Hybrid One-Dimensional CNN–SVM and Attention-Based BI-Gru Models for Drinking-Water Potability Classification with Missing-Data Imputation

***KUSUMA SUNDARA KUMAR[1], Dr. JULURU PAVANU SAI [2]**

1PG Student, Department of CSE, VIKAS COLLEGE OF ENGINEERING AND TECHNOLOGY, Nunna, Vijayawada Rural, A.P
2 Associate Professor, Department of CSE, VIKAS COLLEGE OF ENGINEERING AND TECHNOLOGY, Nunna, Vijayawada Rural, A.P
*Corresponding author: skkusuma123@gmail.com

Abstract

Classifying drinking-water potability from routinely measured physicochemical parameters can reduce dependence on slow laboratory testing, but practical datasets are incomplete and indicator relationships are nonlinear. The present study develops and compares two learning pipelines for binary potability classification on a public water-quality dataset of 3,276 samples described by nine parameters (pH, hardness, total dissolved solids, chloramines, sulphate, conductivity, organic carbon, trihalomethanes and turbidity). Missing values, which affect 14.99% of pH, 23.84% of sulphate and 4.95% of trihalomethane records, are handled by an interpolation-based Missing Values Mutation (MVM) scheme and by a K-nearest-neighbour (KNN) imputer. The first pipeline couples a one-dimensional convolutional neural network (1D-CNN) feature extractor with a support vector machine (SVM) classifier; the second employs a stacked bidirectional gated recurrent unit (Bi-GRU) with an attention layer. Both are benchmarked against MLP, LSTM and GRU networks and against logistic regression, random forest, Gaussian naive Bayes and decision-tree classifiers. On the held-out test set, the CNN–SVM achieved an accuracy of 0.999 and an F1-score of 0.999, compared with 0.77 for logistic regression and 0.63 for random forest. In the sample-size analysis, CNN–SVM accuracy increased from 84% at 100 samples to 92% at 500 samples, and its recall reached 99.56% at 800 samples, marginally above Bi-GRU (99.34%) and well above MLP (83.45%). The radial basis function kernel gave the highest test correlation coefficient (0.801) among four SVM kernels. The findings indicate that combining learned convolutional features with a margin-based classifier, together with explicit missing-data treatment, is an effective strategy for potability screening.

Keywords: *Water quality; potability classification; convolutional neural network; support vector machine; bidirectional GRU; attention mechanism; missing-data imputation; KNN imputer*

1. INTRODUCTION

Access to safe drinking water is fundamental to public health, agriculture and industry. Groundwater and surface-water sources are under increasing pressure from population growth, urbanisation, industrial discharge, agricultural runoff and climate variability, and contaminants such as dissolved salts, disinfection by-products and suspended solids can render water unfit for consumption. Water quality is conventionally assessed by laboratory analysis of physical, chemical and biological parameters—for example pH, hardness, total dissolved solids (TDS), conductivity, organic carbon and turbidity—and by comparing the measured values with guideline limits. Although accurate, laboratory testing is labour-intensive, costly and slow, which limits the spatial and temporal coverage of monitoring programmes and delays the detection of deteriorating water quality.

Each of the commonly monitored parameters reflects a different aspect of water quality. pH indicates the acid–base balance of the water and influences the solubility of metals and the efficiency of disinfection; hardness arises mainly from dissolved calcium and magnesium salts; TDS and electrical conductivity reflect the concentration of dissolved ions; chloramines and trihalomethanes are associated with chlorine-based disinfection; organic carbon indicates the organic load that can react with disinfectants; and turbidity measures suspended matter that can shield microorganisms from treatment. Because these quantities are physically and chemically coupled—an increase in dissolved salts raises both TDS and conductivity, for example—the suitability of water for drinking is determined by their joint behaviour rather than by any single parameter considered in isolation.

Composite indices such as the water quality index (WQI) condense several parameters into a single number, and water-quality classification assigns samples to predefined categories according to such indices or to guideline limits. These

tools facilitate comparison across sites and time, but computing them still requires all constituent parameters to be measured, and the resulting categories depend on fixed weights and thresholds. A predictive classifier that learns the mapping from parameters to potability directly from labelled data can operate when some measurements are missing, can be updated as new data become available, and can serve as the core of an early-warning system that flags samples requiring immediate laboratory confirmation.

Data-driven models offer a complementary route. Once trained on historical measurements, a classifier can estimate the potability of a new sample within milliseconds from parameters that are cheaper or faster to measure, and can therefore support screening, early warning and the prioritisation of confirmatory laboratory tests. Machine learning (ML) methods such as support vector machines (SVMs), artificial neural networks and ensemble trees have accordingly been applied widely to water-quality prediction [1], [2], [3]. More recently, deep learning architectures—convolutional neural networks (CNNs), long short-term memory (LSTM) networks, gated recurrent units (GRUs) and their hybrids with attention mechanisms—have been proposed to capture nonlinear interactions among water-quality indicators [4], [5], [6].

The choice between these model families is not straightforward. Convolutional networks learn local combinations of neighbouring inputs through shared filters and are comparatively data-efficient, whereas recurrent networks model dependencies along a sequence through gated memory. Support vector machines, in turn, provide a well-founded margin-based decision rule with strong generalisation when the number of samples is moderate, but depend on the quality of the input representation. Combining a learned convolutional representation with an SVM decision stage, or complementing a recurrent encoder with an attention mechanism that weights the most informative inputs, are therefore two plausible strategies for improving potability classification. Their relative merits, however, have not been established on a common benchmark.

Two practical difficulties nevertheless limit the accuracy of such models. First, monitoring records are frequently incomplete because of sensor faults, irregular sampling or data-handling errors. Deleting incomplete records reduces the size of the training set and may bias the class distribution, whereas naive substitution by a mean value distorts the joint distribution of the parameters. Second, many approaches evaluate a single indicator or treat the indicators independently, although water quality is a multivariate phenomenon in which, for example, pH, conductivity and dissolved solids are interrelated. A model that learns representations of the joint parameter vector is therefore expected to classify potability more reliably than one based on isolated indicators.

A review of the literature summarised in Section 2 reveals the following gaps. (i) Deep sequence models for water quality are mostly evaluated on river or reservoir time series for regression of future concentrations [4], [5], [7], and fewer studies address binary potability classification from multi-parameter samples. (ii) Missing data are often removed rather than imputed, and the effect of the treatment of missing values is seldom reported together with the classifier. (iii) Unidirectional recurrent models such as the LSTM process the input in one direction only and may not exploit dependencies in the reverse direction, motivating bidirectional architectures. (iv) Hybrid CNN–SVM models, which combine automatically learned convolutional features with the margin-maximising decision boundary of an SVM, have been effective in image-based classification [8], [9] but have received limited attention for tabular water-quality data. (v) Comparative studies rarely examine how the relative performance of competing models changes with the number of training samples and the number of input parameters, which is important when monitoring data are scarce.

To address these gaps, the present study develops and compares two missing-data-aware classification pipelines for drinking-water potability. The main contributions are as follows:

(1) A preprocessing scheme that treats missing values explicitly, using interpolation-based Missing Values Mutation (MVM) and a K-nearest-neighbour (KNN) imputer, followed by z-score standardisation.

(2) A hybrid 1D-CNN–SVM classifier in which a one-dimensional convolutional network learns feature representations of the parameter vector and an SVM with a radial basis function (RBF) kernel performs the final potability decision.

(3) An attention-based stacked Bi-GRU classifier that processes the parameter sequence in both directions and weights the most informative hidden states.

(4) A systematic comparison with MLP, LSTM and GRU networks and with four conventional classifiers, including an analysis of sensitivity to sample size and to the number of input parameters, and a comparison of four SVM kernel functions.

The remainder of the paper is organised as follows. Section 2 reviews related work. Section 3 describes the dataset, preprocessing, model architectures and evaluation protocol. Section 4 presents and discusses the experimental results, and Section 5 summarises the conclusions and directions for future research.

2. LITERATURE REVIEW

2.1 Conventional machine-learning approaches

Machine-learning methods for water-quality prediction have been surveyed by Kalaivanan and Vellingiri [1], who compared the families of techniques applied to the problem. Najwa Mohd Rizal *et al.* [2] compared regression models, SVMs and artificial neural networks (ANNs) for river water-quality prediction, illustrating that nonlinear learners are generally preferred when the relationships among parameters are complex. Kuthe *et al.* [3] applied a range of machine-learning classifiers to water-quality prediction, and Madni *et al.* [10] employed automated machine learning (H2O AutoML) together with explainable-AI techniques to predict water quality and to interpret the contribution of individual parameters. Statistical and decision-analytic approaches have also been used; Yang *et al.* [11], for instance, combined a fuzzy analytic hierarchy process with exponential smoothing to evaluate seawater intrusion and predict water quality in the Dagu River. These studies establish that learning-based models can replace or supplement index-based assessment, but most rely on hand-crafted inputs and on complete records.

A common feature of these conventional approaches is that the quality of the prediction depends heavily on the choice of input variables and on the preprocessing applied to them. Linear models are easy to interpret but cannot represent interactions among parameters unless these are specified explicitly, whereas kernel methods and neural networks can learn such interactions implicitly at the cost of reduced interpretability. Explainable-AI tools, as used in [10], partly bridge this gap by quantifying the contribution of each parameter to the prediction.

2.2 Treatment of missing data

Incomplete records are a recurring problem in water-quality datasets. Juna *et al.* [12] combined a KNN imputer with a multilayer perceptron for water-quality prediction, demonstrating that imputation rather than deletion is a viable preprocessing strategy for this task. KNN imputation estimates a missing value from the most similar complete records, thereby preserving the multivariate structure of the data, whereas interpolation-based schemes estimate missing values from neighbouring observations in an ordered sequence. Despite their importance, imputation strategies are usually treated as a fixed preprocessing step rather than as an integral part of the classification pipeline, and their interaction with deep models has received little attention.

2.3 Deep and hybrid sequence models

Deep learning has been applied to water-quality prediction in several forms. Anand *et al.* [13] used a CNN for water-quality prediction, and Eze *et al.* [14] developed a prediction model for an aquaculture farm. Hybrid architectures that combine convolutional feature extraction with recurrent sequence modelling are particularly common: Yang *et al.* [4] and Mi *et al.* [15] proposed CNN–LSTM models with attention mechanisms, Guo *et al.* [16] a CNN–BiLSTM model with attention, and Zhang *et al.* [6] a PE-CNN-GRU hybrid. Bidirectional recurrent networks with attention were employed by Zhang *et al.* [5] for watershed water-quality prediction. Signal-decomposition approaches have also been explored, including combinations of ANNs, wavelet transforms and LSTM [17], the synchrosqueezed wavelet transform with an improved deep extreme learning machine [18], and signal decomposition with ensemble deep learning [19]. Multi-task learning [7], graph convolutional networks with feature and temporal attention [20], and adaptive graph convolution combined with CNN–LSTM for multi-site prediction [21] extend these ideas to multiple outputs and spatially distributed monitoring stations.

Two design principles recur across these studies. The first is the use of convolutional layers as feature extractors that summarise local patterns in the input before a sequence model or a regressor is applied [4], [16], [6], [21]. The second is the use of attention mechanisms that assign data-dependent weights to hidden states, thereby focusing the model on the most informative inputs or time steps [4], [5], [15], [16], [20]. Bidirectional recurrent layers, which process the input

in both directions, have been adopted to exploit context on both sides of each position [5], [16]. At the same time, these models have been developed predominantly for forecasting continuous water-quality variables at particular monitoring sites, and their behaviour as classifiers of potability from independent multi-parameter samples—where no natural temporal order exists—has received less attention.

2.4 Hybrid CNN–SVM classifiers

In a hybrid CNN–SVM, the convolutional layers act as a trainable feature extractor and the SVM replaces the softmax or sigmoid output layer as the classifier. Because the SVM maximises the margin between classes and can model nonlinear boundaries through kernels, the combination can generalise well when the number of training samples is moderate. Such hybrids have been reported for tumour detection in MRI images [8] and for rice-leaf disease detection [9]. Their application to multi-parameter water-quality data, in which the input is a short one-dimensional vector rather than an image, has been explored far less.

2.5 Summary and research gap

The reviewed studies show a clear trend towards hybrid deep models and attention mechanisms, but they mostly target regression of future concentrations at specific river or reservoir sites. For potability classification from multi-parameter samples, open questions remain regarding (a) the joint treatment of missing values and classification, (b) the relative merits of a convolutional–SVM hybrid and a bidirectional recurrent model with attention, and (c) the robustness of competing models when training data are limited or when fewer parameters are available. The present study addresses these questions within a single experimental framework.

3. MATERIALS AND METHODOLOGY

3.1 Overall framework

The proposed methodology is summarised in Fig. 1. Raw records are first screened for missing values, which are imputed; the parameters are standardised; and the data are divided into training and testing subsets. Two proposed classifiers—a hybrid 1D-CNN–SVM and an attention-based Bi-GRU—are trained alongside deep and conventional baselines, and all models are evaluated with the same metrics. The CNN–SVM pipeline is additionally designed so that samples classified as polluted can be passed to the SVM stage for a further subdivision into slightly and extremely polluted categories.

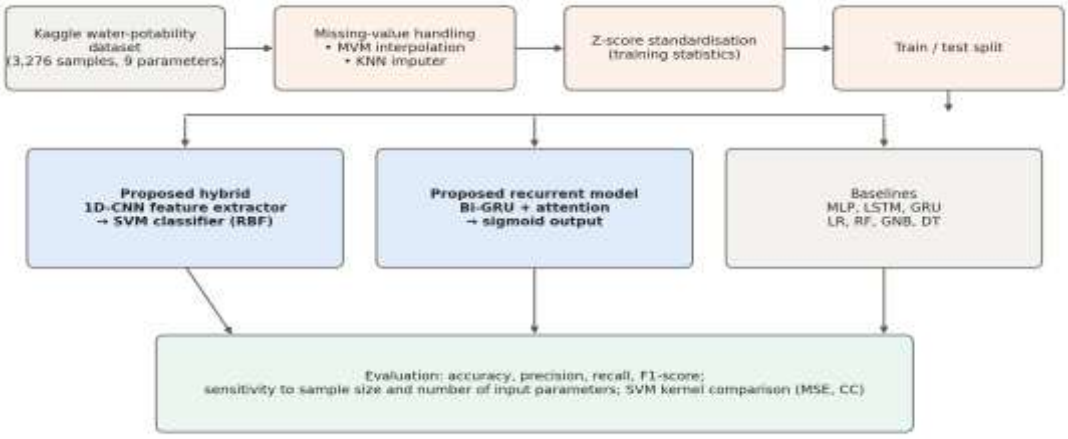


Fig. 1. Overall framework of the proposed methodology: data acquisition, missing-value handling, standardisation, data partitioning, the two proposed classifiers and the baselines, and performance evaluation.

3.2 Dataset

The experiments use the publicly available water-potability dataset hosted on Kaggle, which contains measurements for 3,276 water bodies. Each record comprises nine physicochemical parameters and a binary target, *Potability*, where 1 denotes water that is safe for human consumption and 0 denotes water that is not. Approximately 61% of the records are labelled non-potable and 39% potable, so the classes are moderately imbalanced. The parameters, their units and the guideline limits adopted in the present study are listed in Table 1.

Table 1. Input parameters of the water-potability dataset and the guideline limits adopted in the present study.

Parameter	Description / unit	Guideline limit
pH	Acid–base balance (0–14)	6.5–8.5
Hardness	Soap-precipitating capacity due to Ca and Mg salts (mg/L)	200 mg/L
Solids (TDS)	Total dissolved solids (ppm)	1000 ppm
Chloramines	Chloramine concentration (ppm)	4 ppm
Sulphate	Dissolved sulphate concentration (mg/L)	1000 mg/L
Conductivity	Electrical conductivity (µS/cm)	400 µS/cm
Organic carbon	Total organic carbon (ppm)	10 ppm
Trihalomethanes	Chlorination by-products (µg/L)	80
Turbidity	Light-scattering by suspended matter (NTU)	5 NTU
Potability	Target: 1 = potable, 0 = non-potable	–

The distributions of the nine parameters are shown in Fig. 2. All parameters exhibit approximately unimodal, bell-shaped distributions without strong skewness, although their ranges differ by several orders of magnitude—for example, TDS spans roughly 0–60,000 ppm whereas turbidity lies between about 1.5 and 7 NTU. Most pH values are concentrated around neutral conditions, between approximately 5 and 10. These differences in scale motivate the standardisation step described in Section 3.3, without which parameters with large numerical ranges would dominate distance-based methods such as the SVM and the KNN imputer.

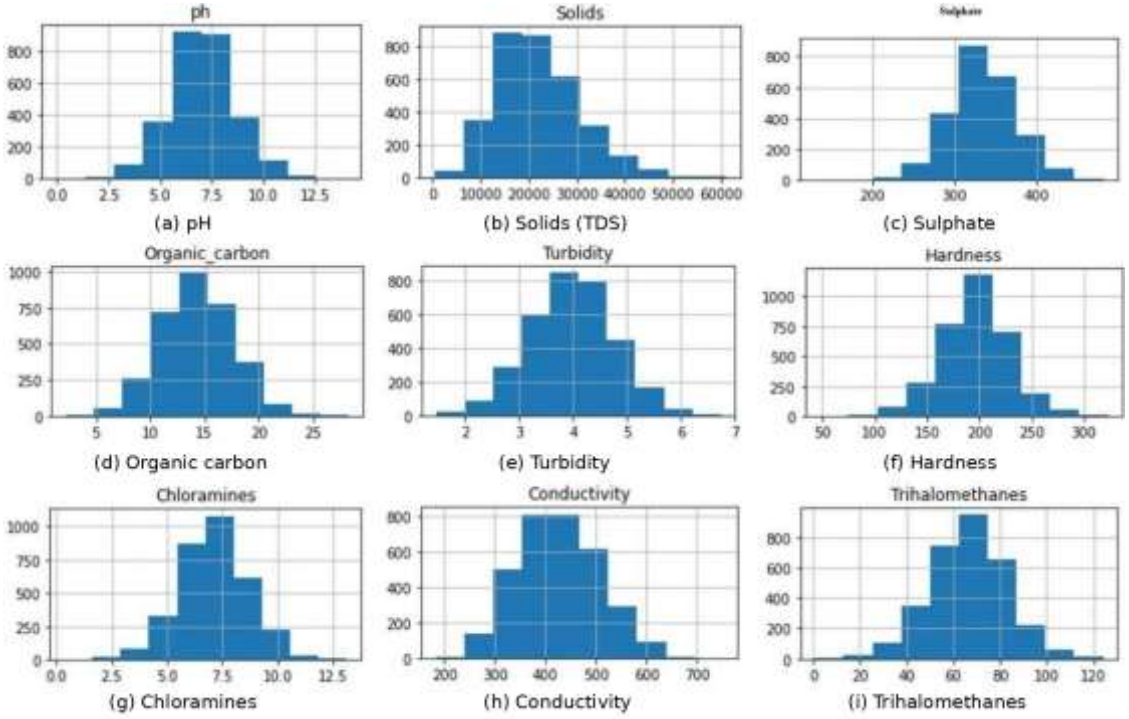


Fig. 2. Histograms of the nine input parameters: (a) pH, (b) solids (TDS), (c) sulphate, (d) organic carbon, (e) turbidity, (f) hardness, (g) chloramines, (h) conductivity and (i) trihalomethanes.

3.3 Missing-value treatment and standardisation

Table 2 summarises the missing values in the dataset. Three parameters are incomplete: pH (491 records, 14.99%), sulphate (781 records, 23.84%) and trihalomethanes (162 records, 4.95%). In total, 1,434 of the 29,484 parameter values (3,276 × 9) are missing, i.e. 4.86% of the feature matrix. Because sulphate alone is missing in almost one quarter of the records, deleting every incomplete record would discard a substantial part of the dataset; imputation was therefore adopted.

Table 2. Missing values in the water-potability dataset.

Parameter	Missing values	Missing (%)
pH	491	14.99
Sulphate	781	23.84
Trihalomethanes	162	4.95
Hardness, solids, chloramines, conductivity, organic carbon, turbidity, potability	0	0.00

Missing Values Mutation (MVM). For the CNN–SVM pipeline, each parameter *i* is treated as an ordered series of observations $(y_{i,1}, T_1), \ldots, (y_{i,n}, T_n)$. If the value $y_{i,v}$ at position T_v is missing, the two nearest positions T_u and T_w with available values are located, and the missing value is estimated by linear interpolation:

$$\hat{Y}_{i,v} = y_{i,u} + [(y_{i,u} - y_{i,w}) / (T_u - T_w)] \cdot (T_v - T_u) \quad (1)$$

KNN imputer. For the Bi-GRU pipeline and the conventional classifiers, missing values were imputed with the KNN imputer of scikit-learn, which replaces a missing entry by the mean of the corresponding parameter in the *k* nearest complete neighbours. Distances are computed only over the coordinates present in both records and are rescaled to compensate for the missing coordinates:

$$D_{xy} = [\, w \cdot \Sigma_{j \in present} (x_j - y_j)^2 \,]^{1/2}$$
$$w = (\text{total number of coordinates}) / (\text{number of present coordinates}) \quad (2)$$

Standardisation. After imputation, each parameter was standardised using the mean μ and standard deviation σ computed over the training samples:

$$X_n = (X - \mu) / \sigma \quad (3)$$

so that every parameter has zero mean and unit variance. For the models compared in the sample-size analysis, all input and output variables were additionally min–max scaled to the range [0, 1].

3.4 Hybrid 1D-CNN–SVM classifier

The architecture of the hybrid classifier is shown in Fig. 3(a). The nine standardised parameters of a sample are arranged as a one-dimensional sequence and passed to a 1D-CNN consisting of one input layer, five hidden layers and one output layer. The first convolutional layer contains eight filters that scan all parameters; it is followed by a max-pooling layer, a second convolutional layer, a flatten layer and a fully connected layer with sigmoid activation. Each convolutional layer computes

$$h_t = f(x_t * k_t + b_t), \qquad f(z) = \max(0, z) \quad (4)$$

where x_t is the input, k_t the convolution kernel, b_t the bias, $*$ the convolution operator and *f* the rectified linear unit (ReLU). The ReLU accelerates training and mitigates vanishing gradients, and max pooling retains the dominant responses while reducing the dimensionality of the feature maps. The CNN was trained with the mean squared error (MSE) loss using stochastic gradient descent (SGD) with a learning rate of 0.001, a mini-batch size of 16 and 1000 epochs.

The representations learned by the convolutional layers are then classified by an SVM. For training samples (x_i, y_i) with $y_i \in \{-1, +1\}$, the soft-margin SVM minimises $\tfrac{1}{2}\|z\|^2 + C \Sigma \varepsilon_i$ subject to $y_i(z^T x_i + b) \geq 1 - \varepsilon_i$, $\varepsilon_i \geq 0$, where C controls the trade-off between margin width and training error. The resulting decision function with kernel K is

$$f(x) = \text{sign}(\, \Sigma_{i=1}^{n} \alpha_i^* \, y_i \, K(x_i, x) + b^* \,), \qquad K_{RBF}(x_i, x_j) = \exp(-\gamma \, \|x_i - x_j\|^2) \quad (5)$$

where α_i^* are the Lagrange multipliers (non-zero only for support vectors) and $\gamma > 0$ controls the width of the RBF kernel. Linear, polynomial ($(x_i \cdot x_j + 1)^d$), RBF and sigmoid kernels were evaluated, and the kernel parameters (C, d, γ) were selected by trial and error with 10-fold cross-validation. The kernel comparison used the mean squared error and the correlation coefficient $CC = \Sigma(m - \bar{m})(p - \bar{p}) / \sqrt{[\Sigma(m - \bar{m})^2 \, \Sigma(p - \bar{p})^2]}$ between observed values *m* and predicted values *p*.

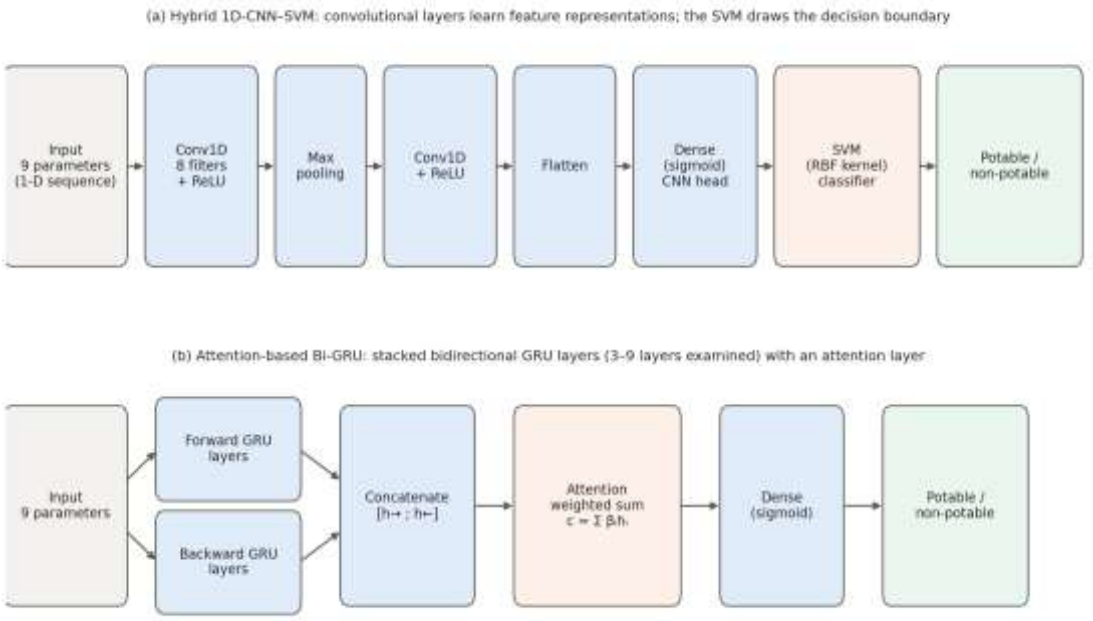


Fig. 3. Architectures of the proposed classifiers: (a) hybrid 1D-CNN–SVM; (b) attention-based bidirectional GRU.

3.5 Attention-based Bi-GRU classifier

The second proposed model, shown in Fig. 3(b), processes the parameter sequence with stacked bidirectional GRU layers. A GRU updates its hidden state g_t through a reset gate R_t and an update gate v_t; a forward GRU and a backward GRU process the sequence in opposite directions, and their hidden states are concatenated. An attention layer then forms a weighted sum of the Bi-GRU outputs so that the most informative positions contribute most to the final representation:

$$R_t = \sigma(P_r y_t + V_r g_{t-1}), \quad v_t = \sigma(P_u y_t + V_u g_{t-1}), \quad \tilde{g}_t = \tanh(P_c y_t + V(R_t \odot g_{t-1}))$$
$$g_t = (1 - v_t) \odot g_{t-1} + v_t \odot \tilde{g}_t, \qquad g_t^{Bi} = [g_t^{\rightarrow}\,;\, g_t^{\leftarrow}], \qquad c = \Sigma_{i=1}^{S} \beta_i\, g_i^{Bi} \qquad (6)$$

where σ is the sigmoid function, ⊙ the element-wise product, P and V the trainable weight matrices, and β_i the attention weights. The context vector *c* is passed to a sigmoid output layer. Architectures with three to nine stacked layers were examined, and the nine-layer configuration was retained. The attention mechanism suppresses irrelevant features and emphasises correlated, informative ones, which is expected to benefit a classifier whose inputs are heterogeneous physicochemical parameters.

3.6 Baselines, data partitioning and evaluation

The deep baselines were an MLP, a unidirectional LSTM and a unidirectional GRU. The conventional baselines were logistic regression (LR), random forest (RF), Gaussian naive Bayes (GNB) and a decision tree (DT); additional scikit-learn classifiers (SVC, KNN, SGD classifier and XGBoost) were used in the exploratory stage of the Bi-GRU pipeline. For the CNN–SVM experiments the dataset was randomly divided into two equal halves for training and testing, whereas the Bi-GRU pipeline used a 70:30 train–test split. To examine robustness, the deep models were evaluated with training subsets of 100 to 800 samples and with 1 to 10 input parameters. Performance was quantified by accuracy = (TP + TN)/(TP + TN + FP + FN), precision = TP/(TP + FP), recall = TP/(TP + FN) and F1-score = 2 × precision × recall/(precision + recall), where TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives. All experiments were implemented in Python 3 (Jupyter Notebook, scikit-learn) on a Windows 10 workstation with a seventh-generation Intel Core i7 processor. The main experimental settings are listed in Table 3.

Table 3. Experimental configuration of the proposed models.

Item	Setting
Dataset	Kaggle water potability; 3,276 samples; 9 parameters; binary target
Missing-value handling	MVM linear interpolation (CNN–SVM); KNN imputer (Bi-GRU, ML baselines)
Scaling	z-score standardisation; min–max scaling to [0, 1] for sample-size study
Train–test split	50:50 (CNN–SVM experiments); 70:30 (Bi-GRU pipeline)

Item	Setting
1D-CNN	1 input, 5 hidden, 1 output layer; 8 filters in first Conv1D; ReLU; max pooling; flatten; dense (sigmoid)
CNN training	MSE loss; SGD; learning rate 0.001; batch size 16; 1000 epochs
SVM	Linear, polynomial, RBF and sigmoid kernels; (C, d, γ) tuned by 10-fold cross-validation; RBF selected
Bi-GRU	Stacked bidirectional GRU (3–9 layers; 9 retained) + attention + sigmoid output
Baselines	MLP, LSTM, GRU; LR, RF, GNB, DT
Metrics	Accuracy, precision, recall, F1-score; MSE and CC for kernel comparison

4. RESULTS AND DISCUSSION

4.1 Comparison with conventional classifiers

Table 4 compares the hybrid CNN–SVM with four conventional classifiers on the test set. The CNN–SVM achieved the highest values of all four metrics (accuracy 0.999, precision 0.997, recall 0.998 and F1-score 0.999). Relative to logistic regression (accuracy 0.77, F1-score 0.76), this corresponds to improvements of 0.229 in accuracy and 0.239 in F1-score, i.e. 29.7% and 31.4% in relative terms [(0.999 − 0.77)/0.77 = 0.297]. The decision tree (0.998) and Gaussian naive Bayes (0.996) also performed strongly, and the margins between them and the CNN–SVM (0.001 and 0.003 in accuracy) are small; the principal distinction of the hybrid model therefore lies not in a large accuracy gain over the best conventional learners but in the consistently highest precision, recall and F1-score across all metrics.

Table 4. Test-set performance of the hybrid CNN–SVM and conventional classifiers. The best value in each column is shown in bold.

Model	Accuracy	Precision	Recall	F1-score
Logistic regression	0.770	0.740	0.750	0.760
Random forest	0.630	0.390	0.630	0.480
Gaussian naive Bayes	0.996	0.994	0.992	0.997
Decision tree	0.998	0.996	0.997	0.998
CNN–SVM (proposed)	**0.999**	**0.997**	**0.998**	**0.999**

The random forest performed worst (accuracy 0.63, precision 0.39, F1-score 0.48). Its accuracy is close to the proportion of the majority (non-potable) class, approximately 0.61, and its precision is close to the value 0.61 × 0.61 ≈ 0.37 that a classifier assigning every sample to the majority class would obtain under class-weighted averaging. This pattern suggests that, with the settings used, the random forest collapsed largely towards majority-class prediction, a known failure mode on imbalanced data when the splitting criterion finds little separable structure. Logistic regression, which can represent only a linear decision boundary, reached 0.77. The large gap between these two learners and the CNN–SVM indicates that the boundary between potable and non-potable samples in the standardised parameter space is strongly nonlinear, which is consistent with the interrelated nature of the parameters discussed in Section 1.

The balance between precision and recall provides further insight. For Gaussian naive Bayes, precision (0.994) exceeds recall (0.992), whereas for the CNN–SVM and the decision tree recall is equal to or slightly higher than precision (0.998 versus 0.997 and 0.997 versus 0.996). In the CNN–SVM the convolutional stage produces a compact representation in which the two classes are better separated, and the SVM then places the decision boundary so as to maximise the margin to the nearest samples of both classes; this symmetric treatment tends to avoid a systematic bias towards either false positives or false negatives. The naive Bayes classifier, by contrast, assumes conditional independence of the parameters, an assumption that is violated for coupled quantities such as TDS and conductivity, and its slightly lower recall is consistent with this simplification.

4.2 Effect of training-sample size

Fig. 4 and Table 5 show how the deep models behave as the number of training samples increases. All models improve with more data, but the ranking MLP < LSTM < GRU < Bi-GRU < CNN–SVM is preserved at every sample size reported. With only 100 samples, the CNN–SVM reached 84% accuracy, 2 percentage points above Bi-GRU (82%) and

29 points above the MLP (55%). At 500 samples, the accuracies were 67.56% (MLP), 76% (LSTM), 85% (GRU), 91.2% (Bi-GRU) and 92% (CNN–SVM). At 800 samples, recall reached 99.56% for CNN–SVM and 99.34% for Bi-GRU, compared with 93.4% for GRU, 86% for LSTM and 83.45% for MLP.

Table 5. Performance of the deep models for different training-sample sizes (values in %).

Model	Accuracy, N = 100	Accuracy, N = 500	Gain 100→500 (pp)	Recall, N = 800
MLP	55.00	67.56	+12.56	83.45
LSTM	68.00	76.00	+8.00	86.00
GRU	75.00	85.00	+10.00	93.40
Bi-GRU (proposed)	82.00	91.20	+9.20	99.34
CNN–SVM (proposed)	**84.00**	**92.00**	+8.00	**99.56**

pp = percentage points.

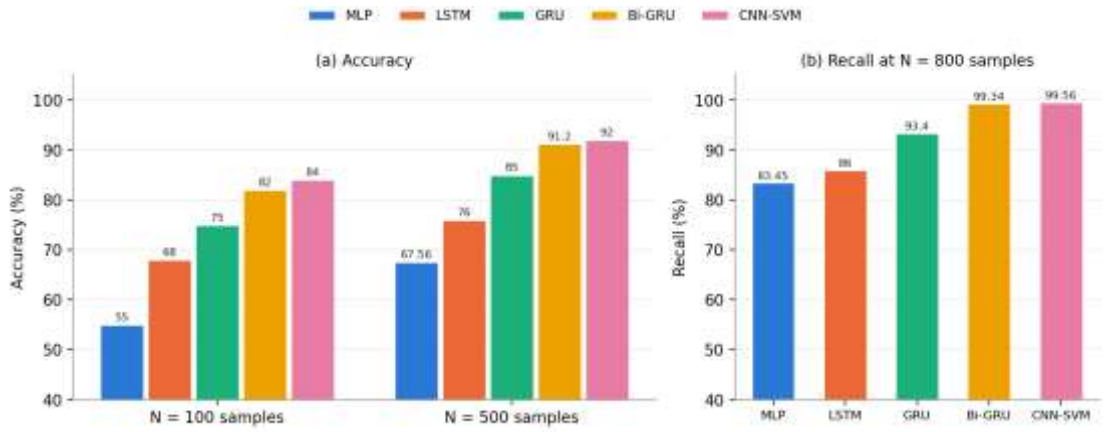


Fig. 4. Performance of the deep models versus training-sample size: (a) accuracy at 100 and 500 samples; (b) recall at 800 samples. Values as reported in Table 5.

Two trends are noteworthy. First, the models that performed worst with small samples gained most from additional data: the MLP improved by 12.56 percentage points (22.8% relative) between 100 and 500 samples, whereas the CNN–SVM improved by 8 points (9.5% relative). The weight-sharing convolutional filters of the CNN and the margin-based SVM decision rule impose a strong inductive bias that reduces the number of effective parameters to be estimated, so the hybrid model reaches a high accuracy with fewer samples; the fully connected MLP, by contrast, must learn every interaction from data. Second, the gap between CNN–SVM and Bi-GRU narrowed as data increased (2 points at 100 samples, 0.8 at 500 and 0.22 in recall at 800). Expressed as error rates, however, the advantage remains meaningful: at 500 samples the error of CNN–SVM (8.0%) is 9.1% lower than that of Bi-GRU (8.8%), 46.7% lower than that of GRU (15.0%) and 75.3% lower than that of the MLP (32.44%).

The consistent ordering among the recurrent models is also informative. The GRU outperformed the LSTM at every sample size (e.g. 85% versus 76% at 500 samples), which may be attributed to its smaller number of gates and parameters and hence its lower data requirement. The bidirectional GRU further improved on the unidirectional GRU by 7 points at 100 samples and 6.2 points at 500 samples. Because the parameter vector has no natural temporal direction, processing it in both directions allows every hidden state to depend on all other parameters, and the attention layer can then emphasise the most discriminative ones. In practical terms, these results indicate that the CNN–SVM and Bi-GRU models are preferable when only a few hundred labelled samples are available, as is typical for newly established monitoring networks.

4.3 Effect of the number of input parameters

Fig. 5 shows the accuracy of the deep models as the number of input parameters increases from 1 to 10. Accuracy increases monotonically with the number of parameters for every model, confirming that potability depends jointly on several indicators and cannot be inferred reliably from a single measurement. With a single parameter, the CNN–SVM reached 87% accuracy, and its recall increased from 87% with one parameter to 98% with the full parameter set; its precision increased from 87% to 95% as the number of parameters grew from 1 to 8. The two proposed models remained close to each other over the entire range, whereas the MLP was the most sensitive to the number of inputs, rising from

below 50% with one parameter to approximately 73% with ten (values read from Fig. 5). The precision, recall and F1-score followed the same pattern as accuracy.

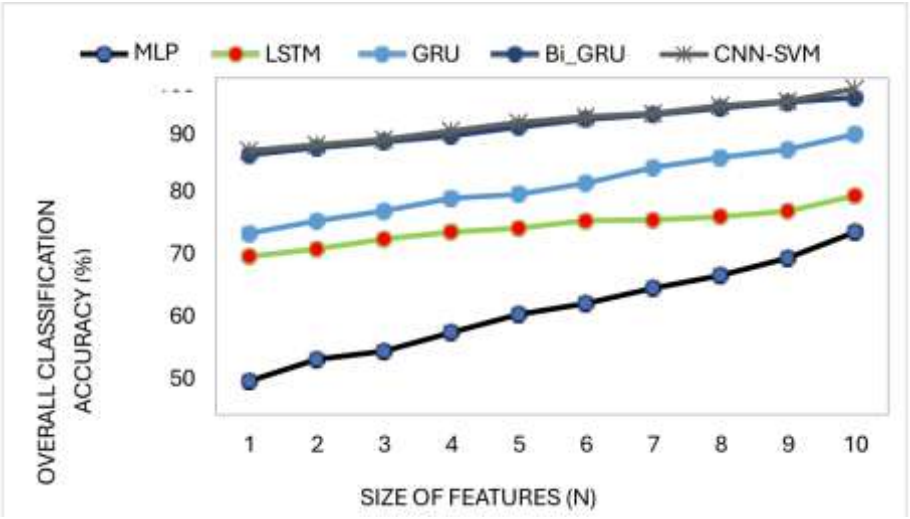


Fig. 5. Overall classification accuracy of MLP, LSTM, GRU, Bi-GRU and CNN–SVM as a function of the number of input parameters.

The weaker dependence of the proposed models on the number of inputs has practical significance. In field deployments, not all parameters are always available—for example, sulphate and trihalomethane analyses require laboratory equipment, whereas pH, conductivity and turbidity can be measured with low-cost probes. A model that retains high accuracy with a reduced parameter set can therefore support rapid screening with inexpensive sensors, reserving full laboratory analysis for samples that are flagged as suspicious.

The shape of the curves in Fig. 5 is also consistent with the physical interpretation of the parameters. The steady rather than saturating increase in accuracy up to the full parameter set suggests that each additional indicator contributes information that is not fully redundant with the others; this is expected because the parameters capture distinct processes—dissolved ions, disinfection by-products, organic load and suspended solids. For the MLP, the gain obtained by adding parameters (roughly 24 percentage points between one and ten inputs, read from Fig. 5) is several times larger than for the proposed models, which indicates that the convolutional and attention-based architectures extract more discriminative information from each individual input.

4.4 Effect of the SVM kernel function

Table 6 compares the four SVM kernel functions in terms of MSE and correlation coefficient (CC). The RBF kernel achieved the highest CC on both the training (0.998) and testing (0.801) data. Its testing CC exceeds those of the polynomial (0.496), linear (0.438) and sigmoid (0.336) kernels by 61.5%, 82.9% and 138.4%, respectively [e.g. (0.801 − 0.496)/0.496 = 0.615]. The RBF kernel maps the inputs into an infinite-dimensional feature space and can therefore represent the localised, nonlinear decision regions implied by the results of Section 4.1. The sigmoid kernel performed worst on every criterion, with a training MSE of 1.155 and a training CC of only 0.105, indicating that it failed to fit the data.

Table 6. Performance of the SVM with four kernel functions. The best value in each column is shown in bold.

Kernel	Mean squared error		Correlation coefficient	
	Training	Testing	Training	Testing
Linear	0.229	**0.189**	0.564	0.438
Polynomial	**0.177**	0.217	0.732	0.496
RBF	**0.177**	0.322	**0.998**	**0.801**
Sigmoid	1.155	0.418	0.105	0.336

The two criteria do not, however, rank the kernels identically. The linear kernel yielded the lowest testing MSE (0.189), and the testing MSE of the RBF kernel (0.322) is 82% higher than its training MSE (0.177), while its CC falls from 0.998 to 0.801. This gap between training and testing performance indicates a degree of overfitting of the RBF model,

which is expected when the kernel width γ allows the decision surface to follow individual training samples closely. Conversely, the linear kernel generalises with small absolute error but captures the trend of the data less faithfully, as reflected in its lower CC. The RBF kernel was retained because it best preserved the correlation between observed and predicted outputs, which is most relevant to classification; nevertheless, the training–testing gap suggests that stronger regularisation (a smaller C or larger kernel width) could further improve generalisation.

4.5 Practical implications for water-quality monitoring

From an operational standpoint, the two types of classification error have unequal consequences. A non-potable sample that is classified as potable (a missed contamination) poses a direct health risk, whereas a potable sample classified as non-potable leads only to an unnecessary confirmatory test. The high recall of the proposed models at larger sample sizes (99.56% for CNN–SVM and 99.34% for Bi-GRU at 800 samples) is therefore particularly relevant, provided that the positive class is defined as the class whose omission is most costly; in deployment, the decision threshold of the classifier can be adjusted to favour the detection of non-potable water at the expense of additional false alarms.

Predictive classification is best viewed as a complement to, rather than a replacement for, physical monitoring. Monitoring data remain the most reliable source of information on polluted water bodies, but models can be used where monitoring is impractical, can extend sparse monitoring networks by estimating the condition of unsampled sources, and can prioritise the samples that should be analysed first. Models of this kind can also support scenario analysis—for example, estimating how improved treatment or changes in discharges would alter the proportion of samples classified as potable. The ability of the proposed models to operate with imputed values and with reduced parameter sets makes them suitable for such screening roles in resource-constrained settings.

4.6 Discussion and limitations

Taken together, the results support three observations. First, explicit treatment of missing data allows all 3,276 records to be used even though sulphate is missing in 23.84% of them, which avoids the loss of information and the potential class-distribution bias associated with listwise deletion. Second, combining learned convolutional features with a margin-based classifier yields the best performance among the models compared, and this advantage is most pronounced when training data or input parameters are limited. Third, among the recurrent models, bidirectional processing with attention consistently outperforms unidirectional GRU and LSTM networks, making the attention-based Bi-GRU a competitive alternative to the hybrid model.

The two imputation strategies embody different assumptions. The MVM scheme estimates a missing value from the two nearest available observations of the same parameter and therefore exploits continuity along the ordering of the records, whereas the KNN imputer estimates it from the most similar records across all parameters and therefore exploits the multivariate correlation structure. For independent samples without temporal ordering, as in the present dataset, the KNN approach is conceptually better aligned with the data, while MVM is better suited to sensor time series collected at regular intervals. Both approaches retain the full sample size, which is particularly important for sulphate, the parameter with the highest missing rate.

Several limitations should be acknowledged. All experiments were conducted on a single public dataset without geographic or temporal metadata, so the generalisability of the models to other water sources, climates and measurement protocols remains to be established. The dataset contains no timestamps, so the ordering used by the interpolation-based MVM scheme and the "sequence" processed by the convolutional and recurrent models are orderings of samples and parameters rather than true time series. The near-perfect test scores of several classifiers in Table 4 warrant confirmation through repeated cross-validation with strict separation of the imputation and scaling steps from the test data, and through evaluation on an independent dataset. In addition, the MVM and KNN strategies were applied in different pipelines, so their individual contributions to classification accuracy were not isolated; a controlled comparison of imputation methods, including listwise deletion, would clarify how much of the performance is attributable to the treatment of missing data. Finally, the subdivision of polluted samples into slightly and extremely polluted categories is supported by the architecture but was not evaluated quantitatively, because the dataset provides only a binary potability label.

5. CONCLUSIONS

The present study developed two missing-data-aware deep-learning pipelines for drinking-water potability classification and compared them with deep and conventional baselines on a public dataset of 3,276 samples and nine physicochemical parameters. The main conclusions are as follows:

(1) Missing values affected three parameters (pH 14.99%, sulphate 23.84% and trihalomethanes 4.95%; 4.86% of all parameter values). Interpolation-based MVM and KNN imputation allowed the complete dataset to be used for training and evaluation.

(2) The hybrid 1D-CNN–SVM achieved the highest test-set performance (accuracy 0.999, precision 0.997, recall 0.998 and F1-score 0.999), outperforming logistic regression by 0.229 in accuracy, while the random forest collapsed largely towards majority-class prediction (accuracy 0.63, precision 0.39).

(3) The ranking MLP < LSTM < GRU < Bi-GRU < CNN–SVM was preserved for all sample sizes. The CNN–SVM reached 84% accuracy with 100 samples and 92% with 500 samples, and a recall of 99.56% with 800 samples; the attention-based Bi-GRU followed closely (82%, 91.2% and 99.34%).

(4) Accuracy increased with the number of input parameters for all models, while the proposed models were the least sensitive to reduced inputs (CNN–SVM: 87% with one parameter, recall 98% with all parameters), which favours deployment with low-cost sensor subsets.

(5) The RBF kernel yielded the highest testing correlation coefficient (0.801) among four SVM kernels, although its larger training–testing gap in MSE indicates scope for stronger regularisation.

SCOPE FOR FUTURE WORK: Future work will validate the models on independent, multi-site datasets with temporal information, evaluate the imputation strategies quantitatively against listwise deletion, address the class imbalance through balanced sampling, extend the classifier to multi-level pollution categories, and integrate the models with real-time sensor streams for continuous water-quality screening.

Data availability: The water-potability dataset used in this study is publicly available on the Kaggle platform.

REFERENCES

[1] K. Kalaivanan and J. Vellingiri, "Survival study on different water quality prediction methods using machine learning," Nature Environment and Pollution Technology, 2022.

[2] N. Najwa Mohd Rizal, G. Hayder, M. Mnzool, B. M. Elnaim, A. O. Mohammed, and M. M. Khayyat, "Comparison between regression models, support vector machine (SVM), and artificial neural network (ANN) in river water quality prediction," Processes, 2022.

[3] A. Kuthe, C. Bhake, V. Bhoyar, A. Yenurkar, V. Khandekar, and K. Gawale, "Water quality prediction using machine learning," International Journal of Computer Science and Mobile Computing, 2023.

[4] Y. Yang, Q. Xiong, C. Wu, Q. Zou, Y. Yu, H. Yi, and M. Gao, "A study on water quality prediction by a hybrid CNN-LSTM model with attention mechanism," Environmental Science and Pollution Research, vol. 28, pp. 55129–55139, 2021.

[5] Q. Zhang, R. Wang, Y. Qi, and F. Wen, "A watershed water quality prediction model based on attention mechanism and Bi-LSTM," Environmental Science and Pollution Research International, vol. 29, pp. 75664–75680, 2022.

[6] L. Zhang, J. Xie, X. Liu, W. Zhang, and P. Geng, "Research on water quality prediction based on PE-CNN-GRU hybrid model," E3S Web of Conferences, 2023.

[7] X. Wu, Q. Zhang, F. Wen, and Y. Qi, "A water quality prediction model based on multi-task deep learning: A case study of the Yellow River, China," Water, 2022.

[8] M. O. Khairandish, M. Sharma, V. Jain, J. M. Chatterjee, and N. Z. Jhanjhi, "A hybrid CNN-SVM threshold segmentation approach for tumor detection and classification of MRI brain images," IRBM, vol. 43, no. 4, pp. 290–299, 2022.

[9] D. J. Chaudhari and K. Malathi, "Detection and prediction of rice leaf disease using a hybrid CNN-SVM model," Optical Memory and Neural Networks, vol. 32, pp. 39–57, 2023.

[10] H. A. Madni, M. Umer, A. Ishaq, N. Abuzinadah, O. Saidani, S. Alsubai, M. Hamdi, and I. Ashraf, "Water-quality prediction based on H2O AutoML and explainable AI techniques," Water, 2023.

[11] H. Yang, C. Jia, X. Li, F. Yang, C. Wang, and X. Yang, "Evaluation of seawater intrusion and water quality prediction in Dagu River of North China based on fuzzy analytic hierarchy process exponential smoothing method," Environmental Science and Pollution Research, vol. 29, pp. 66160–66176, 2022.

[12] A. Juna, M. Umer, S. Sadiq, H. Karamti, A. A. Eshmawi, A. E. Mohamed, and I. Ashraf, "Water quality prediction using KNN imputer and multilayer perceptron," Water, 2022.

[13] V. Anand, C. Sohitha, G. N. Saraswathi, and G. Lavanya, "Water quality prediction using CNN," Journal of Physics: Conference Series, vol. 2484, 2023.

[14] E. C. Eze, S. J. Halse, and T. Ajmal, "Developing a novel water quality prediction model for a South African aquaculture farm," Water, 2021.

[15] Z. Mi, Q. Li, Y. Sha, and Z. Wu, "CNN-LSTM-attention water quality prediction hybrid model," in Proc. Conf., 2023.

[16] Z. Guo, R. Gai, S. Qin, and P. Wang, "CNN-BiLSTM for water quality prediction method based on attention mechanism," in Proc. 2023 IEEE Smart World Congress (SWC), 2023, pp. 1–6.

[17] J. Wu and Z. Wang, "A hybrid model for water quality prediction based on an artificial neural network, wavelet transform, and long short-term memory," Water, 2022.

[18] C. Song and L. Yao, "Application of artificial intelligence based on synchrosqueezed wavelet transform and improved deep extreme learning machine in water quality prediction," Environmental Science and Pollution Research, vol. 29, pp. 38066–38082, 2022.

[19] J. Dong, Z. Wang, J. Wu, J. Huang, and C. Zhang, "A water quality prediction model based on signal decomposition and ensemble deep learning techniques," Water Science and Technology, vol. 88, no. 10, pp. 2611–2632, 2023.

[20] Q. Ni, X. Cao, C. Tan, W. Peng, and X. K. Kang, "An improved graph convolutional network with feature and temporal attention for multivariate water quality prediction," Environmental Science and Pollution Research, vol. 30, pp. 11516–11529, 2022.

[21] S. Yang, S. Zhong, and K. Chen, "W-WaveNet: A multi-site water quality prediction model incorporating adaptive graph convolution and CNN-LSTM," PLOS ONE, vol. 19, 2024.