**Assessment of Water Quality Prediction in Coastal Areas Using Remote Sensing and Artificial Intelligence**

1. **INTRODUCTION**

**1.1 Motivation**

Coastal regions like Rameswaram play a vital role in marine ecosystems and need regular monitoring to track environmental changes and human impact. Key water quality indicators like Chlorophyll-a, SST, and turbidity can reveal pollution and algal blooms, but traditional methods are often costly and time-consuming. Remote sensing offers a faster, large-scale alternative, and by combining it with AI-driven models, this project seeks to improve water quality predictions, aiding early warning systems and sustainable coastal management.

**1.1.1** **Importance in Real-world Applications**

* Enables **large-scale, real-time monitoring** of coastal water quality, reducing dependence on costly and time-consuming fieldwork.
* Supports **data-driven decision-making** for environmental monitoring, climate change studies, and coastal resource management.
* Facilitates **early detection of anomalies** in marine ecosystems, aiding in the mitigation of climate-related impacts.
* Enhances **predictive accuracy** for key oceanographic parameters like Chlorophyll-a and SST, improving conservation and policy strategies.
* Promotes **efficient marine resource management** through scalable integration of remote sensing and machine learning technologies.

**1.2 Applications**

* **Coastal Water Quality Monitoring:** Continuous tracking of parameters like Chlorophyll-a and Sea Surface Temperature using satellite data.
* **Environmental Management:** Supports sustainable planning and conservation of coastal and marine ecosystems.
* **Climate Change Research:** Helps study long-term oceanographic trends and their relation to climate variability.
* **Policy-making Support:** Provides accurate data for governments and environmental agencies to implement effective marine policies.
* **Disaster & Anomaly Detection:** Early identification of environmental changes such as algal blooms, pollution, or thermal anomalies.
* **Resource Management:** Enhances decision-making in fisheries, tourism, and marine biodiversity protection through data-driven insights.

**2.0 LITERATURE SURVEY**

**2.1 Gaps**

* **Inconsistent Input Parameters:** Lack of standardized water quality parameters across studies affects model comparability and generalization.
* **Limited Integration of Climate and Anthropogenic Factors:** Many models exclude climate change, land use, and population growth variables due to data complexity.
* **Data Scarcity for Long-Term Predictions:** Many models perform well short-term but lack robustness when applied to sparse or long-term datasets.
* **Limited Use of Advanced AI Techniques:** There's a need to explore newer architectures (e.g., hybrid deep learning, transfer learning) for improved accuracy and adaptability.

**2.2 Problem Statement**

Traditional water quality assessment methods are costly, time-consuming, and require constant human involvement, making large-scale and real-time monitoring difficult. This limits effective management of coastal water resources. Using satellite-based remote sensing combined with AI offers a more scalable, efficient, and accurate way to monitor key water quality parameters like Chlorophyll-a and Sea Surface Temperature (SST), improving how we manage and protect coastal ecosystems.

**2.3 Objectives**

* To Enhance the Predictions, Incorporation of Time Series Deep Learning Algorithms.
* To Include few more Water Quality Parameters to assess the water quality.
* To develop an AI-driven process for assessing coastal water quality using satellite-based remote sensing data.
* To implement AI-Based algorithms that enhance the accuracy of water quality estimation from MODIS-Aqua satellite imagery.
* To analyze key metrics such as Chlorophyll-a and SST to assess water health.
* To compare AI-based predictions with traditional water quality assessment techniques to evaluate effectiveness and reliability.

1. **METHODOLGY**

**3.1 Dataset Characteristics**

**1. Name**: Ocean Color SMI: Standard Mapped Image MODIS Terra

**Available**: Google Earth Engine (GEE).

**Number of rows**: 1876

**Number of columns**: 4

**Columns**: 1. Longitude, 2. Latitude, 3. Chlorophyll-a, 4. SST.

**2. Name**: Global Ocean Colour (Copernicus-GlobColour)

**Available**: Copernicus.

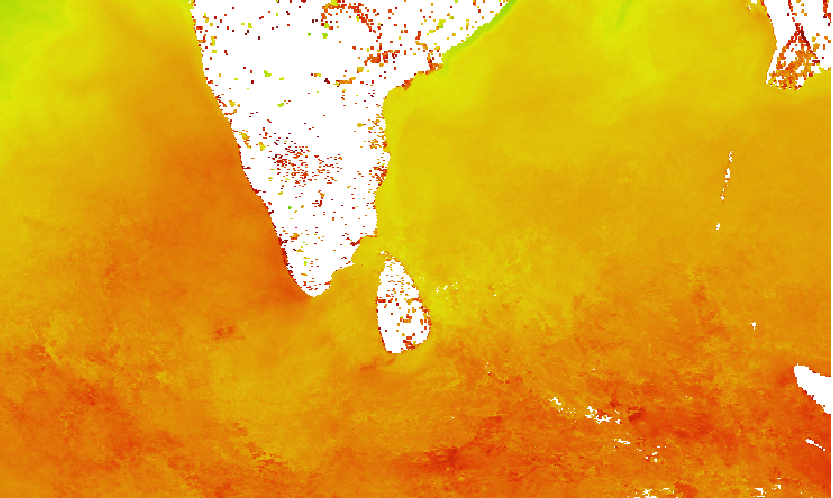
**Number of rows**: 2209

**Number of columns**: 4

**Columns**: 1. Time, 2. Depth, 3. Chlorophyll-a, 4. SST.

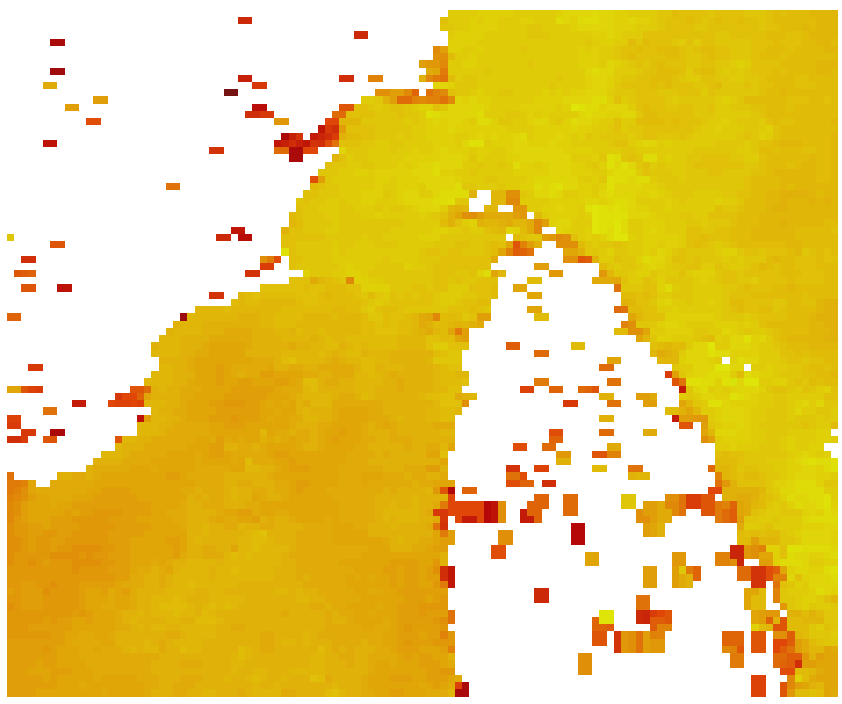
**3.1.1 Dataset Acquisition and Transformation**

The information is obtained from the MODIS-Terra satellite through Google Earth Engine (GEE), which provides access to spectral band imagery of coastal regions. The imagery contains oceanographic parameters of interest for Chl-a and SST prediction. The spectral information is processed and exported from GEE in TIFF format, which is then processed using SeaDAS to obtain numerical values and transform them into a formatted CSV format for subsequent processing. The image (Fig-1) represents the **Sea Surface Temperature (SST)** distribution across the **Arabian Sea and Bay of Bengal**, captured using MODIS-Terra satellite data. The colour variations indicate temperature differences, with warmer regions appearing in shades of yellow and red, while cooler areas are depicted in lighter shades. This data is essential for understanding oceanic thermal patterns and climate variability in the region.



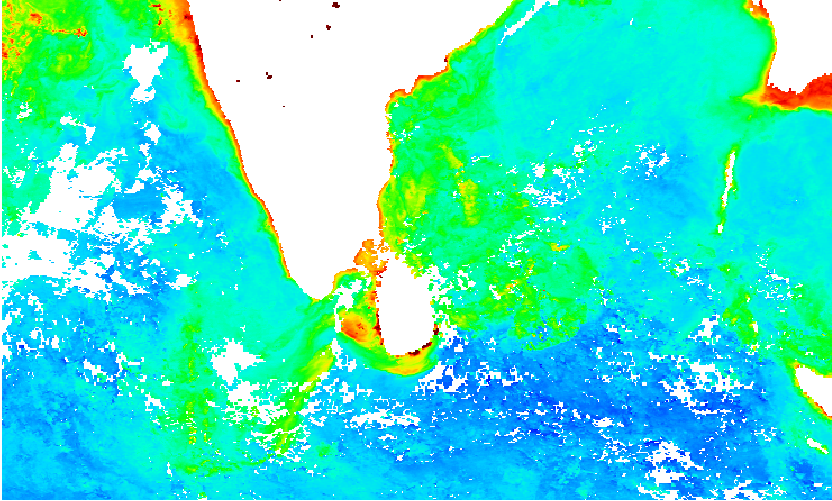
**Fig. 1.** Arabian Sea & Bay of Bengal subset of SST

The image (Fig-2) highlights the **SST distribution specifically in the Rameswaram coastal area**, a key study region. The dataset has been processed to extract localized SST variations, which help in analyzing thermal dynamics in nearshore waters. The image also shows missing data points, which need preprocessing for accurate model predictions



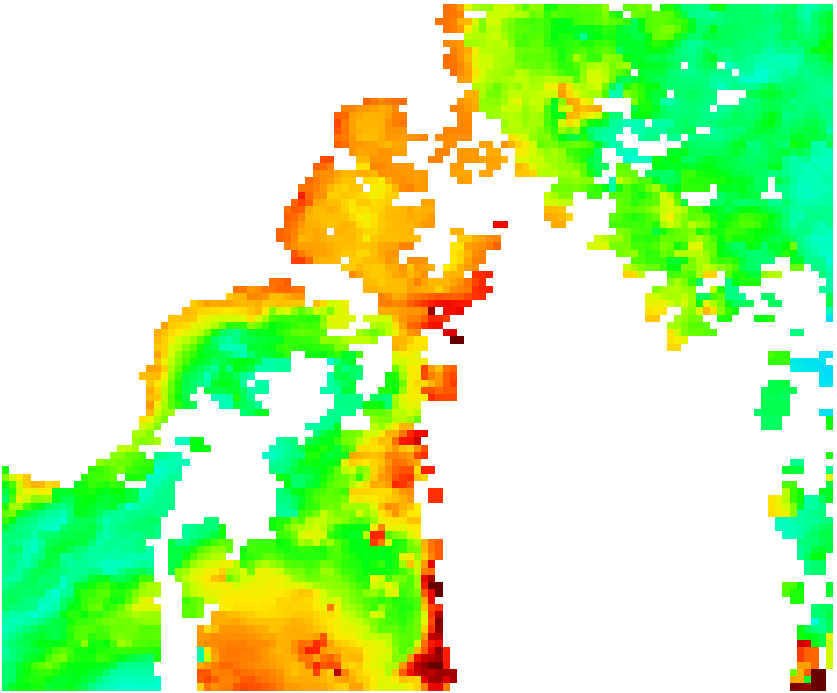
**Fig. 2.** Study area- Rameswaram subset of SST

The image (Fig-3) represents **Chlorophyll-a (Chl-a) concentration** across the **Arabian Sea and Bay of Bengal**, derived from satellite observations. High concentrations of Chl-a, indicated by green and blue hues, correspond to regions with significant phytoplankton presence, while lower concentrations appear in lighter colours. This dataset is crucial for studying **marine ecosystem productivity, algal blooms, and biogeochemical cycles**.

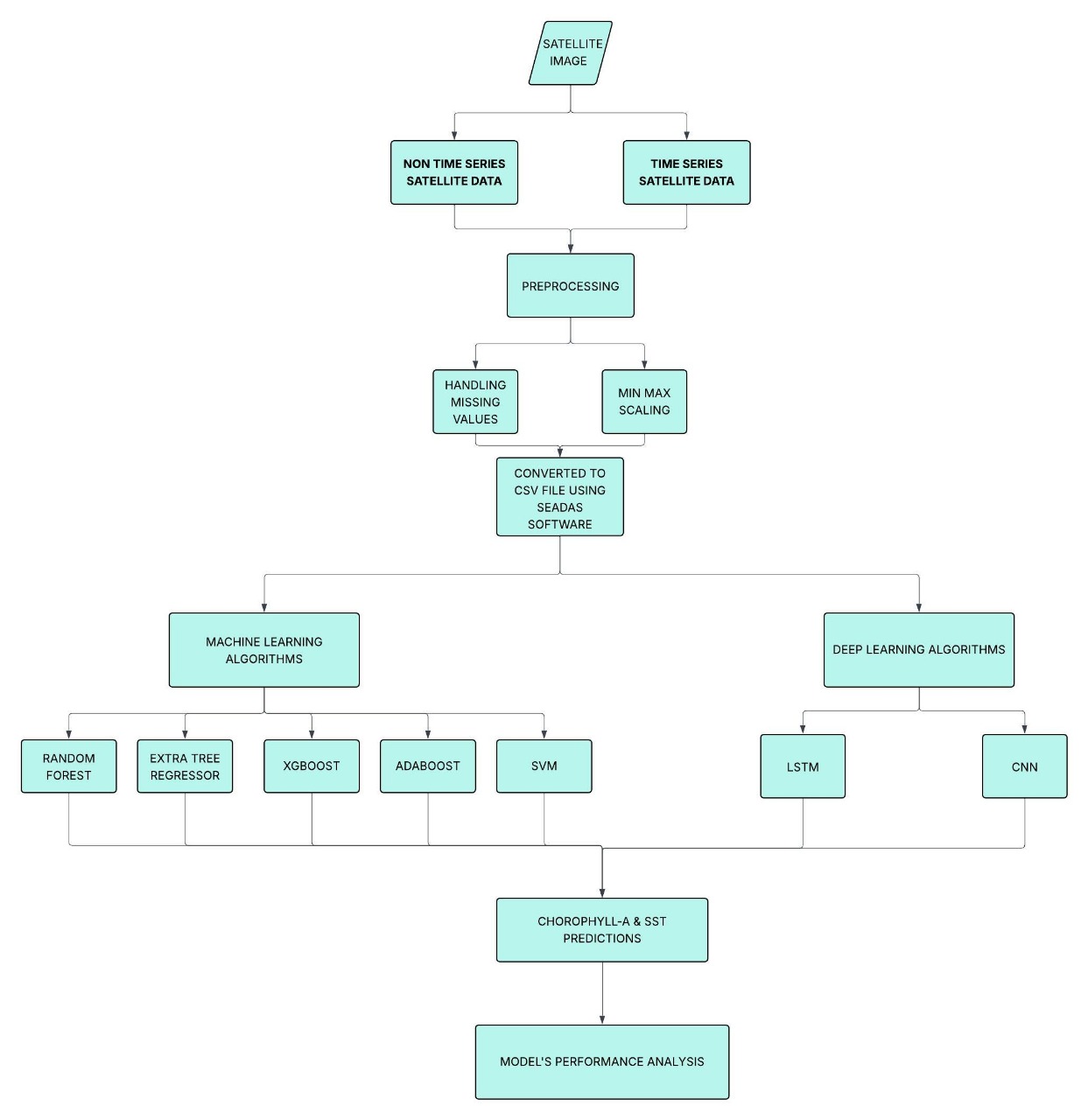


**Fig. 3.** Arabian Sea & Bay of Bengal subset of Chlorophyll-a

This image (Fig-4) focuses on the **Chl-a distribution in the Rameswaram coastal region**, extracted from the broader dataset. It provides insights into **local phytoplankton dynamics and water quality**, which are essential for coastal resource management. Preprocessing steps, such as noise reduction and outlier removal, are necessary to refine the dataset for accurate machine learning predictions.



**Fig. 4.** Study area- Rameswaram subset of Chlorophyll-a

* 1. **Pre-Processing**
* **Handling Missing Values:**Removed or replaced missing data caused by atmospheric interference, sensor failure, or cloud cover using statistical methods.
* **Noise Reduction:**Applied filtering techniques to reduce distortions and measurement errors from environmental conditions.
* **Outlier Detection:**Identified and addressed outlier values to prevent negative impacts on model performance.
* **Normalization and Scaling:**Standardized feature values, especially spectral band data, to ensure uniformity and improve model learning efficiency.
  1. **Proposed Architecture**

**Fig. 5.** Workflow diagram for water quality analysis.

**3.4 Feature Modelling**

**3.4.1. Feature Extraction from Copernicus and Modis Aqua**

1. Using Google Earth Engine (GEE) to extract satellite imagery data from MODIS-Aqua.
2. Copernicus data was taken from <https://marine.copernicus.eu/>

**3.4.2. Feature Selection**

Choosing the most relevant bands and environmental variables that have strong correlations with:

1. Chlorophyll-a concentration
2. SST

**3.4.3. Feature Engineering**

Creating new features or transforming raw data into more informative inputs, such as:

1. Normalizing or scaling values
2. Handling missing values

**3.4.4. Modelling and Input to ML and DL Algorithms**

The processed features are utilized as input variables for both non-time series and time series machine learning and deep learning models:

1. Non-Time Series Modelling (ML Models):
2. Random Forest (RF): An ensemble learning algorithm where many decision trees are created and their outputs are combined to enhance accuracy and avoid overfitting.
3. Extra Trees Regressor (ETR): An extension of Random Forest with more randomness incorporated during feature selection, leading to better generalization. This model achieved the minimum error in this study.
4. Support Vector Machine (SVM): A supervised learning algorithm to project data into a high-dimensional space for regression. It did suffer from a performance problem due to the complexity of the dataset.
5. XGBoost: A gradient boosting method that builds models in sequence and gradually enhances the errors of the earlier iterations.
6. AdaBoost: A boosting algorithm that is adaptive and combines the weak learners to predict more accurately. But it had more error than ensemble-based tree models.
7. Time Series Modelling:

* Machine Learning Models:

1. ARIMA: A classical statistical method for time series forecasting that combines autoregression, differencing (to make data stationary), and moving average components. It is effective for linear patterns but less adaptive to complex, non-linear trends.

* Deep Learning Models:

1. Convolutional Neural Network (CNN): Traditionally used for image processing, CNNs can also be adapted for time series by capturing local temporal patterns through convolutional filters. Effective in identifying spatial features and short-term dependencies in sequential data.
2. Long Short-Term Memory (LSTM): A specialized type of recurrent neural network (RNN) designed to capture long-range dependencies in time series data. It effectively handles vanishing gradient issues and is well-suited for modeling sequential temporal relationships and non-linear trends in water quality parameters.

**3.5 Regression**

The regression was carried out using satellite data from MODIS-Aqua (via Google Earth Engine) and Copernicus (downloaded directly) to predict key coastal water quality parameters such as Chlorophyll-a (Chl-a) and Sea Surface Temperature (SST). Following data preprocessing and an 80/20 train-test split, both non-time series and time series models were employed. For non-time series data, machine learning algorithms including Random Forest, XGBoost, AdaBoost, SVM, and Extra Trees Regressor were applied. Time series analysis utilized both traditional ARIMA models and deep learning approaches like CNN and LSTM, capturing temporal trends and spatial features. Model performance was assessed using MSE, RMSE, and R², demonstrating that the combination of remote sensing and AI provides a reliable and scalable method for monitoring coastal water quality.

1. **PERFORMANCE ANALYSIS**
   1. **Experimental Setup**

**4.1.1 Environment Configuration**

* **Programming Language:** Python
* **IDE:** Google collab
* **Libraries Used:**

1. scikit-learn (For implementing machine learning algorithms (e.g., Random Forest, SVM, AdaBoost, etc.))
2. tensorflow (For any deep learning experimentation or model extension)
3. Matplotlib (For visualizing results and plotting performance metrics)
4. Numpy (For numerical computations and array handling)
5. Xarray (For reading and processing .nc (NetCDF) satellite data files)

**4.1.2 Hardware Specifications**

* **Processor:** Cloud-based virtual machine
* **RAM:** ~25 GB (Google Colab High-RAM instance)
* **Operating System:** Linux-based (Managed by Colab backend)
  1. **Evaluation Metrics**
* **Mean Squared Error (MSE):**Measures the average of the squared differences between predicted and actual values.

**Formula:**

* **Root Mean Square Error (RMSE):**The square root of MSE; gives error in the same unit as the target variable.**Formula:**
* **Coefficient of Determination (R²):**Indicates the proportion of variance in the dependent variable predictable from the independent variables.**Formula:**

**• Training Loss:**Represents the error calculated on the training dataset after each epoch. It reflects how well the model is learning during training—lower training loss typically indicates better model fit on training data.

**• Accuracy (%):**Indicates the percentage of correct predictions made by the model compared to actual values. In regression contexts, it’s often derived from error-based accuracy calculations, where a higher percentage implies better prediction quality.

* 1. **Results & Analysis**

Performance of different machine learning algorithms in predicting Chlorophyll-a (Chl-a) concentration and Sea Surface Temperature (SST) was compared in terms of Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R²).

Performance of different machine learning models was cross-validated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared. Extra Trees Regressor gave the best prediction for Chlorophyll-a with MSE = 0.2809, RMSE = 0.5570, and R² = 0.8426, and Random Forest was closely followed with MSE = 0.2941 and R² = 0.8036. Support Vector Machine (SVM) and AdaBoost showed the weakest performance with greater MSE and lesser R², and therefore lesser predictive power which mentioned in the Table I.

|  |  |  |  |
| --- | --- | --- | --- |
| **Machine Learning**  **Algorithms** | **Mean Squared Error**  **(MSE)** | **Root Mean Squared Error (RMSE)** | **R-squared (R²) values** |
| **Random Forest** | 0.2941 | 0.5423 | 0.8036 |
| **Extra Trees Regressor** | **0.2809** | **0.5570** | **0.8426** |
| **Support Vector Machine** | 2.1647 | 1.5206 | 0.1724 |
| **XGBoost** | 0.3064 | 0.6130 | 0.8094 |
| **AdaBoost** | 1.3519 | 1.231 | 0.2309 |

**Table 1. CHLOROPHYLL-A RESULTS OF ML Models**

For SST prediction, Extra Trees Regressor had the best performance with MSE = 0.3399, RMSE = 0.5830, R² = 0.8593, followed by XGBoost with MSE = 0.4562, R² = 0.8113 and Random Forest (MSE = 0.7932, R² = 0.6719). SVM and AdaBoost had larger error and lower in R² which mentioned in the Table II.

|  |  |  |  |
| --- | --- | --- | --- |
| **Machine Learning**  **Algorithms** | **Mean Squared Error**  **(MSE)** | **Root Mean Squared Error (RMSE)** | **R² values** |
| **Random Forest** | 0.7932 | 0.8906 | 0.6719 |
| **Extra Trees Regressor** | **0.3399** | **0.5830** | **0.8593** |
| **Support Vector Machine** | 2.2335 | 1.4944 | 0.0761 |
| **XGBoost** | 0.4562 | 0.6754 | 0.8113 |
| **AdaBoost** | 1.6843 | 1.2978 | 0.3033 |

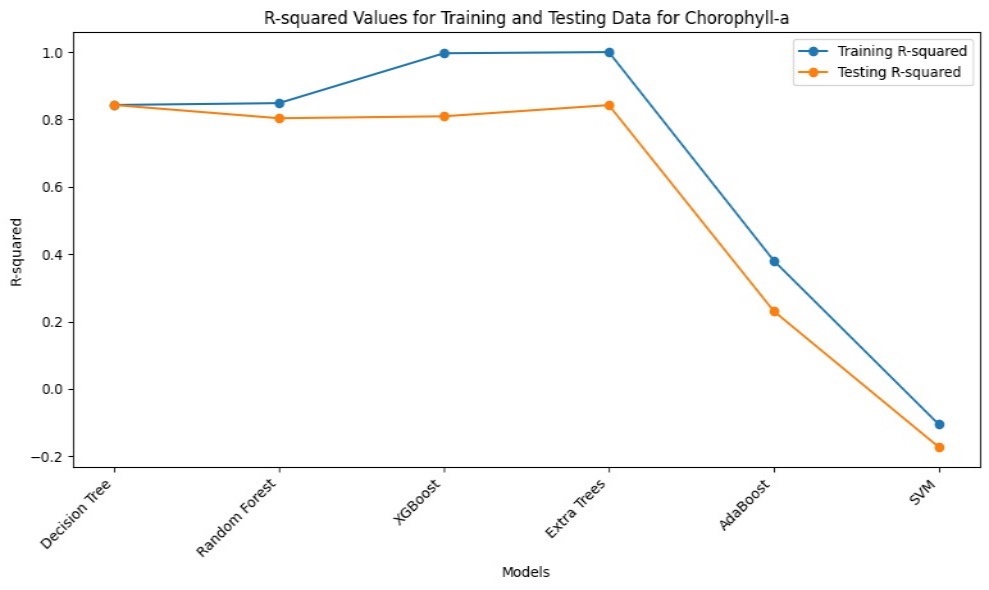
**Table 2. SST RESULTS OF ML Models**

The performance of deep learning models—CNN and LSTM—was assessed using evaluation metrics including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), training loss, and accuracy. Among the two models, the Convolutional Neural Network (CNN) demonstrated superior performance in predicting Chlorophyll-a and Sea Surface Temperature (SST), achieving the highest accuracy with relatively lower prediction error. Although the Long Short-Term Memory (LSTM) model exhibited a slightly lower training loss, it lagged behind in overall accuracy and prediction reliability. These results indicate that CNN is better suited for handling spatial features in satellite data for water quality estimation. A comparative summary of the model performance is provided in Table 3.

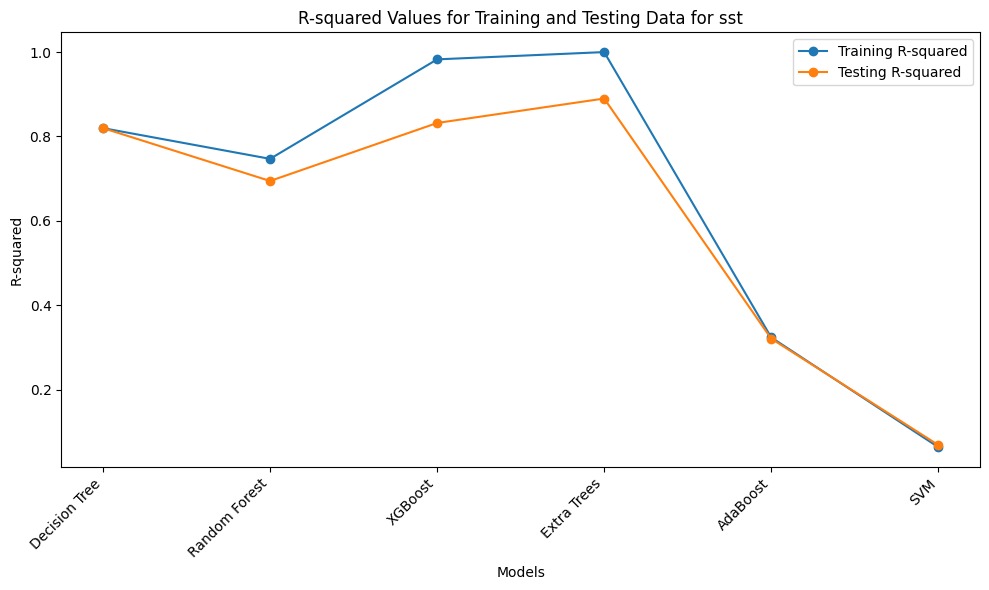
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Deep Learning**  **Algorithms** | **Mean Squared Error**  **(MSE)** | **Root Mean Squared Error (RMSE)** | **Training loss** | **Accuracy**  **(%)** |
| **CNN** | 0.13 | 0.52 | 0.007 | 91.11 |
| **LSTM** | **0.25** | **0.49** | **0.002** | **88.74** |

**Table 3. Chlorophyll-a & SST RESULTS OF DL Models**

The graph (Fig-6) displays the R-squared (R²) measures for training and testing sets over various machine learning algorithms employed to predict chlorophyll-a concentration. Extra Trees Regressor had the maximum training R² (~1.0), implying a good fit to the training set, yet a lower test R² (~0.84) implies a bit of overfitting. Random Forest and XGBoost also had a good performance with high R² values for training and testing, implying good generalization. Conversely, AdaBoost and SVM had much lower R² values, with SVM even having a negative R² for test data, reflecting poor predictive accuracy. The significant difference between training and test R² for some models, especially Extra Trees and AdaBoost, reflects overfitting, whereas ensemble models such as Extra Trees and Random Forest had the best predictive ability.

**Fig.6.** R-Squared values of training & Testing data for Chlorophyll-a

The graph (Fig-7) shows the training and testing R-squared (R²) values for various machine learning models applied to predict sea surface temperature (SST). Extra Trees Regressor had the highest training R² (~1.0) and a very high testing R² (~0.86), reflecting good predictive power with minor overfitting. XGBoost and Random Forest also showed good performance with comparatively high and well-balanced R² values for both datasets. Conversely, AdaBoost and Support Vector Machine (SVM) had much lower R² values, and SVM gave the worst performance, implying that it has difficulty interpreting relationships between data points. The easily apparent difference between training and testing R² in some models, mostly Extra Trees and AdaBoost, suggests different levels of overfitting, whereas the ensemble models, such as Extra Trees and XGBoost, worked best for SST prediction.

**Fig. 7.** R-Squared values of training & Testing data for SST

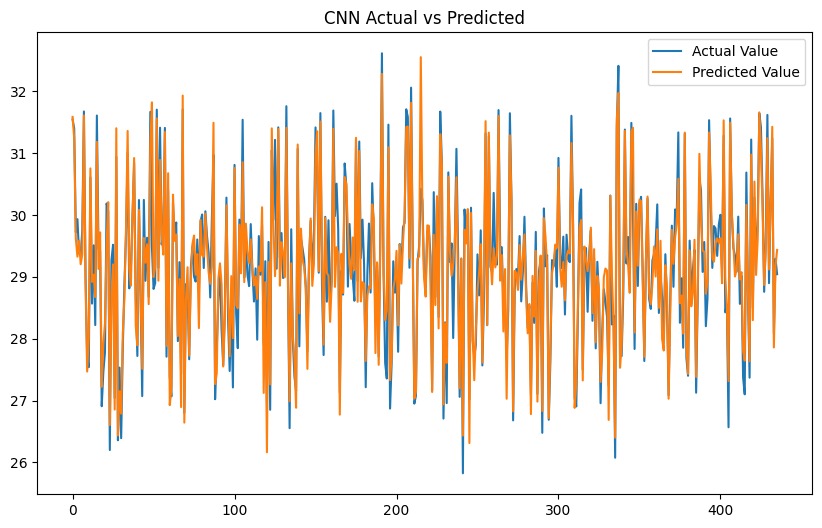
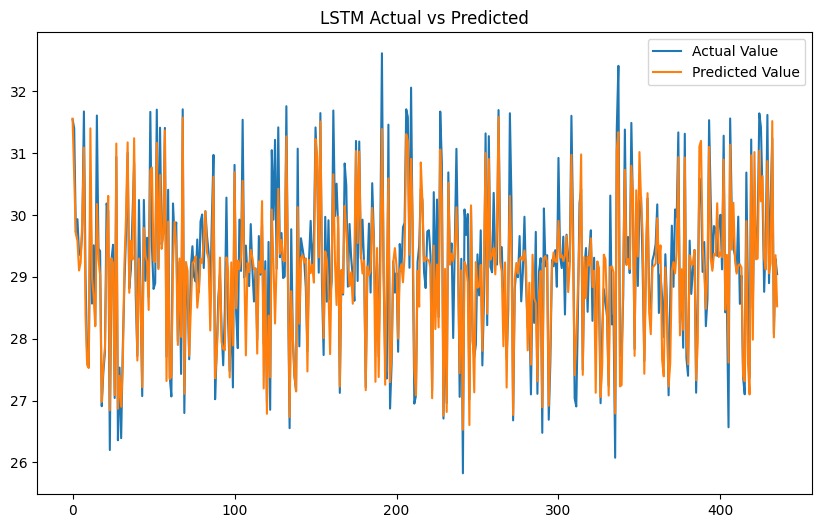
These results show that ensemble models such as Extra Trees and Random Forest worked better than other algorithms in the prediction of SST and Chl-a. The weaker performance of SVM and AdaBoost suggests that these models are not optimally adapted to the complexity of the remote sensing data used in this study.

Fig. 8. Actual Value Vs Predicted Value in CNN Model

The graph (Fig-8) illustrates the performance of a Convolutional Neural Network (CNN) model in predicting values from a time series of remote sensing data. The original value and predicted value show strong alignment, demonstrating that the model successfully captures the temporal dynamics present in the dataset. Despite some fluctuations, the predicted values closely follow the trend and variations of the original values, indicating a high level of accuracy. This close match suggests that the CNN model is well-suited for modeling and forecasting complex patterns in remote sensing time series data.

The graph (Fig-9) illustrates the performance of the Long Short-Term Memory (LSTM) model in predicting values from the time series of remote sensing data. The original value and predicted value exhibit a strong correspondence, indicating that the model effectively learns the temporal dependencies in the dataset. Although some deviations are present, particularly during periods of rapid change, the predicted values generally track the overall pattern and variability of the original values. This result demonstrates that the LSTM model is capable of accurately modelling complex sequential behaviour in the remote sensing time series data.

**Fig. 9. Actual Value Vs Predicted Value in CNN Model**

The CNN model outperformed the LSTM model in predicting Chlorophyll-a and SST, demonstrating higher accuracy and better adaptability to spatial features in satellite data. While LSTM is suited for temporal patterns, it was less effective in handling the spatial complexity of remote sensing inputs. Overall, CNN proved to be the more robust model for this application.

1. **CONCLUSION & FUTURE WORK**

**5.1 CONCLUSION**

In this project, we explored the use of satellite remote sensing and artificial intelligence to predict two key indicators of coastal water quality—Chlorophyll-a and Sea Surface Temperature. These parameters play a crucial role in understanding marine ecosystems, detecting algal blooms, and tracking the effects of climate change. By incorporating time series deep learning models, the system became more effective at recognizing seasonal patterns and long-term variations, resulting in more accurate and meaningful predictions. This approach not only improved prediction performance but also enhanced the system’s ability to adapt to changing environmental conditions. The results highlight the potential of combining Earth observation data with advanced AI techniques to create scalable, real-time monitoring tools. Such tools can support timely decision-making, improve resource management, and contribute to the sustainable development of coastal.

**5.2 FUTURE WORK**

Going forward, the focus will be on incorporating climate-related factors such as rainfall, air temperature, and wind patterns to make the predictions more accurate and responsive to environmental changes. Transfer learning will also be used to adapt the models to other coastal regions with minimal retraining, making the system more flexible and widely applicable. In addition, future work will explore uncertainty estimation techniques to better quantify the confidence of model predictions, helping users assess the reliability of outputs, especially in critical decision-making scenarios.

**REFERENCES**

[1] Ahn, Y.-H., Shanmugam, P., “Derivation and analysis of the fluorescence algorithms to estimate phytoplankton pigment concentrations in optically complex coastal waters,” Journal of Optics A: Pure and Applied Optics, vol. 9, pp. 352–362, 2007, doi: 10.1088/1464-4258/9/4/008.

[2] Akhtar, N., Ishak, M.I.S., Ahmad, M.I., Umar, K., Md Yusuff, M.S., Anees, M.T., Qadir, A., Ali Almanasir, Y.K., “Modification of the water quality index (WQI) process for simple calculation using the multi-criteria decision-making (MCDM) method: a review,” Water, vol. 13, no. 7, p. 905, 2021, doi: 10.3390/w13070905.

[3] Al-Mutairi, N., Abahussain, A., El-Battay, A., “Spatial and temporal characterizations of water quality in Kuwait Bay,” Marine Pollution Bulletin, vol. 83, pp. 127–131, 2014, doi: 10.1016/j.marpolbul.2014.04.009.

[4] Ansari, A., Kumar, A., “Water quality assessment of Ganga River along its course in India,” Innovative Infrastructure Solutions, vol. 7, pp. 1–9, 2022, doi: 10.1007/s41062-021-00668.

[5] Balasubramanian, W., Pahlevan, N., Smith, B., Binding, C., Schalles, J., Loisel, H., Gurlin, D., Greb, S., Alikas, K., Randla, M., Bunkei, M., Moses, W., Nguyễn, H., Lehmann, M.K., O’Donnell, D., Ondrusek, M., Han, T.H., Fichot, C.G., Moore, T., Boss, E., “Robust algorithm for estimating total suspended solids (TSS) in inland and nearshore coastal waters,” Remote Sensing of Environment, vol. 246, p. 111768, 2020, doi: 10.1016/j.rse.2020.111768.

[6] Banerjee, S., Shanmugam, P., “Novel method for reconstruction of hyperspectral resolution images from multispectral data for complex coastal and inland waters,” Advances in Space Research, vol. 67, pp. 266–289, 2021, doi: 10.1016/j.asr.2020.09.045.

[7] Bar, A.R., Mondal, I., Das, S., et al., “Mapping of tide-dominated Hooghly estuary water quality parameters using Sentinel-3 OLCI time-series data,” Environmental Monitoring and Assessment, vol. 195, p. 975, 2023, doi: 10.1007/s10661-023-11552-8.

[8] Geetha, T.S., Gandhimathi, G., Chellaswamy, C., Selvan, P.T., “Comprehensive River water quality monitoring using convolutional neural networks and gated recurrent units: A case study along the Vaigai River,” Journal of Environmental Management, vol. 365, p. 121567, 2024, doi: 10.1016/j.jenvman.2024.121567.

[9] Karthick, M., Shanmugam, P., Kumar, G.S., “Long-term water quality assessment in coastal and inland waters: An ensemble machine-learning approach using satellite data,” Marine Pollution Bulletin, vol. 209, pt. B, p. 117036, 2024, doi: 10.1016/j.marpolbul.2024.117036.

[10] Krishnaraj, A., Honnasiddaiah, R., “Remote sensing and machine learning based framework for the assessment of spatio-temporal water quality in the Middle Ganga Basin,” Environmental Science and Pollution Research, vol. 29, pp. 64939–64958, 2022, doi: 10.1007/s11356-022-20386-9.

[11] Mamun, M., Hasan, M., An, K.-G., “Advancing reservoirs water quality parameters estimation using Sentinel-2 and Landsat-8 satellite data with machine learning approaches,” Ecological Informatics, vol. 81, p. 102608, 2024, doi: 10.1016/j.ecoinf.2024.102608.

[12] Nagaraj, R., Kumar, L.S., “Surface water body extraction and change detection analysis using machine learning algorithms: A case study of Vaigai Dam, India,” 2023 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT), 2023, doi: 10.1109/IConSCEPT57958.2023.10170342.

[13] Nazeer, M., Bilal, M., Alsahli, M.M.M., Shahzad, M.I., Waqas, A., “Evaluation of empirical and machine learning algorithms for estimation of coastal water quality parameters,” ISPRS International Journal of Geo-Information, vol. 6, no. 11, p. 360, 2017, doi: 10.3390/ijgi6110360.

[14] Prakash, P., S. K. S, Bhattacharjee, S., P. U., Gangadharan, K. V., “Predictive model for enhancing water quality monitoring leveraging satellite data,” 2024 IEEE Space, Aerospace and Defence Conference (SPACE), 2024, doi: 10.1109/SPACE63117.2024.10667767.