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

SUMMARY OF DISSERTATION ON INFORMATION SYSTEM

MAJOR: Information Systems

CODE: 9 48 01 04

Hanoi - 2025

The Dissertation is completed at: Graduate University of Science and Technology, Vietnam Academy of Science and Technology.

Supervisors:

1. Supervisor 1: Assoc. Prof. NGUYỄN LONG GIANG, VAST
2. Supervisor 2: Assoc. Prof. TRẦN THỊ NGÂN, VNUIS

Referee 1: _____

Referee 2: _____

Referee 3: _____

The dissertation is examined by Examination Board of Graduate University of Science and Technology, Vietnam Academy of Science and Technology at..... (time, date.....)

The dissertation can be found at:

1. Graduate University of Science and Technology Library
2. National Library of Vietnam

INTRODUCTION

1. Urgency of the Thesis : Southeast Asia's groundwater is under pressure from population growth, urbanization, and climate change. This thesis focuses on drinkability classification using ML/DL (PSO-SCNN, CNN-GIS, AI-LGBM) and GIS for improved accuracy.

2. Research Objectives of the Thesis :Enhance groundwater classification in Vietnam and Odisha using AI-LGBM, PSO-SCNN, and CNN-GIS, benchmarking for higher accuracy and generalization.

3. Research Subjects and Scope : Focus on Mekong Delta and Odisha with physicochemical and spatial data, split 70/15/15 (train/val/test) and labeled as "Excellent," "Good," "Poor." Validation via k-fold and baseline comparison.

4. Methodology and Research Content : Develop and compare AI-LGBM, PSO-SCNN, and CNN-GIS with DT/SVM/RF using datasets from Vietnam and India; evaluate with accuracy, precision, recall, F1, AUC, and map outputs in GIS.

5. Contributions of the Thesis : Present AI-LGBM, CNN-GIS, and PSO-SCNN models optimized with spatial clustering and hyperparameter tuning; integrate GIS for groundwater quality mapping in Odisha and Mekong Delta.

6. Layout of the Thesis :The structure of the thesis includes Introduction, three chapters and Conclusion. In which,

Chapter 1: Groundwater Drinkability Classification and Background knowledge.

Chapter 2 presents proposed Ensemble Spatial Machine Learning Methods.

Chapter 3 shows the results of AI-LGBM, PSO-SCNN improves robustness (ANOVA), CNN–spatial maps risk; system architecture.

Chapter 1

Groundwater Drinkability Classification and Background knowledge

1.1 Groundwater Drinkability classification

Context and Motivation. Groundwater sustains billions but faces risks from heavy metals, nitrates, and pesticides. Traditional assessments are slow and costly, while AI promises timely, scalable classification yet still struggles with accuracy, scalability, and interpretability. This work focuses on three problems: multi-class drinkability classification, robust hyperparameter optimization, and spatial visualization for decision support.

Problem 1: Groundwater Drinkability Classification

Goal. Classify each sample as *Excellent*, *Good*, *Moderate*, *Poor*, or *Unsuitable for Drinking* using physicochemical and spatial features (e.g., pH, TDS, nitrate, latitude, longitude).

Formulation. Let $X = \{x_i\}_{i=1}^{n}$, $x_i \in \mathbb{R}^m$, and labels $y_i \in \{1, \ldots, k\}$. A model $f(\cdot\,; W)$ outputs class scores; the predicted class is

$$\hat{y}_i = \arg \max_{c \in \{1,\ldots,k\}} f_c(x_i; W).$$

We train by minimizing empirical risk

$$\min_{W} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i, f(x_i; W)),$$

with $\mathcal{L}$ typically multi-class cross-entropy. Evaluation uses accuracy, precision, recall, F1, and AUC. Unlike fixed-threshold WQI methods, the model learns nonlinear relationships and scales to large, heterogeneous datasets.

Problem 2: Hyperparameter Optimization for GWQC Models

Goal. Select hyperparameters (e.g., tree depth, learning rate, estimators, regularization) that maximize out-of-sample performance while controlling compute cost.

Approach. Use black-box search via AIO , Optuna and Particle Swarm Optimization (PSO) over a search space $\mathcal{W}$. Let $g(W)$ be a cross-validated score (e.g., macro-F1). The optimizer solves

$$W^* = \arg\max_{W \in \mathcal{W}} g(W),$$

optionally with resource-aware constraints (e.g., time or FLOPs budgets). This improves accuracy, stability, and generalization on diverse data from Vietnam and India, including noisy and high-dimensional settings.

Problem 3: Spatial Visualization of Classified Labels

Goal. Map predictions to geography for risk communication and planning.

Formulation. Let $G = \{(lat_i, lon_i)\}_{i=1}^n$ be sample coordinates and $\hat{y} = \{\hat{y}_i\}_{i=1}^n$ model outputs. A GIS pipeline produces a thematic map

$$M = \mathrm{GIS}(G, \hat{y}),$$

optionally using interpolation or areal aggregation. To couple classification and spatial coherence, we consider a composite objective

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

where $L_{\text{classification}}$ is cross-entropy and L_{spatial} penalizes implausible spatial discontinuities or misalignment with known spatial priors; $\lambda > 0$ tunes this trade-off. The output is an interpretable map of drinkability classes, highlighting hotspots and priority zones for monitoring.

Contribution and Impact. The pipeline replaces rigid WQI thresholds with a flexible, data-driven classifier; uses principled hyperparameter search to ensure reliable deployment; and delivers spatial products that support policy decisions. Together these components enable faster, scalable, and region-aware groundwater quality assessment..

1.2 Literature Review

1.2.1 Classical Methods

Traditional groundwater quality methods are labor-intensive and rely on manual sampling and analysis. The Water Quality Index (WQI) offers simple classification but is limited by subjectivity and expert thresholds.

Limitations of Classical Approaches

Classical water quality methods are slow, subjective, and lack real-time data. Key gaps include addressing non-linearity and improving data integration for real-time analysis.

1.2.2 Machine Learning (ML) Methods

ML methods like SVM, RF, and LightGBM handle large datasets and non-linear patterns, improving accuracy. However, overfitting, data quality, and interpretability remain challenges.

1.2.3 Deep Learning (DL) Methods

DL models (CNNs, RNNs) excel in automatic feature extraction and handling large datasets, but require high computational power and large datasets, with limited interpretability.

1.2.4 Hybrid Spatial Machine Learning Models

Hybrid models combining ML, DL, and GIS enhance groundwater classification by utilizing spatial data for regional variations and real-time insights.

Spatial Clustering and GIS Integration

Spatial clustering (e.g., K-means, DBSCAN) and ML improve classification by capturing spatial patterns, with GIS integrating spatial data. DL models like LightGBM and CNNs improve accuracy but require significant computational resources.

A Hybrid RainNet and GA Model for Hyperparameter Tuning

The RainNet and Genetic Algorithm (GA) hybrid model reduces MAE compared to models like Unet and Segnet, improving rainfall prediction accuracy.

1.3 Limitations and Research Gaps

Despite advances in ML and DL for groundwater quality classification, challenges like data sparsity, overfitting, and interpretability persist. Future research should focus on improving model generalization, reducing computational costs, and enabling real-time monitoring, while exploring hybrid approaches for better robustness and scalability.

1.4 Conclusion

This chapter reviewed groundwater quality classification methods, from traditional approaches to advanced ML and DL techniques. While traditional methods are limited in capturing complex patterns, ML and DL improve accuracy and handle large datasets. Hybrid models combining spatial data with ML and DL offer a promising solution, but challenges like overfitting and interpretability remain, requiring further research.

Chapter 2

Proposed Ensemble Spatial Machine Learning Methods

This chapter introduces ensemble spatial machine learning for groundwater quality classification, centered on two models: an AI-enhanced Light Gradient Boosting Machine (AI-LGBM) and a Particle Swarm Optimization–Spatial Convolutional Neural Network (PSO-SCNN).

2.1 AI-LGBM

2.1.1 Overview of the Proposed AI-LGBM Framework

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.

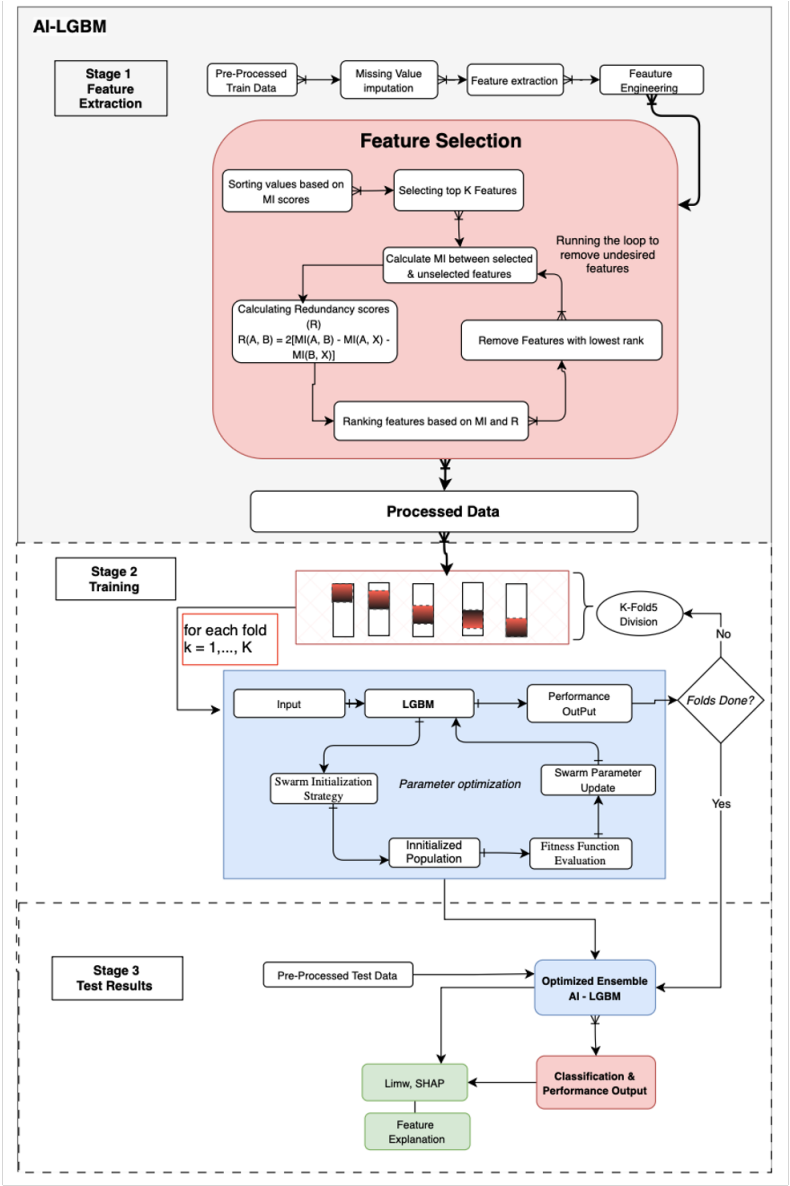


Figure 2.1: Proposed AI-LGBM Methodological Flowchart

Mathematical Formulation of AI-LGBM with MIFS

Setup. Given samples $X = \{x_i\}_{i=1}^{n}$ with $x_i \in \mathbb{R}^m$ (physicochemical + spatial features) and labels $y_i \in \{1, \ldots, k\}$, the AI-LGBM model $f(\cdot; W)$ produces class scores. Prediction and training:

$$\hat{y}_i = \arg\max_c f_c(x_i; W), \qquad \min_W \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\big(y_i, f(x_i; W)\big).$$

Performance / hyperparameters. Select hyperparameters by cross-validated score $g(W)$ (e.g., macro-F1, AUC):

$$W^* = \arg\max_W g(W).$$

We use a hybrid search with **Optuna** (surrogate/TPE proposals), **PSO** (swarm refinement), and **AIO** (adaptive mutation):

$$W^{(s)} \sim \pi_\phi(W \mid \mathcal{H}_{s-1}), \qquad v_{t+1} = \omega v_t + c_1 r_1 (pbest - W_t) + c_2 r_2 (gbest - W_t), \quad W_{t+1} = W_t + v_{t+1},$$

$$W_{t+1} \leftarrow W_{t+1} + \eta\, \mathcal{A}(W_{t+1};\, \mathcal{H}_t),$$

with early stopping and K-fold CV to ensure stable generalization.

Feature selection (MIFS). Rank features by mutual information with the label and control redundancy; select k features S_k by

$$S_k = \arg\max_{S:\,|S|=k} J(S), \quad J(S) = \sum_{x_j \in S} I(x_j; Y) \;-\; \lambda \sum_{\substack{x_j, x_\ell \in S \\ j<\ell}} I(x_j; x_\ell),$$

where $I(\cdot\,;\cdot)$ is mutual information. (Equivalently, $I(X;Y) = H(X) + H(Y) - H(X,Y)$.)

Summary objective. MIFS reduces dimensionality before training; AI-LGBM then optimizes W to minimize loss and maximize $g(W)$ under cross-validation.

Mathematical Foundations

$$\text{Classification: } \min_W \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\big(y_i, f(x_i; W)\big), \quad \hat{y}_i = \arg\max_c f_c(x_i; W). \tag{2.1}$$

$$\text{Hyperparameters: } W^* = \arg\max_W g(W) \text{ (e.g., macro-F1/AUC via CV).} \tag{2.2}$$

$$\text{Feature selection: } S_k = \arg\max_{S:\,|S|=k} J(S). \tag{2.3}$$

Hypotheses (AI-LGBM + MIFS)

$$H_0 : \mathbb{E}[g(\text{AI-LGBM+MIFS})] = \mathbb{E}[g(\text{Baselines})],$$

$$H_1 : \mathbb{E}[g(\text{AI-LGBM+MIFS})] > \mathbb{E}[g(\text{Baselines})].$$

2.1.2 Experimental Setup and Learning Strategy for AI-LGBM

Stratified 70/15/15 split on groundwater data; preprocessing: imputation, Z-score normalization, IQR outliers. Supervised: MIFS selection, SMOTE balancing, LightGBM with Optuna/AIO optimization (5-fold CV, max weighted F1). Metrics: accuracy, precision, recall, F1, AUC; SHAP interpretability.

Mathematical Formulation

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, $x_i \in \mathbb{R}^m$, $y_i \in \{1, \ldots, K\}$.

Feature Selection (MIFS). Select top-K features: $S^\star = \arg\max_{S \subset \{1,\ldots,m\}, |S|=K} \sum_{j \in S} \mathcal{I}(X_j; Y)$, then $X \leftarrow X_{S^\star}$.

SMOTE Balancing. For minority x_i and kNN neighbor $x_i^{(nn)}$: $\tilde{x} = x_i + \lambda(x_i^{(nn)} - x_i)$, $\lambda \sim \mathcal{U}(0,1)$, $\tilde{y} = y_i$, yielding $\mathcal{D}_{\text{train}}^{\text{smote}}$.

Boosted Additive Model. LightGBM fits $F_m(x) = F_{m-1}(x) + \eta \sum_{j=1}^{J_m} \gamma_{jm} \mathbb{I}(x \in R_{jm})$, with $\gamma_{jm} = \arg\min_\gamma \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma)$.
Multiclass loss: $\ell_i = -\sum_{k=1}^{K} y_{ik} \log p_{ik}$, $\mathcal{L} = \sum_i \omega_i \ell_i$. Leaf update: $w_j^\star = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}$, Gain as in original.

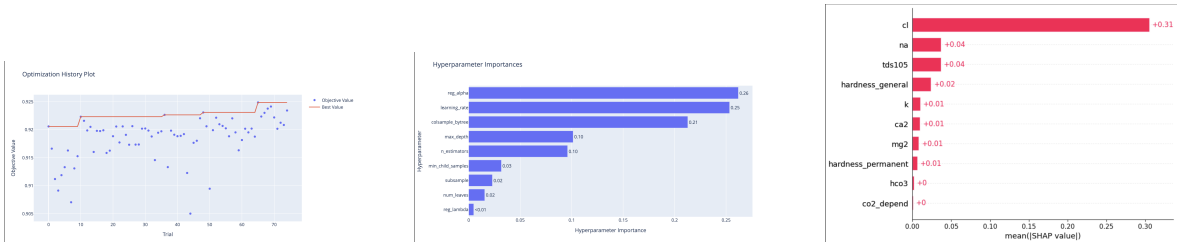
Hyperparameter Optimization. Optimize $\theta^* = \arg\max_\theta \frac{1}{K} \sum_{k=1}^{K} \text{F1-score}_k(\theta)$.

2.1.3 Model Optimization and Performance

AI–LGBM was tuned with AIO/Optuna under 5-fold CV to (*learning_ rate*=0.05, *num_ leaves*=32, *max_ depth*=8, *n_ estimators*=150, *subsample*=0.8, *colsample_ bytree*=0.7). Compared with the default, accuracy rose from 0.812 to 0.865 and weighted F1 from 0.801 to 0.864 (7.9%), with similar gains in precision and recall.

2.1.4 Feature Importance and Visualization

Optuna history/importance in Figs. 2.2a–2.2b. SHAP highlights tds105, na, cl (Fig. 2.2c).



(a) Optuna History (b) HP Importance (c) SHAP Summary

Pros. High accuracy, handles high-dimensional data, models linear/nonlinear relations.

Cons. Computationally expensive, requires tuning, limited interpretability (mitigated by SHAP).

2.2 PSO-SCNN

2.2.1 Overview of PSO-SCNN Framework

Proposes PSO-SCNN, a hybrid DL model building on AI-LGBM, to capture geospatial dependencies via spatial embeddings, Haversine encoding, multi-head attention, and CNNs.

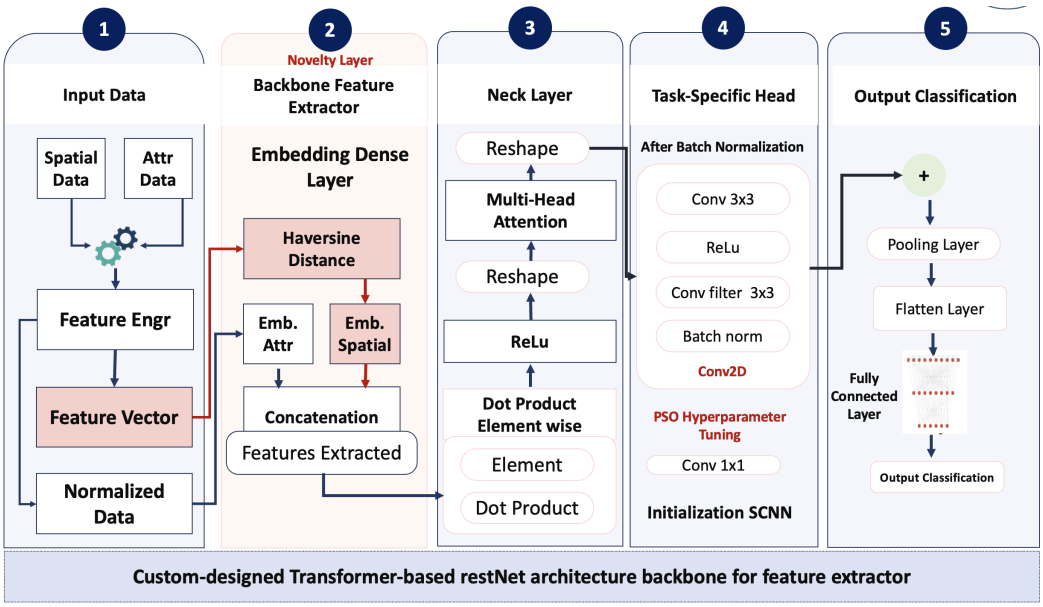


Figure 2.3: PSO-SCNN Spatial Model Architecture

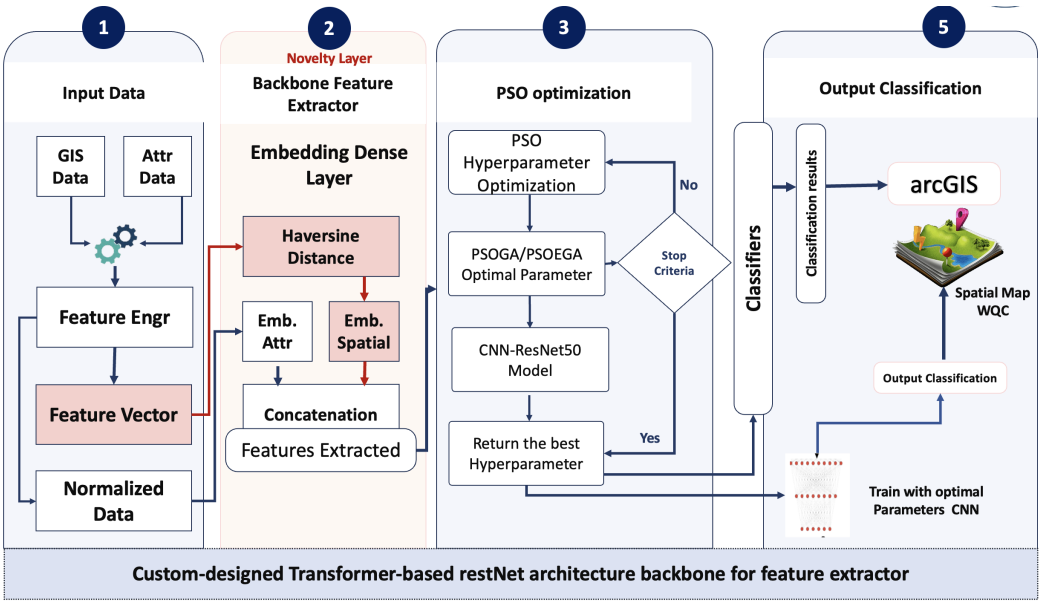


Figure 2.4: Extended for Spatial Map Visualization

2.2.2 Mathematical Formulation of PSO-SCNN and CNN-GIS

PSO-SCNN Formulation

The PSO-SCNN model optimizes a Spatial Convolutional Neural Network (SCNN) using Particle Swarm Optimization (PSO) to classify groundwater quality, incorporating spatial data for enhanced geospatial dependency modeling. Let $\mathcal{D} = \{(x_i, y_i, (lat_i, lon_i))\}_{i=1}^{n}$ be the dataset, where $x_i \in \mathbb{R}^m$ are physicochemical features, $y_i \in \{0, 1\}$ is the binary drinkability label, and (lat_i, lon_i) are coordinates.

Preprocessing and Spatial Encoding. Features are normalized: $x_i^* = \frac{x_i - \mu_x}{\sigma_x}$. Spatial features are encoded via Haversine distance from the centroid $(\bar{lat}, \bar{lon})$:

$$d_i = 2R \arcsin\left(\sqrt{\sin^2\left(\frac{lat_i - \bar{lat}}{2}\right) + \cos(\bar{lat})\cos(lat_i)\sin^2\left(\frac{lon_i - \bar{lon}}{2}\right)}\right), \quad (2.4)$$

yielding augmented inputs $\tilde{x}_i = [x_i^*; d_i]$. SMOTE balances classes by generating synthetic $\tilde{x}_j$ for minorities.

PSO Optimization. PSO searches hyperparameters $\theta = \{\text{filters}, \text{kernel size}, \text{learning rate}\}$ in swarm $\{p_k\}_{k=1}^{P}$. Fitness is negative AUC: $\text{Fit}(p_k) = -\text{AUC}(\text{SCNN}_\theta)$. Updates:

$$v_k^{t+1} = wv_k^t + c_1 r_1 (pbest_k - p_k^t) + c_2 r_2 (gbest - p_k^t), \quad (2.5)$$

$$p_k^{t+1} = p_k^t + v_k^{t+1}, \tag{2.6}$$

where w is inertia, c_1, c_2 are coefficients, $r_1, r_2 \sim \mathcal{U}(0,1)$, converging to optimal θ^*.

SCNN Architecture. SCNN processes $\tilde{x}_i$ through convolutional layers: $h_l = \sigma(W_l * h_{l-1} + b_l)$, pooling, and dense layers, outputting $\hat{y}_i = \sigma(W_f h_L + b_f)$. Trained with binary cross-entropy: $\mathcal{L} = -\sum_i [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)]$.

CNN-GIS Formulation

CNN-GIS extends PSO-SCNN for geospatial visualization, mapping predictions $\hat{y}_i$ to geographic coordinates (lat_i, lon_i) via GIS integration. The goal is to generate a thematic map M highlighting quality classes and hotspots.

Spatial Prediction Mapping. Predictions are interpolated over a grid $G = \{(lat_g, lon_g)\}_{g=1}^{G}$ using inverse distance weighting (IDW):

$$\hat{y}(lat_g, lon_g) = \frac{\sum_i w_i \hat{y}_i}{\sum_i w_i}, \quad w_i = \frac{1}{d((lat_g, lon_g), (lat_i, lon_i))^p}, \tag{2.7}$$

where $d(\cdot)$ is Haversine distance and $p > 0$ controls decay. The composite loss couples classification with spatial regularization:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{class}} + \lambda \sum_{i,j} \|\hat{y}_i - \hat{y}_j\| \cdot d((lat_i, lon_i), (lat_j, lon_j))^{-1}, \tag{2.8}$$

ensuring spatial coherence. Outputs are exported to GeoTIFF for ArcGIS visualization of drinkability hotspots.

2.2.3 Rationale for Hybrid Model Selection and Evaluation Criteria

Hybrids (AI-LGBM, PSO-SCNN) chosen for accuracy, interpretability, and scalability in spatial groundwater classification; evaluation focuses on robustness, efficiency, and utility, outperforming traditional methods.

2.2.4 Learning Strategy of PSO-SCNN

The PSO-SCNN learning strategy integrates initialization, optimization, spatial feature extraction, and training/validation. PSO initializes a swarm of particles for hyperparameter search, optimizing to minimize error. SCNN extracts spatial features from groundwater data. Training uses K-fold cross-validation for robustness, with feature fusion of physicochemical, categorical, and spatial embeddings: $FV_5 = \text{concat}(FV_1, FV_4)$, where FV_4 is from SCNN layers.

Supervised Setup. Objective: $\theta^* = \arg\max_\theta \frac{1}{K} \sum_{k=1}^{K} \text{F1}_\text{w} \left(f_\theta^{(-k)}, \mathcal{D}^{(k)} \right)$.

PSO Loop. Updates particles via velocity/position equations, evaluating on validation AUC to select θ^*.

Training. Uses Adam with early stopping on validation F1 for generalization and spatial mapping.

2.2.5 Comparison of Learning Algorithms

Optimizer choice impacts convergence. Table 2.1 compares Adam, AdamW, and AdaGrad. Adam is selected for efficiency, adaptive rates, and minimal tuning, ideal for SCNN on high-dimensional groundwater data. AdamW suits large-scale models; AdaGrad for sparse data but may converge slowly.

Table 2.1: Comparison of Optimizers

Optimizer	Speed	Adaptivity	Generalization	Tune Need	Use Case
Adam	Fast	Yes	Very Good	Low	Deep networks
AdamW	Fast	Yes	Excellent	Low	Large-scale models
AdaGrad	Medium	Yes	Good early	Medium	Sparse data

Table 2.2: Key PSO-SCNN Hyperparameters

Hyperparameter	Description	Values
Particle Size	Swarm size	10-50
Inertia Weight	Previous velocity impact	0.5-0.9
C1/C2	Personal/global influences	1.5-2.0
Max Iterations	PSO iterations	50-200
Kernel Size	Convolution kernel	3×3, 5×5
Stride	Convolution stride	1-2

2.2.6 Impact of PSO Hyperparameters

PSO parameters balance exploration/exploitation. Used: $n_{\text{particles}} = 3$, $w = 0.9$, $c_1 = 0.5$, $c_2 = 0.3$. Table 2.3 shows impacts on performance.

Table 2.3: PSO Parameter Effects on PSO-SCNN

Config	w	**AUC**	**F1**	**Convergence**
High w (Explor.)	0.9	0.965	0.945	Slow
Balanced (Study)	0.9	0.988	0.965	Moderate
Low w, High c_2	0.4	0.972	0.950	Fast, Risky

Higher w aids exploration but slows convergence; adaptive strategies enhance robustness.

2.2.7 Pros and Cons

Pros: Handles spatial data well, optimizes for stability, models complex dependencies. *Cons:* Computationally intensive, resource-heavy for large datasets, challenging interpretability.

2.3 Chapter Conclusion

This chapter combines AI-LGBM (MIFS, SMOTE, AIO+Optuna, SHAP) with PSO-SCNN (spatial embeddings, Haversine, PSO optimization via AUC).

Key Contributions

- Synergy of tabular ensembles and spatial DL.
- Dual optimization for generalization.
- SHAP and spatial visuals for explainability.

Trade-offs and Limitations High compute cost, tuning complexity, spatial interpretability challenges; mitigated by early stopping, dynamic PSO, compression, uncertainty.

Outlook Next chapter evaluates results, ablations, ANOVA, and spatial visualizations for accuracy and robustness on Vietnam/India data.

Chapter 3

Results and Evaluations

3.1 Performance Evaluation and Comparison

Traditional ML: results in 3.1 are published in *Earth Science Informatics*, 16(2), 1701–1725. Springer. [DOI: https://doi.org/10.1007/s12145-023-00977-x].

Table 3.1: Performance Metrics for Various Models in Odisha Dataset

Model	**Avg. Accuracy**	**Avg. Precision**	**Avg. F1-Score**	**Avg. Recall**
Logistic Regression	0.7051	0.72	0.6275	0.6025
Polynomial SVM	0.9012	0.9175	0.9025	0.8925
Decision Tree	0.8989	0.885	0.8900	0.8850
AdaBoost	0.5445	0.465	0.4950	0.4650
CNN	0.9766	0.9877	0.9877	0.9877
AI-LGBM	0.94	0.95	0.92	0.93

Table 3.2: Performance Metrics for Various Models in Vietnam Dataset

Model	**Avg. Accuracy**	**Avg. Precision**	**Avg. F1-Score**	**Avg. Recall**
Logistic Regression	0.9672	0.5333	0.5517	0.5714
Polynomial SVM	0.9766	0.9950	0.9926	0.9950
Decision Tree	0.9696	0.9877	0.9889	0.9877
AdaBoost	0.9696	0.9901	0.9877	0.9901
CNN	0.9766	0.9877	0.9913	0.9877
AI-LGBM	0.94	0.95	0.92	0.93

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

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**
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 with Baseline and advance models

Table 3.4: 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 3.5: 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

Conclusion of AI-LGBM

Re-evaluated on Odisha and Vietnam, AI–LGBM consistently outperformed KNN, SVM, Decision Trees, and XGBoost across accuracy, precision, recall, and F1. Against deep models (MLP/CNN/Transformer) on a Kaggle set and the Vietnam set, AI–LGBM led on F1 and recall, achieving 99.53% accuracy and 0.0191 log loss on Vietnam.

3.2 Validation of PSO–SCNN

Validated PSO–SCNN using accuracy, precision, recall, and F1.

Optimizer comparison. Grid Search achieved 1.0000 accuracy in 4.56 s; PSO reached 0.9948 in 3.70 s; GA matched 0.9948 but took 11.54 s—PSO offers the best speed–accuracy trade-off, Grid Search the peak accuracy, GA the slowest.

3.2.1 PSO-SCNN Performance Results

Results in Sec. 3.2.1 accepted to *Proc. ICIIT 2025* (Hanoi; in press); hybrid method submitted to *Journal of the Indian Society of Remote Sensing* (SCIE, IF 2.2).

Table 3.6: *Model Performance Vietnam (Testing Set)*

Model	**Precision**	**Recall**	**Accuracy**	**F1 Score**	**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
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 3.7 compares proposed models (AI-LGBM, PSO-SCNN, and CNN-GIS) against conventional machine learning models.

Table 3.7: Comparison of Proposed Models with Baseline Models

Model	**Avg. Accuracy**	**Avg. Precision**	**Avg. F1-Score**	**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 Mapping (proposed)	**0.9700**	**0.9650**	**0.9750**	**0.9800**

3.2.2 Appended (Post- Optimization) PSO-SCNN Results

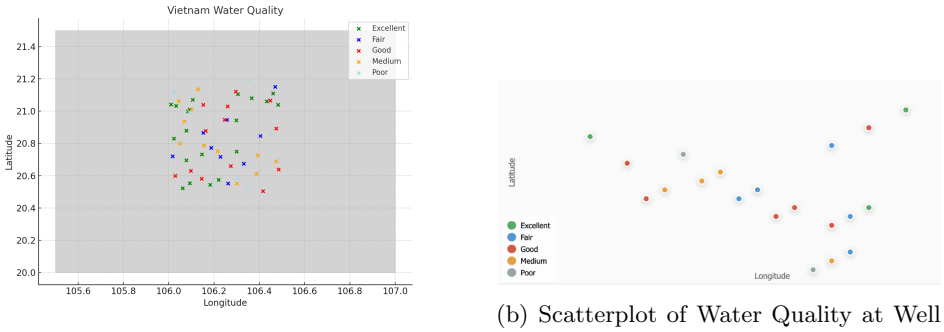
Post-optimization, the PSO-SCNN was re-evaluated on both datasets, yielding notable gains in precision, recall, F1 score, and AUC.

Table 3.8: *Advance Model Performance Vietnam (Testing Set)*

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 3.9: 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



(a) Vietnam - Mekong Region

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

Figure 3.1: Vietnam and Odisha Water Quality Visualizations

Training and Validation Performance

The figures below show the training/validation loss, accuracy, and comparison with baselines, indicating effective training and good generalization.

Table 3.11: 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

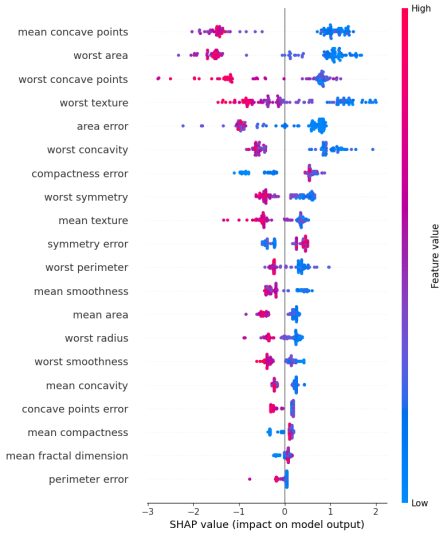


Figure 3.2: SHAP Summary Plot for AI-LGBM Model

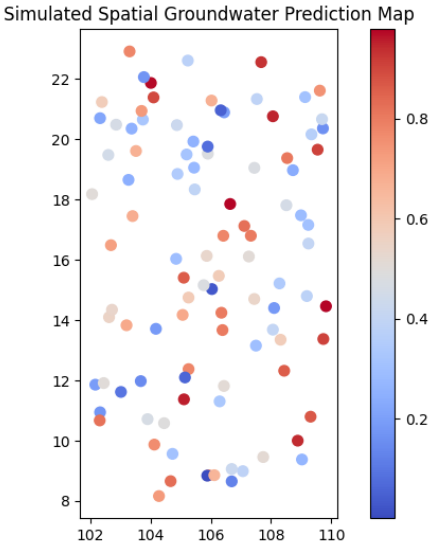
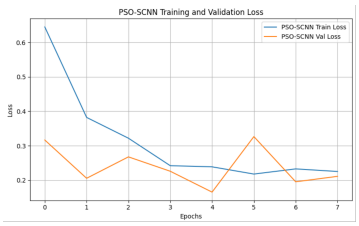
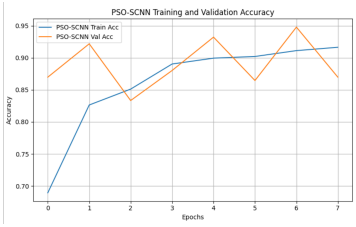


Figure 3.3: Overlay of predicted unsafe zones with actual contamination areas

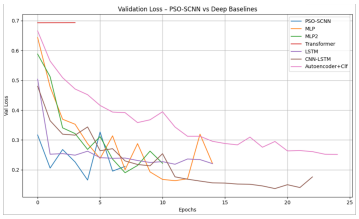
Figure 3.4: SHAP Feature Importance and Spatial Contamination View



(a) PSO-SCNN Training and Validation Loss



(b) PSO-SCNN Training and Validation Accuracy



(c) Validation Loss Comparison - PSO-SCNN vs Deep Baselines

Figure 3.5: Performance Evaluation: Training/Validation Loss, Accuracy, and Comparison

Table 3.10: 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.

Table 3.12 presents the quantitative results of the ablation study, summarizing the precision, recall, F1 score, AUC, and training time for each model variant.

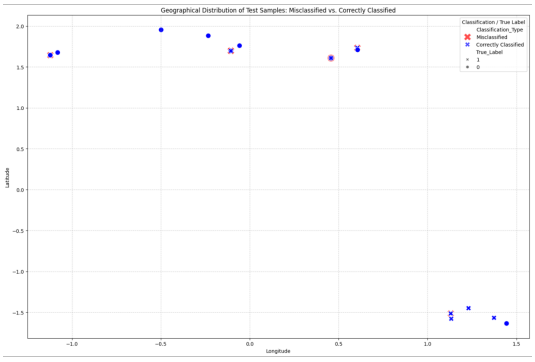
Table 3.12: Ablation Study: Quantitative Impact of Components removal

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

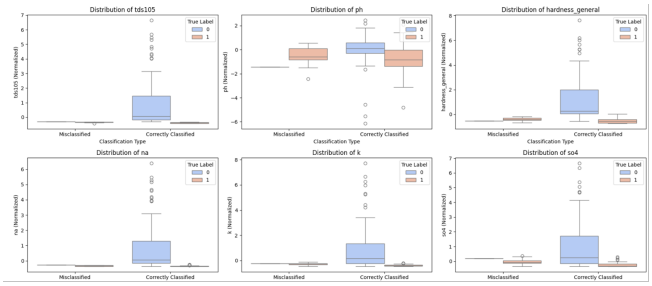
Table 3.13: Training Time and Memory Consumption Comparison for AI-LGBM and PSO-SCNN Models

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

Feature Ranges Where Models Underperform



(a) Misclassification Hotspots (PSO-SCNN)



(b) Feature Distribution for Misclassified vs Correctly Classified Samples

Figure 3.6: Feature Range, Confusion Matrix, Misclassification Hotspots, and Feature Distribution

Section Associated Publications

Peer-reviewed outputs include CNN–GIS optimization in *Proc. ICIIT 2025* and PSO–SCNN in *Journal of the Indian Society of Remote Sensing.*

Chapter Conclusion

AI–LGBM: VN ≥98%, Odisha 92–93%, Prec >0.92, Rec >0.90, F1 >0.91; AIO/Optuna improved F1 by 15–20%.

PSO–SCNN: Superior F1 on spatial tasks; PSO improved convergence by 25–30%, reduced overfitting.

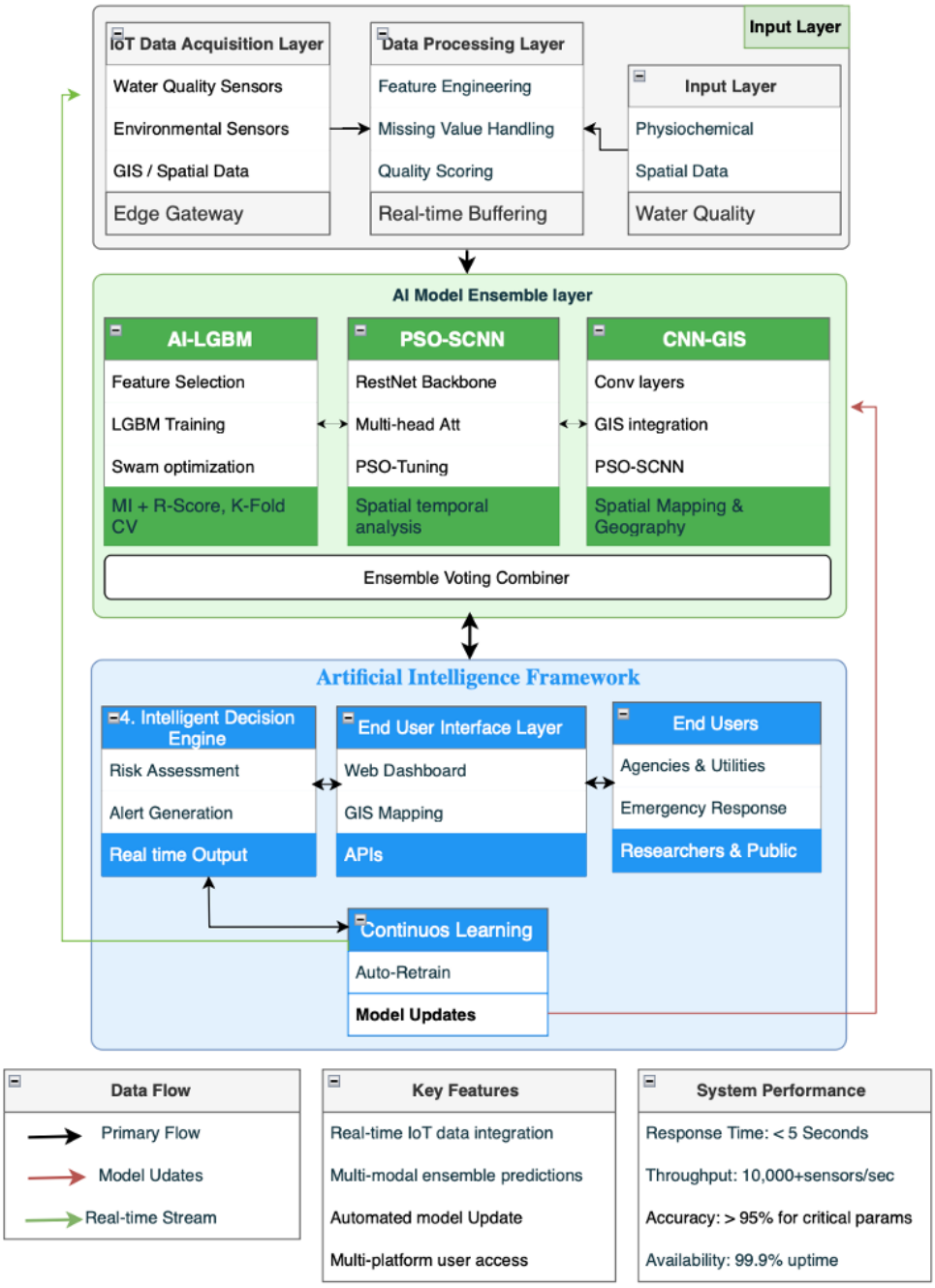


Figure 3.7: Proposed System Architecture for Artificial Intelligence Framework

Conclusion and Future Development

Core Contributions and Novelty

Hybrid spatial ensemble (AI–LGBM, PSO–SCNN, CNN–GIS); explicit geographic feature integration; PSO-based hyperparameter tuning; XAI (SHAP/LIME) for transparent decisions.

Scientific and Theoretical Significance

Advances spatial ML for hydroinformatics; couples PSO with DL; embeds XAI in monitoring; demonstrates cross-regional scalability.

Limitations

Data representativeness limits generalization; PSO–SCNN is compute-intensive; real-time IoT integration pending.

Future Research Directions

Add DL feature extraction for unstructured data; expand to longitudinal, multi-regional datasets; integrate IoT/remote sensing for real time; include socio-economic/climate covariates; release an open-source platform.

APPENDIX A: CODE AND DATA AVAILABILITY

A1 - REPRODUCIBILITY

This section provides details for the reproducibility of this study, including code, dataset, software dependencies, and random seed values.

Code Availability

The code is available at: https://github.com/MichaelOmar24/PSO-SCNN-model, which includes all scripts, Jupyter notebooks, and resources for replication.

Dataset Access

The dataset is available upon request. Contact: omar2@fe.edu.vn:. Preprocessing instructions are in the Methodology and Colab sections.

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.

LIST OF PUBLICATIONS RELATED TO THE THESIS

[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.