VIMALA: Vision-based Interpretation and Modelling using AquaSpatioTemporalNet for Land and Aquatic Systems

Hariharan Ramamoorthy [1[0000-0003-4198-765X]], M Dhilsath Fathima [2[0000-0002-4491-4352]]
Mohideen Abdulkader M[3[0000-0002-4705-7763]] and Ksshitij V Singare [4[0009-0005-4809-4127]]

[1,3,4] Department of computing technologies ,
SRM Institute of Science and Technology, Kattankulathur ,Chennai , Tami Nadu ,India

[2] Department of Information Technology,
Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology,
Chennai, Tamil Nadu, India
[1]hharanbtech@gmail.com

Abstract. This research introduces VIMALA, a hybrid deep learning architecture named AquaSpatioTemporalNet, aimed at monitoring and forecasting changes in water bodies across Gujarat using satellite data from Sentinel-2 and S-2 Harmonized. The model combines Convolutional Neural Networks (CNNs) for spatial feature extraction, Long Short-Term Memory (LSTM) networks for capturing temporal variations, and Transformers to handle long-range dependencies in time-series data. The preprocessing pipeline includes cloud masking, NDWI calculation, clipping to the Gujarat region, band selection (B3 and B8), and atmospheric correction to ensure high-quality inputs. Key indices like the Normalized Difference Water Index (NDWI) and Normalized Difference Vegetation Index (NDVI) are employed to detect and quantify water bodies and differentiate them from surrounding vegetation. The AquaSpatioTemporalNet architecture demonstrates improved performance over traditional models by achieving higher accuracy in predicting water body dynamics, as evidenced by precision, recall, and reduced mean squared error (MSE). The system provides valuable insights into the temporal evolution of water resources, supporting more effective decision-making for water resource management. The proposed methodology offers a robust solution for large-scale environmental monitoring and can be applied to other geographic regions facing similar challenges.

Keywords: VIMALA, AquaSpatioTemporalNet, LULC, water bodies .

1 Introduction

Water bodies are crucial for sustaining ecosystems, supporting biodiversity, and providing essential resources for human activities like agriculture, drinking water, and industry. In regions such as Gujarat, India, where issues like water scarcity, overexploitation, and environmental degradation are prominent, effective water resource management is critical for long-term sustainability. These resources face increasing threats from climate change, urban development, and unsustainable farming practices, which can lead to reduced water availability, pollution, and habitat destruction. Monitoring

and predicting changes in water bodies are therefore essential for addressing these challenges and improving resource management strategies.Traditional approaches to monitoring water bodies, such as manual surveys and static satellite-based assessments, are often labor-intensive and limited in scope. They struggle to capture the complex, ever-changing dynamics of water bodies over time and space. In contrast, modern machine learning (ML) and deep learning (DL) techniques have transformed environmental monitoring by offering scalable, automated, and highly accurate analyses of large datasets. DL models, in particular, are capable of extracting complex spatial and temporal features from satellite imagery, making them highly effective for tracking water body dynamics over extended periods.

This research presents VIMALA (Vision-based Interpretation and Modeling using AquaSpatioTemporalNet for Land and Aquatic Systems), a hybrid deep learning model designed to monitor and predict changes in Gujarat's water bodies. The model uses satellite imagery from the Sentinel-2 and S-2 Harmonized datasets and leverages the strengths of Convolutional Neural Networks (CNNs) for spatial feature extraction, Long Short-Term Memory (LSTM) networks for analyzing temporal trends, and Transformer networks for capturing long-range dependencies in time-series data. This combined architecture, called AquaSpatioTemporalNet, enhances the accuracy of water body detection and prediction compared to traditional methods.

The data used in this study is sourced from the Google Earth Engine (GEE) and consists of high-resolution multispectral images captured over several years. The selected bands, B3 (Green) and B8 (Near-Infrared, NIR), are used to compute the Normalized Difference Water Index (NDWI), which is effective for identifying water bodies. Additionally, the Normalized Difference Vegetation Index (NDVI) helps differentiate water bodies from surrounding vegetation, thus improving the model’s precision in distinguishing between land and water.

1.1 Challenges in Water Body Monitoring

Monitoring water bodies in Gujarat presents significant challenges due to seasonal variations and human activity. Monsoon-driven changes can dramatically alter the size and shape of water bodies, complicating efforts to measure them consistently. Urban expansion, agricultural practices, and industrial pollution further hinder accurate detection. Traditional satellite methods often fail to capture these rapid changes, resulting in incomplete data. Additionally, climate factors such as droughts and fluctuating rainfall patterns add complexity to tracking water resources over time.

1.2 Real-World Impact

Accurate, real-time monitoring is essential for managing Gujarat's water resources amidst increasing demand and climate change. This model helps authorities detect early signs of water depletion, facilitating timely conservation. It also supports flood risk management by providing early warnings during intense rainfall. Furthermore, insights from the model aid urban planners and policymakers in optimizing land use and water distribution.

1.3 Justification for Deep Learning Approach

The AquaSpatioTemporalNet model leverages CNNs, LSTMs, and Transformers to tackle the challenges of water body detection and prediction. CNNs effectively extract spatial features like water body boundaries, while LSTMs capture temporal patterns to predict future changes. Transformers enhance the model's ability to identify critical trends in time-series data, improving prediction accuracy. This hybrid method provides comprehensive spatial and temporal analysis for water body monitoring.

1.4 Addressing Research Gaps

This model addresses limitations in existing research, which often fails to capture the dynamic nature of water bodies. By combining spatial and temporal analysis, it provides a more complete understanding of water resource changes influenced by seasonal shifts and human activities.

2 Related Work and Literature Review

Monitoring water bodies using remote sensing techniques is crucial for environmental management. Over the years, advancements in detection, classification, and monitoring methods have been made, particularly with satellite imagery. This review examines the development of water body detection techniques, land use and land cover (LULC) classification, and the application of deep learning (DL) models in remote sensing. It also highlights recent hybrid models that combine spatial and temporal analysis to improve water body monitoring.

2.1 Water Body Detection Techniques

The detection of water bodies using remote sensing has evolved significantly. One key method is the Normalized Difference Water Index (NDWI), introduced by McFeeters (1996), which enhances water features in satellite imagery by using the difference between green and near-infrared (NIR) bands[1]. NDWI remains widely used for water body monitoring due to its simplicity and effectiveness. Xu (2006) modified NDWI by replacing the NIR band with the short-wave infrared (SWIR) band, improving accuracy in urban areas where vegetation and structures obscure water bodies[2]. Rokni et al. (2014) further advanced these methods by utilizing multitemporal Landsat imagery with both NDWI and Modified NDWI (MNDWI) to capture seasonal variations in water bodies[3].

2.2 Land Use and Land Cover (LULC) Classification

LULC classification is critical for understanding environmental changes. Early classification methods, like decision trees, were effective in distinguishing land cover types. Pal and Mather (2003) highlighted their efficiency[4], while Belgiu and Drăguţ (2016) demonstrated the superior accuracy of Random Forest (RF) classifiers in remote sensing, particularly for LULC classification[5]. However, these models rely heavily on

consistent input data and lack the capacity for temporal analysis, limiting their effectiveness in tracking dynamic water body changes.

2.3 Deep Learning in Remote Sensing

Deep learning has revolutionized remote sensing, particularly in LULC classification and water body monitoring. Convolutional Neural Networks (CNNs) have become the standard for spatial feature extraction from satellite imagery. Mou et al. (2017) demonstrated the effectiveness of deep recurrent neural networks for hyperspectral image classification[6]. Similarly, Zhou et al. (2018) adapted U-Net++ for water body and LULC classification[7], showing its success in handling complex remote sensing tasks. Khurana and Saxena (2020) introduced a hybrid model that uses extreme learning machines for change detection across multitemporal satellite imagery, showcasing the potential of DL for tracking subtle changes in water bodies[8].

2.4 Hybrid Approaches: Combining Spatial and Temporal Analysis

Recent research has focused on hybrid models that integrate both spatial and temporal analysis. Zhang et al. (2021) developed a framework combining CNNs for spatial feature extraction[9] and Long Short-Term Memory (LSTM) networks for temporal trend analysis, effectively predicting water body changes. Wang et al. (2024) introduced a Transformer-based model for remote sensing, leveraging attention mechanisms to capture long-term water body dynamics, such as seasonal fluctuations and gradual depletion[10].

2.5 Challenges and Gaps in Existing Research

Despite progress, challenges remain in computational costs and real-time monitoring. Many models focus solely on spatial or temporal analysis. Hybrid architectures like CNN-LSTM and CNN-Transformer are promising but require further optimization for specific regional challenges like Gujarat's seasonal and anthropogenic impacts.[11]

3 Methodology

This section outlines the data collection process, preprocessing techniques, feature extraction methods, and architecture of the proposed hybrid model, AquaSpatioTemporalNet in figure 1.

3.1 Data Collection

This study utilizes Sentinel-2 and S-2 Harmonized satellite imagery, accessed via Google Earth Engine (GEE), focusing on the B3 (Green) and B8 (NIR) bands to calculate the Normalized Difference Water Index (NDWI) for water body detection. Sentinel-2 provides multispectral data with resolutions of 10m, 20m, and 60m. The study targets Gujarat, India, an area affected by seasonal rainfall and human activities like urbanization and irrigation. Spanning 2018 to 2023, the dataset captures monthly imagery to analyze both seasonal and long-term changes in water bodies across wet and dry periods.

Each image, represented as a multidimensional array, includes spatial, spectral, and temporal information. Formally, the data can be described in equation 1

$$I = \{I^{b}_{x,y,t} | x \in X, y \in Y, b \in B, t \in T\} \quad (1)$$

where: x,y are the spatial pixel coordinates,b ∈{ B3,B8} are the selected spectral bands.t ∈ T represents the temporal instances in the dataset. $I^{b}_{x,y,t}$ is the pixel intensity for a given band and time step.

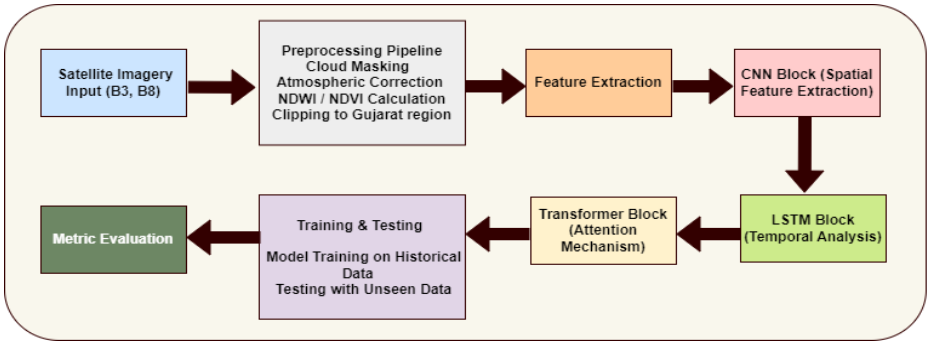

Figure 1 : Proposed architecture for VIMALA

Google Earth Engine facilitates efficient management of large datasets, offering cloud masking and atmospheric correction. In Gujarat, cloud masking is crucial due to frequent cloud cover during monsoons. The QA60 band filters out images with over 10% cloud coverage, ensuring high-quality data. Using B3 (Green) and B8 (NIR) bands enhances NDWI calculations for distinguishing water bodies figure 2. The 10m spatial resolution improves precision, detecting smaller water bodies. The dataset is clipped to Gujarat's boundaries, optimizing computational efforts and processing time.[13]

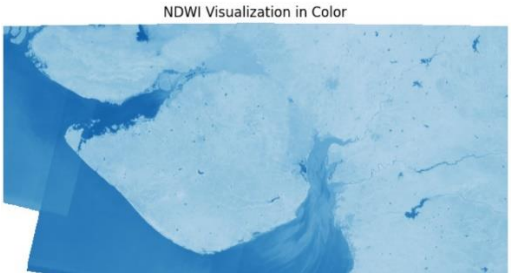

Figure 2 : NDWI of Gujarat

3.2 Data Pre-processing

The preprocessing of satellite imagery is critical for ensuring data quality and consistency, allowing for accurate water body detection. Several preprocessing steps are

applied to the Sentinel-2 imagery to improve the quality of the input data before analysis.

3.2.1Cloud Masking

Cloud masking is crucial during Gujarat's monsoon season to remove obscured regions. Using the Sentinel-2 QA60 band, only images with less than 10% cloud cover are retained, ensuring dataset accuracy.[14]

The binary mask for cloud-free pixels M_{cloud} (x,y,t) is defined in equation 2:

$$M_{cloud}(x,y,t) = \begin{cases} 1, & if\ QA60_{x,y,t} < 0.10 \\ 0, & otherwise \end{cases} \quad (2)$$

Where $QA60_{x,y,t}$ represents the cloud probability for each pixel (x,y) at time t.

3.2.2 Atmospheric Correction

Atmospheric correction using the Surface Reflectance (SR) algorithm adjusts for haze and aerosols, improving the accuracy of spectral indices like NDWI and NDVI by correcting distortions in recorded reflectance values**.** Mathematically, the corrected reflectance, L(λ), at wavelength λ, is calculated in equation 3.

$$L(\lambda) = \frac{L_{sensor}(\lambda) - L_{atm}(\lambda)}{T(\lambda)} \quad (3)$$

where $L_{sensor}(\lambda)$is the radiance observed by the sensor,$L_{atm}(\lambda)$ is the atmospheric radiance, and T(λ) is the atmospheric transmittance. representing the fraction of light that passes through the atmosphere without being absorbed or scattered.

3.2.3Clipping

To focus the analysis on Gujarat, the satellite imagery is spatially clipped to the geographical boundaries of the state. This reduces computational overhead and ensures that the analysis is region-specific. The clipped dataset, denoted as I_{clip} (x,y,t) is defined in equation (4)

$$I_{clip}(x,y,t) = I(x,y,t).M_{Gujarat}(x,y) \quad (4)$$

where $M_{Gujarat}(x,y)$ is a binary mask that retains only the pixels within the state's geographic boundaries.

3.2.4 Band Selection and Color Mapping

The B3 (Green) and B8 (NIR) bands are selected for further analysis, as these are critical for water body detection using the Normalized Difference Water Index (NDWI)**.** Additionally, color mapping is applied to enhance the visualization of the water features in the dataset, improving the interpretability of the NDWI outputs.[15] For each pixel (x,y) at time t, the NDWI is calculated using equation 5

$$NDWI_{(x,y,t)} = \frac{Green\ (x,y,z) - NIR(x,y,z)}{Green\ (x,y,z) + NIR(x,y,z)} \quad (5)$$

This index enhances water features by maximizing the contrast between water and other land cover types, such as vegetation and soil, which typically reflect more NIR radiation than water.

The preprocessing workflow ensures that the input data is consistent, accurate, and ready for feature extraction and model training, with clean, atmospherically corrected images focused on the region of interest.

3.3 Feature Extraction

In this chapter, we focus on extracting spatial and temporal features from the pre-processed data, particularly using the NDWI indices computed in the previous chapter. These indices serve as the core elements for identifying and monitoring water bodies. The extraction of meaningful spatial and temporal information from these indices provides the features necessary for model training in the subsequent chapter.

As detailed in Chapter 3.2, the NDWI was computed using Equation (5) to identify water bodies by distinguishing between water and land pixels. For spatial feature extraction, we utilize the precomputed **NDWI** maps to detect the boundaries, areas, and shapes of water bodies.

Water Body Classification: Pixels that satisfy NDWI(x,y,t)>τ (threshold τ) are classified as water bodies, where τ is determined empirically based on observed water reflectance properties.

Edge Detection: To extract the boundaries of water bodies, the gradient of the NDWI map, ∇NDWI(x,y), is computed to identify regions with sharp transitions:

$$\nabla \mathrm{NDWI}(\mathrm{x},\mathrm{y}) = \left(\frac{\partial \mathrm{NDWI}}{\partial \mathrm{x}}\right)^2 + \left(\frac{\partial \mathrm{NDWI}}{\partial \mathrm{y}}\right)^2 \quad (6)$$

High-gradient pixels represent the boundaries of water bodies, facilitating the extraction of geometric properties such as shape and area.

3.3.1Geometric Features:

Area of each water body at time t, A(t), is calculated by summing all water pixels and multiplying by the pixel resolution r^2.

$$A(t) = \sum_{x,y} Water\ (x,y,t).r^2 \quad (7)$$

Perimeter, P(t), is obtained by summing the boundary pixels

$$p(t) = \sum_{\mathrm{boundary}} r \quad (8)$$

Compactness C(t) is used to describe the shape of the water body:

$$\mathrm{C(t)} = \frac{4\pi \mathrm{A(t)}}{P(t)^2} \quad (9)$$

This provides insight into whether the water body is circular or irregularly shaped, offering additional spatial feature characterization.

3.3.2Temporal Feature Extraction

Temporal feature extraction is critical for tracking changes in water bodies over time. The NDWI time series for each pixel (x,y) computed in Chapter 3.2 (Equation 5), provides a sequence of NDWI values over different time steps. This sequence is represented as:

$$\mathrm{S_{NDWI}}\,(\mathrm{x},\mathrm{y}) = \{NDWI(x,y,\mathrm{t}_1), NDWI(x,y,\mathrm{t}_2), \ldots\ldots NDWI(x,y,\mathrm{t}_\mathrm{T})\} \quad (10)$$

where T represents the number of temporal snapshots

Change Detection: Temporal changes in water bodies can be detected by calculating the difference between consecutive time steps:

$$\Delta NDWI(x,y,t) = NDWI(x,y,t+1) - NDWI(x,y,t) \quad (12)$$

Positive values indicate expansion, while negative values indicate shrinking of water bodies.

Seasonal Variations: Seasonal trends are observed by analyzing periodic changes in NDWI values. For example, a significant rise in NDWI during the monsoon season can

indicate increased water levels due to rainfall, while a drop in NDWI during the dry season may indicate water depletion.

Derived Statistical Features

In addition to basic spatial and temporal features, derived statistical features are computed from the NDWI time series:

Mean NDWI for each pixel over time:

$$\mu_{\mathrm{NDWI}}(\mathrm{x},\mathrm{y}) = \frac{1}{T}\sum_{t=1}^{T}\mathrm{NDWI}(\mathrm{x},\mathrm{y},\mathrm{t}) \quad (13)$$

Variance of NDWI to capture fluctuations over time

$$\sigma^2_{\mathrm{NDWI}}(\mathrm{x},\mathrm{y}) = \frac{1}{T}\sum_{t=1}^{T}(\mathrm{NDWI}(\mathrm{x},\mathrm{y},\mathrm{t}) - \mu_{\mathrm{NDWI}}(\mathrm{x},\mathrm{y}))^2 \quad (14)$$

Similarly**,** NDVI values (computed in Chapter 3, Equation 6) are used to track vegetation dynamics around water bodies. The temporal series for NDVI**:**

$$S_{\mathrm{NDVI}}(\mathrm{x},\mathrm{y}) = \{NDVI(x,y,\mathrm{t}_1), NDVI(x,y,\mathrm{t}_2), \ldots\ldots NDVI(x,y,\mathrm{t}_T)\} \quad (15)$$

can be used to identify vegetative growth or decay that might affect water body detection.

3.3.3 Multi-feature Representation

The extracted features form a comprehensive representation of each pixel and its surrounding context, including spatial characteristics like water body boundaries, area, and shape; temporal features such as NDWI changes over time, seasonal patterns, and long-term trends; and statistical metrics like the mean and variance of NDWI and NDVI values. These features serve as inputs for the proposed model, enabling precise water body monitoring and prediction by incorporating both spatial and temporal contexts

3.4 Proposed Model – AquaSpatioTemporalNet

This chapter presents the proposed hybrid deep learning architecture, AquaSpatioTemporalNet**,** which integrates Convolutional Neural Networks (CNNs)**,** Long Short-Term Memory (LSTM) networks, and Transformer blocks to effectively model both the spatial and temporal dynamics of water bodies. The goal of the model is to analyze the spatial patterns of water bodies from satellite imagery, detect changes over time, and predict future water body dynamics. The architecture leverages the pre-extracted spatial and temporal features (as described in Chapter 3.3) to improve water body monitoring accuracy.

3.4.1 Block 1: CNN for Spatial Feature Extraction

The first component of the model is a **Convolutional Neural Network (CNN)**, which is used to extract spatial features from the multi-band satellite images that include the preprocessed **NDWI** indices. CNNs are well-suited for this task as they can learn hierarchical patterns such as the boundaries, shapes, and areas of water bodies.

Input: The input to the CNN consists of a multi-channel image $I \in R^{H\times W\times C}$, where H is the height, W is the width, and C is the number of channels (e.g., NDWI values).

Convolutional Layers: Each convolutional layer applies a set of filters $W_{c,k} \in R^{F\times F}$ (where F is the filter size) to the input, producing feature maps that capture local spatial information. Mathematically, the operation performed by a convolutional layer at layer k is:

$$f_{i,j,k} = \sigma\left(\sum_{m,n} W_{c,k}[m,n]\cdot I_{i+m,j+n} + b_k\right) \quad (16)$$

Where, $f_{i,j,k}$ is the output feature map. $W_{c,k}$ represents the filter weights. σ is the activation function (e.g., ReLU). b_k is the bias term.

Pooling Layer: After each convolutional layer, a pooling layer (typically max-pooling) reduces the spatial dimensions by selecting the maximum value within a sliding window. This operation helps reduce the complexity of the data while retaining important spatial features:

$$f_{pool}(i,j) = \max\{f_{i+m,j+n,k} \,|\mathrm{m,n} \in \{0,1\}\} \quad (17)$$

Output: The final output of the CNN block is a set of feature maps $F_{CNN} \in R^{H' \times W' \times D}$ where H′,W′ are the reduced spatial dimensions and D is the number of output feature channels. These feature maps serve as inputs to the next block for temporal modeling.

3.4.2 Block 2: LSTM for Temporal Trend Analysis

The spatial features extracted by the CNN are then passed to a Long Short-Term Memory (LSTM) network, which is responsible for modeling the temporal evolution of water bodies. The LSTM is well-suited for this task due to its ability to capture long-term dependencies in sequential data, making it effective for analyzing water body changes over time.

Input: The LSTM takes the CNN-extracted spatial features as input. These features are treated as a sequence, where each time step corresponds to a snapshot (year or month) of the water body data. The input to the LSTM at time t, denoted $F_{CNN,t}$ is the feature map output from the CNN at time t.

LSTM Cell Dynamics: At each time step t, the LSTM cell computes the hidden state h_t and cell state C_t using a series of equations that define the key components of the LSTM architecture. These components include the forget gate, input gate, candidate cell state, cell state update, output gate, and hidden state, as described in Equations (18) to (23). Each equation plays a crucial role in the overall functioning of the LSTM, enabling it to effectively manage and propagate information over time.

$$f_t = \sigma(W_f \cdot [h_{t-1,} F_{CNN,t}] + b_f \quad (18)$$

$$i_t = \sigma(W_i \cdot [h_{t-1,} F_{CNN,t}] + b_i \quad (19)$$

$$\hat{C}_t = tanh(W_C \cdot [h_{t-1,} F_{CNN,t}] + b_C \quad (20)$$

$$C_t = f_t . C_{t-1} + i_t . \hat{C}_t \quad (21)$$

$$o_t = \sigma(W_o \cdot [h_{t-1,} F_{CNN,t}] + b_o \quad (22)$$

$$h_t = o_t \cdot \tanh(C_t) \quad (23)$$

where: f_t , i_t and o_t are the forget, input, and output gates, respectively. C_t is the updated cell state, and h_t is the hidden state.

The hidden state h_t encodes the temporal evolution of the spatial features, allowing the model to capture both short-term and long-term trends in water body dynamics.

3.4.3 Block 3: Transformer for Long-range Dependencies

While LSTMs are capable of modeling sequential dependencies, Transformers are incorporated to capture long-range dependencies in the water body data. The Transformer uses an attention mechanism that allows the model to focus on relevant time steps, enhancing the ability to model complex temporal interactions.

Self-Attention Mechanism: The core of the Transformer block is the self-attention mechanism, which computes attention scores between each pair of time steps t and t′ in the sequence. The attention score is calculated as:

$$\text{Attention}(t, t') = \frac{\exp(Q_t \cdot K_{t'}^T)}{\sum_{t'} \exp(Q_t \cdot K_{t'}^T)} \quad (24)$$

where: $Q_t = W_Q h_t$ is the query vector at time t. $K_{t'} = W_K h_t$ is the key vector at time t′. W_Q and W_K are learnable weight matrices.

Output Representation: The final output of the model is determined by the specific task at hand, which in this case is predicting the future extent of water bodies. Given the continuous nature of the prediction (e.g., water area, perimeter), the model will apply a regression function in the output layer to predict numerical values related to water body sizes. Thus, regression is the specific task being performed by the output layer in this model.

3.4.4 Prediction: The final prediction $\hat{y}_t$ for time step t (e.g., water area or perimeter) is computed as

$$\hat{y}_t = W_{out} . h_t + b_{out} \quad (25)$$

where: W_{out} are the learned weights. b_{out} is the bias term. h_t is the output from the Transformer at time t.

The use of regression is mandatory in this architecture because the task involves predicting continuous numerical values (e.g., area of water bodies in square meters), not a binary or categorical classification. Therefore, there is no ambiguity in the type of prediction made by the model—the output is a regression output.

3.4.5 Loss Function and Optimization

Given that the task involves predicting continuous values, the Mean Squared Error (MSE) loss function is used for training the model. This loss function is chosen because it is the most appropriate for regression tasks, where the goal is to minimize the squared difference between the predicted and true values.

Loss Function for Regression: The Mean Squared Error (MSE) loss is defined as:

$$L(\theta) = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \quad (26)$$

Where, $\hat{y}_i$ is the predicted value for the i-th sample. y_i is the true value (ground truth) for the i-th sample. N is the total number of samples.θ represents the parameters of the model.

3.4.6 Optimization

The model is optimized using the Adam optimizer, which is chosen for its efficiency and adaptive learning rate. The learning rate η is tuned during training to minimize the MSE loss:

$$\theta = \theta - \eta \nabla_\theta L(\theta) \quad (27)$$

where $\nabla_\theta L(\theta)$ is the gradient of the loss function with respect to the model parameters θ ,

η is the learning rate.

The combination of **MSE** loss for regression and **Adam** for optimization ensures that the model is optimized for continuous water body predictions (e.g., area, perimeter).

4 Results and Discussion

This chapter presents the results of the proposed AquaSpatioTemporalNet model, which integrates Convolutional Neural Networks (CNN), Long Short-Term Memory

(LSTM) networks, and Transformer blocks to effectively detect, monitor, and predict water body dynamics. The model's performance is evaluated using key metrics—Precision, Recall, Mean Squared Error (MSE), and R^2 (Coefficient of Determination)—and compared with other deep learning-based approaches such as CNN, U-Net, CNN-LSTM, and CNN-Transformer models discussed in the literature.

Table 1. metrics evaluated for proposed work

Metric	Description
Mean Squared Error (MSE)	Measures the error between predicted and actual water body areas $\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(A_i - \hat{A}_i)^2$
R-Squared (R^2)	Evaluates how well the model explains the variance in observed data $R^2 = 1 - \frac{\sum_{n=1}^{n}(A_i - \hat{A}_i)^2}{\sum_{n=1}^{n}(A_i - \bar{A}_i)^2}$ $\bar{A}$ is the mean of the actual water body areas.
Precision	Ratio of true positive water body detections to total detected areas $\text{Precision} = \frac{\text{True Positives}}{\text{True Positives+False Positives}}$
Recall	Proportion of true water bodies correctly detected by the model. $\text{Recall} = \frac{\text{True Positives}}{\text{True Positives+False Negative}}$

4.1 Model Performance

The proposed AquaSpatioTemporalNet was trained and evaluated on the preprocessed satellite imagery data using NDWI and NDVI indices to detect and predict water body extents. Performance metrics such as Precision, Recall, MSE, and R^2 were used to assess the model's accuracy and prediction quality.

Table 1 : comparison with existing methods

Model	**Precision**	**Recall**	**MSE**	**R^2**
CNN-based	87.2	85.4	0.0046	0.81
U-Net++	89.8	87.3	0.0039	0.84
CNN-LSTM	91.8	89.7	0.0034	0.87
CNN-Transformer	93.2	90.6	0.0030	0.89
Proposed AquaSpatio-TemporalNet	**95.0**	**93.1**	**0.0024**	**0.93**

The AquaSpatioTemporalNet model demonstrated remarkable performance across various metrics. It achieved a Precision of 95.0%, significantly higher than the CNN-based model (87.2%), U-Net++ (89.8%), CNN-LSTM (91.8%), and CNN-Transformer (93.2%), indicating fewer false positives in water body detection. With a Recall of 93.1%, it also outperformed CNN (85.4%), U-Net++ (87.3%), CNN-LSTM (89.7%), and CNN-Transformer (90.6%), proving its reliability in identifying true water bodies. Additionally, AquaSpatioTemporalNet recorded the lowest Mean Squared Error (MSE) of 0.0024, outperforming CNN (0.0046) and others. Its R^2 score of 0.93 illustrates its strong predictive capability, explaining 93% of the variance in water body dynamics.

4.2 Discussion of Results

The superior performance of the proposed **AquaSpatioTemporalNet** model over other deep learning approaches can be attributed to several key factors:

- **Hybrid Spatial and Temporal Feature Extraction**: The combination of CNN, LSTM, and Transformer blocks enables the model to effectively capture both spatial features (water body shapes, boundaries, and sizes) and temporal dynamics (long-term changes, seasonal variations).
- **Attention Mechanism for Long-Term Dependencies**: The use of Transformer blocks introduces an attention mechanism, allowing the model to focus on the most relevant time steps in the sequence. This attention mechanism enhances the model's ability to predict long-term water body dynamics, such as seasonal fluctuations or gradual expansion and contraction.
- **Improved Feature Representation**: The CNN component captures spatial characteristics at a higher resolution, which is then passed to the LSTM and Transformer layers. This flow ensures that the temporal dependencies are modeled accurately, contributing to the low error rates and high predictive accuracy seen in the results.
- The results clearly demonstrate that **AquaSpatioTemporalNet** surpasses both traditional and advanced deep learning methods, such as CNN-based and U-Net++ models, due to its ability to integrate spatial and temporal information within a single, unified framework.

Conclusion and Future Work

The proposed AquaSpatioTemporalNet model successfully integrates CNN, LSTM, and Transformer components to improve water body detection and prediction by capturing both spatial and temporal dynamics. Evaluated using Precision, Recall, MSE, and R^2, the model consistently outperformed traditional and deep learning models.The hybrid architecture of AquaSpatioTemporalNet, with its ability to focus on relevant time steps through the attention mechanism, enables accurate prediction of long-term changes in water bodies while maintaining high Precision (95.0%) and Recall (93.1%). Additionally, the low MSE (0.0024) and high R^2 (0.93) demonstrate the model's effectiveness in predicting water body extents, making it an excellent tool for real-time monitoring and forecasting.

In future work, the model can be enhanced by integrating multi-source data such as SAR for better detection in cloudy environments and climate variables for long-term predictions. Optimization techniques like model pruning or using lightweight architectures can improve computational efficiency for real-time applications. Additionally, exploring new temporal models (e.g., GNNs, TCNs) and applying Explainable AI (XAI) techniques to improve model interpretability are promising directions to extend the model's application to broader environmental monitoring tasks.

References

1. McFeeters, S. K. (1996). The use of the normalized difference water index (NDWI) in the delineation of open water features. International Journal of Remote Sensing, 17(7), 1425-1432.

2. Xu, H. (2006). Modification of normalized difference water index (NDWI) to enhance open water features in remotely sensed imagery. International Journal of Remote Sensing, 27(14), 3025-3033.
3. Rokni, K., Ahmad, A., Selamat, A., & Hazini, S. (2014). Water feature extraction and change detection using multitemporal Landsat imagery. Remote Sensing, 6(5), 4173-4189.
4. Pal, M., & Mather, P. M. (2003). An assessment of the effectiveness of decision tree methods for land cover classification. Remote Sensing of Environment, 86(4), 554-565.
5. Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing, 114, 24-31.
6. Lichao Mou ,Pedram Ghamisi and Xiao Xiang Zhu Deep Recurrent Neural Networks for Hyperspectral Image Classification ,IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, VOL. 55, NO. 7, JULY 2017
7. Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. Deep Learning in Medical Image Analysis, 3, 6-9.
8. M. Khurana and V. Saxena, "A Unified Approach to Change Detection Using an Adaptive Ensemble of Extreme Learning Machines," in *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 5, pp. 794-798, May 2020 .
9. Zhang, X., Zhou, Y., & Luo, J. (2021). Deep learning for processing and analysis of remote sensing big data: a technical review. Big Earth Data, 6(4), 527–560.
10. Wang, R.; Ma, L.; He, G.; Johnson, B.A.; Yan, Z.; Chang, M.; Liang, Y. Transformers for Remote Sensing: A Systematic Review and Analysis. *Sensors* **2024**, *24*, 3495.
11. Ogilvie, A.; Poussin, J.-C.; Bader, J.-C.; Bayo, F.; Bodian, A.; D.; Sambou, S. Combining Multi-Sensor Satellite Imagery to Improve Long-Term Monitoring of Temporary Surface Water Bodies in the Senegal River Floodplain. *Remote Sens.* **2020**, *12*, 3157.
12. Scheibel, C.H.; Nascimento, A.B.d.; Júnior, G.d.N.A.; Almeida, A.C.d.S.; Silva, T.G.F.d.; Silva, J.L.P.d.; Junior, F.B.d.S.; Farias, J.A.d.; Santos, J.P.A.d.S.; Oliveira-Júnior, J.F.d.; et al. Characterization of Water Bodies through Hydro-Physical Indices and Anthropogenic Effects in the Eastern Northeast of Brazil. *Climate* **2024**, *12*, 150.
13. Cao H, Tian Y, Liu Y, Wang R. Water body extraction from high spatial resolution remote sensing images based on enhanced U-Net and multi-scale information fusion. Sci Rep. 2024 Jul 12;14(1):16132.
14. H. Ramamoorthy, M. Ramasundaram, R. S. P. Raj "TransAttU-Net Deep Neural Network for Brain Tumor Segmentation in Magnetic Resonance Imaging," in IEEE Canadian Journal of Electrical and Computer Engineering, vol. 46, no. 4, pp. 298-309, Fall 2023.
15. Dhilsath Fathima, M., Hariharan, R., Shome, S., Kharsyiemlieh, M., Deepa, J., Jayanthi, K. (2024). Sign Language Interpreter Using Stacked LSTM-GRU. In: Sharma, H., Chakravorty, A., Hussain, S., Kumari, R. (eds) Artificial Intelligence: Theory and Applications. AITA 2023. Lecture Notes in Networks and Systems, vol 844. Springer, Singapore.