International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 11 | Nov – 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Leveraging Machine Learning for the Identification of Artificial Groundwater Recharge Potential Zones in Shivamogga District.

Navaneesh Raj[1], Prajwal B[2], Rakesh S[3], Ullas N P[4], Ms. Madhu D Naik[5]

[1]*Department of Computer Science and Engineering, PES Institute of Technology and Management*
[2]*Department of Computer Science and Engineering, PES Institute of Technology and Management*
[3]*Department of Computer Science and Engineering, PES Institute of Technology and Management*
[4]*Department of Computer Science and Engineering, PES Institute of Technology and Management*
[5]*Department of Computer Science and Engineering, PES Institute of Technology and Management*

Abstract - Groundwater depletion has become a major challenge in India due to extensive extraction and limited natural recharge. Identifying suitable locations for artificial groundwater recharge (AGR) is essential for ensuring long-term water sustainability. This study presents a hybrid machine learning framework that integrates K-Means clustering, Convolutional Neural Networks (CNN), and XGBoost to delineate AGR potential zones. The method combines unsupervised spatial zoning, deep feature extraction, and ensemble classification to capture non-linear interactions among hydrogeological, topographic, and climatic factors. Using nine thematic layers, including rainfall, geology, geomorphology, slope, and drainage density, the framework was applied to the Shivamogga district of Karnataka, India. The proposed model achieved an overall accuracy of 99.67%, outperforming conventional and two-stage hybrid approaches. High recharge potential was observed in lateritic and valley regions with favorable infiltration characteristics, while low-potential zones were found in steep and less permeable terrains. The results demonstrate the robustness of multi-stage integration for reliable groundwater recharge mapping and provide a scalable approach for sustainable water resource management in data-rich regions.

Key Words: groundwater recharge, machine learning, neural networks, water management, hybrid approach, spatial analysis

1. INTRODUCTION

Groundwater is one of the most essential water resources globally, providing a vital lifeline for domestic, agricultural, and industrial usage. In India, it accounts for approximately 63% of irrigation water and nearly 80% of rural drinking water supply. However, continuous over-extraction has caused water table levels to decline by 1–2 meters annually across several regions, threatening sustainable water availability and agricultural productivity.

Artificial Groundwater Recharge (AGR) has emerged as a practical solution to counter this decline by deliberately directing surface water into subsurface aquifers. AGR effectiveness largely depends on identifying suitable recharge zones with favorable hydrogeological and geomorphological conditions. Factors such as permeability, slope, drainage density, lithology, land use, and rainfall play critical roles in determining recharge potential.

Traditional recharge site selection methods, such as the Analytical Hierarchy Process (AHP) and Fuzzy-AHP, depend heavily on expert judgment and linear factor weighting [1], [2]. These methods are limited by subjective bias and over simplified assumptions, often achieving only 70–80% mapping accuracy [1]. Recent studies highlight the potential of data driven and machine learning (ML) approaches that automatically learn relationships between variables, capturing complex spatial patterns and non-linear dependencies [3], [4].

In this study, we develop a hybrid ML framework combining K-Means clustering, Convolutional Neural Networks (CNN), and XGBoost to identify AGR potential zones. The framework integrates unsupervised zoning, deep feature extraction, and ensemble decision-making to enhance mapping accuracy and interpretability. Compared to conventional techniques, the proposed approach eliminates subjectivity, improves adaptability across regions, and enhances predictive precision [3], [5].

2. RELATED WORK AND LITERATURE REVIEW

Groundwater recharge assessment has evolved from field based and hydrogeological methods to remote sensing and GIS-driven modeling. Earlier approaches primarily relied on geophysical surveys, borehole data, and manual interpretation [6]. Modern studies increasingly utilize

satellite imagery and GIS integration for large-scale and data-driven map ping [3], [7].

The Analytical Hierarchy Process (AHP) and Fuzzy-AHP (FAHP) have been extensively applied to delineate groundwater recharge zones [1], [2]. These techniques combine thematic layers using pairwise comparison matrices, but they assume linear relationships and rely on subjective weighting, which limits consistency and accuracy [1]. In contrast, data-driven and machine learning methods such as Random Forest (RF), Support Vector Machines, and gradient boosting models can automatically capture hidden dependencies among multiple spatial variables [4], [6].

Recent hybrid and deep learning models have achieved strong predictive performance. Al-Ruzouq et al. [3] implemented CNN–XGBoost for groundwater potential mapping with 90.8% accuracy, while Sachdeva and Kumar [6] achieved 91.67% using Random Forest. Zaresefat et al. [8] reported 97% accuracy using artificial neural networks, and Sivakumar et al. [9] integrated genetic algorithms with regression for improved optimization. Comprehensive reviews further highlight increasing adoption of AI-based methods for hydrogeological applications [4].

Complementary research on GIS–MCDA frameworks has explored managed aquifer recharge (MAR) and rainwater harvesting (RWH) site identification [10]. Ajayakumar and Reghunath [7] demonstrated the use of multi-layer GIS integration for recharge mapping in Kerala, aligning closely with the thematic layer design adopted in this study. The collective literature underscores the shift toward hybrid data intelligent systems that integrate GIS, remote sensing, and ML for more accurate, scalable, and interpretable groundwater recharge prediction.

3. PROPOSED METHODOLOGY

We propose a three-stage hybrid pipeline that combines K Means clustering, Convolutional Neural Networks (CNN), and XGBoost. This integration captures non-linear relationships and spatial patterns that linear frameworks cannot represent. Stage 1 (K-Means) discovers natural hydrogeological zones without manual specification. Stage 2 (CNN) automatically extracts spatial features from multi-band raster data. Stage 3 (XGBoost) models complex factor interactions through gradient-boosted decision trees [3], [6].

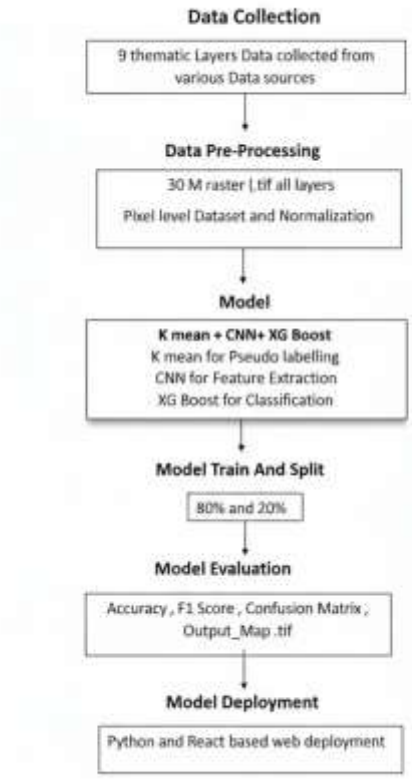

Fig -1: Overall workflow and model architecture.

The study was conducted in Shivamogga district, Karnataka (Western Ghats, India), located at coordinates 13.93◦N, 75.57◦E, covering approximately 3,960 km2. The district exhibits diverse hydrogeological, climatic, and topographic zones, ranging from river valleys to upland plateaus. The region experiences a tropical monsoon climate with pronounced seasonality, where the Southwest monsoon (June–September) contributes about 60–70% of annual rainfall. Western areas receive 3,000–4,000 mm annually, while eastern regions receive 1,500–2,500 mm. Subsurface geology comprises metamorphic formations (gneisses and schists) with primary aquifers in fracture networks; lateritic and alluvial deposits form secondary and tertiary aquifers with higher permeability.

Nine thematic layers were prepared using open and government geospatial datasets. Rainfall (mm/year) data were obtained from the India Meteorological Department (IMD). Distance from residential areas was derived from BBBike extracts based on OpenStreetMap (OSM) shapefiles. Drainage density (km/km2) was calculated using data from the Shivamogga Annual Groundwater Report. Geology (rock type distribution) and Geomorphology (landform classification) were sourced from the Bhukosh geospatial portal (BISAG-N, ISRO).

Groundwater quality in terms of Total Dissolved Solids (TDS) and Groundwater Level (GWL) data were obtained from the Central Ground Water Board (CGWB). Elevation and slope information were derived from the OpenTopography platform using Shuttle Radar Topography Mission (SRTM) 30 m digital elevation data. Lineament density (km/km2), representing fracture and structural features, was extracted from satellite data available on the ISRO Bhuvan portal. These parameters align with typical GIS-based recharge mapping practices [7], [10].

Each thematic layer influences groundwater recharge differently. For instance, rainfall and lineament density positively contribute to infiltration, while high slope, dense drainage networks, or impervious surfaces near settlements reduce recharge potential. This factor-level understanding ensures that the selected layers capture both hydrogeological and anthropogenic influences.

All thematic layers were converted to a uniform 30 m raster grid and clipped to the district boundary. Quantitative parameters underwent StandardScaler normalization:

$$z_i = \frac{x_i - \mu}{\sigma}$$

where z_i is the normalized value, x_i is the raw parameter, μ is the mean, and σ is the standard deviation. The normalized layers were stacked into a multi-band array (H × W ×9) for model training.

Missing and invalid pixels were masked to avoid bias, and all rasters were spatially aligned to a uniform 30 m grid using the study boundary as reference. This ensured comparability across heterogeneous geospatial inputs.

K-Means clustering segmented the study area into k = 5 hydrogeological zones by minimizing:

$$J = \sum_{i=1}^{k} \sum_{x_j \in C_i} |x_j - \mu_{i|}^2$$

where c_i represents the ith cluster, x_j is a data point, and μ_i is the cluster centroid. The elbow method determined k = 5, balancing interpretability and separation.

The K-Means zone map was one-hot encoded (five chan nels) and concatenated with the nine input bands, providing spatial zoning cues to the CNN and XGBoost stages.

The CNN architecture used 256 × 256 raster patches with 9 input bands and 5 K-Means zone maps (14 channels). The architecture comprised three convolutional layers (32, 64, and 128 filters) with ReLU activations, max-pooling, and dense layers producing 512-dimensional feature vectors. Training employed the Adam optimizer (learning rate 0.001), cross entropy loss, batch size 32, and early stopping with patience of 10 epochs.

The CNN acted primarily as a spatial feature extractor, learning non-linear relationships among the nine input layers and the K-Means zoning pattern. The extracted embeddings were subsequently used as inputs to the XGBoost classifier.

XGBoost integrated CNN features, K-Means cluster assignments, and the normalized layers (526 total features) for AGR suitability classification:

$$\hat{Y}_i = \sum_{k=1}^{K} f_k(x_i)$$

where $\hat{Y}_i$ is the predicted suitability score, f_k is the kth tree, and K is the total number of trees. Model tuning used 5-fold cross-validation with 200 estimators, max depth 6, learning rate 0.1, subsample 0.8, and L1/L2 regularization.

Predicted suitability maps were exported as GeoTIFF files for visualization and integration with GIS platforms, enabling practical use by water resource planners.

Data were split 80:20 for training and testing using stratified random sampling to ensure representative spatial coverage. Model performance was evaluated using Accuracy, Precision, Recall, and F1-Score metrics to ensure balanced performance across recharge classes.

4. RESULT AND DISCUSSION

Table -1: Performance Comparison of Hybrid Models

Model	Accuracy	Precision	Recall
K-Means + Random Forest	0.9070	0.9200	0.9200
XGBoost + K-Means	0.9803	0.9800	0.9700
CNN + XGBoost + K-Means	**0.9967**	**0.9960**	**0.9965**

The proposed three-stage hybrid framework achieved an overall accuracy of 99.67%, outperforming the two-stage base lines XGBoost+K-Means (98.03%) and K-Means+Random Forest (90.70%). This improvement demonstrates that combining unsupervised clustering, spatial deep feature extraction, and ensemble classification captures complex inter-layer relationships more effectively than conventional approaches. Random Forest's prior use in recharge estimation [5] validates its role as a suitable baseline for comparison.

The notable performance gain motivated a class-wise analysis of the hybrid model. Table II presents the precision, recall, and F1-score for each recharge potential class, providing insight into model consistency across varying hydrogeological conditions.

The hybrid CNN+XGBoost+K-Means framework achieved an overall accuracy of 99.67% with a macro-averaged F1-score of 0.9962, demonstrating consistent classification across all five recharge classes. All classes achieved precision and recall above 99%, confirming balanced performance, consistent with other high-accuracy CNN-based groundwater mapping studies [3], [8].

Table -2: Classification Metrics for CNN+XGBoost+K Means

Class	Precision	Recall	F1-Score	Support
Very Low	0.997	0.996	0.997	6729
Low	0.995	0.996	0.995	1649
Moderate	0.998	0.999	0.998	7640
High	0.994	0.996	0.995	3262
Very High	0.996	0.996	0.996	6687
Overall Accuracy	**0.9967**			**25,967**

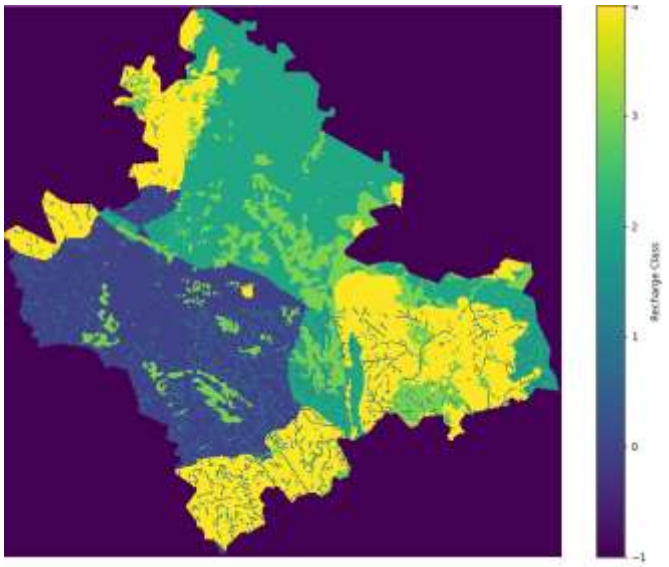

Fig -2: Predicted AGR Suitability Map using CNN+XGBoost+K-Means (Proposed Model).

The resulting spatial map (Figure 2) clearly delineates zones from very low to very high recharge potential. High and very high zones align with valley fills and lateritic terrains, while low-potential zones coincide with steep slopes and hard rock formations.

Compared with previously published approaches, the proposed framework achieved higher accuracy while maintaining strong spatial coherence. Al-Ruzouq et al. reported 90.8% with CNN–XGBoost [3], Sachdeva and Kumar obtained 91.67% using Random Forest [6], and Zaresefat et al. achieved 97% with ANN [8]. Our model exceeds these results, reaching 99.67% accuracy and producing smoother class boundaries.

Overall, these findings confirm the reliability and superiority of the proposed hybrid framework and establish a strong foundation for its application to groundwater recharge potential mapping in diverse hydrogeological settings.

5. CONCLUSION AND FUTURE WORKS

The proposed K-Means + CNN + XGBoost hybrid framework significantly improves groundwater recharge potential mapping accuracy, achieving 99.67%, compared to XGBoost + K-Means (98.03%) and K-Means + Random Forest (90.70%). The integration effectively captures non-linear interactions and spatial patterns; feature importance analysis confirmed CNN and K Means components contribute substantially to predictive value.

The model identifies approximately 30% of Shivamogga district (High + Very High zones) as suitable for future artificial recharge development, offering actionable guidance for sustainable groundwater management.

Future enhancements include incorporating seasonal and climatic variability, validating across additional districts to assess transferability. Expansion toward deep transfer learning architectures may further improve generalization for large scale national applications.

ACKNOWLEDGEMENT

The authors express their sincere gratitude to Ms. Madhu D. Naik, Assistant Professor, Department of Computer Science and Engineering, PES Institute of Technology and Management, Shivamogga, for her valuable guidance and support throughout this project. The authors acknowledge data sources from the India Meteorological Department (IMD), BBBike OpenStreetMap Extracts, Shivamogga Annual Groundwater Report, Bhukosh (BISAG-N, ISRO), Central Ground Water Board (CGWB), OpenTopography, and ISRO Bhuvan. Their open geospatial datasets made this research possible.

REFERENCES

1. P. R. Shekar and A. Mathew, "Delineation of groundwater potential zones and identification of artificial recharge sites in the kinnerasani watershed, india, using remote sensing-gis, ahp, and fuzzy-ahp techniques," AQUA– Water Infrastructure, Ecosystems and Society, 2023.
2. Q. Song et al., "Assessing groundwater artificial recharge suitability in the mi river basin using gis, rs, and fahp: a comprehensive analysis with seasonal variations," Applied Water Science, 2025.
3. R. Al-Ruzouq et al., "Hybrid deep learning and remote sensing for the delineation of artificial groundwater recharge zones," The Egyptian Journal of Remote Sensing and Space Science, 2024.
4. D. Singh, "Review of groundwater potential storage and recharge zone map delineation using statistics based hydrological and machine learning based artificial intelligent models," in Somaiya International Conference on Technology and Information Management (SICTIM). IEEE, 2023.
5. P. Sihag et al., "Estimation of the recharging rate of groundwater using random forest technique," Applied Water Science, 2020.
6. S. Sachdeva, "Groundwater potential mapping using machine learning models for northeastern karbi anglong district, assam, india," in 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN). IEEE, 2020.
7. A. Ajayakumar and R. Reghunath, "Delineation of groundwater recharge zones in lateritic terrains using geospatial techniques," Discover Geo science, 2025.
8. M. Zaresefat et al., "Using artificial intelligence to identify suitable artificial groundwater recharge areas for the iranshahr basin," Water, 2023.
9. V. L. Sivakumar, "Prediction of groundwater potential mapping using linear regression and genetic algorithm," in 2nd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS). IEEE, 2024.
10. A. K. Seif et al., "Identifying managed aquifer recharge and rainwater harvesting sites and structures for storing non-conventional water using gis-based multi-criteria decision analysis approach," Applied Water Science, 2024.