RESEARCH PAPER

Impact of Feature Extraction Techniques on Machine Learning Efficiency: An Empirical Study of Model Performance Improvement

Shubha G. Sanu[1*], Mallikarjun M. Math[2], Santosh L. Deshpande[3], Rudragoud Patil[4]

[1]Department of Computer Science and Engineering (AI & ML), KLS Gogte Institute of Technology, Affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India. ORCID: 0000-0003-4789-0795

[2]Department of Computer Science and Engineering, KLS Gogte Institute of Technology, Affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India. ORCID: 0000-0003-2731-9655

[3]Department of Computer Science and Engineering, Visvesvaraya Technological University, Belagavi, Karnataka, India. ORCID: 0000-0001-5152-0952

[4]Department of Computer Science and Engineering, KLS Gogte Institute of Technology, Affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India. ORCID: 0000-0001-6374-5200

***Corresponding author: Shubha G. Sanu, Department of Computer Science and Engineering (AI & ML), KLS Gogte Institute of Technology, Affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India**

Email: sgsanu@git.edu | ORCID: 0000-0003-4789-0795

Co-author emails: mmmath@git.edu, sld@vtu.ac.in, rspatil@git.edu

Received: 31st May, 2026; **Revised:** 8th June, 2026; **Accepted:** 10th June, 2026; **Available Online:** 12th June, 2026

ABSTRACT

Reliable prediction of groundwater level is essential for aquifer monitoring, seasonal water-resource planning, and early identification of hydrological stress. Groundwater response is nonlinear, spatially heterogeneous, and temporally delayed because observed levels are shaped by rainfall variability, well characteristics, physiographic position, and previous aquifer states. This study develops a leakage-aware feature-extraction and model-selection framework for monthly groundwater-level prediction in Belagavi and Khanapur taluks, Karnataka, India, using integrated well-water-level and rainfall observations from 2015 to 2024. The workflow combines official groundwater records with rain-gauge observations through spatial nearest-neighbor matching, followed by preprocessing, exploratory statistical analysis, feature extraction, model benchmarking, and generalization-gap evaluation. Feature extraction includes cyclic month encoding, water-level and rainfall lag features at 1, 2, 3, 4, 6, and 12 months, and dimensionality-reduction components derived from PCA, SVD, and ICA. The current target water level is explicitly excluded during feature generation, and scaling is fitted only on training data to reduce leakage risk. Classical regression, tree-based models, boosting algorithms, SARIMAX, autoencoder-based regressors, CNN, GRU, LSTM, and BiLSTM are evaluated under a common train-test protocol and time-series cross-validation. Results show that spatial variables, well depth, altitude, and lagged groundwater information are more informative than rainfall alone. Among the evaluated sequence-learning models, LSTM provides the most favorable accuracy-generalization balance, with substantially lower average generalization gaps than GRU and many boosting baselines. The proposed framework contributes a reproducible pathway for robust groundwater prediction where model selection is guided not only by test accuracy but also by stability between cross-validation and independent testing.

Keywords: Groundwater level prediction, LSTM, feature extraction, time-series cross-validation, generalization gap, rainfall, Belagavi, Khanapur, leakage-aware machine learning, PCA, SVD, ICA.

How to cite this article: Sanu SG, Math MM, Deshpande SL, Patil R. Impact of Feature Extraction Techniques on Machine Learning Efficiency: An Empirical Study of Model Performance Improvement. Int J Drug Deliv Technol. 2026;16(58s): 1492-1508. DOI: 10.25258/ijddt.16.58s.160

Source of support: Nil

Conflict of interest: None

1. Introduction

1.1 Background and motivation

Groundwater is a strategic water resource for drinking-water supply, irrigation, drought buffering, and rural livelihood security[1]. In semi-arid and monsoon-influenced regions, groundwater-level dynamics are controlled by interacting processes: recharge from seasonal rainfall, delayed subsurface response, spatial heterogeneity of aquifer material, terrain controls, well construction, and anthropogenic abstraction[2]. Because the water table does not respond instantaneously to rainfall, prediction models must represent both current hydro-climatic conditions and the memory carried by previous groundwater states[3]. Groundwater-level prediction is therefore a spatio-

temporal learning problem rather than a simple regression exercise[4].

The growing availability of monitoring data has encouraged the use of machine-learning and deep-learning models for groundwater-level forecasting[4], [5]. Ridge regression, decision trees, gradient boosting, convolutional neural networks, gated recurrent units, and long short-term memory networks have all been used to learn nonlinear patterns from hydro-meteorological data[6], [7]. Despite this progress, accuracy alone is insufficient for operational groundwater management. A model may perform well during cross-validation yet degrade on an independent test set, particularly when feature sets are large, seasonality is strong, or target leakage occurs during feature construction. Robust groundwater prediction therefore requires high test accuracy and reliable generalization to unseen observations[8], [9].

This study addresses this need through a leakage-aware feature-extraction and generalization-gap evaluation framework. The framework is demonstrated for monthly groundwater-level prediction in Belagavi and Khanapur taluks using an integrated dataset of well-water-level records, rainfall observations, rain-gauge locations, terrain attributes, and well characteristics[10]. The central hypothesis is that engineered spatio-temporal features, particularly lagged groundwater and rainfall variables, improve model efficiency and reduce the difference between cross-validation and independent test performance. The study further evaluates whether LSTM, because of its gated temporal memory, provides a more stable prediction structure than static machine-learning models and alternative deep-learning architectures[11], [12].

1.2 Research gap

Existing groundwater-prediction studies have demonstrated the potential of machine learning and recurrent neural networks[6]. However, several methodological gaps remain important for high-quality hydrological modeling. First, many studies focus on final accuracy without clearly separating the effects of feature extraction, cross-validation performance, and independent test performance. Second, rainfall is often included as an explanatory variable without explicit spatial assignment from rain gauges to observation wells. Third, construction of lag features can introduce data leakage when the current target variable is inadvertently included among predictors. Fourth, feature sets with different numbers of predictors are sometimes compared using ordinary R2 alone, although adjusted R2 and error-based metrics provide a fairer assessment under changing model complexity. Finally, practical model selection should consider whether the model generalizes reliably beyond the samples used for tuning[13], [14]. This study responds to these gaps by combining spatial rainfall-well integration, leakage-aware feature extraction, time-series cross-validation, independent testing, and generalization-gap analysis. The resulting evaluation framework distinguishes models that merely fit well from models that remain reliable when applied to unseen groundwater observations.

1.3 Objectives and novelty

The objective of this study is to evaluate whether spatio-temporal feature extraction improves groundwater-level prediction and to identify the model-feature combination that generalizes most reliably on unseen monthly observations. The novelty of the work lies in four connected contributions: (i) integration of official groundwater-level and rainfall datasets for Belagavi and Khanapur taluks; (ii) leakage-aware generation of cyclic, lagged, and dimensionality-reduction features; (iii) fair benchmarking of statistical, classical machine-learning, boosting, autoencoder, convolutional, and recurrent models under a common data split; and (iv) model selection based on both independent test accuracy and generalization gap.

The paper is organized as follows. Section 2 describes the study area and data sources. Section 3 explains preprocessing and exploratory analysis. Section 4 presents the modeling methodology, feature-set design, model architectures, validation protocol, and evaluation metrics. Section 5 reports the spatial, statistical, feature-extraction, and model-generalization results. Section 6 discusses hydrological meaning, model behavior, practical implications, and limitations. Section 7 concludes the study and identifies future research directions.

2. Study Area and Data Sources

2.1 Study area and monitoring network

The case study is located in Belagavi and Khanapur taluks of Karnataka, India. The region is suitable for evaluating groundwater-level prediction because it exhibits spatial variability in elevation, rainfall distribution, well characteristics, and seasonal groundwater response. The monitoring network includes well locations and rain-gauge stations distributed across the two taluks. The combined spatial layout is important because monthly well-water-level observations must be interpreted relative to nearby rainfall inputs rather than by taluk-level rainfall averages alone.

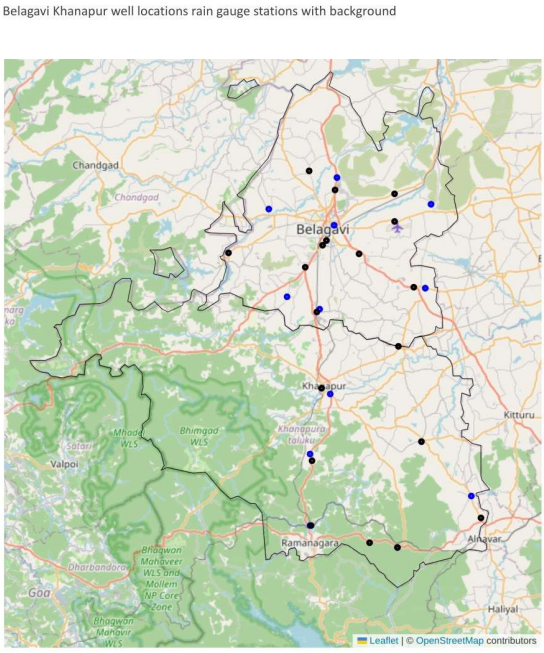

Figure 1. Study-area monitoring network showing observation wells and rain-gauge stations in Belagavi and Khanapur taluks.

2.2 Groundwater-level data

The groundwater-level archive covers monthly observations from 2015 to 2024. The original annual records contain multiple sheets and include wells from the broader district archive. For the present analysis, records were filtered to Belagavi and Khanapur taluks and structured into a machine-learning-ready format. Each groundwater observation is associated with year, month, taluk, well code, village, well type, well depth, latitude, longitude, altitude, and measured water level.

2.3 Rainfall data and rain-gauge coordinates

Rainfall records for Belagavi and Khanapur were consolidated into a monthly station-level rainfall table. Station names were standardized, years and months were extracted from source files, and rain-gauge coordinates were converted into decimal degrees for spatial analysis. The rainfall dataset was then prepared for joining with the groundwater-well dataset using station location and monthly temporal indexing.

2.4 Spatial integration of rainfall and groundwater observations

Rainfall was spatially linked to observation wells using nearest-neighbor assignment. The Haversine formula was used to estimate spherical distance between well coordinates and rain-gauge coordinates, while nearest-neighbor search identified the closest gauge for each well. Rainfall values were merged only when the nearest rain-gauge station was within a 30 km influence radius. This threshold reduces the risk of assigning rainfall observations from stations that are spatially distant from the monitored well. The final merged table includes groundwater attributes, rainfall totals, nearest station coordinates, and station-well distance.

Table 1. Data sources and acquisition details.

Data component	Institutional source	Temporal coverage	Spatial coverage	Primary variables
Groundwater level	Office of the Senior Geologist, District Groundwater Office, Groundwater Directorate, Belagavi; Karnataka Groundwater Authority, Government of Karnataka	Monthly observations, 2015-2024	Belagavi and Khanapur taluk wells	Water level, well depth, well type, coordinates, altitude, village and taluk identifiers
Rainfall	No. 3 Irrigation Investigation Sub Division, Belagavi, Water Resources Department, Government of Karnataka	Monthly station rainfall, 2015-2024	Rain-gauge stations in and around Belagavi and Khanapur	Total monthly rainfall and station coordinates

Administrative and map layers	DataMeet community maps, GADM administrative boundaries, OpenStreetMap and CARTO basemap services	Static spatial context	District and taluk boundaries	Taluk boundaries, base map, monitoring-network visualization

Table 2. Dataset summary after spatial and temporal integration.

Dataset descriptor	Value used in the study
Study period	2015-2024
Taluks	Belagavi and Khanapur
Merged monthly groundwater-rainfall records	1,961 records in the final merged CSV file
Analysis-ready observations used for descriptive feature statistics	1,566 observations after preparation of the modeling and analytics dataset
Unique wells in the merged taluk-specific modeling file	19
Unique villages in the merged taluk-specific modeling file	19
Nearest rainfall stations represented in the merged file	13
Nearest-station distance range	0.010 to 23.667 km
Mean nearest-station distance	3.595 km

Table 3. Variables and modeling roles.

Variable	Role in modeling	Unit or scale
Year	Temporal index and long-term trend proxy	Calendar year
Month	Seasonality indicator and cyclic feature source	1-12
Latitude and longitude	Spatial identity and location-dependent hydrogeological variation	Decimal degrees
Altitude	Terrain-related explanatory variable	Meters above mean sea level
Well depth	Well-structure and aquifer-access variable	Meters
Water level	Prediction target and source of lagged groundwater-memory features	Recorded water-level scale
Total rainfall	Monthly hydrometeorological input	Millimeters
Nearest rainfall coordinates	Spatial metadata for assigned rain-gauge station	Decimal degrees
Station distance	Distance between monitoring well and assigned rain gauge	Kilometers

3. Data Preprocessing and Exploratory Analysis

3.1 Groundwater-data preprocessing

Groundwater records were harmonized across annual sheets and converted into a normalized monthly format. Non-numeric field descriptions were resolved before modeling. Records marked as dry were represented using maximum well depth, reflecting the practical condition that water level was below the observable depth of the well. Textual entries such as not available, box filled with mud, and lock jammed were treated as missing observations and replaced using monthly averages where possible. Rows without usable numerical information were removed. This preprocessing created a consistent monthly time-series structure suitable for downstream feature extraction and model comparison.

3.2 Rainfall-data preprocessing

Rainfall preprocessing transformed multiple station files into a consolidated monthly rainfall dataset. The process included extraction of year and month information, standardization of station names,

correction of inconsistent month labels, conversion of station coordinates from degree-minute-second notation into decimal degrees, and chronological sorting by year, month, and station. This step ensured that rainfall totals could be reliably aligned with groundwater observations by month and assigned spatially to the nearest gauge.

3.3 Spatial and distributional patterns

The spatial visualizations show that both groundwater level and rainfall vary across the two taluks. The mean water-level heatmap highlights localized zones of deeper or higher recorded levels, while the rainfall heatmap indicates that rainfall intensity is not spatially uniform. Such heterogeneity supports the inclusion of location, altitude, well depth, rainfall, and distance-to-station information in the prediction framework.

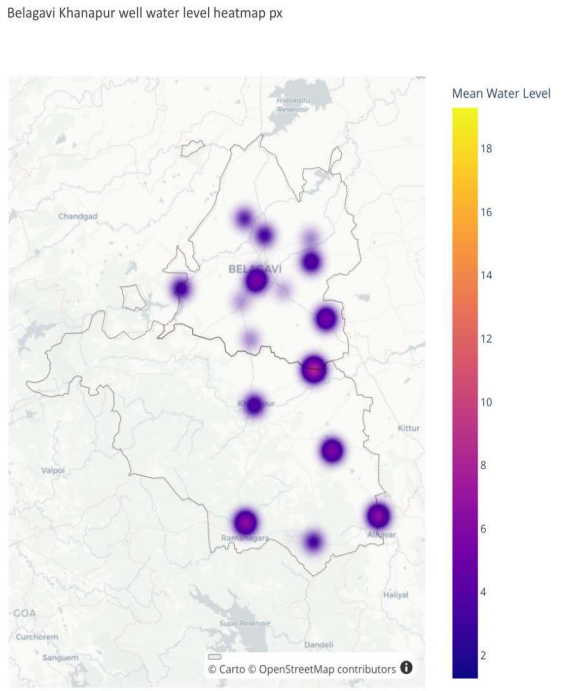

Figure 2. Spatial heatmap of mean groundwater level across observation-well locations.

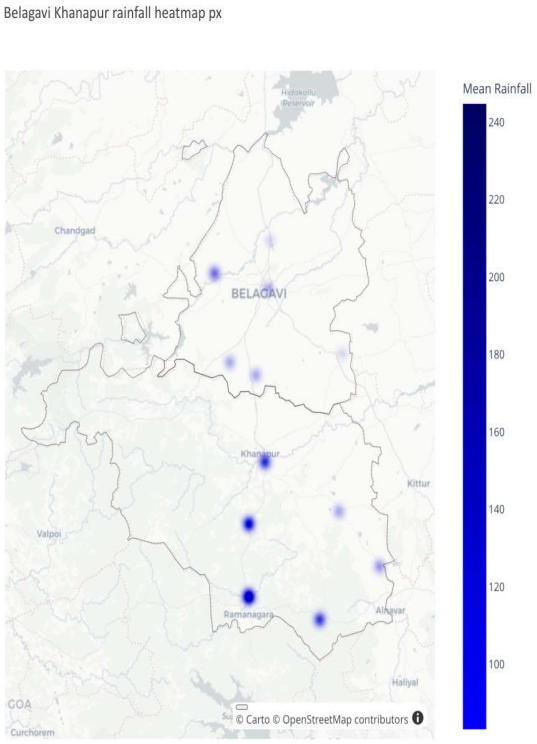

Figure 3. Spatial heatmap of mean rainfall associated with rain-gauge stations and assigned well locations.

Table 4. Descriptive statistics of numerical variables used in the analysis-ready dataset.

Feature	Mean	Median	Std. dev.	Variance	Skewness	Kurtosis
Year	2018.504	2019	2.316	5.362	0.013	-1.202
Month	6.425	6	3.453	11.924	0.025	-1.217
Well Depth	36.585	30	25.846	668.004	0.655	-0.559
Latitude	15.700	15.742	0.179	0.032	-0.324	-1.369
Longitude	74.563	74.559	0.071	0.005	0.382	-0.169
Altitude	704.582	719	50.640	2564.371	-0.767	-0.392
Water	8.103	6.700	6.064	36.773	0.924	0.305

Level						
Total Rainfall	125.513	51.200	199.770	39908.215	3.192	14.113
Nearest Rainfall Latitude	15.712	15.775	0.184	0.034	-0.405	-1.383
Nearest Rainfall Longitude	74.577	74.527	0.076	0.006	0.215	-1.371
Station Distance in km	3.284	2.715	2.860	8.182	1.156	2.827

3.4 Correlation and information-theoretic feature relevance

Pearson correlation and information-theoretic measures were used to examine the relationship between candidate predictors and water level before model training. The correlation analysis indicates that well depth has the strongest positive linear association with water level, followed by longitude and nearest rainfall longitude. Month, altitude, latitude, and rainfall show negative direct correlations with water level, but these correlations should not be interpreted as causal effects because groundwater response can be delayed, nonlinear, and spatially mediated.

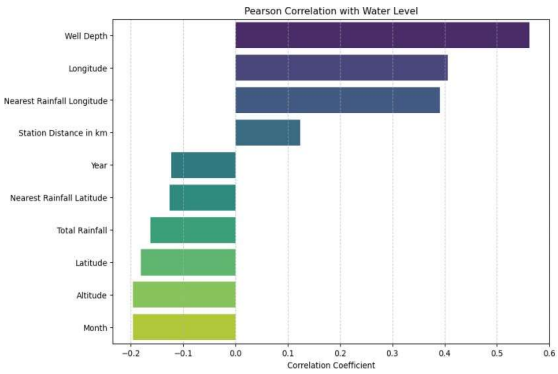

Figure 4. Pearson correlation coefficients between numerical predictors and groundwater level.

Table 5. Target-correlation summary for groundwater level.

Feature	Pearson correlation with water level	Interpretation
Well Depth	0.562	Strongest positive linear association among available predictors
Longitude	0.406	Positive spatial association
Nearest Rainfall Longitude	0.391	Positive association linked to rain-gauge spatial position
Station Distance in km	0.124	Weak positive association
Year	-0.123	Weak decreasing trend across the study period
Nearest Rainfall Latitude	-0.126	Weak negative spatial association
Total Rainfall	-0.163	Weak direct negative association, suggesting delayed or nonlinear

		rainfall response
Latitude	-0.181	Weak to moderate negative spatial association
Altitude	-0.196	Weak to moderate negative terrain association
Month	-0.196	Seasonality-related negative association

Information-theoretic measures provide a complementary view because they can capture non-linear dependence after discretization. Altitude, latitude, longitude, and well depth have higher mutual information with water level than rainfall alone. This pattern indicates that spatial and structural attributes are central to explaining groundwater-level variability, while rainfall contributes through delayed and interacting effects rather than through a strong instantaneous linear relationship.

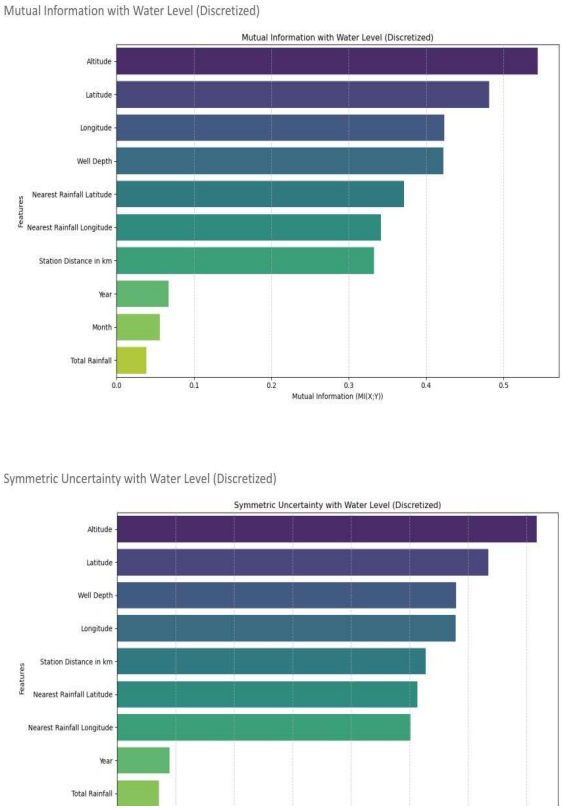

Figure 5. Mutual information and symmetric uncertainty between predictors and groundwater level after discretization.

Table 6. Information-theoretic feature relevance with groundwater level.

Feature	Entropy H(X)	Mutual information MI(X;Y)	Conditional entropy H(Y|X)	Conditional entropy H(X|Y)	Symmetric uncertainty
Altitude	3.104642	0.544566	2.421141	2.560076	0.179418
Latitude	3.106709	0.481561	2.484145	2.625148	0.158606
Longitude	2.891340	0.423485	2.542222	2.467856	0.144607
Well Depth	2.864212	0.422183	2.543524	2.442030	0.144833
Nearest Rainfall	2.822136	0.371633	2.594073	2.450503	0.128419

Latitude					
Nearest Rainfall Longitude	2.487674	0.341735	2.623972	2.145939	0.125329
Station Distance in km	2.079884	0.332873	2.632834	1.747011	0.131946
Year	3.050104	0.067116	2.898590	2.982988	0.022313
Month	3.584592	0.055921	2.909785	3.528671	0.017074
Total Rainfall	1.375403	0.038566	2.927141	1.336837	0.017768

4. Methodology

4.1 Overall modeling workflow

The full workflow consists of eight stages: data acquisition, preprocessing and normalization, spatial integration, exploratory analytics and visualization, feature engineering, training setup, model benchmarking, and result evaluation. The workflow is designed to maintain separation between training and testing operations, reduce leakage during feature generation, and evaluate models through both performance metrics and generalization stability.

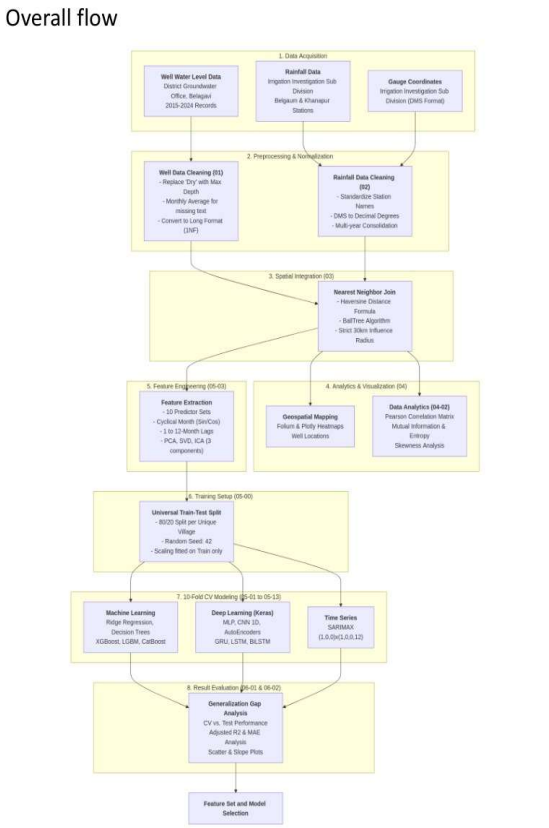

Figure 6. Overall workflow for leakage-aware feature extraction, model benchmarking, and generalization-gap evaluation.

4.2 Leakage-aware feature extraction

Feature extraction was organized around hydrological interpretability and leakage prevention. Month was transformed into cyclic sine and cosine variables to represent annual seasonality. Lagged features were generated for water level and rainfall at 1, 2, 3, 4, 6, and 12 months, allowing models to learn short-term, seasonal, and annual memory effects. For missing lag values, a hierarchical imputation strategy was applied using village-month averages, village averages, and then global averages. PCA, SVD, and ICA components were generated from numerical predictor variables to examine whether dimensionality-reduction features improved predictive stability. During feature extraction, the current water-level target was excluded from the predictor matrix. This step is essential because including the current target, even indirectly, would inflate model performance and reduce the credibility of validation results.

Table 7. Feature-set definitions used for model benchmarking.

Feature set	Description	Hydrological or modeling rationale
SpcTm	Year, month, latitude, longitude	Baseline space-time identity
SpcTmAlt	SpcTm plus altitude	Adds terrain-related variability
SpcTmWDpth	SpcTm plus well depth	Adds well-structure and aquifer-access information
SpcTmRn	SpcTm plus total rainfall	Adds monthly hydrometeorological forcing
SpcTmRn_MntSes	SpcTmRn plus Month_sin and Month_cos	Represents cyclic seasonality
SpcTmRn_WlLag	SpcTmRn plus 1, 2, 3, 4, 6, and 12 month water-level lags	Represents groundwater memory and delayed aquifer response
SpcTmRn_RfLag	SpcTmRn plus 1, 2, 3, 4, 6, and 12 month rainfall lags	Represents delayed rainfall-recharge effects
SpcTmRn_PCA	SpcTmRn plus three PCA components	Reduces correlated predictor structure
SpcTmRn_SVD	SpcTmRn plus three SVD components	Captures dominant low-rank predictor directions
SpcTmRn_ICA	SpcTmRn plus three ICA components	Extracts statistically independent components

4.3 Model families and hyperparameters

A broad model set was selected to compare interpretable baselines, nonlinear machine-learning models, boosting methods, statistical time-series modeling, and deep-learning architectures. This design makes it possible to determine whether recurrent neural networks provide real gains over simpler alternatives. Model training used a consistent feature-set naming convention, where each model was trained and evaluated under each candidate feature set.

Table 8. Model architectures and hyperparameter design.

Model family	Model	Core configuration	Overfitting control
Linear baseline	Ridge regression	L2-regularized linear regression; alpha values 0.01, 0.1, 1.0, 10.0, 100.0	L2 penalty
Statistical time series	SARIMAX	Order (1,0,0), seasonal order (1,0,0,12), fitted for unique locations	Seasonal specification and location-wise modeling
Tree model	Decision tree regressor	max_depth, min_samples_leaf, min_samples_split, ccp_alpha considered	Pruning and split constraints
Gradient boosting	XGBoost	learning rate 0.03, subsample 0.7, reg_alpha 0.1 and reg_lambda 1.0	Early stopping and L1/L2 regularization
Gradient boosting	LightGBM	subsample 0.8, colsample_bytree 0.8, reg_alpha 0.1, reg_lambda 0.1	Early stopping and row/column bagging

Gradient boosting	CatBoost	iterations 1000, learning_rate 0.05, depth 6, l2_leaf_reg 3.0	L2 leaf regularization
Dense neural network	ANN	Dense layers with 39, 24, and 12 units; ReLU activations; Adam optimizer, learning rate 0.001	Batch normalization, dropout 0.2, L2 regularization, early stopping
Representation learning	Autoencoder and stacked autoencoder	Symmetric encoder-decoder and stacked dense architecture with 39, 24, and 12 unit layers	Dropout, L2 regularization, batch normalization, early stopping
Convolutional model	1D CNN	Conv1D with 64 filters, kernel size 3; flatten layer; dense layers 39, 24, 12	Dropout, L2 regularization, early stopping
Recurrent model	GRU	Three GRU layers with 39, 24, and 12 units; ReLU activations; Adam optimizer	Dropout 0.2, L2 regularization, early stopping
Recurrent model	LSTM	Three LSTM layers with 39, 24, and 12 units; ReLU activations; linear output; Adam optimizer	Dropout 0.2, L2 kernel regularization, early stopping
Bidirectional recurrent model	BiLSTM	Bidirectional LSTM layer with 39 units followed by dense layers with 24 and 12 units	Dropout, recurrent dropout, L2 regularization, early stopping

4.4 Training, validation, and testing protocol

All models were evaluated using common training and testing files to ensure fair comparison. The data were split using an 80:20 train-test structure with village representation preserved across the split. Within training, 10-fold time-series cross-validation was used so that validation subsets respect temporal order. Training-only scaling was used: scalers were fitted on training data and then applied to test data. This prevents information from the test distribution from entering model training. A fixed random seed of 42 was used where applicable. Negative model predictions were clipped to 0.0 because negative groundwater levels are physically invalid in the modeled target representation.

The independent test set was used only after model training and cross-validation. This separation is necessary because cross-validation performance describes internal stability, whereas test performance describes expected behavior on unseen records. The difference between these two quantities is therefore treated as a model reliability indicator.

4.5 Performance metrics and generalization gap

The evaluation uses multiple complementary metrics: explained variance, maximum error, mean squared error, root mean squared error, mean absolute error, normalized mean squared error, symmetric mean absolute percentage error, R2, and adjusted R2. Because candidate feature sets contain different numbers of predictors, adjusted R2 is emphasized over ordinary R2 when comparing feature sets. Error metrics such as RMSE and MAE remain essential because they preserve the magnitude of prediction error in the water-level scale.

The generalization gap is defined as the difference between cross-validation performance and independent test-set performance. For error metrics, smaller absolute differences indicate more stable generalization. For adjusted R2, smaller differences

likewise indicate that the explanatory capacity observed during validation is not lost on unseen data. The final model-selection logic prioritizes models that combine low test error, high adjusted R2, low SMAPE, and small generalization gaps.

Table 9. Principal evaluation metrics and interpretation.

Metric	Formula or definition	Interpretation
RMSE	sqrt(mean((y - y_hat)^2))	Penalizes large prediction errors; lower is better
MAE	mean(abs(y - y_hat))	Average absolute error; lower is better
SMAPE	mean(2 abs(y - y_hat) / (abs(y) + abs(y_hat)))	Scale-aware symmetric percentage error; lower is better
Adjusted R2	1 - (1 - R2)(n - 1)/(n - p - 1)	Goodness of fit penalized by number of predictors; higher is better
Generalization gap	CV metric - test metric	Difference between validation and independent test behavior; smaller magnitude indicates stronger reliability

5. Results

5.1 Descriptive and spatial results

The descriptive statistics indicate strong variability in both groundwater level and rainfall. Water level has a mean of 8.103 and median of 6.700, with a standard deviation of 6.064, indicating substantial variation across months and locations. Rainfall is highly skewed, with a mean of 125.513, median of 51.200, and standard deviation of 199.770. The rainfall skewness and kurtosis confirm that rainfall inputs contain extreme seasonal events rather than a uniform monthly distribution. This distributional pattern supports the use of nonlinear models and lag features, because groundwater response to extreme rainfall is unlikely to be instantaneous or linear.

Spatially, the water-level and rainfall heatmaps show local clusters instead of uniform taluk-wide behavior. Therefore, location and station-distance information are not merely descriptive metadata; they are necessary components of the predictive design. The exploratory analysis also explains why rainfall alone is insufficient. Direct rainfall-water-level correlation is weak, while spatial and well-structural variables show stronger relationships with the target.

5.2 Effect of feature extraction on generalization

Feature extraction changed model behavior substantially. The comparison of feature-set generalization gaps shows that lagged water-level features provide the most stable engineered representation among the evaluated feature sets. The SpcTmRn_WlLag feature set has the lowest average RMSE gap among the engineered feature sets and also maintains a low SMAPE gap. This result is hydrologically plausible because previous groundwater levels summarize antecedent recharge, aquifer storage, delayed drainage, and local pumping effects better than rainfall alone. Dimensionality-reduction features, especially PCA, also reduce gaps compared with several raw or expanded feature sets, indicating that compact feature representations can reduce instability.

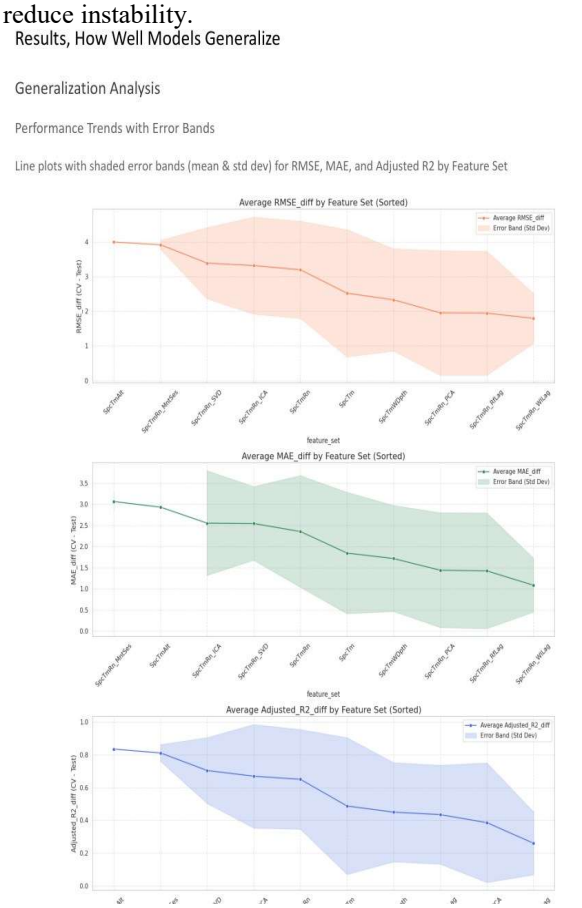

Figure 7. Average RMSE, MAE, and adjusted R2 performance trends by feature set with uncertainty bands.

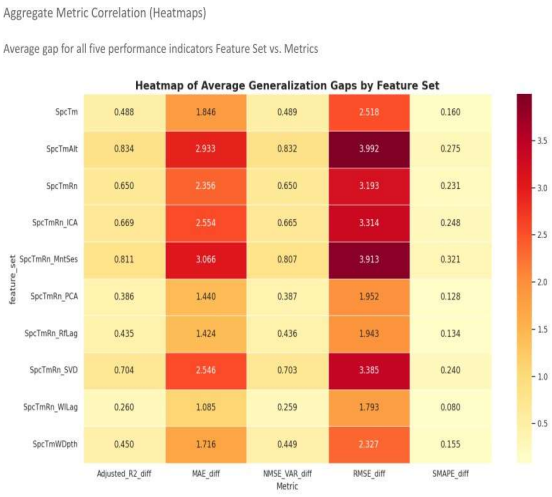

Figure 8. Average generalization gaps across performance indicators by feature set.

Table 10. Average generalization gaps by feature set across major performance indicators.

Feature set	Adjusted R2 gap	MAE gap	NMSE gap	RMSE gap	SMAPE gap
SpcTm	0.488	3.840	0.499	2.518	0.160
SpcTmAlt	0.834	2.931	0.832	3.992	0.275
SpcTmRn	0.650	2.356	0.650	3.193	0.231
SpcTmRn_ICA	0.669	2.554	0.665	3.314	0.248
SpcTmRn_MntSes	0.811	2.066	0.807	3.913	0.321
SpcTmRn_PCA	0.386	1.440	0.387	1.952	0.128
SpcTmRn_RfLag	0.835	4.294	0.836	3.961	0.134
SpcTmRn_SVD	0.704	2.548	0.703	3.385	0.240
SpcTmRn_WlLag	0.260	1.085	0.259	1.793	0.080
SpcTmWDpth	0.450	1.716	0.449	2.327	0.155

5.3 Model-family comparison

The model-family comparison indicates that sequence-learning models are more competitive when temporal memory is present in the feature space. Among the recurrent and deep-learning models, LSTM shows the most favorable average generalization behavior, with lower RMSE, MAE, NMSE, adjusted R2, and SMAPE gaps than GRU and several boosting methods. Decision tree and ridge regression exhibit small gaps for some metrics, but small gap alone is not sufficient for final selection; a model can generalize stably while remaining less accurate. The final interpretation therefore considers both the accuracy trend and the location of model-feature combinations in the accuracy-generalization trade-off space.

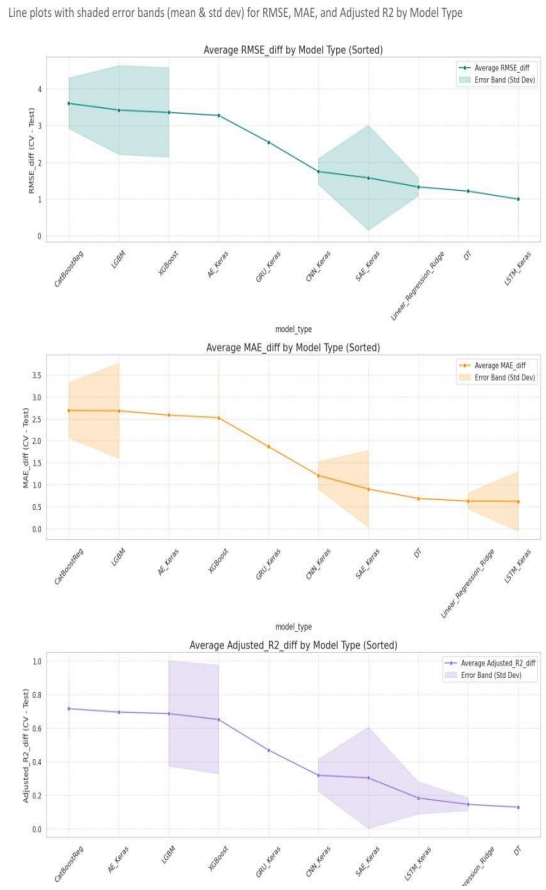

Figure 9. Average RMSE, MAE, and adjusted R2 performance trends by model type with uncertainty bands.

Figure 10. Average generalization gaps across performance indicators by model type.

Table 11. Average generalization gaps by model type across major performance indicators.

Model type	Adjusted R2 gap	MAE gap	NMSE gap	RMSE gap	SMAPE gap
AE_Keras	0.095	2.582	0.080	3.277	0.218
CNN_Keras	0.319	1.207	0.323	1.741	0.071
CatBoostReg	0.716	2.688	0.713	3.605	0.267
DT	0.129	0.082	0.128	1.207	0.035
GRU_Keras	0.469	1.863	0.465	2.547	0.159
LGBM	0.687	2.660	0.683	3.422	0.279
LSTM_Keras	0.183	0.617	0.183	0.989	0.060
Linear_Regression_Ridge	0.145	0.624	0.146	1.593	0.021
SAE_Keras	0.303	0.897	0.302	1.570	0.062
XGBoost	0.651	2.522	0.648	3.356	0.259

5.4 Accuracy-generalization trade-off and LSTM selection

The accuracy-generalization scatter plot is the central model-selection result. The ideal region is characterized by low independent test RMSE and low RMSE generalization gap. Models outside this region are less attractive because they either produce larger test error, exhibit a larger difference between validation and test behavior, or both. The LSTM-based configurations are positioned in a favorable trade-off region compared with many competing models, supporting the selection of LSTM as the preferred architecture for this dataset and validation design.

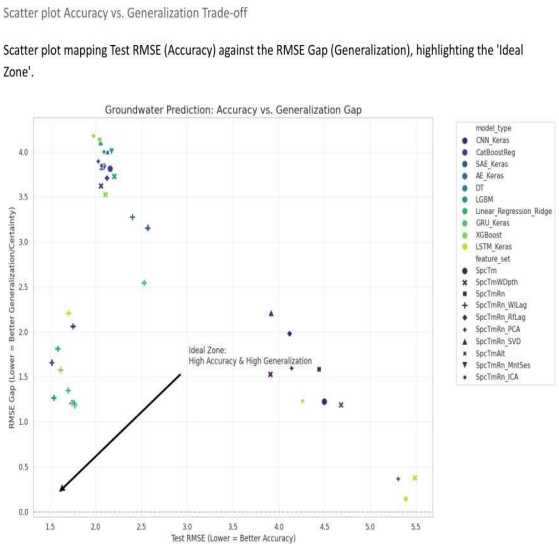

Figure 11. Accuracy-generalization trade-off between test RMSE and RMSE generalization gap.

Table 12. Final model-selection rationale based on accuracy, feature extraction, and generalization.

Selection criterion	Evidence from the analysis	Implication for final model choice
Temporal dependence	Groundwater level contains antecedent memory and delayed rainfall response, represented by lagged water-level and rainfall features.	Sequence-learning models are appropriate candidates.
Feature-set stability	SpcTmRn_Wl Lag shows low RMSE and SMAPE gaps among engineered feature sets.	Lagged groundwater memory is a critical feature-extraction contribution.
Model-family gap	LSTM_Keras shows RMSE gap 0.989 and SMAPE gap 0.060, lower than GRU, LightGBM, XGBoost, CatBoost, and autoencoder baselines.	LSTM offers strong generalization among deep-learning alternatives.
Accuracy-generalization balance	The trade-off plot favors low test RMSE combined with low gap rather than a single metric.	The final claim is based on balanced reliability, not only accuracy.
Interpretability requirement	Correlation and information-theoretic analysis identify physically meaningful drivers.	Model results are supported by exploratory hydrological evidence.

6. Discussion

6.1 Why feature extraction improved groundwater prediction

The results demonstrate that groundwater-level prediction benefits from representing hydrological memory explicitly. Lagged water-level features improve stability because they summarize antecedent aquifer storage, recent recharge history, local extraction effects, and delayed drainage. Rainfall lag features are also hydrologically meaningful, but rainfall alone does not fully explain groundwater variability because recharge is filtered by soil, geology, slope, land cover, vadose-zone storage, and pumping. The low direct correlation between rainfall and water level is therefore not contradictory; it reflects the fact that rainfall influence is delayed, spatially mediated, and nonlinear.

Cyclic month features provide a compact encoding of annual seasonality. Unlike integer month values, sine and cosine transformations preserve the circular relationship between December and January. This is important in monsoon-influenced groundwater systems where recharge and recession cycles repeat annually. PCA, SVD, and ICA components provide alternative representations of correlated numerical predictors. These components can reduce redundancy and improve stability, but they are less directly interpretable than lagged physical variables. For applied groundwater management, lagged and spatially interpretable features therefore remain preferable when their performance is comparable to or better than abstract components.

6.2 Why LSTM performed better than alternative deep-learning models

LSTM is designed to retain information over sequential inputs through gated memory. This makes

it suitable for groundwater-level prediction, where current level may depend on previous levels, previous rainfall, and seasonally delayed recharge processes. Compared with GRU, the LSTM architecture provides separate input, forget, and output gates, allowing more flexible control over retained and discarded temporal information. Compared with CNN, which emphasizes local feature extraction, LSTM is more directly aligned with sequential dependencies. Compared with static boosting models, LSTM can represent temporal state transitions more naturally when lagged features are included.

The results should be interpreted in a balanced way. LSTM is not universally superior for all groundwater systems, all feature sets, or all validation designs. In this study, LSTM provides the strongest combined accuracy-generalization behavior among deep-learning alternatives and competing nonlinear baselines. The methodological contribution is therefore not simply the use of LSTM; rather, it is the combination of leakage-aware feature extraction, time-series validation, and generalization-gap-based model selection that justifies LSTM as the preferred architecture for the examined dataset.

6.3 Hydrological interpretation of feature relevance

Well depth, altitude, latitude, and longitude emerge as important variables because they encode aspects of aquifer setting, terrain position, and spatial heterogeneity. The stronger mutual information of altitude and spatial coordinates relative to rainfall indicates that groundwater levels are not controlled only by month-to-month rainfall totals. Instead, the water-level response is conditioned by where the well is located and how the well interacts with the aquifer system. This interpretation is consistent with the spatial heatmaps and with the feature-set results showing that location-aware and lag-aware representations are more stable than rainfall-only expansions.

6.4 Practical implications

The proposed framework can support groundwater monitoring agencies by providing a reproducible method for transforming routine well and rainfall records into predictive information. Because the framework explicitly evaluates generalization gap, it helps avoid adopting models that appear accurate during validation but are unreliable on independent observations. The method can also help identify which features are worth collecting and maintaining. For example, well depth, coordinates, altitude, monthly rainfall, and historical water-level records all have clear predictive roles. In operational settings, such models could be used to flag wells likely to experience abnormal declines, prioritize monitoring locations, and support seasonal groundwater planning.

6.5 Limitations and future work

Several limitations should be recognized. First, the study is limited to Belagavi and Khanapur taluks, and model behavior may differ in aquifers with different geology, pumping intensity, and recharge mechanisms. Second, rainfall assignment is based on the nearest rain gauge within a 30 km threshold; although this is spatially defensible, gridded rainfall or radar-derived precipitation could further improve representation of local rainfall variability. Third, the dataset does not explicitly include groundwater extraction, irrigation demand, land use, soil type, aquifer lithology, canal influence, or groundwater-management interventions. Fourth, imputation of missing water-level observations introduces uncertainty, particularly where field conditions such as dry wells or inaccessible wells are frequent. Finally, LSTM models are less directly interpretable than linear or tree-based methods. Future work should add permutation importance, SHAP analysis, feature ablation, and seasonal error diagnostics for the final LSTM model.

7. Conclusion

This study developed a leakage-aware feature-extraction and model-selection framework for groundwater-level prediction using monthly rainfall-well data from Belagavi and Khanapur taluks. The analysis integrated groundwater observations, rainfall records, well attributes, spatial coordinates, and station-distance information into a consistent modeling dataset. Exploratory analysis showed that groundwater level is more strongly related to spatial and well-structural variables than to instantaneous rainfall alone, emphasizing the need for spatio-temporal and lagged feature representations.

Feature extraction improved the reliability of model behavior. In particular, lagged water-level features produced the most stable engineered feature set, indicating that groundwater memory is central to monthly prediction. The comparison across model families showed that LSTM provides the most favorable accuracy-generalization balance among recurrent and deep-learning architectures and outperforms several boosting baselines in generalization-gap behavior. The main contribution of the work is therefore a reproducible framework in which model selection is guided by independent test performance and generalization gap, not by a single accuracy score.

Future extensions should incorporate pumping data, land use, aquifer lithology, soil properties, groundwater abstraction records, and gridded climate products. Model interpretability should also be

strengthened through SHAP, permutation importance, and error analysis by season and location. Such extensions would further improve the suitability of the framework for groundwater planning and decision support.

Declarations

Data availability

The groundwater-level and rainfall datasets were obtained from the acknowledged government departments and processed into a monthly modeling table for the present analysis. Access to raw administrative datasets is subject to the policies of the respective data-providing offices.

Code availability

The computational workflow was implemented through reproducible preprocessing, feature-extraction, model-training, and generalization-analysis notebooks. Code can be made available with the final publication package subject to repository and data-permission requirements.

Conflict of interest

The authors declare no conflict of interest.

Funding

No specific funding statement is included in this draft. The final version should be updated according to the authors institutional and project funding details.

Acknowledgments

The authors acknowledge the Office of the Senior Geologist, District Groundwater Office, Groundwater Directorate, Belagavi, the Groundwater Directorate and Karnataka Groundwater Authority, Government of Karnataka, for groundwater-level records. The authors also acknowledge No. 3 Irrigation Investigation Sub Division, Belagavi, Water Resources Department, Government of Karnataka, for rainfall records. Administrative and mapping support is acknowledged from DataMeet community maps, GADM administrative boundaries, OpenStreetMap contributors, Leaflet, Folium, Plotly, and CARTO basemap services.

References

[1] Ahmadi *et al.*, "Groundwater Level Modeling with Machine Learning: A Systematic Review and Meta-Analysis," *Water*, vol. 14, no. 6, p. 949, Mar. 2022, doi: 10.3390/w14060949.

[2] L. Guillaumot, L. Longuevergne, J. Marçais, N. Lavenant, and O. Bour, "Frequency domain water table fluctuations reveal impacts of intense rainfall and vadose zone thickness on groundwater recharge," *Hydrology and earth system sciences*, vol. 26, no. 22, p. 5697, Nov. 2022, doi: 10.5194/hess-26-5697-2022.

[3] Y. Ma, C. Montzka, B. Bayat, and S. Kollet, "Using Long Short-Term Memory networks to connect water table depth anomalies to precipitation anomalies over Europe," *Hydrology and earth system sciences*, vol. 25, no. 6, p. 3555, Jun. 2021, doi: 10.5194/hess-25-3555-2021.

[4] M. L. Taccari, H. Wang, J. Nuttall, X. Chen, and P. K. Jimack, "Spatial-temporal graph neural networks for groundwater data," *Scientific Reports*, vol. 14, no. 1, p. 24564, Oct. 2024, doi: 10.1038/s41598-024-75385-2.

[5] Zhuang, L. Cui, and Y. Cui, "Enhancing groundwater level prediction with a hybrid deep learning model in Jinan City, China," *Scientific Reports*, vol. 15, no. 1, p. 44535, Dec. 2025, doi: 10.1038/s41598-025-28200-5.

[6] M. Sit, B. Z. Demiray, Z. Xiang, G. J. Ewing, Y. Sermet, and İ. Demir, "A comprehensive review of deep learning applications in hydrology and water resources," *Water Science & Technology*, vol. 82, no. 12. Pergamon Press, p. 2635, Aug. 05, 2020. doi: 10.2166/wst.2020.369.

[7] H. Afzaal, A. A. Farooque, F. Abbas, B. Acharya, and T. J. Esau, "Groundwater Estimation from Major Physical Hydrology Components Using Artificial Neural Networks and Deep Learning," *Water*, vol. 12, no. 1, p. 5, Dec. 2019, doi: 10.3390/w12010005.

[8] T. Kim, D. D. Thiem, T. T. N. Quynh, and T. Q. Nguyen, "Enhancing Prediction Accuracy and Data Handling for Environmental Applications in Innovative Modeling of Groundwater Level Fluctuations Based on the Tree Ensembles Technique," *Journal of Hydrologic Engineering*, vol. 30, no. 4, May 2025, doi: 10.1061/jhyeff.heeng-6395.

[9] A. Hussein, C. Thron, M. Ghaziasgar, A. Bagula, and M. Vaccari, "Groundwater Prediction Using Machine-Learning Tools," *Algorithms*, vol. 13, no. 11, p. 300, Nov. 2020, doi: 10.3390/a13110300.

[10] P. R. Patil, Y. D. Bafna, P. S. Khandikar, V. V. Nhayade, and S. D. Wani, "Predictive Modeling of Groundwater Resources Using Machine Learning and Spatial Analysis," *Zenodo (CERN European Organization for Nuclear Research)*, Jan. 2026, doi: 10.5281/zenodo.18257744.

[11] Metwally, P. S. Yu, D. Reiman, Y. Dai, P. W. Finn, and D. L. Perkins, "Utilizing

longitudinal microbiome taxonomic profiles to predict food allergy via Long Short-Term Memory networks," *PLoS Computational Biology*, vol. 15, no. 2, Feb. 2019, doi: 10.1371/journal.pcbi.1006693.

[12] J. Zhao *et al.*, "Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction," *Scientific Reports*, vol. 9, no. 1, Jan. 2019, doi: 10.1038/s41598-018-36745-x.

[13] S. Okser, T. Pahikkala, A. Airola, T. Salakoski, S. Ripatti, and T. Aittokallio, "Regularized Machine Learning in the Genetic Prediction of Complex Traits," *PLoS Genetics*, vol. 10, no. 11, Nov. 2014, doi: 10.1371/journal.pgen.1004754.

[14] G. Vishwakarma, A. Sonpal, and J. Hachmann, "Metrics for Benchmarking and Uncertainty Quantification: Quality, Applicability, and Best Practices for Machine Learning in Chemistry," *Trends in Chemistry*, vol. 3, no. 2, p. 146, Jan. 2021, doi: 10.1016/j.trechm.2020.12.004.