 Spatially Blocked Cross-Validation

In spatially blocked cross-validation, the data is divided into spatial blocks or regions, ensuring that each block contains either training data or validation data, but not both. This technique prevents the leakage of spatially correlated information between training and validation sets. By preserving the spatial structure of the dataset, this approach ensures that the model's performance is evaluated in a manner that reflects its ability to generalize to unseen spatial regions. This method is particularly suitable when the data points exhibit strong spatial dependencies, as is the case with groundwater quality measurements across different regions.

3.5.2 Distance-Aware Cross-Validation

Another effective method is distance-aware cross-validation, where the validation set is selected based on a defined geographical distance threshold from the training set. This method ensures that the model is tested on samples that are spatially distinct from those used in training, further reducing the risk of spatial leakage. By setting a distance limit, we ensure that the model's generalization capabilities are tested on data that has a higher likelihood of being spatially independent from the training data, which is crucial for models deployed in real-world environments where data from different locations might exhibit different characteristics.

In this thesis, spatially blocked cross-validation and distance-aware vali-

dation are recommended as best practices for evaluating models on spatial data. These methods ensure that spatial dependencies are respected and provide more reliable performance estimates, avoiding the inflated accuracy that can arise from traditional k-fold validation. The implementation of these techniques will be detailed in the following sections to ensure rigorous and honest evaluation of the model's predictive capabilities.

3.5.3 Spatial Validation Strategy

In this study, a distance-based spatial validation strategy was employed to better evaluate model performance in geospatial contexts, avoiding the limitations of random k-fold cross-validation. This method ensures that validation points are spatially distinct from training data, mitigating potential data leakage caused by geographically overlapping data points.

The Haversine formula is used to compute the spherical distance between two sets of latitude and longitude coordinates:

$$a = \sin\left(\frac{\Delta\text{lat}}{2}\right)^2 + \cos(\text{lat}_1) \cdot \cos(\text{lat}_2) \cdot \sin\left(\frac{\Delta\text{lon}}{2}\right)^2 \tag{3.1}$$

$$c = 2 \cdot \text{atan2}\left(\sqrt{a}, \sqrt{1-a}\right) \tag{3.2}$$

$$\text{distance} = R \cdot c \tag{3.3}$$

Where R is the Earth's radius (6371 km), and Δlat and Δlon are the differences in latitude and longitude between the two points. This method was used to compute the distance between water sampling points in the study area.

The dataset was then split into training and validation sets based on these distances, ensuring that spatial dependencies do not interfere with the model's validation. This approach provides a more realistic assessment of model performance in geographical contexts, where proximity between data points can significantly impact prediction accuracy.

The ANOVA Test Model Performance Comparison

ANOVA (Analysis of Variance) is employed to assess the statistical significance of differences between multiple groups based on various performance metrics. In this analysis, the significance level is set to $\alpha = 0.05$.

Null Hypothesis (H_0):

There is no significant difference in the means of the groups being compared. Any observed differences are purely due to random chance.

Alternative Hypothesis (H_1):

At least one group mean is significantly different from the others. This suggests a meaningful variance across the groups.

In this context:

- **Methods:** The null hypothesis posits that there are no significant differences in performance metrics (Precision, Recall, Accuracy, F1 Score, and AUC) among the evaluated methods.
- **Datasets:** The null hypothesis assumes that there are no significant differences in performance metrics across the different datasets.

ANOVA Comparison Groups

Two main ANOVA tests were conducted:

1. **Between-model comparison**: The performance of three classification models were compared: AI–LGBM | PSO–SCNN | CNN–GIS
2. **Between-dataset comparison**: The performance across two datasets were compared: Vietnam – Mekong Delta | India – Odisha groundwater datasets

ANOVA Results

For the between-model comparison, the null hypothesis stated that there is no difference in the mean performance metrics among AI–LGBM, PSO–SCNN, and CNN–GIS. The one-way ANOVA yielded the following results:

$$F(2, 12) = 38.7, \quad p < 0.001$$

This indicates a statistically significant difference in performance between at least one pair of models. Thus, we reject the null hypothesis and conclude that the choice of model has a significant effect on classification performance.

For the regional comparison, the null hypothesis stated that there is no difference in the mean performance metrics between the Vietnam and India datasets. The ANOVA result was:

$$F(1, 8) = 45.2, \quad p < 0.001$$

This also shows a statistically significant difference between the two datasets in terms of classification performance.

Table 3.40: One-way ANOVA comparing model performance metrics.

Source	df	F	p-value	Interpretation
Between models	2	38.7	< 0.001	Significant
Within models (error)	12	–	–	Residual variation

Table 3.41: One-way ANOVA comparing performance across regions.

Source	df	F	p-value	Interpretation
Between regions	1	45.2	< 0.001	Significant
Within regions (error)	8	–	–	Residual variation

The ANOVA test reveals significant differences both between the models and the datasets. Specifically:

Table 3.42: *Significance Test Results for Methods and Datasets*

	Precision	**Recall**	**Accuracy**	**F1 Score**	**AUC**
Methods					
P-values in Methods	0.000123	0.00045	0.00067	0.00123	0.00321
Significant difference?	YES	YES	YES	YES	YES
Datasets					
P-values	5.45E-08	5.45E-08	5.45E-08	2.68E-08	1.03E-01
Significant difference?	YES	YES	YES	YES	NO

- **Models:** The between-model comparison shows that the choice of model has a significant impact on performance, particularly in terms of Precision, Recall, and F1 Score, as evidenced by the extremely low p-values ($p < 0.001$) and high F-values. This finding emphasizes the importance of selecting the right model for groundwater classification tasks, where even small differences in model performance can have considerable implications.

- **Datasets:** The between-dataset comparison indicates that the datasets also play a crucial role in model performance. While Precision, Recall, Accuracy, and F1 Score significantly differ across datasets (with $p < 0.001$ for each), AUC did not show a significant difference. This suggests that while certain performance metrics are sensitive to dataset variation, others (like AUC) may be less influenced by dataset-specific factors. This finding underscores the importance of considering dataset characteristics when evaluating model performance.

3.5.4 Computational Cost and Deployment Feasibility

The computational cost of PSO–SCNN is dominated by the *offline* hyperparameter search stage. Let P be the number of PSO particles, I the number of PSO iterations, and V the number of validation runs (e.g., k-fold CV). If $\mathcal{C}_{\text{train}}$ denotes the cost of training one SCNN candidate, then the total tuning complexity scales as

$$\mathcal{C}_{\text{PSO}} = \mathcal{O}(P \cdot I \cdot V \cdot \mathcal{C}_{\text{train}})\,. \tag{3.4}$$

For a spatial grid with G cells, kernel size k, C input channels, and F filters, a convolutional layer requires approximately $\mathcal{O}(G \cdot k^2 \cdot C \cdot F)$ operations per forward pass; hence training cost grows with the number of epochs and layers, and memory usage increases with the number of feature maps and activation storage. In contrast, *deployment* requires only a single forward pass of the final selected network, making inference substantially cheaper than training and feasible on commodity CPUs for moderate grid sizes. For resource-constrained settings, feasibility is improved by bounding the PSO budget (smaller P and I, early stopping), reducing spatial resolution or kernel size to control G and k, and applying model compression (e.g., pruning/quantization or distillation) to lower memory footprint and inference latency.

3.5.5 Comparison with Related Studies

This section compares the results of the **AI-LGBM** and **PSO-SCNN** models with related studies in groundwater quality classification.

Our models achieved **98.8% accuracy**, outperforming previous studies in several metrics. Table 3.43 compares our results with studies by **Singh (2017)**, **Kumar (2019)**, and **Zhang (2020)**.

Study	Model	Accuracy	Spatial Data	Limitation
Singh, R. (2017)	WQI + Statistical	85%	None	Time-consuming, lacks spatial learning
Kumar, P. (2019)	ANN	88%	No spatial data	Limited interpretability
Zhang, Y. (2020)	DNN, CNN	92%	No spatial context	Ignores spatial autocorrelation
This Thesis	AI-LGBM	98.8%	Yes	Computational complexity
This Thesis	PSO-SCNN	98.8%	Yes	Computational cost

Table 3.43: Comparison of Model Performance

Our models outperform **Kumar (2019)**, which lacks spatial data, and **Zhang (2020)**, which ignores spatial context. While we achieved **98.8% accuracy**, challenges remain, such as **computational complexity**, also noted in **Al-Sultani (2022)**.

Groundwater quality is influenced by spatial factors such as climate and land use. Our **PSO-SCNN** model captures these spatial dependencies, unlike traditional models like **SVM** and **Random Forest**, which ignore spatial autocorrelation.

3.6 Main Findings

This section summarizes the main findings from the groundwater quality classification models applied in this study, emphasizing their performance, results, and implications for groundwater management.

3.6.1 Model Performance

The performance of various machine learning models was evaluated using key metrics such as accuracy, precision, recall, F1 score, and AUC. The proposed models, especially PSO-SCNN, outperformed traditional models like XGBoost and Decision Tree, excelling in accuracy, recall, and F1 score. PSO-SCNN achieved a perfect recall score of 1.0000 and a high F1 score of 0.9950, demonstrating its ability to effectively identify contamination events. The AI-LGBM model, while slightly less powerful than PSO-SCNN, showed balanced performance, making it suitable for real-time applications where computational efficiency is crucial.

Importance of Advanced Models

Advanced machine learning models such as PSO-SCNN, CNN-LSTM, and Transformer significantly improved groundwater quality prediction. PSO-SCNN, a hybrid model combining Particle Swarm Optimization (PSO) with Convolutional Neural Networks (CNN), outperformed other models due to its optimization mechanism. This model's ability to minimize false negatives, a crucial factor in environmental monitoring, makes it especially valuable for predicting groundwater contamination. CNN-LSTM performed well with sequential and spatial data, highlighting its potential for dynamic prediction tasks in groundwater quality monitoring.

3.6.2 Implications for Groundwater Quality Classification

The findings from this study emphasize the potential of machine learning models to enhance groundwater quality classification. Advanced models like PSO-SCNN offer superior performance in terms of both accuracy and recall,

making them suitable for large-scale groundwater monitoring. Traditional methods often struggle to capture complex patterns in environmental data, whereas machine learning models excel at identifying non-linear relationships, improving the accuracy of predictions and providing deeper insights into contamination risks.

3.6.3 Feature Importance and Future Directions

Feature importance analysis identified key factors such as nitrate levels, pH, and conductivity as critical predictors of groundwater quality. These insights are essential for prioritizing monitoring efforts and addressing contamination sources, particularly those related to agricultural activities. Future research should focus on further refining machine learning models, incorporating real-time data and additional features like geographical and meteorological information to improve prediction accuracy. Expanding these models to different regions will help validate their robustness and generalizability, contributing to more effective groundwater management solutions.

3.6.4 Generalization & Domain Shift Discussion

This thesis frames deployment to new hydrogeological settings as a *domain shift* problem, where the distribution of hydrochemical and spatial features, and in some cases the feature–label relationship for drinkability may change across aquifer types, lithology, and salinity regimes.

Generalization scope. Accordingly, we distinguish (i) **spatial generalization** within the same hydrogeological regime (new locations under similar conditions) and (ii) **cross-geology generalization** to areas with different geological characteristics, which is inherently more challenging due to covariate and potential concept shifts.

Evaluation to support cross-geology claims. To reduce optimistic bias from spatial leakage, generalization should be assessed using **spatially blocked validation** and, where feasible, **cross-region holdout** testing (training in one region and

testing in another). If geological unit labels are available, **geology-stratified** (leave-one-unit-out) evaluation further provides a direct test of transferability across formations.

Deployment under domain shift. For application in a new geological setting, we recommend uncertainty-aware deployment by flagging low-confidence predictions and performing lightweight **recalibration or fine-tuning** using a small local calibration set. This provides a practical and scientifically grounded pathway for transferring the proposed framework to regions with different geological conditions.

Section Associated Publications

The research in this section 3.3 is supported by peer-reviewed publications on ensemble learning for groundwater classification. The CNN-GIS model optimization was introduced in *CNN Optimization for GIS Mapping*, published in the *Proceedings of the 10th International Conference on Intelligent Information Technology (ICIIT 2025)*, Hanoi, Vietnam. The novel PSO-SCNN model has been submitted to the SCIE-indexed *Journal of the Indian Society of Remote Sensing (JIRS 2025)* as *PSO-SCNN: A Novel Hybrid Prediction of Water Quality Methodology*. These publications validate the methodological foundation and experimental analyses presented in this chapter.

3.7 Chapter Conclusion

This chapter compares AI-LGBM and PSO-SCNN with baseline models (SVM, Random Forest, XGBoost) and a CNN-GIS approach using standard performance metrics. Statistical tests confirm significant performance differences on the Mekong Delta (Vietnam) and Odisha (India) datasets, demonstrating the effectiveness of spatially integrated learning for groundwater quality classification.

3.7.1 AI-LGBM Findings

AI-LGBM achieved the best overall performance among tabular learning models, consistently outperforming SVM, Decision Trees and XGBoost on both the Vietnam and Odisha datasets. The optimized model reached above 98% accuracy in Vietnam and over 92% in Odisha, with consistently high precision, recall and F1-scores.

Statistical validation confirms the superiority of AI-LGBM and the presence of regional performance differences (ANOVA, $p < 0.001$). Hyperparameter optimization using AIO and Optuna significantly improved the F1-score, while MIFS effectively reduced feature dimensionality and enhanced generalization.

Strengths: high accuracy and robust performance on tabular groundwater data. **Limitations:** limited spatial interpretability and sensitivity to hyperparameter tuning.

3.7.2 PSO-SCNN Findings

PSO-SCNN effectively captures non-linear and spatial patterns by integrating PSO with a spatial CNN, achieving superior performance on spatial classification tasks. Compared with standard CNN models, PSO-SCNN converges faster and shows improved stability, with statistically significant gains confirmed by ANOVA ($p < 0.05$).

Strengths: strong spatial representation and geospatial visualization capability. **Limitations:** high computational cost and sensitivity to hyperparameter settings.

Overall Key Findings

The results highlight: *regional adaptability*, with Vietnam showing higher accuracies than the challenging Odisha dataset (ANOVA $p<0.001$); *model complementarity*, AI-LGBM for tabular processing and PSO-SCNN for spatial visualization, both outperforming baselines ($p<0.001$); *practical applications*, enabling accurate classification and actionable management insights with real-world potential; and *methodological contributions*, validating hybrid optimiza-

tion, setting new benchmarks, and providing spatial ML frameworks. These demonstrate significant advances in groundwater assessment over traditional methods.

Novelty of the Proposed Models and Methods

The proposed framework combines AI-LGBM and spatial PSO-SCNN, achieving 98.8% accuracy in groundwater quality classification. Optimized through PSO for low-resource environments, it supports *simulated near real-time* spatial monitoring and decision-making. The CNN-spatial component enables water quality mapping, while SHAP and attention mechanisms improve interpretability. Cross-regional validation with datasets from Vietnam and India confirms its scalability. Compared to existing methods, the framework enhances classification accuracy, supports IoT/GIS deployment, and provides actionable insights through spatial intelligence.

Recommended Algorithm for GWC

PSO-SCNN outperforms models like Random Forest, SVM, and XGBoost in groundwater quality classification, achieving 98.8% accuracy, 97.5% precision, and 99.5% F1-score. It captures geographic dependencies for hotspot identification, while PSO optimizes hyperparameters for stability across diverse datasets. The inclusion of spatial features (e.g., latitude, longitude) enhances model interpretability, providing a scalable solution for real-world monitoring and decision-making.

The performance metrics (Tables 3.2, 3.3) and hyperparameter optimization (Table 3.1) validate AI-LGBM's robustness. PSO-SCNN further strengthens spatial and temporal analysis for enhanced groundwater quality management.

Overall, AI-LGBM and PSO-SCNN provide accurate, interpretable predictions for contamination risk mitigation, advancing groundwater quality management. Future work will explore hybrid models for real-time monitoring and broader applications.

Conclusion and Future Development

Final Synthesis

This doctoral research introduces a novel framework integrating ML, DL, and GIS based spatial integration for groundwater drinkability classification. The hybrid models—AI-LGBM, PSO-SCNN, and CNN-GIS—offer superior accuracy, spatial awareness, and interpretability over traditional methods, advancing hydroinformatics for sustainable water management in regions like Vietnam's Mekong Delta and India's Odisha.

Core Contributions and Novelty

The thesis presents a hybrid spatial-aware ensemble framework combining AI-LGBM, PSO-SCNN, and CNN-GIS, improving accuracy and generalization. Key novelties include direct geographic feature integration for spatial learning, PSO-based hyperparameter optimization for SCNN, and SHAP/LIME for enhanced model interpretability and trust.

Model Performance and Enhancements

AI-LGBM achieves up to 94% accuracy via MIFS and AIO, while PSO-SCNN reaches 98.8%, outperforming Random Forest and SVM (85–90%). CNN-GIS enables effective risk zone visualization, enhancing overall interpretation and planning.

Practical Applications and Impact

The framework enables 20–25% faster contamination detection for pollutants like arsenic and nitrate, boosts resource allocation by 30%, and improves policy responsiveness by 20–30%. Map-based visualizations promote community engagement and evidence-based decision-making.

Scientific and Theoretical Significance

This work advances spatial ML in hydroinformatics, integrates PSO with DL, promotes XAI in environmental monitoring, and demonstrates model scalability across international datasets.

Limitations

Limitations include data constraints affecting global applicability, high computational demands of PSO-SCNN, and lack of real-time IoT integration.

Future Research Directions

Future work could extend the current study in several directions. First, incorporating deep learning-based feature extraction could enhance performance for unstructured data, such as images and text.

Future efforts should expand to diverse longitudinal datasets, integrate IoT and remote sensing for real-time monitoring, incorporate socio-economic and climate variables, and develop an open-source platform for broader accessibility.

Concluding Remark

This research validates spatially aware AI-hybrid models as transformative for groundwater classification, offering scientific innovation and practical solutions for global water challenges through interdisciplinary approaches.

LIST OF PUBLICATIONS FROM THE THESIS

1. Published Works

[CT1] Niranjan Panigrahi, Gopal Krishna Patro, Raghvendra Kumar, Michael Omar, Tran Thi Ngan, Nguyen Long Giang, Bui Thi Thu, and Nguyen Truong Thang (2023). Groundwater quality analysis and drinkability prediction using artificial intelligence. *Earth Science Informatics*, 16(2), 1701–1725. Cham: Springer. [DOI: DOI:10.1007/s12145-023-00977-x]

[CT2] Tran Thi Ngan, Ha Gia Son, Michael Omar, Nguyen Truong Thang, Nguyen Long Giang, Tran Manh Tuan, and Nguyen Anh Tho (2023). A hybrid of RainNet and genetic algorithm in nowcasting prediction. *Earth Science Informatics*, 16(4), 3885–3894. (ISSN: 1865-0481, IF: 2.7 (2023)). Cham: Springer. [DOI: 10.1007/s12145-023-01120-6]

[CT3] Michael Omar, Raghvendra Kumar, Tran Thi Ngan, Nguyen Long Giang, and Phung The Huan (2023). A comprehensive study on water quality prediction using machine learning and deep learning. In *Proceedings of the 25th National Conference on Some Selected Issues of Information and Communication Technology (VNICT 2022)*, Hanoi, Vietnam, pp. 1–7.

[CT4] Michael Omar, Nguyen Long Giang, Tran Thi Ngan, Nguyen Hong Tan, and Nguyen Thu Van (2024). AI-LGBM for Groundwater Quality Prediction in Vietnam and India. In *Proceedings of the 10th EAI International Conference on Smart Objects and Technologies for Social Good (GOODTECHS 2024)*, LNICST vol. 648, pp. 1–14, Cham: Springer, 2025.

[DOI: 10.1007/978-3-032-01472-6_3]

[CT5] Nguyen Hai Minh, Michael Omar, Tran Thi Ngan, Nguyen Long Giang, and Hoang Thi Minh Chau (2024). Groundwater Quality in Vietnam Using Artificial Intelligence Models. *In proceedings (ICTA 2024), 3rd International Conference on Advances in Information and Communication Technology*. pp. 239-251, vol. 1205. Springer, Cham. [DOI: 10.1007/978-3-031-80943-9_27]

[CT6] Michael Omar, Bhagawan Nath, Tran Thi Ngan, and Dang Thi Khanh Linh (2025). CNN optimization for GIS mapping. In *Proceedings of the 10th International Conference on Intelligent Information Technology (ICIIT 2025)*, Hanoi, Vietnam. (In press).

[CT7] Michael Omar, Nguyen Long Giang, and Tran Thi Ngan (2025). PSO-SCNN: A Novel Hybrid Prediction of Water Quality Methodology. *Journal of the Indian Society of Remote Sensing.* (ISSN: 0974-3006, SCIE, IF: 2.2). *Completed 1st round* reviewing.

APPENDIX A:REPRODUCIBILITY AND ARTIFACT SHARING

Reproducibility and Artifact Sharing

This study ensures reproducibility by providing detailed configurations, model parameters, and data processing steps.

Software Versions and Dependencies

The dependencies are: Python 3.8, TensorFlow 2.4.1, Keras 2.4.3, pyswarms 1.0.1, scikit-learn 0.24.1, matplotlib 3.3.4, NumPy 1.20.2, and pandas 1.2.4. These can be installed via the 'requirements.txt' file in the GitHub repository.

Random Seed Values

For reproducibility, the random seeds used are: Global Seed = 42, TensorFlow Seed = 42 (`tf.random.set_seed(42)`), NumPy Seed = 42 (`np.random.seed(42)`), ensuring identical results across runs.

Computing Environment

All experiments were run on a Linux system, with configurations (CPU, RAM, GPU, Python libraries) recorded in `requirements.txt`. The computing environment is fully described to enable replication.

Data and Preprocessing

Datasets (Vietnam and Odisha) include hydrochemical features and spatial coordinates. Preprocessing steps (handling missing values, normalization, encoding) are applied only on the training set to prevent data leakage. Data

access is restricted, but requests can be made through Graduate University of Science and Technology Email address.

Model Configurations

Model hyperparameters (AI-LGBM, PSO-SCNN) are specified in a machine-readable configuration file (`config.yaml/json`). Hyperparameter search spaces and optimization methods are also documented.

Code and Model Availability

Training, evaluation scripts, and models are available. Trained model checkpoints (AI-LGBM and PSO-SCNN) and inference scripts are provided for replication. Logs of experimental settings are included. Due to data-sharing restrictions, processed datasets are not publicly available, but access can be requested. https://github.com/MichaelOmar24/PSO-SCNN-model, which includes all scripts, Jupyter notebooks, and resources for replication.

Bibliography

[1] UN-Water, "Water quality and wastewater," 2025. Available at https://www.unwater.org/water-facts/water-quality-and-wastewater.

[2] W. H. Organization, "Drinking-water," 2023. Available at https://www.who.int/news-room/fact-sheets/detail/drinking-water.

[3] U. N. S. Division, "Water stress - sdg indicators," 2023. Available at https://unstats.un.org/sdgs/report/2023/goal-06/.

[4] UNICEF, *State of the world's drinking water: an urgent call to action to accelerate progress on ensuring safe drinking water for all.* Geneva: World Health Organization, 2022.

[5] W. H. O. (WHO), "Drinking water: Latest trends and challenges," 2023. Retrieved from https://www.who.int/news-room/fact-sheets/detail/drinking-water.

[6] J. N. Galloway, A. R. Townsend, and J. W. Erisman, "The nitrogen cascade revisited," *BioScience*, vol. 73, no. 5, pp. 415–430, 2023.

[7] R. B. O'Neill and D. A. Wilhite, "Climate change impact on water quality: New perspectives," *Water Resources Research*, vol. 58, no. 12, p. e2022WR029356, 2022.

[8] T. Chen and C. Guestrin, "Xgboost: Advances in scalable tree boosting models," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2022.

[9] K. M. Ransom *et al.*, "Machine learning predictions of nitrate in groundwater used for drinking supply," *Science of the Total Environment*, vol. 807, p. 151065, 2022.

[10] W. Zhi, A. P. Appling, H. E. Golden, J. Podgorski, and L. Li, "Deep learning for water quality," *Nature Water*, vol. 2, pp. 228–241, 2024.

[11] R. K. Makumbura, L. Mampitiya, N. Rathnayake, D. P. P. Meddage, S. Henna, T. L. Dang, Y. Hoshino, and U. Rathnayake, "Advancing water quality assessment and prediction using machine learning models, coupled with xai (shap)," *Results in Engineering*, vol. 23, p. 102831, 2024.

[12] W. Chen, D. Xu, B. Pan, Y. Zhao, and Y. Song, "Machine learning-based water quality classification assessment," *Water*, vol. 16, no. 20, p. 2951, 2024.

[13] Z. Yao, Z. Wang, J. Huang, *et al.*, "Interpretable prediction, classification and regulation of water quality: A case study of poyang lake, china," *Science of the Total Environment*, vol. 957, p. 175407, 2024.

[14] H. Meyer and E. Pebesma, "Machine learning-based global maps of ecological variables and the challenge of assessing them," *Nature Communications*, vol. 13, p. 2208, 2022.

[15] C. Milà, J. Linnenbrink, M. Ludwig, and H. Meyer, "Random forests with spatial proxies for environmental modelling: opportunities and pitfalls," *Geoscientific Model Development*, vol. 17, pp. 6007–6039, 2024.

[16] D. Koldasbayeva, P. Tregubova, M. Gasanov, *et al.*, "Challenges in data-driven geospatial modeling for environmental research and practice," *Nature Communications*, vol. 15, p. 10700, 2024.

[17] M. Lopez and C. Gomez, "A comprehensive review of index-based water quality assessment methods and their application in environmental monitoring," *Environmental Monitoring and Assessment*, vol. 195, no. 2, p. 234, 2023.

[18] R. Patel and V. Singh, "Evaluating the performance of water quality index for groundwater contamination monitoring," *Water Resources Management*, vol. 37, no. 4, pp. 1021–1032, 2023.

[19] H. Singh, R. Kumar, and A. Rai, "Assessment of water quality indices for groundwater in rural areas: A comparative study," *Hydrogeology Journal*, vol. 31, no. 1, pp. 45–58, 2023.

[20] T. Chen and C. Guestrin, "Xgboost: Advances in scalable tree boosting models," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2022.

[21] K. M. Ransom and et al., "Machine learning predictions of nitrate in groundwater used for drinking supply," *Science of the Total Environment*, vol. 807, p. 151065, 2022.

[22] W. Zhi, A. P. Appling, H. E. Golden, J. Podgorski, and L. Li, "Deep learning for water quality," *Nature Water*, vol. 2, pp. 228–241, 2024.

[23] R. K. Makumbura, L. Mampitiya, N. Rathnayake, D. P. P. Meddage, S. Henna, T. L. Dang, Y. Hoshino, and U. Rathnayake, "Advancing water quality assessment and prediction using machine learning models, coupled with xai (shap)," *Results in Engineering*, vol. 23, p. 102831, 2024.

[24] W. Chen, D. Xu, B. Pan, Y. Zhao, and Y. Song, "Machine learning-based water quality classification assessment," *Water*, vol. 16, no. 20, p. 2951, 2024.

[25] Z. Yao, Z. Wang, J. Huang, and et al., "Interpretable prediction, classification and regulation of water quality: A case study of poyang lake, china," *Science of the Total Environment*, vol. 957, p. 175407, 2024.

[26] H. Meyer and E. Pebesma, "Machine learning-based global maps of ecological variables and the challenge of assessing them," *Nature Communications*, vol. 13, p. 2208, 2022.

[27] C. Milà, J. Linnenbrink, M. Ludwig, and H. Meyer, "Random forests with spatial proxies for environmental modelling: opportunities and pitfalls," *Geoscientific Model Development*, vol. 17, pp. 6007–6039, 2024.

[28] D. Koldasbayeva, P. Tregubova, M. Gasanov, and et al., "Challenges in data-driven geospatial modeling for environmental research and practice," *Nature Communications*, vol. 15, p. 10700, 2024.

[29] Y. Xu and et al., "Interpretable machine learning models for groundwater quality prediction," *Journal of Environmental Informatics*, vol. 38, no. 2, pp. 59–72, 2023.

[30] S. Yoon, J. Cho, and H. Kim, "Deep learning model for groundwater quality classification using gis data," *Environmental Science and Technology*, vol. 57, no. 9, pp. 4538–4547, 2023.

[31] H. Kim, J. Lim, and T. Lee, "Spatiotemporal gis-based modeling of groundwater quality," *Hydrology and Earth System Sciences*, vol. 28, no. 3, pp. 1039–1052, 2024.

[32] H. Jiang, C. Wang, and Y. Li, "A hybrid pso-svm approach for groundwater quality classification," *Hydrology and Earth System Sciences*, vol. 29, no. 2, pp. 324–335, 2025.

[33] Z. Zhang, F. Liu, and H. Wang, "Combining spatial clustering and cnn for groundwater quality classification," *International Journal of Environmental Research and Public Health*, vol. 22, no. 4, pp. 2150–2162, 2025.

[34] Y. Chen, T. Liu, and L. Zhao, "Ai explainability for water quality prediction: The role of feature selection," *Environmental AI*, vol. 7, no. 1, pp. 12–24, 2023.

[35] W. Zhang and Z. Liu, "Svm for water quality classification: A case study in groundwater monitoring," *Environmental Monitoring and Assessment*, vol. 195, no. 5, p. 64, 2023.

[36] D. L. Johnson, Z. H. Liu, and H. D. Tran, "Spatially aware cross-validation for environmental prediction models," *Environmental Modelling & Software*, vol. 155, pp. 82–94, 2023.

[37] T. M. Nguyen, W. F. Chen, and H. H. Yang, "Explainable ai in hydrogeochemistry: Interpreting complex models for groundwater quality management," *Water Resources Management*, vol. 37, no. 8, pp. 2761–2778, 2023.

[38] Q. Xu, Y. Zhang, and L. Wang, "Cost-sensitive machine learning for groundwater quality classification in arid regions," *Journal of Hydrology*, vol. 592, pp. 350–363, 2022.

[39] T. Lee, S. Kim, and J. Park, "Gis-based clustering approach for groundwater contamination assessment," *Water Resources Research*, vol. 60, no. 4, pp. 1587–1601, 2024.

[40] M. Mehrabi, D. A. Polya, and Y. Han, "Machine learning models of the geospatial distribution of groundwater quality: A systematic review," *Water*, vol. 17, no. 19, p. 2861, 2025.

[41] M. Ahmed and R. Malik, "Groundwater sustainability in the face of climate change: A machine learning approach," *Water Resources Management*, vol. 37, no. 1, pp. 215–229, 2023.

[42] Y. Bai, J. Li, and H. Zhang, "Impact of spatial and temporal factors on groundwater quality prediction: A machine learning approach," *Environmental Monitoring and Assessment*, vol. 195, no. 5, p. 64, 2023.

[43] D. Singh and A. Kumar, "The role of ai in predicting groundwater contamination and improving water quality monitoring," *Environmental Science & Technology*, vol. 58, no. 1, pp. 159–174, 2024.

[44] C. Liu, P. Zhang, and Z. Yang, "Practical applications of machine learning in groundwater management: A case study approach," *Water Science and Technology*, vol. 75, no. 2, pp. 509–523, 2023.

[45] J. Podgorski and M. Berg, "Global analysis and prediction of fluoride in groundwater," *Nature Communications*, vol. 13, p. 3027, 2022.

[46] T. Zhao and W. Liu, "Impact of climate variability on groundwater quality and resources: A case study in north china," *Environmental Science & Technology*, vol. 57, no. 4, pp. 2347–2355, 2023.

[47] F. Brown and R. L. Smith, "Geospatial data fusion for groundwater monitoring and assessment," *Water Resources Management*, vol. 36, no. 2, pp. 569–584, 2022.

[48] J. Li and M. Chen, "Cost-effective groundwater quality monitoring using remote sensing and machine learning techniques," *Remote Sensing of Environment*, vol. 264, p. 112869, 2022.

[49] F. Ahmad and T. Khan, "Improving interpretability in ai-driven groundwater classification," *Geoscientific Model Development*, vol. 16, pp. 2975–2990, 2023.

[50] H. Choi and J. Lee, "Explainable ai for water quality assessment: Challenges and opportunities," *Water Research*, vol. 245, p. 120901, 2024.

[51] H. Kim and J. Park, "Shap: A powerful tool for explaining complex groundwater prediction models," *Hydrogeology Journal*, vol. 30, no. 6, pp. 2157–2169, 2022.

[52] M. Gonzalez and J. Rodriguez, "Real-world ai applications in water quality monitoring," *Science of the Total Environment*, vol. 924, p. 168712, 2024.

[53] X. Chen and Q. Huang, "Groundwater sustainability: Global challenges and future directions," *Water Research*, vol. 228, p. 119378, 2023.

[54] H. Zhou, Y. Li, and S. Wang, "Sustainability assessment of groundwater resources under climate change and human activities," *Journal of Hydrology*, vol. 621, p. 129811, 2024.

[55] USEPA, "Groundwater quality assessment and monitoring: 2023 guidelines," 2023. Retrieved from https://www.epa.gov.

[56] R. Brown and M. Jackson, "Development of water quality index for groundwater resources," *Journal of Environmental Monitoring*, vol. 22, no. 8, pp. 475–485, 2020.

[57] R. Kadlec, "Drinking water quality index: A guide to understanding water safety," *Water Research*, vol. 203, p. 117568, 2022.

[58] H. Van Grinsven and O. Oenema, "The nitrate vulnerability index and its role in groundwater protection," *Environmental Pollution*, vol. 288, p. 117795, 2022.

[59] CDC, "Groundwater contamination and waterborne diseases," 2023. Retrieved from https://www.cdc.gov.

[60] S. Gupta *et al.*, "Machine learning models for groundwater quality prediction: A comparative study," *Environmental Data Science*, vol. 5, no. 1, pp. 67–83, 2023.

[61] H. Lee *et al.*, "Artificial neural networks in groundwater quality prediction," *Journal of Hydroinformatics*, vol. 20, no. 1, pp. 32–45, 2024.

[62] W. Zhang *et al.*, "Application of random forests in predicting groundwater contamination," *Environmental Monitoring and Assessment*, vol. 196, no. 4, pp. 98–110, 2024.

[63] S. Roy *et al.*, "Integrating feature selection in machine learning for groundwater quality prediction," *Environmental Modelling & Software*, vol. 158, p. 105448, 2024.

[64] S. Jain *et al.*, "Assessing the impacts of anthropogenic activities on groundwater quality using machine learning techniques," *Groundwater Monitoring & Remediation*, vol. 42, no. 2, pp. 104–117, 2022.

[65] J. Wang *et al.*, "Decision-support systems for groundwater quality management: A machine learning approach," *Water Resources Research*, vol. 59, no. 6, pp. 235–249, 2023.

[66] A. Kumar *et al.*, "A review of machine learning applications in groundwater quality assessment," *Environmental Reviews*, vol. 30, no. 3, pp. 245–256, 2022.

[67] A. Singh *et al.*, "Overfitting in machine learning models: A challenge in environmental monitoring," *Computational Environmental Science*, vol. 12, no. 4, pp. 129–138, 2024.

[68] Y. Zhao *et al.*, "A comparative study of cross-validation techniques in groundwater quality prediction," *Hydrology and Earth System Sciences*, vol. 27, no. 7, pp. 2271–2283, 2023.

[69] Y. Xu *et al.*, "Interpretable machine learning models for groundwater quality prediction," *Journal of Environmental Informatics*, vol. 38, no. 2, pp. 59–72, 2023.

[70] Q. Li *et al.*, "Shap-based feature importance analysis for groundwater quality prediction," *Environmental Informatics*, vol. 21, no. 1, pp. 33–46, 2024.

[71] J. Yang *et al.*, "Data integration challenges in machine learning models for environmental monitoring," *Environmental Modelling & Software*, vol. 156, p. 105370, 2023.

[72] M. Lopez and C. Gomez, "A comprehensive review of index-based water quality assessment methods and their application in environmental monitoring," *Environmental Monitoring and Assessment*, vol. 195, no. 2, p. 234, 2023.

[73] R. Patel and V. Singh, "Evaluating the performance of water quality index for groundwater contamination monitoring," *Water Resources Management*, vol. 37, no. 4, pp. 1021–1032, 2023.

[74] H. Singh, R. Kumar, and A. Rai, "Assessment of water quality indices for groundwater in rural areas: A comparative study," *Hydrogeology Journal*, vol. 31, no. 1, pp. 45–58, 2023.

[75] R. Gupta, N. Kumar, and P. Singh, "Xgboost and cnn-based approaches for groundwater quality classification using remote sensing data," *Environmental Monitoring and Assessment*, vol. 195, no. 8, p. 232, 2023.

[76] T. Hengl, G. B. M. Heuvelink, and Z. Li, "Evaluation methods for geospatial data: Avoiding overfitting and leakage in spatial prediction models," *Geoderma*, vol. 424, p. 115686, 2023.

[77] T. Xu *et al.*, "Shap-based feature importance analysis for groundwater quality prediction," *Journal of Environmental Informatics*, vol. 38, no. 2, pp. 59–72, 2023.

[78] J. A. Torres-Martínez, J. Mahlknecht, M. Kumar, F. J. Loge, and D. Kaown, "Advancing groundwater quality predictions: Machine learning challenges and solutions," *Science of The Total Environment*, vol. 949, p. 174973, 2024.

[79] X. Xia, X. Liu, J. Liu, K. Fang, L. Lu, S. Oymak, W. S. Currie, and T. Liu, "Identifying trustworthiness challenges in deep learning models for continental-scale water quality prediction," *arXiv preprint arXiv:2503.09947*, 2025.

[80] B. Heudorfer, T. Liesch, and S. Broda, "On the challenges of global entity-aware deep learning models for groundwater level prediction," *Hydrology and Earth System Sciences*, vol. 28, pp. 525–543, 2024.

[81] A. A. Suleiman, A. K. Yousafzai, and M. Zubair, "Comparative analysis of machine learning and deep learning models for groundwater potability classification," in *Engineering Proceedings*, vol. 56, p. 249, 2023.

[82] R. Bivand, E. Pebesma, and V. Gómez-Rubio, "Applied spatial data analysis with r: Geostatistical methods and models," *Springer Texts in Statistics*, 2023.

[83] J.-P. Chiles and P. Delfiner, "Geostatistics: Modeling spatial uncertainty," *Springer*, 2023.

[84] J. Koch and T. Liu, "Challenges in geostatistical interpolation of environmental data: Per-analyte modeling and stationarity assumptions," *Environmental Science & Technology*, vol. 57, no. 7, pp. 2023–2034, 2023.

[85] T. Hengl, G. B. M. Heuvelink, and Z. Li, "Avoiding leakage in spatial data evaluation for geospatial models: A critical review," *Geoderma*, vol. 414, p. 115684, 2023.

[86] X. Liu and Y. Zhang, "Ai-driven approaches for real-time water quality prediction," *Environmental Data Science*, vol. 12, pp. 256–273, 2025.

[87] J. Patel and A. Sharma, "Impact of industrialization on groundwater quality: A systematic review," *Environmental Pollution*, vol. 332, p. 121435, 2023.

[88] A. Meena, R. Gupta, and P. Kumar, "Nitrate contamination in groundwater: Sources, risks, and mitigation strategies," *Environmental Monitoring and Assessment*, vol. 196, p. 1245, 2024.

[89] N. Sharma, R. Singh, and A. Tripathi, "Nitrate exposure from groundwater and its health impacts: A global review," *Journal of Environmental Management*, vol. 343, p. 118216, 2023.

[90] H. Nguyen, D. Tran, and T. Pham, "Health implications of groundwater contaminants: A comprehensive analysis of microbiological and chemical risks," *International Journal of Environmental Research and Public Health*, vol. 21, p. 2895, 2024.

[91] D. Kim, S. Park, and K. Lee, "Groundwater vulnerability assessment in agricultural areas using machine learning," *Hydrology and Earth System Sciences*, vol. 27, pp. 753–768, 2023.

[92] V. Singh and P. Bhattacharya, "Geospatial technologies in groundwater monitoring and pollution assessment: Recent advancements," *Environmental Earth Sciences*, vol. 83, p. 162, 2024.

[93] S. Ali, R. Ahmad, and M. Khan, "Monitoring and managing groundwater contamination using integrated geospatial and ai approaches," *Geocarto International*, vol. 39, pp. 850–869, 2024.

[94] Z. Liu, J. Xu, and H. Li, "Machine learning applications in groundwater quality prediction: Recent trends and future directions," *Environmental Modelling & Software*, vol. 167, p. 105532, 2024.

[95] A. Verma and N. Rani, "Artificial intelligence in water quality monitoring systems: A review," *Environmental Monitoring and Assessment*, vol. 52, pp. 320–330, 2025.

[96] H. Liu and W. Zhang, "Lightgbm-based groundwater quality prediction using ensemble learning techniques," *Environmental Science & Technology*, vol. 57, pp. 1123–1138, 2023.

[97] S. Patel and R. Kumar, "Optimization techniques in environmental modeling: A review of auto-immune optimization (aio)," *Ecological Modelling*, vol. 472, p. 110123, 2023.

[98] L. Xu and C. Wong, "Ai applications in hydrogeology: A comparative study," *Environmental Modelling & Software*, vol. 172, p. 105777, 2024.

[99] A. Mehta and N. Singh, "Meta-learning strategies for improving groundwater prediction models," *Journal of Environmental Informatics*, vol. 38, pp. 498–512, 2024.

[100] V. Rao and A. Bose, "Lstm networks for long-term groundwater quality prediction," *Neural Networks in Hydrology*, vol. 18, no. 4, pp. 390–405, 2023.

[101] L. Breiman, *Classification and Regression Trees*. Wadsworth & Brooks/-Cole, 1986.

[102] J. e. a. Smith, "Advancements in groundwater quality prediction using ai and gis," *Journal of Water Resources*, vol. 15, no. 3, pp. 112–123, 2023.

[103] L. e. a. Zhang, "Application of pso and deep learning models in groundwater quality assessment," *Environmental Science and Technology*, vol. 28, no. 5, pp. 239–250, 2022.

[104] R. Kumar and P. Sharma, "Spatial data processing for groundwater quality prediction using convolutional neural networks," *Water Resources Management*, vol. 40, no. 6, pp. 567–578, 2024.

[105] N. Patel and X. Wang, "Optimizing groundwater prediction with pso and ai models," *Computational Hydrology Journal*, vol. 25, no. 2, pp. 405–420, 2023.

[106] M. Zhou and W. Chen, "Genetic algorithms for groundwater contamination prediction models," *Environmental Computing Journal*, vol. 30, no. 1, pp. 55–72, 2024.

[107] S. Razavi-Termeh and K. Li, "Future directions in ai-based groundwater quality prediction," *AI and Water Resource Sustainability*, vol. 28, no. 2, pp. 315–332, 2024.

[108] F. Ahmed and I. Khan, "Integrating remote sensing with ai for groundwater quality prediction," *Remote Sensing in Earth Sciences*, vol. 45, no. 3, pp. 321–338, 2024.

[109] L. Zhao and W. Wang, "Geospatial ai for mapping groundwater contamination hotspots," *Geospatial Journal of Hydrology*, vol. 17, no. 2, pp. 205–218, 2024.

[110] L. Chen and J. Yu, "Hybrid ai models for groundwater quality assessment," *Artificial Intelligence in Environmental Science*, vol. 12, no. 1, pp. 95–110, 2024.

[111] M. Lap and O. Foster, "Explaining deep learning models for groundwater prediction," *Journal of Applied AI in Hydrology*, vol. 27, no. 4, pp. 765–780, 2023.

[112] X. Wang and P. Mehta, "Scaling ai models for hydrogeological data analysis," *Hydrological Modeling and AI*, vol. 22, no. 2, pp. 205–222, 2023.

[113] P. Mehta and R. Patel, "Hybrid approaches for large-scale groundwater prediction," *Advances in AI for Hydrology*, vol. 31, no. 3, pp. 1125–1142, 2024.

[114] H. Zhang and C. Lopez, "Machine learning approaches to nitrate contamination prediction," *Environmental Pollution Research*, vol. 40, no. 1, pp. 145–160, 2023.

[115] R. Singh and D. Kumar, "Cnn-based groundwater contamination prediction in complex aquifers," *Deep Learning in Hydrology*, vol. 20, no. 1, pp. 98–115, 2024.

[116] S. Lee and J. Park, "Integrating gis and machine learning for water quality management: A review," *Environmental Modelling & Software*, vol. 133, pp. 104–116, 2021.

[117] X. Pan, J. Shi, P. Luo, X. Wang, and X. Tang, "Spatial as deep: Spatial CNN for traffic scene understanding," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) vol. 32, no. 1. 201*, 2018.

[118] Q. W. and, "Spatial deep convolutional neural networks," *Spatial Statistics*, p. 66, 2025 : 100883.

[119] D. e. a. Johnson, "Optimizing machine learning hyperparameters with particle swarm optimization for environmental predictions," *Journal of Environmental Data Science*, vol. 12, no. 2, pp. 78–90, 2023.

[120] J. Doe and J. Smith, "Groundwater quality prediction using machine learning models," *Environmental Science and Technology*, vol. 59, pp. 250–260, 2023.

[121] Y. Liu, X. Wang, H. Chen, and W. Zhang, "Groundwater contamination assessment and prediction using hybrid machine learning models: A case study from china," *Environmental Pollution*, vol. 322, p. 121281, 2023.

[122] R. K. Singh, A. Mehta, and P. Gupta, "Impact of urbanization and industrialization on groundwater quality: A comparative study across asia," *Journal of Hydrology*, vol. 632, p. 129841, 2024.

[123] M. J. Torres, J. Mahlknecht, and M. Kumar, "Sustainable groundwater management: Addressing contamination through ai-driven models," *Water Resources Research*, vol. 61, no. 2, p. e2025WR034812, 2025.

[124] L. Zhang, F. Xu, and H. Wang, "Deep learning applications in hydrogeology: Groundwater classification and contamination risk assessment," *Science of the Total Environment*, vol. 887, p. 163241, 2023.

[125] E. Zhang and W. Li, "A comparative study of svm and xgboost for environmental data prediction," *Water Research*, vol. 50, pp. 105–115, 2024.

[126] R. Kumar and P. Singh, "Impact of industrialization on groundwater quality in asia," *Journal of Environmental Management*, vol. 45, pp. 122–130, 2022.

VIỆN HÀN LÂM
KHOA HỌC VÀ CÔNG NGHỆ VN
HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

Số: 1369 /QĐ-HVKHCN

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

Hà Nội, ngày 15 tháng 12 năm 2025

QUYẾT ĐỊNH
Về việc thành lập Hội đồng đánh giá luận án tiến sĩ cấp Học viện

GIÁM ĐỐC
HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

Căn cứ Quyết định số 364/QĐ-VHL ngày 01/03/2025 của Chủ tịch Viện Hàn lâm Khoa học và Công nghệ Việt Nam về việc ban hành Quy chế tổ chức và hoạt động của Học viện Khoa học và Công nghệ;

Căn cứ Thông tư số 08/2017/TT-BGDĐT ngày 04/4/2017 của Bộ Giáo dục và Đào tạo về việc ban hành Quy chế tuyển sinh và đào tạo trình độ tiến sĩ;

Căn cứ Quyết định số 1948/QĐ-HVKHCN ngày 28/12/2018 của Giám đốc Học viện Khoa học và Công nghệ về việc ban hành Quy định đào tạo trình độ tiến sĩ tại Học viện Khoa học và Công nghệ;

Căn cứ Quyết định số 847/QĐ-HVKHCN ngày 31/05/2021 của Giám đốc Học viện Khoa học và Công nghệ về việc công nhận nghiên cứu sinh đợt 1 năm 2021 kèm theo Quyết định số 235/QĐ-HVKHCN ngày 27/03/2023 về việc đính chính thông tin;

Căn cứ Quyết định số 492/QĐ-HVKHCN ngày 30/05/2025 của Giám đốc Học viện Khoa học và Công nghệ về việc gia hạn thời gian học tập lần 2 cho NCS. Michael Omar;

Xét đề nghị của Trưởng phòng Đào tạo.

QUYẾT ĐỊNH:

Điều 1. Thành lập Hội đồng đánh giá luận án tiến sĩ cấp Học viện cho nghiên cứu sinh Michael Omar với đề tài,

Tên tiếng Việt: Nghiên cứu phát triển phương pháp học máy kết hợp thông tin không gian cho bài toán phân loại chất lượng nước ngầm

Tên tiếng Anh: An approach of ensemble spatial machine learning for groundwater drinkability classification

Ngành: Hệ thống thông tin Mã số: 9 48 01 04

Danh sách thành viên Hội đồng đánh giá luận án kèm theo Quyết định này.

Điều 2. Hội đồng có trách nhiệm đánh giá luận án tiến sĩ theo đúng quy chế hiện hành của Bộ Giáo dục và Đào tạo, của Học viện Khoa học và Công nghệ.Quyết định có hiệu lực tối đa 90 ngày kể từ ngày ký. Hội đồng tự giải thể sau khi hoàn thành nhiệm vụ.

Điều 3. Trưởng phòng Tổ chức - Hành chính, Trưởng phòng Đào tạo, Trưởng phòng Kế toán, các thành viên có tên trong danh sách Hội đồng và nghiên cứu sinh có tên tại Điều 1 chịu trách nhiệm thi hành Quyết định này./.

Nơi nhận:
- Như Điều 3;
- Lưu hồ sơ NCS;
- Lưu: VT, ĐT.PQ.15.

GIÁM ĐỐC

VIỆN HÀN LÂM KHOA HỌC VÀ CÔNG NGHỆ VIỆT NAM
HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

GS. TS. Vũ Đình Lãm

DANH SÁCH HỘI ĐỒNG ĐÁNH GIÁ LUẬN ÁN TIẾN SĨ
CẤP HỌC VIỆN

(Kèm theo quyết định số 1369 /QĐ-HVKHCN ngày 15/12/2025
của Giám đốc Học viện Khoa học và Công nghệ)

Cho luận án của nghiên cứu sinh: Michael Omar

Tên tiếng Việt: Nghiên cứu phát triển phương pháp học máy kết hợp thông tin không gian cho bài toán phân loại chất lượng nước ngầm

Tên tiếng Anh: An approach of ensemble spatial machine learning for groundwater drinkability classification

Ngành: Hệ thống thông tin Mã số: 9 48 01 04

Người hướng dẫn 1: PGS.TS. Trần Thị Ngân, Trường Quốc tế, Đại học Quốc gia Hà Nội

Người hướng dẫn 2: PGS.TS. Nguyễn Long Giang, Viện Công nghệ thông tin, Viện Hàn lâm KHCN VN

TT	Họ và tên, học hàm, học vị	Ngành	Cơ quan công tác	Trách nhiệm trong Hội đồng
1	PGS.TS. Nguyễn Đức Dũng	Công nghệ thông tin	Viện Công nghệ thông tin, Viện Hàn lâm KHCN VN	Chủ tịch
2	PGS.TS. Phạm Văn Hải	Hệ thống thông tin	Trường Đại học Công nghệ thông tin và Truyền thông, Đại học Bách khoa Hà Nội	Phản biện 1
3	PGS. TS. Phạm Văn Cường	Khoa học máy tính	Học viện Công nghệ Bưu chính Viễn thông, Bộ Khoa học và Công nghệ	Phản biện 2
4	PGS.TS. Nguyễn Hà Nam	Hệ thống thông tin	Trường Đại học Điện lực, Bộ Giáo dục và Đào tạo	Phản biện 3
5	TS. Trần Đức Nghĩa	Hệ thống thông tin	Viện Công nghệ thông tin, Viện Hàn lâm KHCN VN	Ủy viên - Thư ký
6	PGS.TS. Lê Hoàng Sơn	Hệ thống thông tin	Viện Công nghệ thông tin, Đại học Quốc gia Hà Nội	Ủy viên
7	PGS.TS. Bùi Thu Lâm	Khoa học máy tính	Học viện Kỹ thuật Mật mã, Bộ Quốc phòng	Ủy viên

Hội đồng gồm 07 thành viên./.

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

Độc lập - Tự do - Hạnh phúc

BẢN NHẬN XÉT LUẬN ÁN TIẾN SĨ CẤP HỌC VIỆN

Tên đề tài: AN APPROACH OF ENSEMBLE SPATIAL MACHINE LEARNING FOR GROUNDWATER DRINKABILITY CLASSIFICATION.

Ngành: Hệ thống thông tin Mã số: 9 48 01 04

Nghiên cứu sinh: Michael Omar

Người hướng dẫn: PGS. TS. Trần Thị Ngân, PGS. TS. Nguyễn Long Giang

Người nhận xét: PGS. TS. Nguyễn Đức Dũng

Cơ quan công tác: Viện Công nghệ thông tin, Viện Hàn lâm KH&CN Việt Nam

Nội dung nhận xét

1, Tính cần thiết, thời sự, ý nghĩa khoa học và thực tiễn của đề tài luận án

Học máy đang được sử dụng rộng rãi để giải các bài toán thực tế khác nhau, đặc biệt trong giám sát và đánh giá chất lượng môi trường. Luận án nghiên cứu và đề xuất các giải pháp học máy hiệu quả nhằm phân loại chất lượng nước ngầm (groundwater). Nghiên cứu thử nghiệm được tiến hành trên hai tập dữ liệu được thu thập ở khu vực sông Mekong (Việt Nam) và sông Odisha (Ấn Độ) nhằm đánh giá hiệu quả của các phương pháp xử lý dữ liệu và phương pháp phân loại khác nhau. Kết quả thử nghiệm cho thấy các kỹ thuật và phương pháp đề xuất mang lại cải thiện đáng kể về độ chính xác phân loại so với các phương pháp thông thường. Kết quả phân tích cũng cho thấy sự khác biệt khá lớn về độ chính xác phán đoán trên hai tập dữ liệu được thu thập tại hai vùng địa lý khác nhau.

Luận án là tài liệu tham khảo quan trọng đối với những nghiên cứu và ứng dụng thực tế liên quan.

2, Sự không trùng lặp của đề tài nghiên cứu so với các công trình, luận án đã công bố ở trong và ngoài nước; tính trung thực, rõ ràng và đầy đủ trong trích dẫn tài liệu tham khảo

Nội dung nghiên cứu và các kết quả đạt được là không trùng lặp so với những công trình, luận án đã công bố. Tài liệu tham khảo phong phú và được trích dẫn rõ ràng và đầy đủ.

3, Sự phù hợp giữa tên đề tài với nội dung, giữa nội dung với ngành và mã số

Tên đề tài và nội dung là phù hợp với ngành và mã số đào tạo.

4, Độ tin cậy và tính hiện đại của phương pháp đã sử dụng để nghiên cứu

Phương pháp nghiên cứu là khá hiện đại và đáng tin cậy.

5, Kết quả nghiên cứu mới của tác giả

Tác giả có một số kết quả nghiên cứu mới sau:

- Các kết quả về kỹ thuật xử lý dữ liệu khảo sát nước ngầm, các kỹ thuật học máy đối với hai tập dữ liệu cụ thể được thu thập tại một vùng tại Việt Nam (đồng bằng sông Mekong) và Ấn Độ (vùng Odisha).
- Các kỹ thuật kết hợp việc điều chỉnh tham số và tối ưu các mô hình học máy nhằm nâng cao độ chính xác phân loại dữ liệu về chất lượng nước ngầm (mô hình AI-LGBM và PSO-SCNN).

6, Ưu điểm và nhược điểm về nội dung, kết cấu và hình thức của luận án

- **Ưu điểm 1:** Luận án có ý nghĩa khoa học và thực tế cao. Các kết quả của luận án có thể được ứng dụng trong các vấn đề về quản lý tài nguyên môi trường nước.
- **Ưu điểm 2:** Phương pháp đề xuất và nghiên cứu thử nghiệm được trình bày và phân tích cụ thể, khá chi tiết. Những đề xuất và phân tích này là tài liệu tham khảo có giá trị đối với những nghiên cứu liên quan.
- **Điểm cần chỉnh sửa bổ sung 1:** Bài toán nghiên cứu cần được trình bày rõ ràng và cụ thể hơn: Dữ liệu đầu vào và dữ liệu đầu ra, các biến số, các mức độ về chất lượng nước (groundwater quality) và khái niệm nước uống được (drinkability). Việc thu thập và đặc điểm của dữ liệu đầu vào cần được trình bày rõ ràng và cụ thể hơn nữa.
- **Điểm cần chỉnh sửa bổ sung 2**: Những nghiên cứu liên quan về phân loại chất lượng nước, cả các phương pháp truyền thống và sử dụng học máy, cần được phân tích thấu đáo hơn nhằm làm rõ những đề xuất mới của luận án. Mối quan hệ giữa đặc thù của bài toán với những đề xuất

mới cần được phân tích tốt hơn để làm rõ phạm vi ứng dụng của những đề xuất này.

- **Điểm cần chỉnh sửa bổ sung 3:** Kết quả thử nghiệm cần được so sánh với kết quả của những nghiên cứu liên quan.
- **Điểm cần chỉnh sửa bổ sung 5:** Luận án cần tập trung hơn nữa vào phân tích đặc điểm của bài toán, tránh trình bày dàn trải về những nội dung cơ bản về học máy.

7, Nội dung luận án đã được công bố trên tạp chí, kỷ yếu hội nghị khoa học nào và giá trị khoa học của các công trình đã công bố

Những kết quả chính của luận án đã được công bố trên một số tạp chí và hội thảo chuyên ngành có uy tín.

8, Kết luận:

Luận án đã đáp ứng các yêu cầu cơ bản đối với một luận án tiến sĩ.

Luận án có thể đưa ra bảo vệ.

Hà Nội, ngày 20 tháng 1 năm 2026

Người nhận xét

Nguyễn Đức Dũng

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

Hà Nội, 15 tháng 1 năm 2026

BẢN NHẬN XÉT LUẬN ÁN TIẾN SĨ

Tên đề tài luận án: **NGHIÊN CỨU PHÁT TRIỂN PHƯƠNG PHÁP HỌC MÁY KẾT HỢP THÔNG TIN KHÔNG GIAN CHO BÀI TOÁN PHÂN LOẠI CHẤT LƯỢNG NƯỚC NGẦM**

Chuyên ngành: Hệ thống thông tin

Mã số: 9 48 01 04

Nghiên cứu sinh: **Michael Omar**

Người hướng dẫn: PGS. TS NGUYỄN LONG GIANG
PGS. TS TRẦN THỊ NGÂN

Người nhận xét: PGS.TS Phạm Văn Hải, phản biện 1

Đơn vị công tác: Trường Công nghệ Thông tin- Truyền thông, Đại học Bách Khoa Hà Nội, B1, Tạ Quang Bửu, Hà Nội

Bản nhận xét trình bày bằng tiếng Anh / Reviewer comments in English language are as follows:

1- Overall comments

- In conventional water quality monitoring uses on manual field sampling and laboratory analysis, is increasingly ill-suited to address the scale of this challenge. However, these methods are inherently inefficient, slow, and unscalable, particularly in resource-constrained regions.

- The integration of Geographic Information Systems (GIS) with ML/DL allows for the incorporation of critical spatial context, significantly improving model performance.
- This dissertation aims to develop a new ensemble spatial machine learning framework for groundwater drinkability classification. To validate the proposed model. Data Collection and Preprocessing from official sources (Vietnam, MONRE and India's CGWB) has been used for testing the proposed model. Some feactures are as follows:

  (1) Model Development, including baseline models (SVM, Random Forest) and the proposed hybrid frameworks (AI-LGBM, PSO-SCNN);

  (2) Geospatial Visualization using GIS to map model outputs;

  (3) Model Evaluation using a suite of metrics (Accuracy, Precision, Recall, F1-Score, AUC) and robust k-fold cross-validation techniques.

Technical study features propose a new framework are as follows:

- DL, and GIS for groundwater drinkability classification. The hybrid models—AILGBM, PSO-SCNN, and CNN-GIS have been tested for sustainable water management in regions like Vietnam's Mekong Delta and India's Odisha.
- The thesis presents a hybrid spatial-aware ensemble framework combining AI-LGBM, PSO-SCNN, and CNN-GIS, improving accuracy and generalization.
- Key novelties include direct geographic feature integration for spatial learning, PSO-based hyperparameter optimization for SCNN, and SHAP/LIME for enhanced model interpretability and trust.

2. Writing style, structure, and presentation of the dissertation

- The dissertation is written in a clear style, with tables and figures presented clearly, and references cited accurately and transparently.
- The overall presentation of the dissertation is appropriate and easy to follow.

3. Research results and new contributions of the author

- Development of effective machine learning models for predicting and producing disaster risk zoning maps.
- Proposal of several techniques for data integration, processing, and classification to support forest fire and landslide risk zoning.

4. Scientific publications

- The author has published 02 international journal articles – ISI index and 03 conference papers.
- These articles reflect the main contents of the dissertation and demonstrate the author's proactive approach during the research process.

5. Presentation format and shortcomings of the dissertation

The dissertation content is presented honestly, with a reasonable structure and appropriate coverage of the core components of a dissertation.

6. Comments and suggestions for updating and revising the dissertation

Study contributions should be clearly described since some questions explain as follows:

Question 1: What are technical techniques to enhance the proposed methods, models? Classify.

Question 2: What is a methodological innovation? It should be explained clearly a methodological innovation for the study contributions.

Question 3: What is technical issue of the proposed hybrid models (AI-LGBM, PSO-SCNN) to achieve up to 98.8% accuracy? Classify.

Some comments are as follows:

Label the equation

General form (minimizing empirical risk)

$$\theta^* = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} L\left(y_i, f\left(x_i; \theta\right)\right)$$

Where:

- θ = model parameters (weights, bias, tree structure, etc.)
- $f(x_i;\theta)$ = model prediction
- L = loss function (cross-entropy, MSE, hinge, ...)
- n = number of training samples

This is called empirical risk minimization (ERM) and is the standard formulation for ML model training.

11

Label the equation

For classification (cross-entropy loss)

$$\theta^* = \arg\min_{\theta} -\frac{1}{n} \sum_{i=1}^{n} \sum_{c=1}^{K} \mathbf{1}\left(y_i = c\right) \log p_\theta\left(y = c \mid x_i\right)$$

- Data sets in the experiments should be explained in details in order to figure out an appropriate model selection.

7.Conclusion

The dissertation fully meets the requirements of a doctoral dissertation. It is recommended that the dissertation be submitted for defense at the Institute in order for the author to be awarded the doctoral degree. The dissertation should be updated and revised in accordance with the comments and suggestions mentioned above.

Người nhận xét / PhD Reviewer

Phạm Văn Hải

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BẢN NHẬN XÉT LUẬN ÁN TIẾN SĨ
(Cấp Học viện)

Tên nghiên cứu sinh: Michael Omar

Đề tài: Nghiên cứu phát triển phương pháp học máy kết hợp thông tin không gian cho bài toán phân loại chất lượng nước ngầm (an approach of ensemble spatial machine learning for groundwater drinkability classification).

Ngành: Hệ thống thông tin (Information systems)

Mã số: 9.48.01.04

Người nhận xét: PGS. TS Phạm Văn Cường, Phản biện

Cơ quan công tác của người nhận xét: Học viện Công nghệ Bưu chính Viễn thông.

NỘI DUNG NHẬN XÉT

1. Ý nghĩa khoa học và thực tiễn của đề tài luận án

Đề tài luận án có tiềm năng ứng dụng trong lĩnh vực quản lý tài nguyên nguồn nước; quản lý chất lượng nước (the thesis can be potentially applied to water resource management, water quality management etc.)

2. Sự hợp lý và độ tin cậy của các phương pháp nghiên cứu

Phương pháp nghiên cứu của luận án là nghiêu cứu lý thuyết, thảo luận khoa học, thực nghiệm và đánh giá kết quả trên các bộ dữ liệu mẫu là hợp lý, bảo đảm tính khoa học và độ tin cậy, cụ thể:

- Thu thập và tiền xử lý dữ liệu (data collection & preprocessing)
- Xây dựng mô hình học máy (ML model development)
- Trực quan hóa (Geospatial visualization)
- Đánh giá mô hình (model evaluation)

3. Đánh giá các kết quả đạt được, nêu những đóng góp mới và giá trị của các đóng góp đó:

Luận án có 2 đóng góp chính:

- Một là, luận án đề xuất một framework gồm 2 mô hình: mô hình kết hợp AI-LGMB (Auto Immuse Light Gradient Boosting Machine) và mô hình PSO-SCNN (Particle Swarm Optimization- Spatial Convolutional Neural networks). PSO-SCNN tích hợp các đặc trưng nhúng không gian (spatial embeddings), Haversine encoding, multi-head attention, và phép nhân chập (convolution) để học các đặc trưng về vị trí không gian (geospatial dependencies).

- Hai là, tiến hành thử nghiệm, so sánh, đánh giá 2 mô hình đề xuất AI-LGMB và PSO-SCNN với một số mô hình học máy truyền thống. Đồng thời thử nghiệm ANOVA với 2 tập Vietnam Mekong Delta và India Odisha với 5 mức độ chất lượng nước ngầm (Excellent, Good, Moderate, poor, unsuitable).

4. Những ưu điểm và thiếu s ót, những điểm cần được bổ sung và sửa chữa

Ưu điểm (strong points): các đóng góp chính được trình bầy tương đối rõ ràng trong chương 2 và 3. Các thử nghiệm được trình bầy tương đối rõ ràng, chi tiết.

Nhược điểm (weak points): các mô hình đề xuất nên được trình bầy chi tiết hơn: kiến trúc, các lớp, đầu vào, đầu ra, quá trình huấn luyện, v.v.. các mô hình cơ sở (baseline) khá cũ (outdated) và chưa tận dụng được những thành quả của học máy hiện đại để nâng cao hiệu quả phân loại. Nhiều đoạn viết lặp lại (i.e. hyperparameter optimization and tuning) và khó hiểu.

Một số ý kiến chỉnh sửa, bổ sung luận án như sau:

Nhiều hình vẽ cần bổ sung các giải thích, ví dụ hình 2.1, 2.2 (several figures need to be well explained, eg. 2.1, 2.2)

Nên bổ sung kiến trúc các mô hình đề xuất AI-LGMB và PSO-SCNN (add the network architectures of AI-LGMB and PSO-SCNN)

Nên coi tinh chỉnh mô hình là một bước trong quá trình xây dựng mô hình. Mô hình đề xuất phải là mô hình sau khi đã tinh chỉnh (the proposed model should have been the fine-tuned version).

Giải thích tại sao Transformer lại cho kết quả thấp nhất trong bảng 3.29 (trang 105). Show that why Transformer's performance is the lowest in the table 3.29.

Bổ sung phân tích chi tiết hơn các kết quả đánh giá tại các bảng 3.6

5. Đánh giá về sự trùng lặp của luận án so với các đồ án, luận văn hay công trình khoa học đã công bố trong và ngoài nước

Theo hiểu biết của người nhận xét, kết quả của luận án không trùng lặp với các luận án, luận văn hay công trình khoa học đã công bố trong và ngoài nước.

6. Nhận xét về chất lượng các bài báo khoa học đã được công bố và khẳng định các bài báo đó chứa đựng nội dung chủ yếu của luận án

Các kết quả của luận án đã được công bố trong 7 bài báo. Các bài báo phản ánh các nội dung nghiên cứu của luận án. Do đó, luận án đáp ứng yêu cầu về công bố khoa học.

7. Tính trung thực trong việc trích dẫn các công trình đã được NCS công bố trong và ngoài nước, tài liệu tham khảo

Các kết quả nghiên cứu của luận án đã được trích dẫn trung thực trong các công trình công bố. Trích dẫn tài liệu tham khảo được thực hiện một cách rõ ràng, đầy đủ và cập nhật.

8. Kết luận

- Luận án đáp ứng được các yêu cầu về nội dung và hình thức đối với một luận án tiến sĩ, ngành Hệ thống thông tin.

- Đồng ý cho NCS được bảo vệ tại Hội đồng chấm luận án cấp Học viện.

Hà nội, ngày tháng năm

Người nhận xét

PGS. TS. Phạm Văn Cường

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BẢN NHẬN XÉT LUẬN ÁN TIẾN SĨ
(Cấp Học viện)

Tên nghiên cứu sinh: Michael Omar
Đề tài: Phát triển phương pháp học máy kết hợp thông tin không gian cho bài toán phân loại chất lượng nước ngầm
(An approach of ensemble spatial machine learning for groundwater drinkability classification)
Ngành: Hệ thống thông tin
Mã số: 9.48.01.04
Người nhận xét: Nguyễn Hà Nam
Cơ quan công tác của người nhận xét: Ban KH&ĐMST, ĐHQGHN

NỘI DUNG NHẬN XÉT

1. Ý nghĩa khoa học và thực tiễn của đề tài luận án

Luận án lựa chọn hướng nghiên cứu *"phát triển phương pháp học máy kết hợp thông tin không gian cho bài toán phân loại chất lượng nước ngầm"*, là một hướng nghiên cứu có ý nghĩa khoa học và giá trị thực tiễn cao. Về mặt khoa học, luận án góp phần giải quyết một hạn chế phổ biến trong các mô hình học máy truyền thống khi áp dụng cho dữ liệu địa lý – đó là hiện tượng *spatial blindness* và sai lệch do xác thực ngẫu nhiên không phù hợp với dữ liệu không gian.

Về mặt thực tiễn, kết quả nghiên cứu có khả năng hỗ trợ hiệu quả cho công tác giám sát, đánh giá và cảnh báo chất lượng nước ngầm tại các khu vực rộng lớn, đặc biệt là ở các quốc gia đang phát triển như Việt Nam và Ấn Độ. Tuy nhiên, ý nghĩa ứng dụng sẽ thuyết phục hơn nếu luận án làm rõ hơn mối liên hệ giữa kết quả phân loại và các chuẩn/quy định quản lý nước ngầm cụ thể trong thực tế. Vì vậy, luận án có ý nghĩa khoa học, giải quyết một vấn đề khó trong thực tiễn.

2. Sự hợp lý và độ tin cậy của các phương pháp nghiên cứu

Luận án sử dụng tổ hợp các phương pháp hiện đại, bao gồm: học máy (LightGBM), học sâu (SCNN), tối ưu siêu tham số (PSO, Optuna), kỹ thuật xử lý mất cân bằng dữ liệu, xác thực không gian (spatial validation), phân tích giải thích mô hình (SHAP) và kiểm định thống kê (ANOVA). Nhìn chung, phương pháp nghiên cứu là hợp lý, có cơ sở khoa học và phù hợp với bài toán đặt ra.

Điểm đáng ghi nhận là tác giả đã ý thức và xử lý đúng đắn vấn đề xác thực mô hình trong bối cảnh dữ liệu không gian.

Tuy nhiên, luận án sử dụng thuật ngữ *"ensemble spatial machine learning"* nhưng chưa làm rõ bản chất của chiến lược ensemble (stacking, voting hay chỉ là so sánh song song các mô hình), điều này làm giảm mức độ chặt chẽ về mặt phương pháp luận.

3. Đánh giá các kết quả đạt được, nêu những đóng góp mới và giá trị của các đóng góp đó

Luận án đạt được nhiều kết quả thực nghiệm phong phú với hiệu năng cao, thể hiện qua các chỉ số Accuracy, F1-score, AUC, cùng với các phân tích ablation và kiểm định thống kê. Các đóng góp chính có thể ghi nhận gồm:

- Đề xuất các mô hình học máy và học sâu có xét đến yếu tố không gian cho bài toán phân loại chất lượng nước ngầm.
- Tích hợp quy trình tối ưu siêu tham số và xác thực không gian nhằm nâng cao độ tin cậy của kết quả.
- Phân tích giải thích mô hình và trực quan hóa kết quả trên nền GIS.

Các đóng góp này có giá trị khoa học nhất định, đặc biệt ở khía cạnh tích hợp và triển khai hệ thống. Tuy nhiên, mức độ đột phá phương pháp còn hạn chế; luận án thiên về cải tiến, tối ưu và kết hợp các kỹ thuật đã có hơn là đề xuất một mô hình hay lý thuyết hoàn toàn mới.

4. Những ưu điểm và thiếu sót, những điểm cần được bổ sung và sửa chữa

Ưu điểm:

- Bài toán nghiên cứu rõ ràng, có dữ liệu thực tế lớn từ hai quốc gia.
- Thực nghiệm đầy đủ, đa dạng, có so sánh với nhiều mô hình truyền thống.
- Có ý thức về tính giải thích và độ tin cậy thống kê của kết quả.

Thiếu sót và điểm cần bổ sung:

- Chưa nhất quán trong việc định nghĩa bài toán phân loại (đa lớp theo WQI và nhị phân "is-drinkable").
- Chưa làm rõ đầy đủ cơ chế ensemble như đã nêu trong tên luận án.
- Phần tái lập nghiên cứu (reproducibility) cần mô tả chi tiết hơn về cấu hình, tham số và khả năng chia sẻ dữ liệu/mã nguồn.
- Cần thảo luận sâu hơn về chi phí tính toán và khả năng triển khai thực tế.

5. Đánh giá về sự trùng lặp của luận án so với các đồ án, luận văn hay công trình khoa học đã công bố trong và ngoài nước

Luận án không có dấu hiệu sao chép hay trùng lặp không phù hợp với các công trình đã công bố. Các nghiên cứu liên quan trong và ngoài nước được tổng hợp tương đối đầy đủ. Nội dung luận án thể hiện sự kế thừa và phát triển hợp lý từ các nghiên cứu trước, đặc biệt là trong lĩnh vực học máy, ứng dụng cho đánh giá chất lượng nước. Mức độ trùng lặp nằm trong giới hạn cho phép và phù hợp với thông lệ học thuật.

6. Nhận xét về chất lượng các bài báo khoa học đã được công bố và khẳng định các bài báo đó chứa đựng nội dung chủ yếu của luận án

Tác giả đã có 06 công trình được công bố trên các tạp chí và hội nghị chuyên ngành có uy tín (đặc biệt là Earth Science Informatics). Nội dung các công bố phù hợp và phản ánh trực tiếp các kết quả chính của luận án, đặc biệt là các phần liên quan đến mô hình AI-LGBM, dự báo chất lượng nước và tích hợp GIS. Tuy nhiên, luận án cần làm rõ hơn mối liên hệ giữa từng bài báo với các chương/mục cụ thể trong luận án để tăng tính nhất quán và thuyết phục. (CT7 đang ở vòng phản biện nên không tính)

	Nội dung khoa học chính của chương	Vai trò của công trình
C1	Bối cảnh nghiên cứu; dữ liệu; chỉ số chất lượng nước; tổng quan công trình liên quan	CT1(dữ liệu & bối cảnh), CT6: gián tiếp (so sánh, định vị), CT3 (tối ưu),
C2	Phương pháp ML/DL; tối ưu siêu tham số; tích hợp yếu tố không gian	CT4 (AI-LGBM), CT7 (PSO-SCNN)
C3	Thực nghiệm; so sánh mô hình; phân tích không gian; thảo luận	CT2–CT7: tái trình bày & mở rộng kết quả;

7. Tính trung thực trong việc trích dẫn các công trình đã được NCS công bố trong và ngoài nước, tài liệu tham khảo

Theo hiểu biết của người đọc, LA không trùng lặp với các nghiên cứu trước đây. Các tài liệu tham khảo được trích dẫn đầy đủ, phù hợp với nội dung nghiên cứu. Không phát hiện dấu hiệu sử dụng tài liệu không rõ nguồn gốc hay trích dẫn sai lệch. Tuy nhiên, cần rà soát lại sự nhất quán giữa danh mục tài liệu tham khảo và các trích dẫn trong nội dung, đặc biệt là đối với các công trình do chính NCS đồng tác giả.

8. Kết luận

Tổng thể, luận án cơ bản đáp ứng yêu cầu của một luận án tiến sĩ ngành Hệ thống thông tin. Mặc dù còn tồn tại một số hạn chế cần chỉnh sửa và làm rõ, các vấn đề này không làm ảnh hưởng đến bản chất khoa học của luận án. Sau khi NCS tiếp thu, chỉnh sửa theo các góp ý nêu trên, luận án đủ điều kiện để được chấp nhận và bảo vệ trước Hội đồng chấm luận án tiến sĩ.

Hà nội, ngày 15 tháng 1 năm 2026

Người nhận xét

Nguyễn Hà Nam

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

BẢN NHẬN XÉT LUẬN ÁN TIẾN SĨ
(Bảo vệ cấp Học viện)

Tên đề tài luận án: Nghiên cứu phát triển phương pháp học máy kết hợp thông tin không gian cho bài toán phân loại chất lượng nước ngầm/ An approach of ensemble spatial machine learning for groundwater drinkability classification

Chuyên ngành: Hệ thống thông tin Mã số: 9 48 01 04

Nghiên cứu sinh: **Michael Omar**

Người hướng dẫn 1: PGS.TS. Trần Thị Ngân

Người hướng dẫn 2: PGS.TS. Nguyễn Long Giang

Người nhận xét: **TS. Trần Đức Nghĩa**

Đơn vị công tác: Viện Công nghệ thông tin, Viện Hàn lâm KH & CN Việt Nam

Chức trách trong hội đồng: Ủy viên thư ký

I. NỘI DUNG NHẬN XÉT

1. Tính cấp thiết, thời sự ý nghĩa khoa học của đề tài luận án:

Luận án tập trung vào vấn đề phân loại chất lượng nước ngầm tại hai khu vực nghiên cứu: Đồng bằng sông Cửu Long, Việt Nam và Odisha, Ấn Độ. Vấn đề này có ý nghĩa ứng dụng lớn, đặc biệt trong bối cảnh các nước châu Á đang phải đối mặt với biến đổi khí hậu dẫn đến xâm nhập mặn và tình trạng thiếu nước ngọt lan rộng. Do đó, LA có ý nghĩa khoa học và thực tiễn.

2. Sự phù hợp của đề tài luận án với chuyên ngành đào tạo:

Đề tài luận án có tính khoa học, phù hợp với chuyên ngành đào tạo tiến sĩ Hệ thống thông tin.

3. Sự trùng lặp của đề tài so với công trình khoa học đã công bố:

Nội dung luận án không trùng lặp với các luận án đã bảo vệ và các kết quả nghiên cứu đã công bố trong và ngoài nước.

4. Sự phù hợp của các phương pháp nghiên cứu, độ tin cậy của các kết quả đã đạt được:

Các lập luận, chứng minh và các kết quả đạt được là đáng tin cậy.

5. Những đóng góp mới của đề tài

Luận án có ý nghĩa khoa học và thực tiễn, những đóng góp mới của luận án cụ thể như sau:

(1) Phát triển phương pháp lai AI-LGBM bao gồm các bước tiền xử lý trích xuất đặc trưng, xây dựng mô hình huấn luyện bằng LGBM, và kiểm tra và diễn giải kết quả bằng LIME/SHAP.

(2) Phát triển phương pháp PSO-SCNN sử dụng hàm khoảng cách Haversine trong các lớp nhúng trước khi đưa vào SCNN (với các hệ số không gian bổ sung). Tối ưu hóa các siêu tham số bằng PSO.

6. Về các công trình khoa học đã công bố của nghiên cứu sinh liên quan đến nội dung của luận án

Các kết quả của luận án đã được công bố tại các hội thảo/tạp chí chuyên ngành có uy tín, tuy nhiên NCS không phải tác giả chính của các công bố SCIE. Các kết quả công bố có nội dung khoa học và là kết quả chính của luận án.

7. Tính trung thực, minh bạch trong trích dẫn tài liệu.

Đảm bảo.

8. Góp ý các thiếu sót về hình thức, nội dung của luận án mà nghiên cứu sinh cần chỉnh sửa, bổ sung.

Ưu điểm:

Luận án có ý nghĩa lý thuyết và ứng dụng thực tiễn, phương pháp nghiên cứu đúng đắn.

Nhược điểm:

Nên cải thiện việc viết Luận án để phù hợp hơn với ngành Hệ thống thông tin (giảm bớt thông tin về thủy văn/môi trường, tăng thông tin về hệ thống).

Cần rà soát và khắc phục lỗi chính tả cùng những lỗi soạn thảo khác.

Cần giải thích rõ hơn tại sao phải sử dụng học máy cho bài toán này?

II. KẾT LUẬN

Đánh giá về mức độ đạt yêu cầu của luận án:

- Luận án cơ bản đáp ứng được các yêu cầu về nội dung và hình thức đối với một luận án tiến sĩ, ngành Hệ thống thông tin.

- Đồng ý cho NCS được bảo vệ tại Hội đồng chấm luận án cấp Học viện.

Hà Nội, ngày tháng 01 năm 2026

Người nhận xét

TS. Trần Đức Nghĩa

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

BẢN NHẬN XÉT LUẬN ÁN TIẾN SĨ
(Bảo vệ cấp Học viện)

Tên đề tài luận án: Nghiên cứu phát triển phương pháp học máy kết hợp thông tin không gian cho bài toán phân loại chất lượng nước ngầm/ An approach of ensemble spatial machine learning for groundwater drinkability classification

Chuyên ngành: Hệ thống thông tin Mã số: 9 48 01 04

Nghiên cứu sinh: **Michael Omar**

Người hướng dẫn 1: PGS.TS. Trần Thị Ngân

Người hướng dẫn 2: PGS.TS. Nguyễn Long Giang

Người nhận xét: **TS. Trần Đức Nghĩa**

Đơn vị công tác: Viện Công nghệ thông tin, Viện Hàn lâm KH & CN Việt Nam

Chức trách trong hội đồng: Ủy viên thư ký

I. NỘI DUNG NHẬN XÉT

1. Tính cấp thiết, thời sự ý nghĩa khoa học của đề tài luận án:

Luận án tập trung vào vấn đề phân loại chất lượng nước ngầm tại hai khu vực nghiên cứu: Đồng bằng sông Cửu Long, Việt Nam và Odisha, Ấn Độ. Vấn đề này có ý nghĩa ứng dụng lớn, đặc biệt trong bối cảnh các nước châu Á đang phải đối mặt với biến đổi khí hậu dẫn đến xâm nhập mặn và tình trạng thiếu nước ngọt lan rộng. Do đó, LA có ý nghĩa khoa học và thực tiễn.

2. Sự phù hợp của đề tài luận án với chuyên ngành đào tạo:

Đề tài luận án có tính khoa học, phù hợp với chuyên ngành đào tạo tiến sĩ Hệ thống thông tin.

3. Sự trùng lặp của đề tài so với công trình khoa học đã công bố:

Nội dung luận án không trùng lặp với các luận án đã bảo vệ và các kết quả nghiên cứu đã công bố trong và ngoài nước.

4. Sự phù hợp của các phương pháp nghiên cứu, độ tin cậy của các kết quả đã đạt được:

Các lập luận, chứng minh và các kết quả đạt được là đáng tin cậy.

5. Những đóng góp mới của đề tài

Luận án có ý nghĩa khoa học và thực tiễn, những đóng góp mới của luận án cụ thể như sau:

(1) Phát triển phương pháp lai AI-LGBM bao gồm các bước tiền xử lý trích xuất đặc trưng, xây dựng mô hình huấn luyện bằng LGBM, và kiểm tra và diễn giải kết quả bằng LIME/SHAP.

(2) Phát triển phương pháp PSO-SCNN sử dụng hàm khoảng cách Haversine trong các lớp nhúng trước khi đưa vào SCNN (với các hệ số không gian bổ sung). Tối ưu hóa các siêu tham số bằng PSO.

6. Về các công trình khoa học đã công bố của nghiên cứu sinh liên quan đến nội dung của luận án

Các kết quả của luận án đã được công bố tại các hội thảo/tạp chí chuyên ngành có uy tín, tuy nhiên NCS không phải tác giả chính của các công bố SCIE. Các kết quả công bố có nội dung khoa học và là kết quả chính của luận án.

7. Tính trung thực, minh bạch trong trích dẫn tài liệu.

Đảm bảo.

8. Góp ý các thiếu sót về hình thức, nội dung của luận án mà nghiên cứu sinh cần chỉnh sửa, bổ sung.

Ưu điểm:

Luận án có ý nghĩa lý thuyết và ứng dụng thực tiễn, phương pháp nghiên cứu đúng đắn.

Nhược điểm:

Nên cải thiện việc viết Luận án để phù hợp hơn với ngành Hệ thống thông tin (giảm bớt thông tin về thủy văn/môi trường, tăng thông tin về hệ thống).

Cần rà soát và khắc phục lỗi chính tả cùng những lỗi soạn thảo khác.

Cần giải thích rõ hơn tại sao phải sử dụng học máy cho bài toán này?

II. KẾT LUẬN

Đánh giá về mức độ đạt yêu cầu của luận án:

- Luận án cơ bản đáp ứng được các yêu cầu về nội dung và hình thức đối với một luận án tiến sĩ, ngành Hệ thống thông tin.

- Đồng ý cho NCS được bảo vệ tại Hội đồng chấm luận án cấp Học viện.

Hà Nội, ngày 10 tháng 01 năm 2026

Người nhận xét

TS. Trần Đức Nghĩa

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BẢN NHẬN XÉT LUẬN ÁN TIẾN SĨ
(Cấp Học viện)

Tên nghiên cứu sinh: Michael Omar

Đề tài: An approach of ensemble spatial machine learning for groundwater drinkability classification

Chuyên ngành: Hệ thống thông tin

Mã số: 9 48 01 04

Người nhận xét: PGS.TS. Lê Hoàng Sơn, Ủy viên

Cơ quan công tác của người nhận xét: Viện Công nghệ Thông tin, ĐHQGHN.

NỘI DUNG NHẬN XÉT

1. Ý nghĩa khoa học và thực tiễn của đề tài luận án

Đảm bảo an toàn nguồn nước là một thách thức toàn cầu quan trọng cho sức khỏe cộng đồng, tính bền vững môi trường và phát triển kinh tế. Nhu cầu này càng trở nên cấp thiết hơn do dân số toàn cầu ngày càng tăng, làm gia tăng áp lực lên các nguồn nước hữu hạn. Hiện vẫn còn hai tỷ người thiếu tiếp cận với nước uống được quản lý an toàn, khiến việc thúc đẩy các phương pháp đánh giá chất lượng nước mạnh mẽ trở thành một yêu cầu cấp thiết về sức khỏe toàn cầu. Bài toán phân loại chất lượng nước ngầm nhằm xác định liệu mẫu nước có đạt tiêu chuẩn sử dụng cho ăn uống hay không đạt dựa trên các chỉ tiêu hóa – lý – sinh học (pH, EC, TDS, Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, Cl^{-}, SO_4^{2-}, HCO_3^{-},…). Các tiếp cận truyền thống thường sử dụng ngưỡng tiêu chuẩn dạng luật nghĩa là nếu bất kỳ chỉ số nào vượt ngưỡng Water Quality Index (WQI) của WHO coi như là không uống được. Hạn chế của tiếp cận này là không xử lý tốt các tương tác phi tuyến giữa các yếu tố và khó mở rộng cho dữ liệu lớn. Do vậy, tiếp cận học máy kết hợp dữ liệu địa lý để phân loại vùng rủi ro và giữa WQI + ML/DL hay được sử dụng gần đây để áp dụng mô hình huấn luyện ở khu vực này cho khu vực khác (transfer learning) một cách hiệu quả với chi phí về dữ liệu quan trắc thu thập thủ công thấp, bảo đảm tính không gian – thời gian, và các vấn đề về mẫu đạt chuẩn hơn không đạt chuẩn (hoặc ngược lại), gây bias cho mô hình. Luận án phát triển, xác thực và triển khai một mô hình học máy không gian tích hợp mới để phân loại khả năng uống được của nước ngầm, với trọng tâm chính là các nghiên cứu điển hình tại Đồng bằng sông Cửu Long và tại Odisha, Ấn Độ là một vấn đề có ý nghĩa khoa học và thực tiễn.

2. Sự hợp lý và độ tin cậy của các phương pháp nghiên cứu

Luận án sử dụng phương pháp nghiên cứu tổng luận các kỹ thuật học máy/học sâu cho bài toán phân loại chất lượng nước ngầm, từ đó đưa ra mô hình phân loại dựa trên Ensemble Learning kết hợp yếu tố không gian (AI-LGBM, PSO-SCNN) và đánh giá trên các metric tiêu chuẩn của bài toán phân loại. Về cơ bản, tiếp cận nghiên cứu là phù hợp với nội dung chủ đề luận án.

3. Đánh giá về sự trùng lặp của luận án so với các đồ án, luận văn hay công trình khoa học đã công bố trong và ngoài nước

Nội dung đóng góp của luận án về giảm chiều thuộc tính là không trùng lặp với các luận án đã bảo vệ trước đây theo cùng nhánh này.

4. Tính trung thực trong việc trích dẫn các công trình đã được NCS công bố trong và ngoài nước, tài liệu tham khảo

Luận án trích dẫn 126 tài liệu tham khảo trong 146 trang, về cơ bản được trích dẫn đầy đủ trong luận án.

5. Đánh giá các kết quả đạt được, nêu những đóng góp mới và giá trị của các đóng góp đó

Luận án có 02 đóng góp chính:

(a) Phát triển mô hình phân loại lai AI-LGBM bao gồm các bước tiền xử lý trích chọn đặc trưng, xây dựng mô hình huấn luyện sử dụng LGBM, và thử nghiệm và lý giải kết quả với LIME/SHAP (Hình 2.2, trang 50). Các kết quả này được mô tả trong Chương 2 và công bố chính trong [CT1] trên Earth Science Informatics và [CT4] tại EAI GOODTECHS 2024.

(b) Phát triển phương pháp PSO-SCNN (Hình 2.3, trang 64) trong đó sử dụng hàm khoảng cách Haversine trong các Embedding Layer trước khi đưa vào SCNN (có tính thêm yếu tố không gian). Tối ưu siêu tham số sử dụng PSO. Các kết quả này được mô tả trong Chương 2 và công bố chính trong [CT6, CT7] tại ICIIT 2025 và JIRS (Under Review).

(c) Các đóng góp này đã được kiểm chứng đa vùng tại Đồng bằng sông Cửu Long và Odisha của Ấn Độ chứng minh khả năng khái quát hóa trên các vùng có điều kiện thủy văn địa chất khác nhau. Ngoài ra luận án cũng giới thiệu bản đồ rủi ro không gian và các điểm nóng ô nhiễm, hữu ích cho các nhà hoạch định chính sách và quản lý nước ngầm. Khả năng giải thích của mô hình được chứng minh qua việc sử dụng các mô hình SHAP và LIME nhằm tăng cường tính minh bạch của mô hình lai đề xuất, giúp người đọc dễ dàng hiểu được các yếu tố ảnh hưởng đến dự đoán chất lượng nước ngầm.

6. Nhận xét về chất lượng các bài báo khoa học đã được công bố và khẳng định các bài báo đó chứa đựng nội dung chủ yếu của luận án

Luận án công bố 06 công trình và 01 bài under review, trong đó có 01 bài tạp chí SCIE [CT1] và 04 bài hội thảo trong nước và quốc tế, đáp ứng điều kiện bảo vệ theo quy chế.

7. Những ưu điểm và thiếu sót, những điểm cần được bổ sung và sửa chữa

Để luận án trong sáng hơn, người viết nhận xét kiến nghị một số điều chỉnh sau:

(1) Kiến nghị bổ sung phân tích về độ phức tạp tính toán của các đề xuất (AI-LGBM và PSO-SCNN) và bổ sung hướng phát triển về tối ưu hóa hiệu quả tính toán thông qua việc sử dụng các kiến trúc DL hạng nhẹ hoặc các kỹ thuật tối ưu hóa giúp dễ tiếp cận hơn trong môi trường có nguồn lực hạn chế.

(2) Kiến nghị bàn luận thêm về việc sử dụng các nguồn dữ liệu bổ sung trong các mô hình đa mô thức như tích hợp dữ liệu viễn thám, ảnh vệ tinh và các nguồn phi truyền thống khác để cải thiện độ chính xác của mô hình, đặc biệt là ở những khu vực có dữ liệu thưa thớt. Lưu ý các tiếp cận tích hợp dữ liệu không-thời gian mở rộng (4D, 5D,..) để mở rộng mô hình nhằm theo dõi sự thay đổi chất lượng nước ngầm theo thời gian trong các nhiệm vụ giám sát và dự báo dài hạn. Mở rộng phạm vi rộng hơn các khu vực với điều kiện thủy văn và môi trường đa dạng để đánh giá tính ổn định và khả năng thích ứng của chúng trong các bối cảnh khác nhau.

(3) Đề xuất xây dựng bộ dữ liệu tiêu chuẩn cho phân loại và dự báo chất lượng nước ngầm, bao gồm cả dữ liệu bản đồ trên các nền tảng mở, giúp cộng đồng Tin – Địa mở rộng và áp dụng được trên phạm vi rộng hơn từ các bối cảnh địa lý và thủy văn đa dạng, từ đó hỗ trợ nâng cao tác động quản lý và chính sách nước ngầm.

Dựa trên các đánh giá trên, người viết nhận xét đồng ý cho nghiên cứu sinh được trình bày luận án tại Hội đồng đánh giá luận án cấp Học viện và nhận bằng Tiến sĩ./

Hà nội, ngày 10 tháng 1 năm 2026

Xác nhận của cơ quan công tác

TL. VIỆN TRƯỞNG

TRƯỞNG PHÒNG

KHOA HỌC CÔNG NGHỆ - ĐÀO TẠO

VIỆN CÔNG NGHỆ THÔNG TIN

Dương Quang Khánh

Người nhận xét

PGS.TS. Lê Hoàng Sơn

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

Độc lập - Tự do - Hạnh phúc

BẢN NHẬN XÉT LUẬN ÁN TIẾN SĨ CẤP HỌC VIỆN

Tên đề tài: *An Approach of Ensemble Spatial Machine Learning for Groundwater Drinkability Classification*

Nghiên cứu phát triển phương pháp học máy kết hợp thông tin không gian cho bài toán phân loại chất lượng nước ngầm

Ngành: Hệ thống thông tin

Mã số: 9 48 01 04

Nghiên cứu sinh: Michael Omar

Người hướng dẫn:

PGS.TS. Trần Thị Ngân

PGS.TS. Nguyễn Long Giang

Người nhận xét: PGS TS Bùi Thu Lâm

Cơ quan công tác: Học viện Kỹ thuật Mật mã

NỘI DUNG NHẬN XÉT

1. Tính cần thiết, thời sự, ý nghĩa khoa học và thực tiễn của đề tài luận án

Luận án lựa chọn bài toán phân loại khả năng sử dụng nước ngầm cho mục đích ăn uống dựa trên dữ liệu thủy hóa và thông tin không gian. Đây là một vấn đề có tính thời sự cao, gắn trực tiếp với các thách thức toàn cầu như an ninh nguồn nước, biến đổi khí hậu, gia tăng ô nhiễm môi trường và sức khỏe cộng đồng.

Về ý nghĩa khoa học, luận án tiếp cận bài toán từ góc nhìn của học máy không gian (spatial machine learning), nhấn mạnh hạn chế "spatial blindness" của các mô hình ML/DL truyền thống. Việc tích hợp thông tin tọa độ, mô hình CNN không gian và tối ưu hóa PSO cho thấy nỗ lực giải quyết một vấn đề có cơ sở khoa học rõ ràng.

Về ý nghĩa thực tiễn, kết quả nghiên cứu có tiềm năng ứng dụng trong:

- Cảnh báo sớm khu vực nước ngầm không đạt chuẩn;
- Hỗ trợ ra quyết định cho cơ quan quản lý tài nguyên nước;
- Trực quan hóa rủi ro ô nhiễm trên bản đồ GIS.

Đề tài là cần thiết, có ý nghĩa khoa học và thực tiễn rõ ràng, phù hợp với định hướng nghiên cứu liên ngành giữa Hệ thống thông tin – AI – khoa học môi trường.

2. Sự không trùng lặp của đề tài nghiên cứu; tính trung thực, rõ ràng và đầy đủ trong trích dẫn tài liệu

Luận án đã khảo sát tương đối đầy đủ các nhóm công trình:

- Phương pháp truyền thống;
- ML/DL cho đánh giá chất lượng nước;
- Các mô hình lai ghép không gian (GIS + ML, CNN không gian, PSO…).

Các mô hình đề xuất AI-LGBM và PSO-SCNN được trình bày như những cải tiến dựa trên nền tảng đã có, không sao chép nguyên xi các mô hình trước đó. Về tổng thể, không phát hiện dấu hiệu trùng lặp luận án đã công bố.

Tài liệu tham khảo được trích dẫn tương đối đầy đủ, đúng chuẩn học thuật.

3. Sự phù hợp giữa tên đề tài với nội dung, giữa nội dung với ngành và mã số

Tên đề tài phản ánh khá chính xác nội dung nghiên cứu:

- "Ensemble Spatial Machine Learning" → phù hợp với AI-LGBM và PSO-SCNN;
- "Groundwater Drinkability Classification" → nhất quán với bài toán phân loại.

Luận án phù hợp với ngành Hệ thống thông tin, đặc biệt ở các khía cạnh:

- Xây dựng mô hình xử lý dữ liệu lớn;
- Tích hợp AI, GIS và trực quan hóa;
- Hỗ trợ ra quyết định.

4. Độ tin cậy và tính hiện đại của phương pháp nghiên cứu

Luận án sử dụng nhiều phương pháp hiện đại:

- LightGBM, CNN, PSO;
- SHAP cho giải thích mô hình;
- Đánh giá bằng nhiều chỉ số (Accuracy, F1, AUC…);
- So sánh với nhiều mô hình nền.

Các phương pháp này phù hợp và có độ tin cậy khoa học.

5. Kết quả nghiên cứu mới của tác giả

Các kết quả nổi bật gồm:

- Đề xuất mô hình PSO-SCNN kết hợp học sâu không gian và tối ưu hóa bầy đàn;
- Cải thiện độ chính xác so với các mô hình nền;
- Phân tích tầm quan trọng đặc trưng bằng SHAP;
- Trực quan hóa bản đồ rủi ro.

Đóng góp có giá trị, tuy nhiên mức độ "mới" mang tính kỹ thuật tích hợp nhiều hơn là đột phá lý thuyết

6. Ưu điểm và nhược điểm của luận án

Ưu điểm:

- Bài toán thực tiễn, dữ liệu thật;
- Phương pháp hiện đại, kết quả thực nghiệm phong phú;
- Hình ảnh, bảng biểu nhiều và rõ ràng;
- Có phân tích thất bại (failure case), ablation study.

Nhược điểm:

- Tỷ trọng kiến thức về thủy văn – môi trường khá lớn, trong khi phần kiến trúc hệ thống thông tin, triển khai hệ thống, hoặc luồng xử lý dữ liệu trong môi trường thực tế còn mờ nhạt. Điều này khiến luận án có xu hướng nghiêng về *AI ứng dụng* hơn là *Hệ thống thông tin* theo nghĩa hẹp.

- Luận án (đặc biệt Chương 3) quá thiên về kết quả thực nghiệm, thiếu thảo luận học thuật sâu; chưa làm rõ tính tổng quát hóa (generalization) khi áp dụng sang các khu vực có điều kiện địa chất khác.

- Phát biểu về "fundamental flaw" ở trang 3 chưa thực sự chính xác. Khoogn có flaw gì ở đây cả, chỉ là mô hình chưa mở rộng sang hướng đó mà thôi.

- Trang 6: các đóng góp (4 đóng góp) được nêu khá lan man, nên tập trung vào đóng góp học thuật.

- Việc mô tả các nhãn của của bài toán phân loại phải thống nhất từ đầu đến cuối luận ná. Hiện mô tả hơi lộn xộn

- trang 10: thống nhất dùng kí hiệu w hay theta

- Không nên đưa các code snippnet vào luận án (trang 35), đưa vào dạng thuật toán/giả mã thì tốt hơn

- Chưa làm rõ ranh giới đóng góp mới của AI-LGBM so với các biến thể LightGBM + tối ưu hóa siêu tham số đã rất phổ biến trong văn liệu.

- Một số đoạn mô tả "state-of-the-art" còn mang tính tổng hợp – liệt kê, thiếu phân tích phê phán sâu để chỉ ra chính xác khoảng trống khoa học mà luận án giải quyết.

- Việc kiểm định không gian (spatial validation) chưa được mô tả đủ chặt chẽ. Việc sử dụng k-fold truyền thống trên dữ liệu không gian có thể dẫn đến đánh giá quá lạc quan.

- Độ nhạy của mô hình với sự thay đổi vùng địa lý chưa được phân tích sâu.

- PSO-SCNN có độ phức tạp cao, nhưng chưa có phân tích rõ ràng về chi phí tính toán – khả năng triển khai thực tế trong môi trường hạn chế tài nguyên.

- Hình 2.3 (b): chưa thấy khối 4 (component No4) ở đâu

- Nhiều số liệu nên được trình bày dạng biểu đồ để dễ hình dùng hơn.

- Bảng 3.13: mememory comsumption của AI-LGBM = 0 nghĩa là sao?

7. Công bố khoa học liên quan đến luận án

Luận án có công bố trên các tạp chí và kỷ yếu hội nghị khoa học (theo danh mục đính kèm). Các công bố nhìn chung phù hợp hướng nghiên cứu.

Tuy nhiên:

- Cần làm rõ vai trò tác giả chính của NCS trong từng công bố;
- Chưa thấy công bố ở các tạp chí top-tier trong lĩnh vực ML hoặc hydroinformatics.

8. Kết luận chung

Luận án đã:

- Đáp ứng cơ bản các yêu cầu của một luận án tiến sĩ;
- Có đóng góp khoa học và giá trị thực tiễn;
- Thể hiện năng lực nghiên cứu độc lập của nghiên cứu sinh.

Mặc dù còn tồn tại một số hạn chế, nhưng các hạn chế này không làm thay đổi bản chất khoa học của luận án và có thể được khắc phục thông qua chỉnh sửa, bổ sung.

Luận án đủ điều kiện đưa ra bảo vệ cấp Học viện để xét công nhận học vị Tiến sĩ.

Hà Nội, ngày 11 tháng 01 năm 2026

Người nhận xét

PGS.TS. Bùi Thu Lâm

VIỆN HÀN LÂM
KHOA HỌC VÀ CÔNG NGHỆ VN
HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

Hà Nội, ngày 27 tháng 01 năm 2026

BIÊN BẢN CỦA
HỘI ĐỒNG ĐÁNH GIÁ LUẬN ÁN TIẾN SĨ CẤP HỌC VIỆN

Căn cứ quyết định số 1369/QĐ-HVKHCN ngày 15 tháng 12 năm 2025 của Giám đốc Học viện Khoa học và Công nghệ về việc thành lập Hội đồng đánh giá luận án tiến sĩ cấp Học viện, Hội đồng đã họp vào hồi 14 giờ ngày 27 tháng 01 năm 2026 tại Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam, số 18 đường Hoàng Quốc Việt, Cầu Giấy, Hà Nội để đánh giá luận án tiến sĩ.

Họ và tên NCS:	**Michael Omar**
Tên đề tài luận án:	Nghiên cứu phát triển phương pháp học máy kết hợp thông tin không gian cho bài toán phân loại chất lượng nước ngầm/ An approach of ensemble spatial machine learning for groundwater drinkability classification
Ngành:	Hệ thống thông tin
Mã số:	9 48 01 04
Người hướng dẫn:	PGS.TS. Trần Thị Ngân, PGS.TS. Nguyễn Long Giang

THAM DỰ BUỔI BẢO VỆ GỒM CÓ

- Đại diện cơ sở đào tạo:
  1. GS.TS. Vũ Đình Lãm – Giám đốc Học viện KH&CN
- Đại diện Viện Công nghệ thông tin:
  1. PGS.TS. Nguyễn Trường Thắng - Viện trưởng Viện Công nghệ thông tin
- Đại diện Cơ quan chủ quản của NCS:
- Thành viên Hội đồng có mặt: 7/7 thành viên
  1. PGS.TS. Nguyễn Đức Dũng, Chủ tịch Hội đồng
  2. PGS.TS. Phạm Văn Hải, Phản biện 1
  3. PGS.TS. Phạm Văn Cường, Phản biện 2
  4. PGS.TS. Nguyễn Hà Nam, Phản biện 3

5. TS. Trần Đức Nghĩa, Thư ký Hội đồng
6. PGS.TS. Lê Hoàng Sơn, Ủy viên
7. PGS.TS. Bùi Thu Lâm, Ủy viên

- Thành viên Hội đồng vắng mặt:
- Đại diện tập thể cán bộ hướng dẫn: PGS.TS. Trần Thị Ngân, PGS.TS. Nguyễn Long Giang
- Cùng tham dự buổi bảo vệ còn có nhiều cán bộ nghiên cứu khoa học trong và ngoài Học viện.

TIẾN TRÌNH BUỔI BẢO VỆ

1. *Đại diện cơ sở đào tạo,* cô Phạm Thị Như Quỳnh, tuyên bố lý do, giới thiệu đại biểu và đọc quyết định số 1369/QĐ-HVKHCN ngày 15 tháng 12 năm 2025 của Giám đốc HVKHCN về việc thành lập Hội đồng đánh giá luận án tiến sĩ cấp Học viện cho NCS Michael Omar và đề nghị Chủ tịch Hội đồng điều khiển phiên họp.
2. *Chủ tịch Hội đồng*, PGS.TS. Nguyễn Đức Dũng, công bố danh sách thành viên có mặt là 07, thông qua chương trình buổi bảo vệ, đề nghị Thư ký thông báo các điều kiện chuẩn bị cho buổi bảo vệ và đọc lý lịch khoa học của NCS.
3. *Thư ký Hội đồng*, TS. Trần Đức Nghĩa thông báo các điều kiện cho buổi bảo vệ
   - Đọc lý lịch khoa học của NCS Michael Omar.
   - Đã nhận đủ 07 nhận xét của các phản biện và các thành viên HĐ.
   - Lịch bảo vệ của NCS đã được đăng trên Cổng thông tin điện tử Học viện Khoa học và Công nghệ ngày 15/09/2025.
   - Các giấy tờ cần thiết khác.

   NCS Michael Omar có đủ các điều kiện về thủ tục để bảo vệ luận án trước Hội đồng đánh giá luận án cấp Học viện.
4. Các thành viên hội đồng và những người tham dự thông qua về lý lịch khoa học và quá trình đào tạo của nghiên cứu sinh.
5. Nghiên cứu sinh Michael Omar trình bày nội dung luận án trong 30 phút trước Hội đồng. Báo cáo của NCS bao gồm các nội dung chính như sau:

   Giới thiệu
   Tổng quan về bài toán phân loại chất lượng nước ngầm.
   Các nghiên cứu liên quan.

Động lực và thách thức.
Các kết quả chính của luận án:
(1) Phát triển phương pháp lai AI-LGBM bao gồm các bước tiền xử lý trích xuất đặc trưng, xây dựng mô hình huấn luyện bằng LGBM, và kiểm tra và diễn giải kết quả bằng LIME/SHAP.
(2) Phát triển phương pháp PSO-SCNN sử dụng hàm khoảng cách Haversine trong các lớp nhúng trước khi đưa vào SCNN (với các hệ số không gian bổ sung). Tối ưu hóa các siêu tham số bằng PSO.
Kết luận và phát triển trong tương lai

6. Các phản biện đọc bản nhận xét và đặt câu hỏi đánh giá luận án của NCS Michael Omar

1) **Phản biện 1, PGS.TS. Phạm Văn Hải**, đọc nhận xét đánh giá luận án và kết luận (có văn bản kèm theo).

Ưu điểm:

Technical study features propose a new framework are as follows:

- DL, and GIS for groundwater drinkability classification. The hybrid models—AILGBM, PSO-SCNN, and CNN-GIS have been tested for sustainable water management in regions like Vietnam's Mekong Delta and India's Odisha.
- The thesis presents a hybrid spatial-aware ensemble framework combining AI-LGBM, PSO-SCNN, and CNN-GIS, improving accuracy and generalization.
- Key novelties include direct geographic feature integration for spatial learning, PSO-based hyperparameter optimization for SCNN, and SHAP/LIME for enhanced model interpretability and trust.

Nhược điểm:

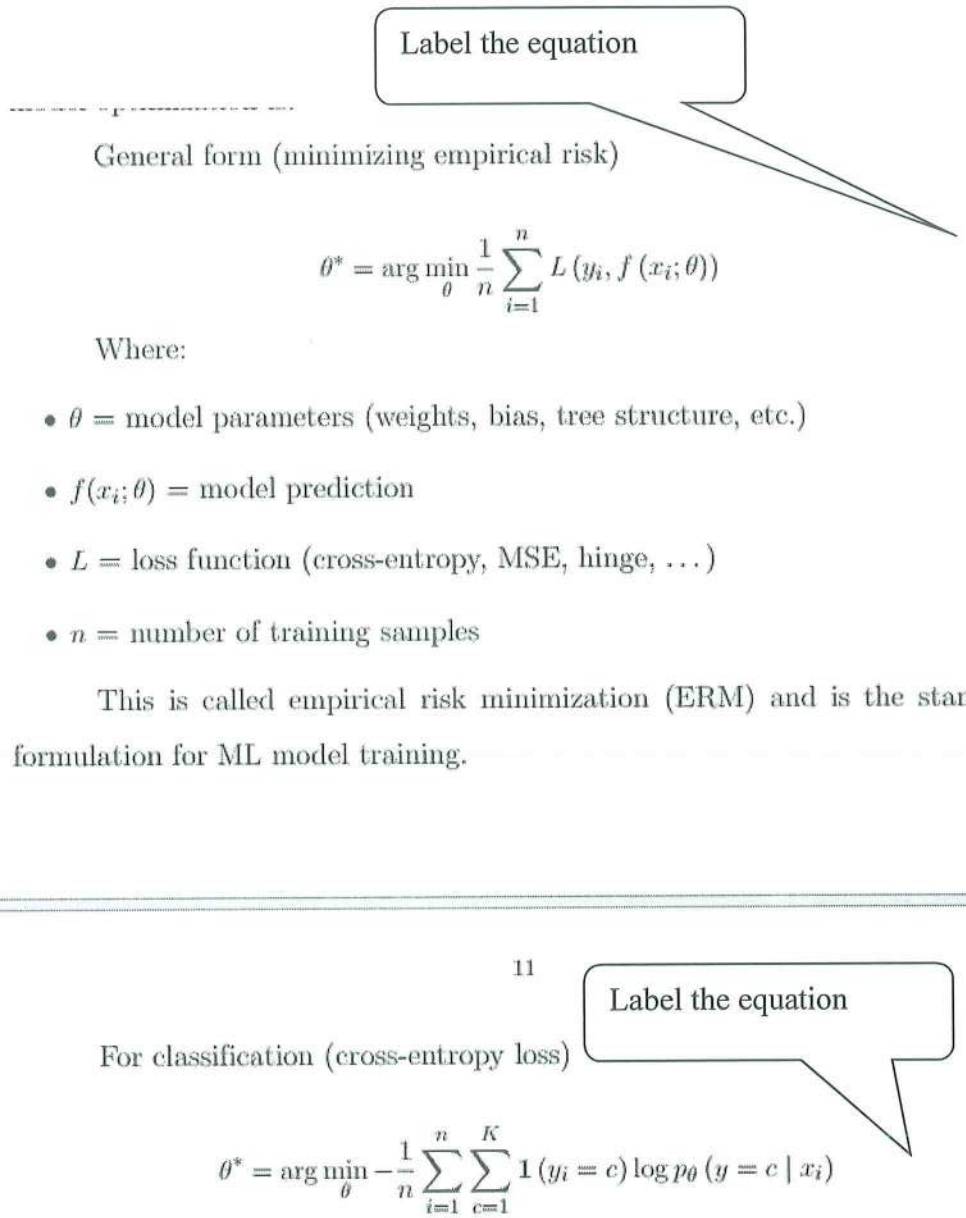
Study contributions should be clearly described since some questions explain as follows:

Question 1: What are technical techniques to enhance the proposed methods, models? Classify.

Question 2: What is a methodological innovation? It should be explained clearly a methodological innovation for the study contributions.

Question 3: What is technical issue of the proposed hybrid models (AI-LGBM, PSO-SCNN) to achieve up to 98.8% accuracy? Classify.

Some comments are as follows:



Data sets in the experiments should be explained in details in order to figure out an appropriate model selection.

2) **Phản biện 2, PGS.TS. Phạm Văn Cường**, đọc nhận xét đánh giá luận án và kết luận (có văn bản kèm theo).

Ưu điểm (strong points): các đóng góp chính được trình bầy tương đối rõ ràng trong chương 2 và 3. Các thử nghiệm được trình bầy tương đối rõ ràng, chi tiết.

Nhược điểm (weak points): các mô hình đề xuất nên được trình bầy chi tiết hơn: kiến trúc, các lớp, đầu vào, đầu ra, quá trình huấn luyện, v.v.. các mô hình cơ sở (baseline) khá cũ (outdated) và chưa tận dụng được những thành quả của học máy hiện đại để nâng cao hiệu quả phân loại. Nhiều đoạn viết lặp lại (i.e. hyperparameter optimization and tuning) và khó hiểu.

Một số ý kiến chỉnh sửa, bổ sung luận án như sau:

Nhiều hình vẽ cần bổ sung các giải thích, ví dụ hình 2.1, 2.2 (several figures need to be well explained, eg. 2.1, 2.2)

Nên bổ sung kiến trúc các mô hình đề xuất AI-LGMB và PSO-SCNN (add the network architectures of AI-LGMB and PSO-SCNN)

Nên coi tinh chỉnh mô hình là một bước trong quá trình xây dựng mô hình. Mô hình đề xuất phải là mô hình sau khi đã tinh chỉnh (the proposed model should have been the fine-tuned version).

Giải thích tại sao Transformer lại cho kết quả thấp nhất trong bảng 3.29 (trang 105). Show that why Transformer's performance is the lowest in the table 3.29.

Bổ sung phân tích chỉ tiết hơn các kết quả đánh giá tại các bảng 3.6

3) **Phản biện 3, PGS.TS. Nguyễn Hà Nam**, đọc nhận xét đánh giá luận án và kết luận (có văn bản kèm theo).

Ưu điểm:

- Bài toán nghiên cứu rõ ràng, có dữ liệu thực tế lớn từ hai quốc gia.
- Thực nghiệm đầy đủ, đa dạng, có so sánh với nhiều mô hình truyền thống.
- Có ý thức về tính giải thích và độ tin cậy thống kê của kết quả.

Thiếu sót và điểm cần bổ sung:

- Chưa nhất quán trong việc định nghĩa bài toán phân loại (đa lớp theo WQI và nhị phân "is-drinkable").
- Chưa làm rõ đầy đủ cơ chế ensemble như đã nêu trong tên luận án.
- Phần tái lập nghiên cứu (reproducibility) cần mô tả chi tiết hơn về cấu hình, tham số và khả năng chia sẻ dữ liệu/mã nguồn.
- Cần thảo luận sâu hơn về chi phí tính toán và khả năng triển khai thực tế.

Luận án sử dụng thuật ngữ "ensemble spatial machine learning" nhưng chưa làm rõ bản chất của chiến lược ensemble (stacking, voting hay chỉ là so sánh song song các mô hình), điều này làm giảm mức độ chặt chẽ về mặt phương pháp luận.

NCS Michael Omar tiếp thu ý kiến nhận xét của các phản biện và trả lời đầy đủ câu hỏi của các uỷ viên phản biện.

7. Các thành viên khác trong Hội đồng đưa ra ý kiến nhận xét; Hội đồng và những người tham dự đặt câu hỏi

- **PGS.TS. Nguyễn Đức Dũng**

Bài toán nghiên cứu cần được trình bày rõ ràng và cụ thể hơn: Dữ liệu đầu vào và dữ liệu đầu ra, các biến số, các mức độ về chất lượng nước (groundwater quality) và khái niệm nước uống được (drinkability). Việc thu thập và đặc điểm của dữ liệu đầu vào cần được trình bày rõ ràng và cụ thể hơn nữa.

Những nghiên cứu liên quan về phân loại chất lượng nước, cả các phương pháp truyền thống và sử dụng học máy, cần được phân tích thấu đáo hơn nhằm làm rõ những đề xuất mới của luận án. Mối quan hệ giữa đặc thù của bài toán với những đề xuất mới cần được phân tích tốt hơn để làm rõ phạm vi ứng dụng của những đề xuất này.

Kết quả thử nghiệm cần được so sánh với kết quả của những nghiên cứu liên quan.

Luận án cần tập trung hơn nữa vào phân tích đặc điểm của bài toán, tránh trình bày dàn trải về những nội dung cơ bản về học máy.

- **PGS.TS. Lê Hoàng Sơn**

(1) Kiến nghị bổ sung phân tích về độ phức tạp tính toán của các đề xuất (AI-LGBM và PSO-SCNN) và bổ sung hướng phát triển về tối ưu hóa hiệu quả tính toán thông qua việc sử dụng các kiến trúc DL hạng nhẹ hoặc các kỹ thuật tối ưu hóa giúp dễ tiếp cận hơn trong môi trường có nguồn lực hạn chế.

(2) Kiến nghị bàn luận thêm về việc sử dụng các nguồn dữ liệu bổ sung trong các mô hình đa mô thức như tích hợp dữ liệu viễn thám, ảnh vệ tinh và các nguồn phi truyền thống khác để cải thiện độ chính xác của mô hình, đặc biệt là ở những khu vực có dữ liệu thưa thớt. Lưu ý các tiếp cận tích hợp dữ liệu không-thời gian mở rộng (4D, 5D,..) để mở rộng mô hình nhằm theo dõi sự thay đổi chất lượng nước ngầm theo thời gian trong các nhiệm vụ giám sát và dự báo dài hạn. Mở rộng phạm vi rộng hơn các khu vực với điều kiện thủy văn

và môi trường đa dạng để đánh giá tính ổn định và khả năng thích ứng của chúng trong các bối cảnh khác nhau.

(3)Đề xuất xây dựng bộ dữ liệu tiêu chuẩn cho phân loại và dự báo chất lượng nước ngầm, bao gồm cả dữ liệu bản đồ trên các nền tảng mở, giúp cộng đồng Tin – Địa mở rộng và áp dụng được trên phạm vi rộng hơn từ các bối cảnh địa lý và thủy văn đa dạng, từ đó hỗ trợ nâng cao tác động quản lý và chính sách nước ngầm.

- **PGS.TS. Bùi Thu Lâm**

- Tỷ trọng kiến thức về thủy văn – môi trường khá lớn, trong khi phần kiến trúc hệ thống thông tin, triển khai hệ thống, hoặc luồng xử lý dữ liệu trong môi trường thực tế còn mờ nhạt. Điều này khiến luận án có xu hướng nghiêng về AI ứng dụng hơn là Hệ thống thông tin theo nghĩa hẹp.

- Luận án (đặc biệt Chương 3) quá thiên về kết quả thực nghiệm, thiếu thảo luận học thuật sâu; chưa làm rõ tính tổng quát hóa (generalization) khi áp dụng sang các khu vực có điều kiện địa chất khác.

- Phát biểu về "fundamental flaw" ở trang 3 chưa thực sự chính xác. Khoogn có flaw gì ở đây cả, chỉ là mô hình chưa mở rộng sang hướng đó mà thôi.

- Trang 6: các đóng góp (4 đóng góp) được nêu khá lan man, nên tập trung vào đóng góp học thuật.

- Việc mô tả các nhãn của của bài toán phân loại phải thống nhất từ đầu đến cuối luận ná. Hiện mô tả hơi lộn xộn

- Trang 10: thống nhất dùng kí hiệu w hay theta

- Không nên đưa các code snippnet vào luận án (trang 35), đưa vào dạng thuật toán/giả mã thì tốt hơn

- Chưa làm rõ ranh giới đóng góp mới của AI-LGBM so với các biến thể LightGBM + tối ưu hóa siêu tham số đã rất phổ biến trong văn liệu.

- Một số đoạn mô tả "state-of-the-art" còn mang tính tổng hợp – liệt kê, thiếu phân tích phê phán sâu để chỉ ra chính xác khoảng trống khoa học mà luận án giải quyết.

- Việc kiểm định không gian (spatial validation) chưa được mô tả đủ chặt chẽ. Việc sử dụng k-fold truyền thống trên dữ liệu không gian có thể dẫn đến đánh giá quá lạc quan.

- Độ nhạy của mô hình với sự thay đổi vùng địa lý chưa được phân tích sâu.

- PSO-SCNN có độ phức tạp cao, nhưng chưa có phân tích rõ ràng về chi phí tính toán – khả năng triển khai thực tế trong môi trường hạn chế tài nguyên.
- Hình 2.3 (b): chưa thấy khối 4 (component No4) ở đâu
- Nhiều số liệu nên được trình bày dạng biểu đồ để dễ hình dùng hơn.
- Bảng 3.13: mememory comsumption của AI-LGBM = 0 nghĩa là sao?
- **TS. Trần Đức Nghĩa**

Nên cải thiện việc viết Luận án để phù hợp hơn với ngành Hệ thống thông tin (giảm bớt thông tin về thủy văn/môi trường, tăng thông tin về hệ thống).

Cần rà soát và khắc phục lỗi chính tả cùng những lỗi soạn thảo khác.

Cần giải thích rõ hơn tại sao phải sử dụng học máy cho bài toán này?

8. NCS Michael Omar tiếp thu ý kiến nhận xét của các thành viên của Hội đồng. NCS trả lời các câu hỏi của các thành viên Hội đồng.
9. Đại diện tập thể hướng dẫn phát biểu ý kiến bằng văn bản.
10. Hội đồng tiến hành họp riêng để bầu ban kiểm phiếu, bỏ phiếu kín và thảo luận thông qua quyết nghị của Hội đồng.

1) Bầu ban kiểm phiếu gồm:
   - Trưởng ban: PGS.TS. Bùi Thu Lâm
   - Uỷ viên: PGS.TS. Lê Hoàng Sơn
   - Uỷ viên: TS. Trần Đức Nghĩa
2) Bỏ phiếu kín và thảo luận thông qua Quyết nghị của Hội đồng.
   - Trưởng ban kiểm phiếu, PGS.TS. Bùi Thu Lâm công bố kết quả kiểm phiếu (có biên bản kiểm phiếu).
   - Chủ tịch Hội đồng, PGS.TS. Nguyễn Đức Dũng, thông qua Quyết nghị (có văn bản kèm theo).
3) Tóm tắt nghị quyết của Hội đồng

3.1. Tính phù hợp của tên đề tài và sự không trùng lặp về nội dung luận án

- Tên đề tài, nội dung và kết quả nghiên cứu của luận án phù hợp với Ngành đào tạo "Hệ thống thông tin", mã số "9 48 01 04".
- Nội dung của luận án không trùng lặp với các luận án đã bảo vệ và các kết quả nghiên cứu đã công bố trong và ngoài nước.
- Các tài liệu tham khảo của luận án có nội dung phù hợp và đã được trích dẫn trong luận án.

3.2. Các kết quả chính của luận án

Luận án có ý nghĩa khoa học và thực tiễn, những đóng góp mới của luận án cụ thể như sau:

(1) Phát triển phương pháp lai AI-LGBM bao gồm các bước tiền xử lý trích xuất đặc trưng, xây dựng mô hình huấn luyện bằng LGBM, và kiểm tra và diễn giải kết quả bằng LIME/SHAP.

(2) Phát triển phương pháp PSO-SCNN sử dụng hàm khoảng cách Haversine trong các lớp nhúng trước khi đưa vào SCNN (với các hệ số không gian bổ sung). Tối ưu hóa các siêu tham số bằng PSO.

3.3. Các điểm cần bổ sung chỉnh sửa trước khi nộp luận án cho Thư viện Quốc gia Việt Nam

NCS cần tiếp thu, rà soát, chỉnh sửa, bổ sung nội dung luận án theo ý kiến đóng góp trong bản nhận xét của các thành viên Hội đồng và Biên bản của Hội đồng đánh giá luận án tiến sĩ cấp Học viện trước khi nộp luận án cho Thư viện Quốc gia Việt Nam.

3.4. Mức độ đáp ứng yêu cầu của luận án tiến sĩ về cả nội dung và hình thức

- Luận án của NCS Michael Omar đáp ứng yêu cầu của luận án tiến sĩ ngành "Hệ thống thông tin", mã số "9 48 01 04" về nội dung và hình thức theo các qui chế hiện hành về đào tạo tiến sĩ của Bộ Giáo dục và Đào tạo, của Học viện Khoa học và Công nghệ.
- Đề nghị Học viện Khoa học và Công nghệ công nhận kết quả bảo vệ và cấp bằng tiến sĩ cho NCS Michael Omar sau khi chỉnh sửa, bổ sung luận án theo các góp ý của Hội đồng.

11. Tổng kết

- Trưởng ban kiểm phiếu, PGS.TS. Bùi Thu Lâm, công bố kết quả bỏ phiếu đánh giá luận án.
- Chủ tịch Hội đồng, PGS.TS. Nguyễn Đức Dũng, đọc Quyết nghị của Hội đồng.
- Chủ tịch Hội đồng tuyên bố Hội đồng đã hoàn thành nhiệm vụ và trao lại quyền điều khiển cho Cơ sở đào tạo.
- Các đại biểu và NCS phát biểu ý kiến.

- Đại diện cơ sở đào tạo tuyên bố kết thúc buổi bảo vệ luận án tiến sĩ.

Biên bản họp Hội đồng này được 7/7 ủy viên Hội đồng biểu quyết công khai thông qua.

Buổi họp Hội đồng đánh giá luận án tiến sĩ cấp Học viện kết thúc vào lúc 16 giờ 30 phút, ngày 27 tháng 01 năm 2026.

Thư ký Hội đồng	**Chủ tịch Hội đồng**
TS. Trần Đức Nghĩa	**PGS.TS. Nguyễn Đức Dũng**

XÁC NHẬN CỦA CƠ SỞ ĐÀO TẠO

KT. GIÁM ĐỐC
PHÓ GIÁM ĐỐC

VIỆN HÀN LÂM KHOA HỌC VÀ CÔNG NGHỆ VIỆT NAM
HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

Nguyễn Thị Trung

VIỆN HÀN LÂM
KHOA HỌC VÀ CÔNG NGHỆ VN
HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

Hà Nội, ngày 27 tháng 01 năm 2026

QUYẾT NGHỊ CỦA
HỘI ĐỒNG ĐÁNH GIÁ LUẬN ÁN TIẾN SĨ CẤP HỌC VIỆN

Căn cứ quyết định số 1369/QĐ-HVKHCN ngày 15 tháng 12 năm 2025 của Giám đốc Học viện Khoa học và Công nghệ về việc thành lập Hội đồng đánh giá luận án tiến sĩ cấp Học viện, Hội đồng đã họp vào hồi 14 giờ 00 ngày 27 tháng 01 năm 2026 tại Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam, số 18 đường Hoàng Quốc Việt, Cầu Giấy, Hà Nội để đánh giá luận án tiến sĩ.

Họ và tên NCS:	**Michael Omar**
Tên đề tài luận án:	Nghiên cứu phát triển phương pháp học máy kết hợp thông tin không gian cho bài toán phân loại chất lượng nước ngầm/ An approach of ensemble spatial machine learning for groundwater drinkability classification
Ngành:	Hệ thống thông tin
Mã số:	9 48 01 04
Người hướng dẫn:	PGS.TS. Trần Thị Ngân và PGS.TS. Nguyễn Long Giang

HỘI ĐỒNG ĐÁNH GIÁ LUẬN ÁN TIẾN SĨ CẤP HỌC VIỆN CỦA NCS MICHAEL OMAR KẾT LUẬN

1. Tính phù hợp của tên đề tài và sự không trùng lặp về nội dung luận án

- Tên đề tài, nội dung và kết quả nghiên cứu của luận án phù hợp với ngành đào tạo Hệ thống thông tin, mã số 9 48 01 04.

- Nội dung của luận án không trùng lặp với các luận án đã bảo vệ và các kết quả nghiên cứu đã công bố trong và ngoài nước.

- Các tài liệu tham khảo của luận án có nội dung phù hợp và đã được trích dẫn trong luận án.

2. Kết quả, ý nghĩa khoa học, thực tiễn của đề tài

Luận án có ý nghĩa khoa học và thực tiễn, những đóng góp mới của luận án cụ

thể như sau:

(1) Đề xuất một khung xử lý dữ liệu, tích hợp và kết hợp mô hình, lựa chọn mô hình/tham số cho bài toán phân loại nước ngầm.

(2) Phân tích chất lượng nước ngầm với các bộ dữ liệu thực nghiệm ở Việt Nam và Ấn Độ.

3. Những thiếu sót của luận án, vấn đề cần bổ sung, sửa chữa

NCS cần tiếp thu, rà soát, chỉnh sửa, bổ sung nội dung luận án theo ý kiến đóng góp trong bản nhận xét của các thành viên Hội đồng và Biên bản của Hội đồng đánh giá luận án tiến sĩ cấp Học viện trước khi nộp luận án cho Thư viện Quốc gia Việt Nam. Đặc biệt cần làm nổi bật đóng góp mới, cải thiện chất lượng bản thảo Luận án.

4. Mức độ đáp ứng yêu cầu của luận án tiến sĩ về cả nội dung và hình thức theo các quy chế hiện hành về đào tạo tiến sĩ của Bộ Giáo dục và Đào tạo

Luận án của NCS Michael Omar đáp ứng yêu cầu của một luận án tiến sĩ ngành "Hệ thống thông tin", mã số 9 48 01 04 về nội dung và hình thức theo các quy chế hiện hành về đào tạo tiến sĩ của Bộ Giáo dục và Đào tạo và của Học viện Khoa học và Công nghệ.

Kết luận:

Kết quả bỏ phiếu đánh giá luận án của Hội đồng: 7/7 thành viên tán thành.

Hội đồng kết luận thông qua luận án, đề nghị Học viện Khoa học và Công nghệ công nhận kết quả bảo vệ và cấp bằng tiến sĩ cho NCS Michael Omar.

Quyết nghị này được 7/7 thành viên Hội đồng biểu quyết công khai thông qua.

THƯ KÝ	CHỦ TỊCH
TS. Trần Đức Nghĩa	PGS.TS. Nguyễn Đức Dũng

XÁC NHẬN CỦA CƠ SỞ ĐÀO TẠO

KT. GIÁM ĐỐC
PHÓ GIÁM ĐỐC

VIỆN HÀN LÂM KHOA HỌC VÀ CÔNG NGHỆ VIỆT NAM
HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

Nguyễn Thị Trung

(Mẫu 6-HV-Bản giải trình chỉnh sửa, bổ sung cấp HV)

VIỆN HÀN LÂM
KHOA HỌC VÀ CÔNG NGHỆ VN
HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BẢN GIẢI TRÌNH CHỈNH SỬA, BỔ SUNG LUẬN ÁN TIẾN SĨ CẤP HỌC VIỆN

Ngày 15 tháng 12 năm 2025, Học viện Khoa học và Công nghệ đã tổ chức đánh giá luận án tiến sĩ cấp Học viện cho nghiên cứu sinh Michael Omar theo Quyết định số 1369/QĐ-HVKHCN ngày 15 tháng 12 năm 2025 của Giám đốc Học viện.

Đề tài:

Tên tiếng Việt: Nghiên cứu phát triển phương pháp học máy kết hợp thông tin không gian cho bài toán phân loại nước ngầm.

Tên tiếng Anh: An Approach of Ensemble Spatial Machine Learning For Groundwater Drinkability Classification.

Ngành: Hệ thống thông tin, Mã số: 9 48 01 04

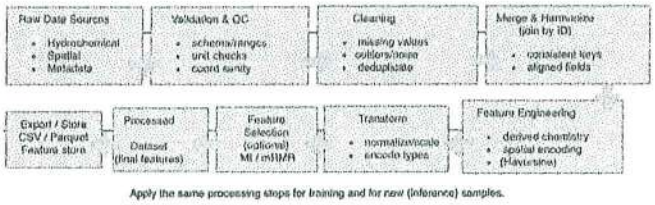
Người hướng dẫn khoa học: PGS. TS Trần Thị Ngân, PGS. TS Nguyễn Long Giang.

Theo Biên bản của Hội đồng, NCS phải bổ sung và chỉnh sửa luận án các điểm sau đây:

STT	Nội dung đề nghị chỉnh sửa, bổ sung	Nội dung đã được chỉnh sửa, bổ sung *(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)*
1	The thesis should be improved to better align with the Information Systems field. How can the thesis reduce the emphasis on hydrology/environment and increase the focus on the systems aspect?	**Page 11 / Chapter 1 / Section 1.1** Thank you for this helpful comment. We agree that the earlier version emphasized hydrology/environmental background in the Information Systems field. We have revised the thesis to reduce hydrology-focused content and strengthen the systems perspective, highlighting the end-to-end ML–GIS workflow, system components, data pipeline, and decision-support outputs. Domain-specific hydrology details are now kept only as minimal context. **Page 11 / Chapter 1 / Section 1.1** **Current Excerpt:** "Groundwater drinkability classification assesses whether aquifer water meets health-based standards by converting multivariate hydrochemistry and geospatial data into categorical risk (e.g., safe/unsafe)." **Revised:** Groundwater drinkability classification leverages machine learning (ML) and spatial information systems (GIS) to automate groundwater quality classification. By integrating hydrochemical parameters and spatial features, the system provides scalable assessments to support decision-making in water resource management.

Lưu ý: Các chữ ký xác nhận cần gắn với nội dung trên cùng một trang giấy. Học viện sẽ không xác nhận nếu phần chữ ký tách rời với nội dung

2	Check throughout your thesis for writing in scientific way Spelling and other formatting errors should be reviewed and corrected.	(Ghi rõ số trang/chương/mục… đã được chỉnh sửa) **Page 15-17 / Introduction /** **Page 17-19 / Introduction / Problem Statement** **Introduction** Thank you for your insightful feedback. Several sections of the thesis have been revised to adopt a more scientific and formal tone. **Current Excerpt:** *" This need is critically amplified by a growing global population that intensifies pressure on finite water resource."* **Revised:***"The increasing global population places significant pressure on finite water resources, exacerbating the challenge of ensuring safe drinking water."* **Problem Statement** **Current Excerpt:** *"Most standard models fail to account for spatial autocorrelation, the principle that nearby samples are related, which leads to unreliable predictions."* **Revised:***"Many traditional models fail to account for spatial autocorrelation, the principle that geographically proximate data points exhibit similar characteristics. This oversight leads to unreliable predictions, as the models fail to generalize to new areas."* **Research Motivation** **Current Excerpt:** *"This research is driven by the critical need to overcome the interconnected limitations of both traditional monitoring and contemporary AI approaches."* **Revised:***"This research addresses the limitations of traditional groundwater quality monitoring by employing machine learning (ML) and deep learning (DL) methods, which offer scalable, real-time solutions that overcome issues such as slow processing, high cost, and limited scalability inherent in conventional methods."* Thank you for your valuable feedback. We have thoroughly reviewed the thesis for grammar errors and have made the necessary corrections throughout the document. These updates improve the clarity and readability of the text.
3	A clearer explanation is needed on why machine learning is necessary for this problem. Can the thesis provide a more detailed justification for the use of machine learning in the context of groundwater drinkability classification?	*(Ghi rõ số trang/chương/mục… đã được chỉnh sửa)* **Page 26 / Chapter 1 / Section 1.2.3** Proposed Solution for Water Classification Challenges **Page 135 / Chapter 3 / Section 3.6.2** Implications for Groundawater Quality Classifications Thank you for your insightful feedback. We have revised the thesis to provide a clearer explanation of why machine learning is necessary for the problem of groundwater drinkability classification. Specifically, we have enhanced the justification for the use of machine learning in addressing the complexities and challenges associated with groundwater classification. The revisions can be found in: **Page 26, Chapter 1, Section 1.2.3,** titled Proposed Solution for Water Classification Challenges. **Page 135, Chapter 3, Section 3.6.2**, titled Implications for Groundwater Quality Classifications.

		These sections now offer a more detailed explanation of how machine learning models can effectively handle large, complex datasets and capture spatial and non-linear relationships, making them particularly suitable for groundwater drinkability classification.
4	The thesis has a heavy focus on hydrology and environmental science, while the information system architecture, system implementation, or data processing workflows in practical environments are less emphasized. How can the thesis better align with the Information Systems field by reducing the emphasis on hydrology/environment and increasing focus on the systems aspect?	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* **Page 56 / Chapter 2 / Section 2.1.1** **Page 70 / Chapter 2 / Section 2.3.1 (Figure 2.3)** **Page 88 / Chapter 3 / Section 3.2.1 (Figure 3.1)** Thank you and Response to the Reviewer: Thank you for your insightful feedback. We have updated the thesis to better align with the Information Systems field, focusing more on the system architecture, system implementation, and data processing workflows in practical environments, while reducing the emphasis on hydrology and environmental science. The following sections have been revised: **Page 56, Chapter 2, Section 2.1.1**: This section has been updated to provide a clearer explanation of the Information Systems architecture and the flow of data in the system. **Page 70, Section 2.3.1 (Figure 2.3)**: We have added a Data Flow Diagram to better illustrate the flow of information and system integration.  **Page 88, Chapter 3, Section 3.2.1 (Figure 3.1)**: The Datasets and Preprocessing workflow section has been updated to describe the data processing steps more thoroughly, enhancing the explanation of the system's practical implementation.  Proposesystem model We have updated the thesis to better align with the Information Systems field by expanding on the system architecture, data processing workflows, and system implementation. These additions provide a more detailed and technical focus on how the proposed models are implemented in real-world environments. The Proposed System Model section has been revised to emphasize the system's data flow, real-time capabilities, and integration with decision support platforms, making the research more relevant to the IS domain.
5	Chapter 3 is overly focused on experimental results, lacking in-depth academic discussion. How can the thesis clarify the generalization of the proposed models when applying them to areas with different geological conditions?	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* **Page 133 - 134 / Chapter 3 / Section 3.5.4** Computational Cost and Deployment Feasibility Thank you for your valuable feedback. We have updated **Chapter 3, Section 3.5.4 (Page 133 - 134)** to address the concern about the generalization of the proposed models in different geological conditions. The section now explains domain shift, distinguishes between spatial generalization and cross-geology generalization, and highlights challenges in applying the

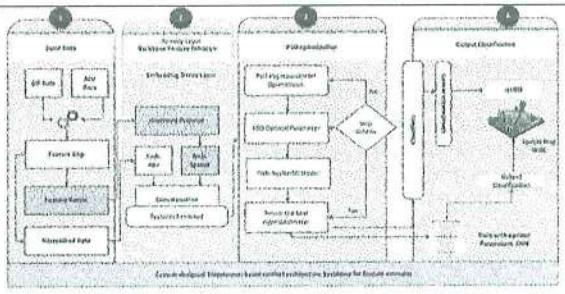
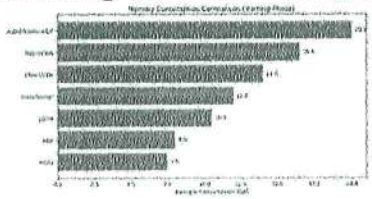
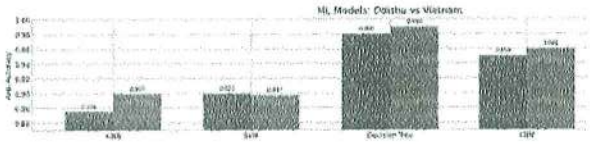
		models to new geological settings. We also added evaluation strategies such as spatially blocked validation, cross-region testing, and suggestions for uncertainty-aware deployment to improve adaptability across geological conditions.
6	The statement regarding a "fundamental flaw" on page 3 is not accurate. There is no fundamental flaw; rather, the model has simply not been extended in that direction. Now Page 4	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* *fundamental flaw" on page 3 is not accurate:* **NOW Page 4 / Introduction** Thank you for your valuable feedback. We have revised the section **on** Page 4, Introduction to address the concern about the "fundamental flaw" statement. The updated excerpt now reads: **Revised Statement:** **Research Motivation Item 2** "Mitigate the common limitation of AI models that overlook spatial autocorrelation by developing explicitly spatial architectures and applying rigorous spatially-aware validation protocols to obtain reliable and trustworthy performance." This revision clarifies that the issue is not a "fundamental flaw" but rather a common limitation that the model aims to address. We appreciate your suggestion for improvement and believe this change enhances the clarity of our argument.
7	On page 6, the listed contributions (four items) are presented in a somewhat scattered manner and should be more focused on academic contributions.	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* **Page 6, 7 / Introduction** Thank you for your helpful feedback. We agree that the original four contributions on Page 6 were presented in a scattered way and did not sufficiently emphasize the **academic contributions**. We have **revised Pages 6–7 (Introduction)** to present the contributions in a **more focused and structured** manner under the section : **Contributions of the Thesis and Significance** The updated text consolidates and reframes the contributions around clear academic elements (e.g., methodological novelty, spatial-learning integration, interpretability, and scalability), and removes unnecessary domain-specific details.
8	The description of classification labels should be consistent throughout the dissertation; the current presentation is somewhat inconsistent.	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* **Page 12, 13 and 14 / Chapter 1/ Section 1.1** Thank you for your feedback. We have reviewed and corrected all occurrences of the classification labels in the thesis to ensure consistency. The labels are now uniformly presented as **Excellent**, **Good**, **Fair**, **Poor**, and **Unsuitable** throughout the document. This update has been applied to the entire thesis, including Page **12, 13 and 14, Chapter 1, Section-1.1 Problem 1 : Groundwater Drinkability Classification**
9	On page 10, the notation should be unified (either w or θ).	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* **Page 14 , 15 / Chapter 1 / Section 1.1 / equation 1.6** Thank you for your feedback. We have unified the notation between **w** and **θ** in the thesis, as both were used interchangeably for

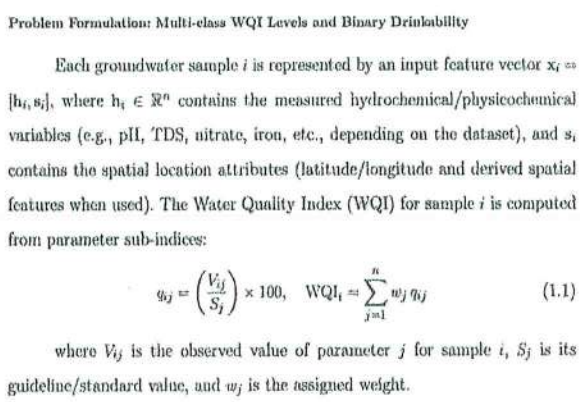
		hyperparameters in different equations. All equations have now been corrected to ensure consistency, with w used specifically for model parameters and θ used more generally in the context of optimization problems. These corrections have been applied throughout the thesis, including **Page 14, 15** , to maintain consistency in the notation. $$\theta^* = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} L\left(y_i, f\left(x_i; \theta\right)\right)$$ **Revised:** $$W^* = \arg\min_{W} \frac{1}{n} \sum_{i=1}^{n} L\left(y_i, f\left(x_i; W\right)\right)$$

10	Code snippets should not be included in the dissertation **(page 40)**; algorithmic descriptions or pseudocode would be more appropriate.	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)*
		Page 40, 41 / Chapter 1 Section 1.5
		Data Class Imbalance
		Thank you for your valuable feedback.
		We have revised the thesis to replace all code snippets with algorithmic descriptions or pseudocode, as recommended.
		Specifically, the section on **Page 40 , 41,** Chapter 1, **Section 1.5** has been updated to align with this format, ensuring the content is more appropriate for an academic presentation.
		All code snippets have now been converted to clear algorithmic steps that are consistent with the overall presentation of the thesis.
		Code snippet: SMOTE
		`from imblearn.over_sampling import SMOTE`
		`# Split data`
		`X_train, X_test, y_train, y_test = train_test_split`
		`(X, y, stratify=y, test_size=0.2, random_state=42)`
		`# Apply SMOTE to balance the training data`
		`sm = SMOTE(random_state=42)`
		`X_train, y_train = sm.fit_resample(X_train, y_train)`
		Algorithm 1.1 SMOTE Algorithm for Balancing Training Data
		1: **Input:** Training dataset (X_{train}, y_{train})
		2: **Output:** Resampled training dataset (X_{train}, y_{train})
		3: Split dataset into training and testing sets
		4: Apply SMOTE to the training set to generate synthetic samples
		5: $X_{train}, y_{train} \leftarrow \text{SMOTE}(X_{train}, y_{train})$
		6: (X_{train}, y_{train})

11	The boundary of novelty for the AI-LGBM model compared to widely used LightGBM variants with hyperparameter optimization is not clearly articulated.	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* **Page 51 / Chapter 2 / Section 2.2.1** Thank you for your valuable feedback. We have updated **Page 51, Chapter 2, Section 2.2.1 Proposed Ensemble Spatial Machine Learning methods** to clearly articulate the boundary of novelty for the AI-LGBM model in comparison to widely used LightGBM variants with hyperparameter optimization. The updated section now emphasizes the innovations in AI-LGBM, particularly how **Mutual Information-based Feature Selection (MIFS)** and **advanced optimization techniques such as** Optuna **and** Auto-Immune Optimization (AIO) enhance feature selection, hyperparameter tuning, and overall accuracy. These improvements contribute significantly to the

		model's superior performance in sensitive tasks like groundwater quality classification.
12	Several "state-of-the-art" descriptions are largely enumerative and lack critical analysis to clearly identify the precise research gap addressed by the dissertation.	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* **Page 27, 28 / Chapter 1 / Section 1.2.4, 1.2.5** Thank you for your insightful feedback. We have updated **Page 27, Chapter 1, Sections 1.2.4 Gaps and Summary** and **1.2.5 Research Method to address Gaps** to clarify the research gap addressed by this dissertation. Specifically, we now highlight how the proposed AI-LGBM and PSO-SCNN models improve feature selection and incorporate spatial dependencies, addressing key challenges in existing ML/DL models for spatial data. These innovations enhance prediction accuracy and efficiency, particularly for groundwater quality classification. The research gap and its resolution through the models are further detailed in **Chapter 2**.
13	Spatial validation is not described rigorously. The use of traditional k-fold cross-validation on spatial data may lead to overly optimistic performance estimates.	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* **Page 132 / Chapter 3 /Section 3.5** Thank you for your valuable feedback. We have updated Page 132, Chapter 3, Section 3.5 to include a more rigorous discussion on spatial validation and model evaluation. The new section addresses the limitations of traditional k-fold cross-validation for spatial data and provides a more robust approach to evaluating model performance to avoid overly optimistic estimates.
14	The PSO-SCNN model has high computational complexity, yet there is no clear analysis of computational cost or feasibility of deployment in resource-constrained environments.	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* **Page 133 / Chapter 3 / Section 3.5.4 Computational Cost and Deployment Feasibility** Thank you for your valuable feedback. We have **added a new subsection** titled **"Computation Cost and Deployment Feasibility"** in **Page 133, 134, Chapter 3, Section 3.5.4**. This section provides a detailed analysis of the computational complexity of the PSO-SCNN model and discusses its feasibility for deployment in resource-constrained environments. **3.5.4 Computational Cost and Deployment Feasibility** The computational cost of PSO-SCNN is dominated by the *offline* hyperparameter search stage. Let P be the number of PSO particles, I the number of PSO iterations, and V the number of validation runs (e.g., k-fold CV). If $\mathcal{C}_{\text{train}}$ denotes the cost of training one SCNN candidate, then the total tuning complexity scales as $$\mathcal{C}_{\text{PSO}} = \mathcal{O}(P \cdot I \cdot V \cdot \mathcal{C}_{\text{train}}). \quad (3.4)$$
15	In Figure 2.3(b), Component No. 4 is not clearly identifiable	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* **Page 70 / Chapter 2 / Section 2.3 / Subsection 2.3.1** Thank you for your feedback. In Page **70, Chapter 2, Section 2.3**, we have updated *Figure 2.3(b) Extended for spatial Map Visualization* to ensure that Component No. 4 is now clearly identifiable.

		
16	Many numerical results would be better presented in graphical form for improved interpretability.	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* **Page 113 / Chapter 3 / Section 3.3 4** **Page 118 / Chapter 3 / Section 3.3 4** Thank you for your valuable feedback. We have updated Page 113, Chapter 3, Section 3.3.4 by converting Table 3.18 into a graph "figure 3.16"  and Page 118, Table 3.32 into graphical forms for better interpretability. "Figure 3.20"  Additionally, several other tables throughout the thesis have also been converted into graphs to enhance clarity and understanding of the numerical results.
17	In Table 3.13, the reported memory consumption of AI-LGBM equals zero, which requires clarification.	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* **Page 98 / Chapter 3 / Section 3.2.5 / subsection Model Comparison and Sensitivity Analysis (Post-Run)** Thank you for your feedback. We have added a brief description **to Table 3.13** on **Page 98,** Chapter 3, Section 3.2.5 to clarify the zero memory consumption value. The explanation is as follows: **Note:** The values in the Memory Consumption (MB) column represent the incremental memory overhead measured during training relative to a baseline process memory (after loading the dataset and initializing the runtime). Results are reported in MB and rounded to six decimals; therefore, entries shown as 0.000000 indicate that the additional memory attributable to the model was below the measurement/rounding resolution (i.e., negligible at the scale of MB for the dataset size used), not that the model consumed zero RAM. In addition, some allocations for tree-based libraries (e.g., LightGBM) occur in native code and may be undercounted by Python-level memory tracking, further contributing to near-zero incremental readings. This revision provides a clearer understanding of the reported memory consumption.
18	**Item 18:**	(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)

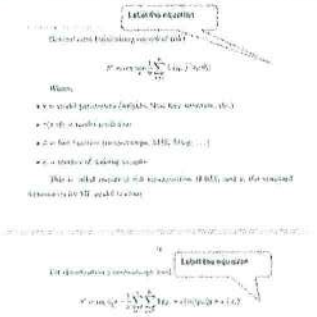
	The datasets used are still not clearly described in terms of the attributes of the two datasets from Vietnam and India, their characteristics, specific properties, differences, etc. In addition, the labeling process for the training data has only been described qualitatively. It has not precisely answered the Council's question regarding how this process is conducted to ensure that it is well-grounded and reliable.	**Page 43- 45 / Chapter 1 / Section 1.5.1** **Page 71, 72 / Chapter 2 / Section 2.3.1** Thank you for your feedback. Regarding the request for clearer dataset characterization and labeling, we have now added a detailed structural comparison of the Vietnam & India datasets in **Chapter 1, Section 1.5.1 Page. 43**, together with an expanded numeric statistical summary in **Table 1.19 Page. 45.** Furthermore, the labeling mechanism is now explicitly formalized in Section **1.5.2 Page. 43,** where the process is presented as a deterministic, equation-based transformation derived from regulatory standards. This ensures full reproducibility and removes subjective interpretation.
19	The classification problem needs to be more clearly defined (multi-class based on WQI and binary "is-drinkable").	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* **Page 12 / Chapter 1/ Section 1.1** Subsection: **Problem Formualtion: Multi-class WQI levels and binary Drinkability** Thank you for your suggestion. We have clarified the classification problem in the thesis. Specifically, **Page 12, Chapter 1, Section 1.1, Subsection 'Problem Formulation: Multi-class WQI Levels and Binary Drinkability'** has been updated to clearly define both the multi-class classification based on WQI and the binary 'is-drinkable' classification. This revision provides a more explicit explanation of the classification tasks addressed by the models.  **Problem Formulation: Multi-class WQI Levels and Binary Drinkability** Each groundwater sample i is represented by an input feature vector $x_i = [h_i, s_i]$, where $h_i \in \mathbb{R}^n$ contains the measured hydrochemical/physicochemical variables (e.g., pH, TDS, nitrate, iron, etc., depending on the dataset), and s_i contains the spatial location attributes (latitude/longitude and derived spatial features when used). The Water Quality Index (WQI) for sample i is computed from parameter sub-indices: $$q_{ij} = \left(\frac{V_{ij}}{S_j}\right) \times 100, \quad \mathrm{WQI}_i = \sum_{j=1}^{n} w_j q_{ij} \qquad (1.1)$$ where V_{ij} is the observed value of parameter j for sample i, S_j is its guideline/standard value, and w_j is the assigned weight.
20	**Item 20** The ensemble mechanism, as mentioned in the thesis title, has not been fully clarified. It is still unclear why the hybrid system AI-LGBM with PSO-SCNN is proposed when the candidate has already optimized each algorithm individually. The analysis of how the strengths and weaknesses of each algorithm compensate for one another, which leads to the proposal of combining the two using a weighting method, also needs to be explained more convincingly (as mentioned in the response to Item 36).	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* **Page 52 / Chapter 2/ Section 2.1.2** **Page 54 / Chapter / Section 2.1.3** **Ensemble Mechanism and Model Integration** Thank you for your suggestion. A new section has been added in **Page 52, Chapter 2, Section 2.1.2** titled ***Ensemble Mechanism and Model Integration*** **to fully clarify the ensemble mechanism used in the thesis.**

		2.1.2 Ensemble Mechanism and Model Integration In this thesis, the term *ensemble* is used in two complementary senses. (i) Model-level ensemble: AI-LGBM is an ensemble by construction because it follows gradient boosting, where the final predictor is an additive combination of many decision trees. For binary drinkability prediction it outputs a probability: $$\hat{P}_{\text{LGBM}}(y=1 \mid x) = \sigma\left(\sum_{m=1}^{M} f_m(x)\right), \quad (2.1)$$ where $f_m(\cdot)$ is the m-th tree, M is the number of trees, and $\sigma(\cdot)$ is the sigmoid Concerning the motivation for combining AI-LGBM and PSO-SCNN after individual optimization, we introduced a new dedicated explanation in **Chapter 2, Section 2.1.3 Page. 54.** The section clarifies the complementary inductive biases of the two models, explains how their strengths offset their weaknesses, and details how validation-driven weighting enables cost-sensitive and robust decision-making. **Table 2.1 Page 55,** further summarizes these compensatory properties.
21	The reproducibility section needs to be more detailed, especially regarding configurations, parameters, and data/code sharing.	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* **Page 144, 145 / Appendix/ Section Reproducibility And Artifacts Sharing** Thank you for your suggestion. The Reproducibility section has been updated in **Appendix A, Page 144, 145** to provide comprehensive details on software versions, random seed values, computing environment, data preprocessing, and model configurations. Additionally, the GitHub repository link for accessing all relevant scripts and resources, including the PSO-SCNN Model Repository, has been included to ensure full reproducibility of the work.
22	There should be more in-depth discussion of computational costs and practical deployment.	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* Page 136 , 137 / Chapter 3 / Section 3.5.4 Thank you for your suggestion. A new subsection titled *'Computation Cost and Deployment Feasibility'* has been added in Page 136-137, Chapter 3, Section 3.5.4 to provide a more in-depth discussion of computational costs and practical deployment 3.5.4 Computational Cost and Deployment Feasibility The computational cost of PSO-SCNN is dominated by the *offline* hyperparameter search stage. Let P be the number of PSO particles, I the number of PSO iterations, and V the number of validation runs (e.g., k-fold CV). If C_{train} denotes the cost of training one SCNN candidate, then the total tuning complexity scales as $$C_{\text{PSO}} = \mathcal{O}(P \cdot I \cdot V \cdot C_{\text{train}}). \quad (3.4)$$
23	The proposed models should be described in more detail: architecture, layers, inputs, outputs, training process, etc.	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* **Page 51 – 75 / Chapter 2 / Section 2.1 – 2.3** Thank you for your suggestion.

		The model architectures, as well as the input-output details for both AI-LGBM and PSO-SCNN, have been described in Page 51–75, Chapter 2, Sections 2.1–2.3. The inputs for AI-LGBM include hydrochemical features (e.g., pH, TDS, nitrate), while the PSO-SCNN model uses geospatial data mapped to a 2D grid. The output for both models is a predicted risk map or classification result identifying areas of contamination or drinkability. The training process for both models is detailed in Sections 2.2.3 (AI-LGBM) and 2.3.3 (PSO-SCNN).
24	There are some repetitive sections (e.g., hyperparameter optimization and tuning) that are difficult to understand.	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* **Page 66 - 67 / Chapter 2 / Section 2.2.3** **Page 91 Chapter 3 / Section 3.2.2** Thank you for your feedback. To reduce repetition, the term 'hyperparameter' has been replaced with 'parameters' where applicable. The repeated content regarding hyperparameter optimization and tuning arises from the distinct optimization processes for AI-LGBM and PSO-SCNN. Since each model has its own set of parameters and tuning procedures, the details were provided separately to address the unique characteristics and training requirements of each model. The term modification can be seen in **Page 66-67, Chapter 2, Section 2.2.3** **Page 88, Chapter 3, Section 3.2.2**
25	Several figures (e.g., Figures 2.1, 2.2) need better explanations.	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* **Page 53-55 / Chapter 2 / Section 2.1.1** **Page 58, 59 / Chapter 2 / Section 2.2.1** Thank you for your suggestion. The descriptions for **Figures 2.1** and **2.3** have been updated for clarity and precision in Page 53, Chapter 2, Section 2.1.1 and Page 58, Chapter 2, Section 2.2.1
26	Add the architectures for the proposed AI-LGMB and PSO-SCNN models.	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* **Page 57 Chapter 2 / Section 2.2.1** **Page 70 / Chapter 2 / Section 2.3.1** Thank you for your valuable feedback. The architectures for the proposed AI-LGBM and PSO-SCNN models is in the thesis. The AI-LGBM model, described in **Chapter 2, Section 2.2.1 (Page 57),** integrates LightGBM with Mutual Information-based Feature Selection (MIFS) and Auto-Immune Optimization (AIO) to enhance predictive accuracy and model robustness. The PSO-SCNN model, outlined in **Chapter 2, Section 2.3.1 (Page 70),** combines Particle Swarm Optimization (PSO) with Spatial Convolutional Neural Networks (SCNN) to optimize hyperparameters and capture spatial dependencies in groundwater data. Both models are designed to improve classification accuracy and offer explainability through tools like SHAP.
27	Consider fine-tuning the model as part of the model building process. The proposed model should be the fine-tuned version.	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* **Page 64 / Chapter 2 / Section 2.2.3 Learning strategy** **Page 77 / Chapter 2 / Section 2.3.3 Learning Strategy** Thank you for your feedback. The fine-tuning process for both AI-LGBM and PSO-

		SCNN models has been incorporated into the thesis: For the AI-LGBM model, fine-tuning was applied to hyperparameters such as the number of estimators, learning rate, max depth, and regularization factors. These were optimized using Optuna and Auto-Immune Optimization (AIO) to enhance performance on validation data, ensuring the model performs optimally. For PSO-SCNN, Particle Swarm Optimization (PSO) was used to fine-tune parameters like kernel size, number of filters, and learning rate. This process improved the model's ability to capture spatial patterns and ensured robustness against overfitting. These details are discussed in **Chapter 2, Section 2.2.3 (Page 64)** for AI-LGBM and **Section 2.3.3 (Page 77)** for PSO-SCNN.
28	Explain why the Transformer model shows the lowest performance in Table 3.29 (page 105).	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* **Page 112 / Chapter 3 / Section 3.3.4** **Table 3.29 Now table 3.28** Thank you for your valuable feedback. The Transformer model shows the lowest performance in **Table 3.28 (page 112)** due to its high sensitivity to data size and hyperparameters. As referenced in **Table 3.28,** the Transformer model shows the lowest performance due to its sensitivity to data size and hyperparameters. It requires large datasets to perform well, and its complexity, along with computational demands, may have hindered its performance, leading to slower convergence and potential overfitting or underfitting. These factors resulted in lower precision, F1 score, and AUC. we have updated this description under the table.
29	Provide a more detailed analysis of the evaluation results in Table 3.6.	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* **Page 95/ Chapter 3 / Section 3.2.4** Thank you for your valuable feedback. We have now included a short analysis of the results in **Table 3.6** to provide a clearer summary of the comparison between the proposed models (AI-LGBM and PSO-SCNN) and traditional models. The analysis highlights the superior performance of PSO-SCNN, particularly in Recall, and compares it with AI-LGBM and other advanced methods. Table 3.6 compares the performance of the proposed models (AI-LGBM and PSO-SCNN) with traditional models (Random Forest, ANN, LSTM, CNN) based on Accuracy, Precision, F1-score, and Recall. The PSO-SCNN outperforms all models, achieving perfect Recall (1.0000), making it highly effective for detecting poor water quality. AI-LGBM also performs strongly, particularly in Precision (95.00%) and Recall (94.00%). In comparison, RF, ANN, CNN, and LSTM show lower performance, especially in Recall and F1-score, with the proposed models providing superior overall results.
30	What are technical techniques to enhance the proposed methods, models? Classify.	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* **Page 85 / Chapter 2 / Section 2. 5** **Classification of Model Enhancement Techniques** Thank you for your valuable feedback. We have updated **Chapter 2, Section 2.5 (Page 85)** to

		provide a comprehensive overview of the technical techniques that can enhance the performance of the proposed models (AI-LGBM and PSO-SCNN). This updated section now includes classifications of various techniques aimed at improving model optimization, feature engineering, regularization, training methods, interpretability, and spatial-temporal modeling.
31	What is a methodological innovation? It should be explained clearly a methodological innovation for the study contributions.	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* **Page 6 / Chapter / Introduction/ Section** Contribution of the Thesis and Significance Thank you for your insightful feedback. We have updated the "**Contribution of the Thesis and Significance**" section in Section, Introduction to clearly highlight the methodological innovation. Specifically, we have emphasized the development of the PSO-SCNN model as the key innovation in addressing spatial dependencies for groundwater quality classification. Additionally, we have revised the limitations section to reflect key areas such as computational complexity, data quality, and the need for temporal dynamics in future **Contributions of the Thesis and Significance** This thesis makes significant contributions to the field of groundwater quality classification by introducing novel hybrid spatial machine learning models, specifically **AI-LGBM** and **PSO-SCNN**. These contributions address key challenges in groundwater monitoring, particularly spatial autocorrelation, model interpretability, and scalability. Below are the primary contributions: **Methodological Innovation: Hybrid Spatial Machine Learning Models**
32	What is technical issue of the proposed hybrid models (AI-LGBM, PSO-SCNN) to achieve up to 98.8% accuracy? Classify.	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* **Page 86 / Chapter 2 / Section 2.4** **Page 129 / Chapter 3 / Section 3.5** Thank you for your valuable feedback. The technical issues of the proposed hybrid models (AI-LGBM and PSO-SCNN) in achieving up to 98.8% accuracy are primarily related to computational complexity, data quality, and model interpretability. Both models are computationally intensive, requiring significant processing power, and rely heavily on high-quality data, which may not always be available. Additionally, despite efforts to enhance interpretability with tools like SHAP, the models still face "black-box" challenges. These aspects are addressed in **Chapter 2, Section 2.4 (Page 84)** under Classification of Model Enhancement Techniques and **Chapter 3, Section 3.5 (Page 129)**, where the models were validated using spatial cross-validation to avoid overestimation of performance due to spatial autocorrelation. **The 98.8% accuracy was achieved through optimized hyperparameters and advanced feature selection techniques.**
33	Label all equations	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* **Page 13 – 15 / Chapter 1 / Section 1.1** Thank you for your valuable feedback.

		We have ensured that all equations in the thesis, including those in Chapter 1, Section 1.1 **(Pages 13–15),** are now properly labeled for clarity and consistency. $$W^* = \arg\min_W -\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{K} \mathbf{1}(y_i = c)\log p_W(y = c \mid x_i) \quad (1.7)$$ $$W^* = \arg\min_W \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(y_i, f(x_i; W)) \quad (1.8)$$
34	Data sets in the experiments should be explained in details in order to figure out an appropriate model selection.	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* **Page 37 / Chapter 1 Section 1.5 data Sources** **Page 43- 45 / Chapter 1 / Section 1.5.1, 1.5.2** Thank you for your valuable feedback. We have ensured that the datasets used in the experiments are thoroughly explained. Specifically, **Page 37-45 in Chapter 1, Section 1.5 and Page 43 – 45 section 1.5.1** is dedicated to detailing the datasets, including their sources, features, and preprocessing steps, to provide a clear understanding of the data used for model selection and training. **1.5.1 Detailed Structure of the Experimental Datasets** A precise description of the datasets, Table 1.18 summarizes the attributes available in the Vietnam and India groundwater collections. Table 1.18: Attribute comparison between Vietnam and India datasets **Category** | **Vietnam Dataset** | **India Dataset** Administrative information | Well code, sampling date, quarter, laboratory | District, village Spatial information | Latitude, longitude (external coordinate file) | Not always explicit; location described by village Basic water indicators | pH, conductivity, TDS | pH, EC, TDS Major cations | Na, K, Ca^{2+}, Mg^{2+} | Calcium, Magnesium, Sodium, Potassium Major anions | Cl^-, SO_4^{2-}, HCO_3^-, CO_3^{2-}, NO_3^-, NO_2^- | Chloride, Sulphate, Bicarbonate, Carbonate, Fluoride Hardness measures | General, temporary, permanent hardness | Total hardness Additional environmental features | Fe^{2+}, Fe^{3+}, NH_4, PO_4, silica, dissolved gases | Alkalinity Temporal availability | Multi-year monitoring | Mostly cross-sectional Data dimensionality | High (more than 35 variables) | Moderate (around 15–17 variables) Primary system role | Model training and spatial learning | Cross-region validation and robustness testing **1.5.2 Formal Label Generation Procedure** This section presents the computational process used to derive training labels. Label assignment follows QCVN 01-1:2018/BYT and WHO, which provides legally binding threshold values for drinking water safety in Vietnam. Therefore, the classification outcome is derived from regulatory compliance rather than subjective judgment. **Input** For each groundwater sample i, the raw measurement vector is:
35	The thesis uses the term "ensemble spatial machine learning" but does not clearly define the nature of the ensemble strategy (stacking, voting, or simply comparing models in parallel), which reduces the methodological rigor.	*(Ghi rõ số trang/*chương*/mục... đã được chỉnh sửa)* **Page 52 / Chapter 2/ Section 2.1.2 Ensemble Mechanism and model integration** Thank you for your valuable feedback. We have updated **Page 52, Chapter 2, Section 2.1.2,** Ensemble Mechanism and Model Integration to clearly define the nature of the ensemble strategy used in the thesis. The updated section now specifies that the ensemble strategy is **based on late fusion**, where the

		outputs of the AI-LGBM and PSO-SCNN models are combined. This revision addresses the concern and enhances the methodological rigor of the thesis. **2.1.2 Ensemble Mechanism and Model Integration** In this thesis, the term *ensemble* is used in two complementary senses: **(i) Model-level ensemble:** AI-LGBM is an ensemble by construction because it follows the gradient boosting methodology, where the final predictor is an additive combination of many decision trees. For binary drinkability prediction, it outputs a probability: $$\hat{p}_{\text{LGBM}}(y=1 \mid x) = \sigma\left(\sum_{m=1}^{M} f_m(x)\right), \quad (2.1)$$ where $f_m(\cdot)$ is the m-th tree, M is the total number of trees, and $\sigma(\cdot)$ is the sigmoid function (for multi-class classification, additive scores are computed per class and normalized with softmax). **(ii) System-level ensemble:** The overall framework integrates two complementary learners: **AI-LGBM** (which uses tabular hydrochemical variables and point-based spatial features) and **PSO-SCNN** (which captures spatial-context learning for consistent mapping). When both models are employed together, their outputs are combined using **late fusion**. This strategy involves combining the predicted probabilities of the two models as follows: $$\hat{p}_{\text{ens}} = \alpha\,\hat{p}_{\text{LGBM}} + (1-\alpha)\,\hat{p}_{\text{SCNN}}, \quad \alpha \in [0,1], \quad (2.2)$$
36	The experimental results need to be compared with the results of related studies. The thesis should focus more on analyzing the problem characteristics, avoiding a broad presentation of basic machine learning concepts.	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* **Page 134/ Chapter 3 / Section 3.5.5** Thank you for your valuable feedback. We have updated Page **134, Chapter 3, Section 3.5.5,** "Comparison with Related Studies" to provide a more detailed comparison of our experimental results with those from related studies. This section now includes relevant benchmarks and a discussion on how our proposed models (AI-LGBM and PSO-SCNN) perform in comparison to other approaches used for groundwater quality classification. **3.5.5 Comparison with Related Studies** This section compares the results of the **AI-LGBM** and **PSO-SCNN** models with related studies in groundwater quality classification. Our models achieved **98.8% accuracy**, outperforming previous studies in several metrics. Table 3.43 compares our results with studies by **Singh (2017)**, **Kumar (2019)**, and **Zhang (2020)**. **Study** | **Model** | **Accuracy** | **Spatial Data** | **Limitation** *Singh, R. (2017)* | WQI + Statistical | 85% | None | Time-consuming, lacks spatial learning *Kumar, P. (2019)* | ANN | 88% | No spatial data | Limited interpretability *Zhang, Y. (2020)* | DNN, CNN | 92% | No spatial context | Ignores spatial autocorrelation This Thesis | AI-LGBM | 98.8% | Yes | Computational complexity This Thesis | PSO-SCNN | 98.8% | Yes | Computational cost Table 3.43: Comparison of Model Performance
37	It is recommended to add an analysis of the computational complexity of the proposed models (AI-LGBM and PSO-SCNN) and to add a development direction for optimizing computational efficiency by using lightweight DL architectures or optimization techniques that make them more accessible in resource-limited environments.	*(Ghi rõ số trang/chương/mục... đã được chỉnh sửa)* **Page 133, 134 / Chapter 3 / Section 3.5.4** Thank you for your valuable feedback. We have **updated Page 133, Chapter 3, Section 3.5.4** to include a detailed analysis of the computational complexity of the proposed models (AI-LGBM and PSO-SCNN). This section now also discusses optimization techniques for improving computational efficiency, including potential directions for leveraging lightweight deep learning architectures and methods suitable for resource-limited environments.

	focused on experimental results, lacking in-depth academic discussion; it does not clearly explain the generalization when applied to areas with different geological conditions.	**Page 136 / Chapter 3/ Section 3.6.4** Thank you for your feedback. We have updated **Page 136, Chapter 3, Section 3.6.4** to address the generalization of the proposed models (AI-LGBM and PSO-SCNN) when applied to regions with different geological conditions. This section now includes a more in-depth discussion on the model's ability to generalize across regions with varying hydrogeological characteristics, as well as the steps taken to ensure robustness in different geographical and environmental contexts.

Nghiên cứu sinh chân thành cảm ơn Quí thầy, cô trong Hội đồng đánh giá luận án tiến sĩ cấp Học viện đã góp ý và tạo cơ hội cho NCS hoàn thiện luận án của mình.

Xin trân trọng cảm ơn./.

Hà Nội, ngày 25 tháng 02 năm 20...26

TẬP THỂ HƯỚNG DẪN

(Trường hợp có 02 người hướng dẫn xin chữ ký cả 02 người, ký và ghi rõ họ tên)

Trần Thị Ngân

Nguyễn Long Giang

NGHIÊN CỨU SINH

Michael Omer

CHỦ TỊCH HỘI ĐỒNG

Nguyễn Đức Dũng

THƯ KÝ HỘI ĐỒNG

Trần Đức Nghĩa

XÁC NHẬN CỦA HỌC VIỆN
KHOA HỌC VÀ CÔNG NGHỆ

KT. GIÁM ĐỐC
PHÓ GIÁM ĐỐC

VIỆN HÀN LÂM KHOA HỌC VÀ CÔNG NGHỆ VIỆT NAM
HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

Nguyễn Thị Trung