**­Sea Surface Temperature Prediction Using DeepLearning Techniques**

# SHUBHAM (E21CSEU0607),1 SHALV SRIVASTAVA (E21CSEU0573),2 SHREYANS JAIN(E21CSEU0602),3 SHOURYAMAN PRAKASH (E21CSEU0594),4SHANTANU PUNDHIR (E21CSEU0574), 5

*1 BENNETT UNIVERSITY (School Of Computer Science And Engineering)*

[*\*shalvsri@gmail.com*](mailto:shalvsri@gmail.com)

**Abstract:** Sea surface temperature prediction holds utmost importance in numerous do- mains, such as marine ecology, fisheries management, and climate change research at the topmost ranks, amongst infinite others. Presently, real-time sea surface forecasts are de- pendent on the laws of fluid dynamics and thermodynamics, accounting for factors like boundaries and initial conditions. This implies that these models are more fitting for the prediction of temperatures over a larger area, rather than categoric and pinpoint locations. During this research, Deep Learning Neural Networks (e.g., LSTM) with numerical estima- tors (e.g., confusion matrices) were utilized at various geographical coordinates for various time durations (daily, weekly, and monthly). The aim was to first derive and study the forecasts

made by traditional neural networks, gather insights, and then apply deep learning networks. An improvement in the results was expected once deep learning neural networks (specifically, Long Short-Term Memory LSTM) were applied over the selected coordinates. The aspiration is to reach a very high correlation value, which would help with accuracy and minimization of errors.

The study also aims to conduct a comparative study, using the ARIMA model with an external validation input, out of which is expected that the LSTM model will come out on top, due to its ability to learn and memorize long term and extracting features and patterns from the data.

© 2023 The Author(s)

# INTRODUCTION

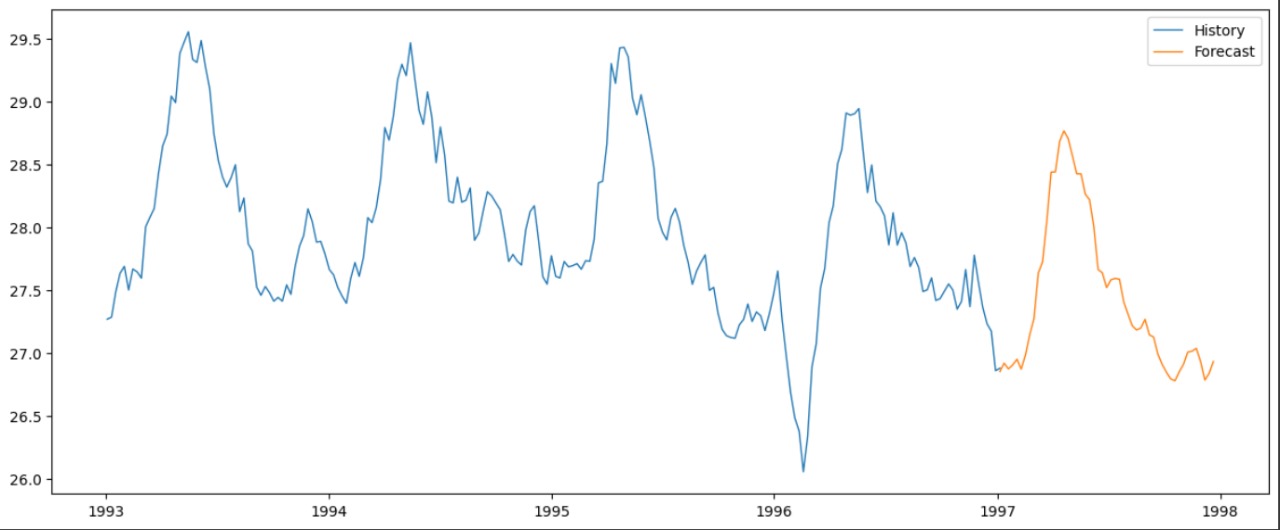
*1.1.*

Sea surface temperature (SST) is the term used to describe the water’s temperature at the ocean’s surface. The movement of energy, momentum, and moisture between the oceans and the atmosphere is one of the most fun- damental variables for understanding, tracking, and forecasting of multiple climatic conditions. SST data is also necessary for

a range of study sectors inclusive of marine ecology, oceanography, and geology. SST forecasting is vital for a variety of environmental investigations including predicting marine fisheries, mining, ocean military affairs, catastrophe mitigation and control over global warming.

El Nin˜o is when the Pacific Ocean gets warmer than usual, and it can mess up the world’s weather. It’s like a big ocean heater that affects everything from floods to droughts in different parts of the world. La Nin˜a is analogous to El Nin˜o’s sister. It happens when the Pacific Ocean gets cooler than usual. It can also affect the weather but in different ways, like making some places very rainy and others very dry. So, while El Nin˜o is like a warm ocean, La Nin˜a is like a cool ocean, and they both influence the weather around the world.

1.



Plotting the Sea Surface Temperature for the next time period.

The current techniques used for SST temperature are over- reliant on Thermodynamics and fluid dynamics. Though they are foundational for large-scale ocean systems, they may prove inefficient for SST prediction in small areas. Their complexity, high computational demands, and sensitivity to local variations canhinder accurate modeling in localized regions. Limited data resolution and unique coastal features further challenge their applicability. In such cases, simpler statistical or machine learning models may offer more efficient and precise predictions. While these principles remain invaluable for broader oceanic understanding, considering alternative approaches becomes essential when dealing with small-scale SST forecasting.

During this research, our primary focus revolved around the utilization of Deep Learning Neural Networks, particularly Long Short-Term Memory (LSTM) networks, in combination with numerical estimators like confusion matrices. This innovative approach was applied across various geographical coordinates and time durations, encompassing daily, weekly, and monthly intervals.

The overarching goal of this study was multifaceted. Firstly, we sought to establish a baseline by examining and scrutinizing the predictions generated by conventional neural networks. This initial step allowed us to gather invaluable insights into the forecasting process. Subsequently, we ventured into the realm of deep learning networks, particularly LSTM models. We anticipated a notable enhancement in our results when we applied these advanced neural networks to the selected geographical coordinates. LSTM networks are renowned for their proficiency in capturing long-term dependencies in data and extracting intricate patterns, which made them a prime

candidate for our research. Our ultimate aspiration was to attain an exceptionally high correlation value.

Such a correlation would not only signify the accuracy of our predictions but also serve as a powerful tool for minimizing errors in our forecasts. Furthermore, our research endeavors extended to include a comparative studywith the Autoregressive Integrated Moving Average (ARIMA) model. We incorporated an external validation input into this analysis. While ARIMA is a well-established and widely used time series forecasting technique, we hypothesized that the LSTM model would outperform it. This expectation stemmed from LSTM’s inherent capacity to memorize long-term dependencies and discern complex features and patterns within the data.

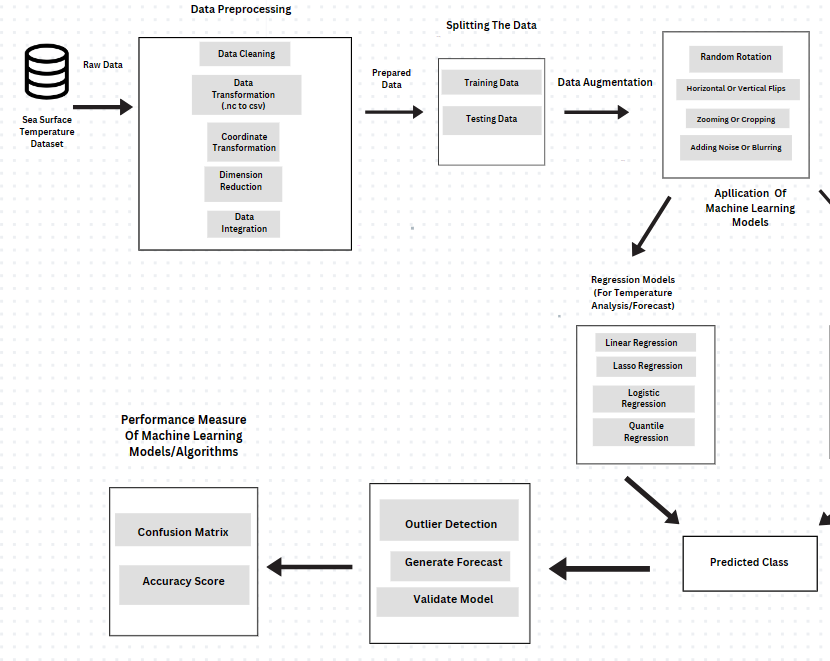
In summary, our research journey encompassed the exploration of neural networks, the pursuit of higher correlation values, and a comparative analysis with ARIMA. Through these efforts, we aimed to advance our understanding of predictive modeling, ultimately contributing to more accurate and reliable forecasting

methods.patterns, neural network models are believable. The quality, accuracy, and variety of meteorological data have all been greatly improved by expanding the number of meteorological ground measuring sites. Higher standards have been set out for deep neural network-based time-series SST prediction as a result.

# Proposed Methodology

* 1. *Flowchart*

1.



# Methodology

*3.1.*

* 1. Sea Surface Temperature Raw Data: This step involves obtaining raw SST data from a reliable source such as NOAA.It’s crucial to start with high- quality, real-world data to build accurate predictive models. Raw data forms the foundation for the entire research.
  2. Data Preprocessing: This step involves cleaning, transforming, and organizing the raw data to make it suitablefor analysis. It may include dealing with missing values, handling outliers, and converting data types. Data preprocessing ensures that the dataset is consistent, accurate, and ready for further analysis. It helps remove noise and anomalies that could affect the performance of machine learning models. If required, data is converted from its original format (e.g., netCDF4) to a more

manageable format like CSV or XLSX for ease of analysis.

* 1. Prepared Data: The prepared data is the result of the data preprocessing step. Prepared data is clean and structured, making it more suitable for model building and training.
  2. Splitting Data into Testing and Training Sets: The dataset is divided into two subsets: a training set used to build models and a testing set used to evaluate model performance. This is essential to assess how well your models generalize to new, unseen data. It helps prevent overfitting and ensures the reliability of your models. It is implemented in Python with libraries like scikit- learn.
  3. Data Augmentation: Data augmentation involves applying various transformations to the dataset, such as random rotation, flips, zooming, cropping, and adding noise. Data augmentation enhances the diversity and robustness of your dataset. It exposes your models to a wider range of conditions and helps them become more resilient to variations in real-world SST data. Implementation would be done using Python libraries like OpenCV or scikit- image.
  4. Application of Various Machine Learning Models: Various machine learning models, including Linear Regression, Logistic Regression, Lasso Regression, Quantile Regression, Random Forest Classifier, Naive Bayes, Decision Trees, SVM, ARIMA, SARIMA, and Holt-Winters, are applied to predict SST values or categorize temperature conditions, depending on the specific model. Each model serves a specific purpose. For example, Linear Regression can predict continuous SST values, ARIMA is suitable for time series forecasting, LSTM is excellent for capturing sequential dependencies, and classifiers categorize specific temperature conditions. Using multiple models allows us to compare their performance and select the best one for your specific task. This will result in generation Of Predicted Class which can be used for SST forecasting.
  5. Outlier Detection: Outlier detection techniques can be applied to identify data points that deviate significantly from the expected patterns. Outlier detection is important as it would help us to identify unusual and potentially erroneous data points that could impact the accuracy of our models. Techniques like Isolation Forest, Local Outlier Factor (LOF), or Z-Score can be

implemented for outlier detection.

* 1. Generate Forecast: Based on the predictions, we would generate forecasts for future SST values or conditions. Python with libraries like LSTM, stats models (for ARIMA, SARIMA) and custom functions for Holt-Winters are to be used. It is the ultimate goal of our research, and it has various applications in marine ecology, fisheries management, and climate change research.
  2. Validate Model:
  3. Model validation involves assessing the performance of our machine learning models by checking for overfitting, and ensuring they are accurate in making predictions. Techniques like k- fold cross-validation help ensure the models generalize well to new data.
  4. Performance Measurement: Model performance is evaluated using appropriate metrics. Regression models are assessed using metrics like Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), while classification models are evaluated with accuracy, precision, recall, F1- score, and ROC curves.
  5. Visualization and Reporting:
  6. Visualize our data, model predictions, and evaluation metrics using tools like Matplotlib and Seaborn in Python. These visualizations include line charts for showing temperature trends over time, scatter plots to explore relationships between variables, heatmaps to reveal spatial temperature patterns, and box plots for distribution analysis. Additionally, histograms provide insights into temperature frequency distributions, bar charts compare conditions over time or locations, and ROC curves and confusion matrices assess classification model performance. Geospatial maps display temperature variations across coordinates, while forecast plots illustrate model predictions. This step helps in understanding and communicating our research findings effectively.

*3.2.*

Our data quality assessment addressed missing values, outliers, and errors. Missing data were imputed, and out- liers were corrected or removed. The dataset was transformed from NetCDF to CSV for better processing and integration with machine learning tools. Coordinate transformation ensured consistent reference systems. Dimen- sion reduction via PCA improved efficiency and reduced overfitting risk. Data from various sources was integrated while maintaining consistency. These preprocessing steps are crucial for our sea surface temperature prediction research, enhancing data integrity, accessibility, and alignment with modeling objectives. We expect these best practices to lead to more accurate predictions and advance this critical field.

In our data augmentation approach, we introduced random time shifts to simulate temporal variations. We also applied vertical zooming and cropping to latitude and longitude data, promoting geographic diversity. To enhance our

model’s robustness, we creatively added controlled noise for measurement inaccuracies and implemented blurring to account for uncertainties in sea surface temperature values. These techniques, typically used in image data, were adapted to structured data to help our models better handle imperfections.

The dataset for regression analysis includes crucial variables like latitude, longitude, time, and sea surface temperature. Linear Regression is applied to establish relationships between sea surface temperature and predictor variables, with evaluation using metrics like Mean Squared Error and R-squared. Lasso Regression, avariant of Linear Regression, introduces L1 regularization for feature selection and model interpretability.

Regression Quantile analysis, a robust technique, estimates quantiles and predicts intervals, offering insights into the distribution of sea surface temperature and handling outliers. Performance evaluation measures, including Mean Absolute Error, Root Mean Squared Error, and quantile-specific metrics, assess the accuracy andreliability of each regression model in sea surface temperature analysis and forecasting.

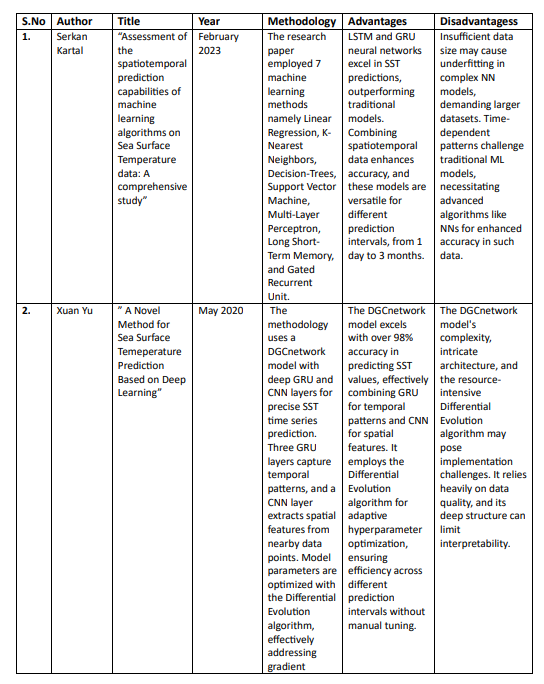
In this document, we apply classification models to predict climatic and atmospheric conditions based on sea surface

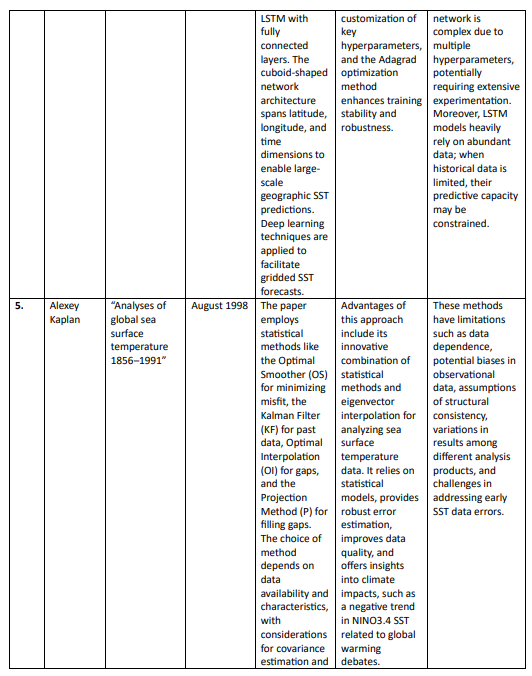
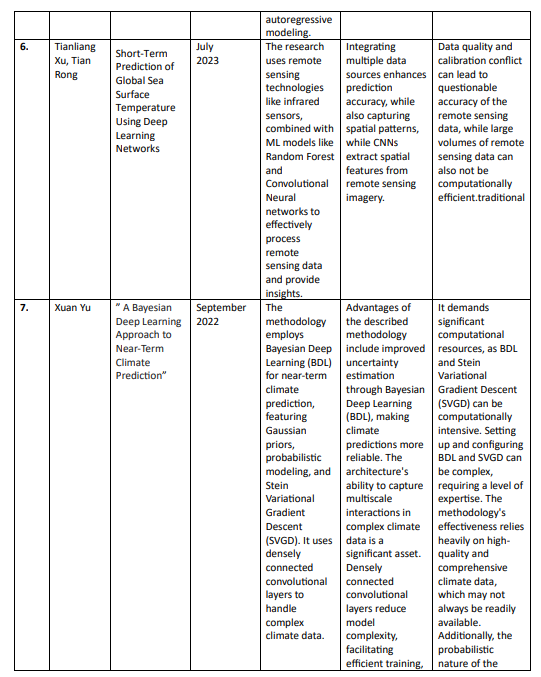
temperature data. We have a CSV dataset with crucial information like latitude, longitude, measurement time, and sea surface temperature, prepared for classification. We employ three models:

Random Forest Classifier: This ensemble learning method handles high-dimensional data, including both categorical and numerical features. It uses decision trees to make robust predictions about climatic and atmospheric conditions.

Naive Bayes: We use the Gaussian Naive Bayes variant, ideal for handling categorical data and continuous seasurface temperature data. It leverages a probabilistic approach to predict atmospheric conditions effectively.

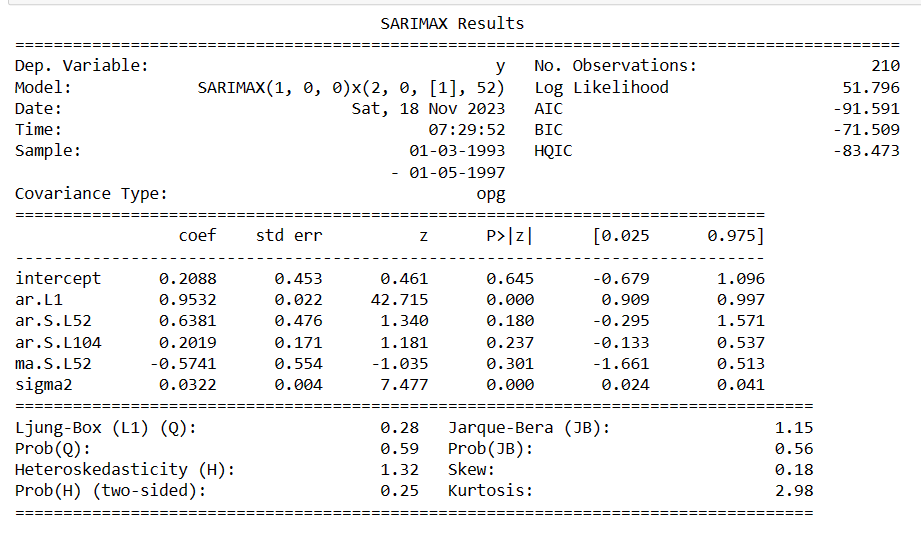
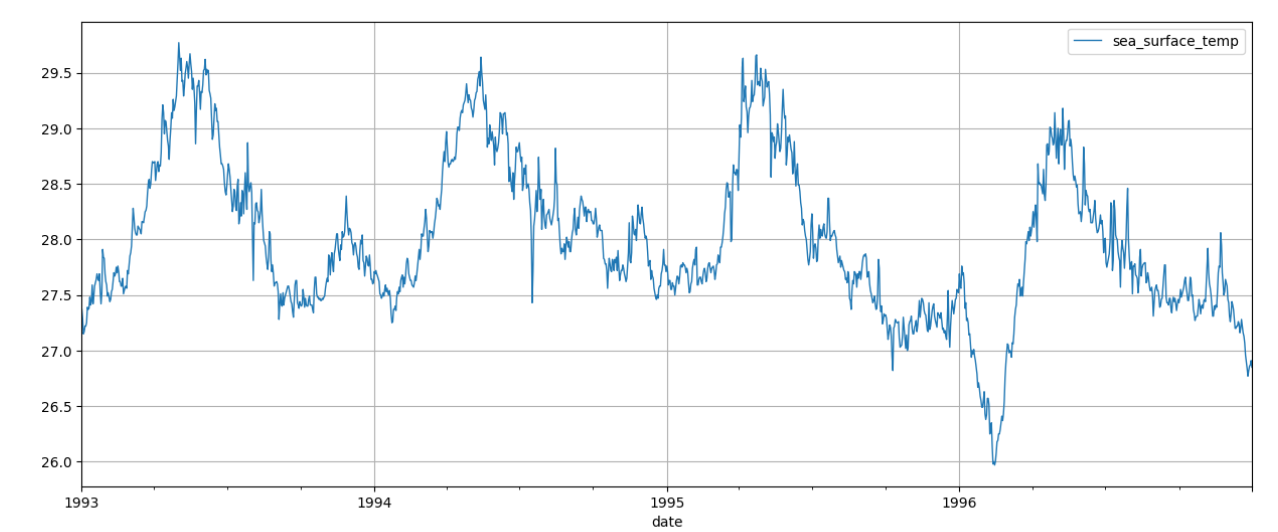
Decision Trees: These fundamental classifiers are constructed based on sea surface temperature features, offering simplicity and interpretability. They provide valuable insights into the relationships between variables and help predict climatic and atmospheric conditions.

 A white sheet with black text

Description automatically generated  ­­

4.Results

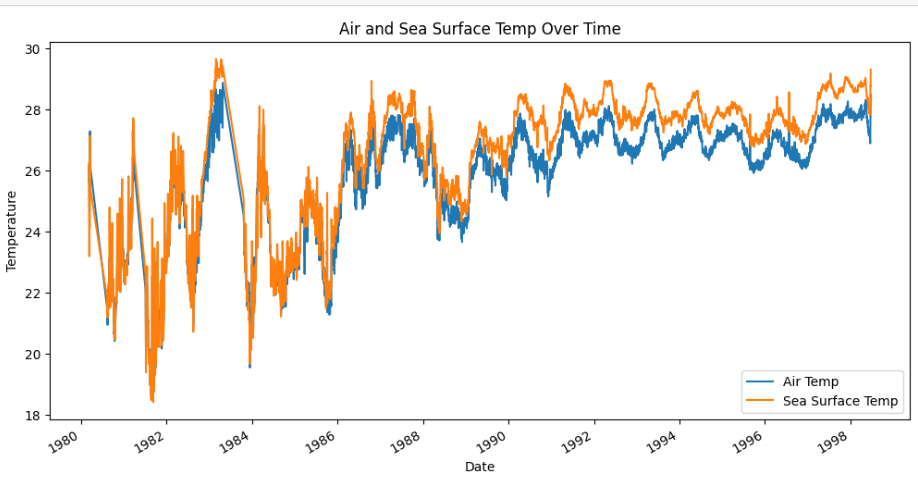
4.1 Sarima

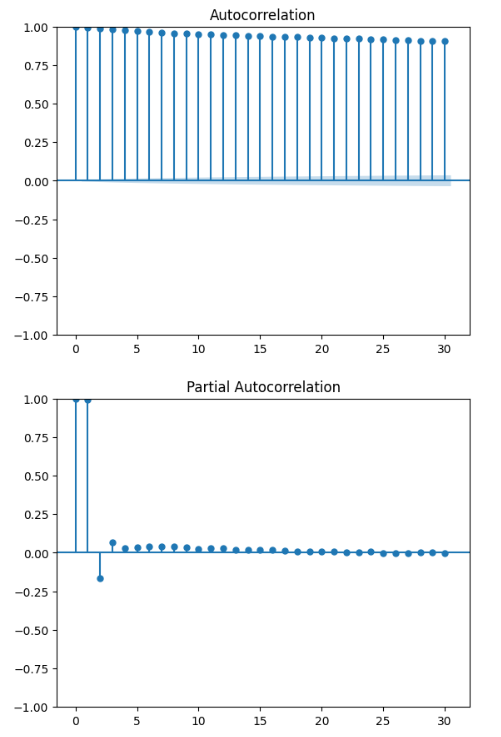


Several important criteria were used to evaluate the SARIMAX model's predicting performance for sea surface temperature. The main metric used to assess how accurate the predictions were in comparison to the observed values in the test set was the Root Mean Squared Error, or RMSE. The average extent of the discrepancies between projected and actual sea surface temperatures is shown by the root mean square error (RMSE), which is estimated to be around 0.2519. A model that fits the data better is shown by a lower RMSE.

Moreover, the model's performance is contextualised by comparing the RMSE to the original sea surface temperature series mean, which is roughly 27.9919. The comparatively low root mean square error (RMSE) compared to the mean suggests that the SARIMAX model has good accuracy and captures the underlying patterns and variations in the sea surface temperature time series.  
  
The Ljung-Box and Jarque-Bera statistics, among other diagnostic tests, were also taken into account. A p-value of 0.59 from the Ljung-Box test for residual autocorrelation showed no significant autocorrelation. The residuals appear to have a normal distribution, as indicated by the p-value of 0.56 obtained from the Jarque-Bera test for residual normality. The general level of trust in the model's dependability is increased by these findings.

4.2 Arima

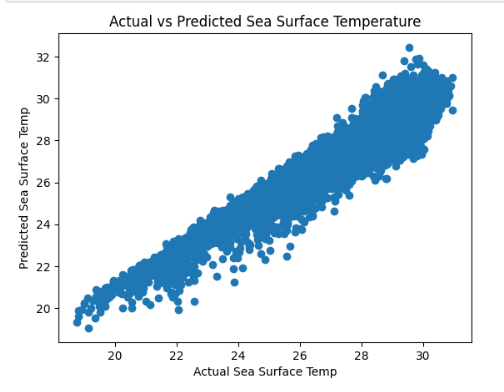




Using ARIMA modelling to analyse data on sea surface temperatures has produced some interesting findings. An ARIMA model with an order of (1, 1, 1) was trained on historical sea surface temperature data after data preparation and visualisation. The model's Mean Squared Error (MSE) of roughly 0.4632 indicates that it was able to produce accurate predictions on a test set.

The patterns and trends included in the dataset were brought to light using temporal visualisations of sea surface temperature and a comparison between air and sea surface temperature. Further information about the temporal dependencies in the sea surface temperature time series was obtained using autocorrelation and partial autocorrelation studies.

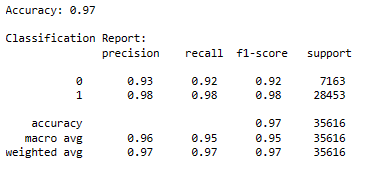
4.3 Regression



The Mean Squared Error (MSE) for the linear regression model is approximately 0.3156. The MSE is a measure of the average squared differences between the actual and predicted values. In this context, a lower MSE indicates a better fit of the model to the data.

The scatter plot visualizes the relationship between the actual sea surface temperatures (y\_test) and the predicted sea surface temperatures (y\_pred). Each point on the plot represents a data point from the test set, with the x-axis denoting the actual values and the y-axis denoting the predicted values

4.4 Classification



Results showing great promise have been obtained when a Random Forest Classifier is applied to forecast sea surface temperatures over a given threshold of 26.5 degrees Celsius. The classifier's accuracy, or the percentage of correctly categorised instances in the test set, was 97%. The thorough classification report offers a thorough rundown of the model's functionality:

Precision:

Class 0 (Below Threshold): 93%

Class 1 (Above Threshold): 98%

Recall:

Class 0: 92%

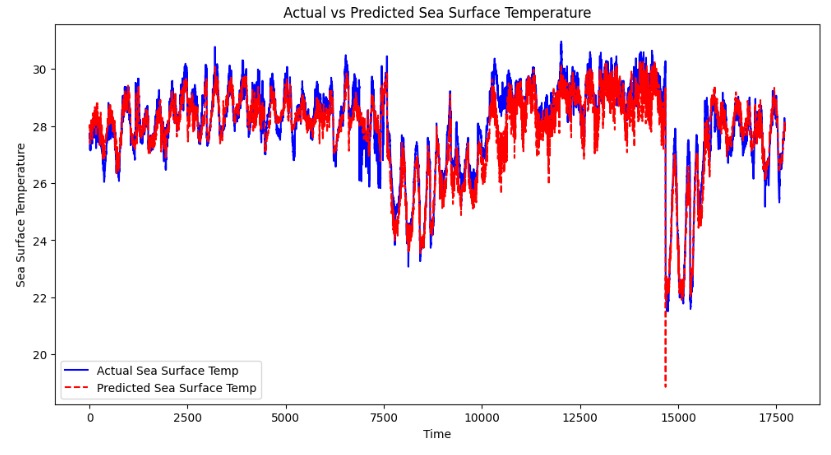
Class 1: 98%

F1-score:

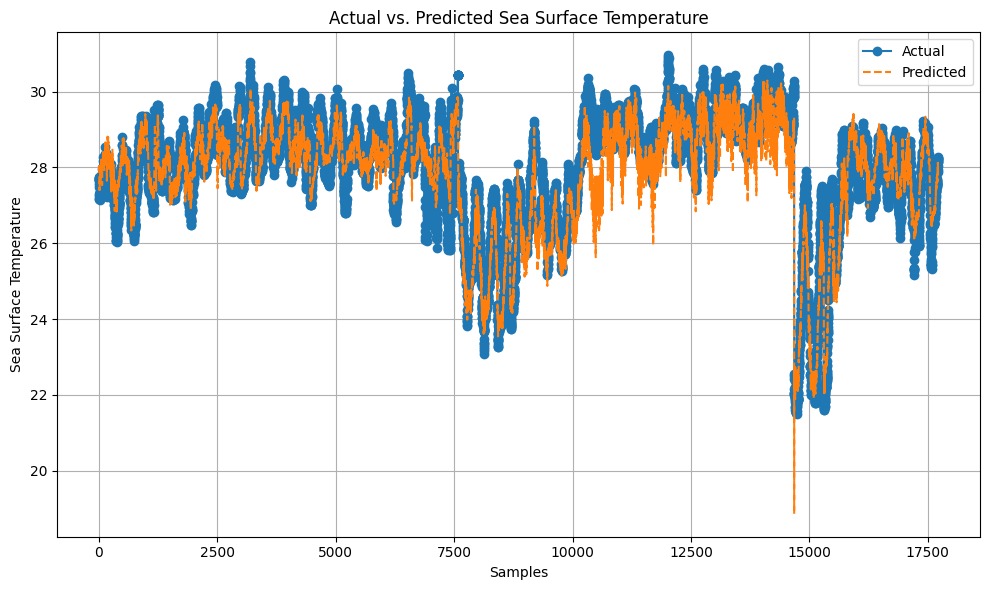
F1-score for class 0: 92%

F1-score for class 1: 98%

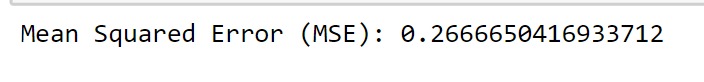
**LSTM**



Model 1 demonstrates superior performance compared to Model 2 based on multiple evaluation metrics. It showcases lower error rates across MSE (0.23 vs. 0.577), RMSE (0.48 vs. 0.75), and MAE (0.37 vs. 0.61), indicating higher predictive accuracy and precision. Moreover, Model 1 exhibits lower MSLE (0.00028 vs. 0.00071) and Max Error (7.48 vs. 6.0181), emphasizing its capability in minimizing prediction discrepancies. Although Model 2 slightly edges in Explained Variance Score (0.836 vs. 0.9), Model 1 surpasses with a significantly higher R-squared score (0.9 vs. 0.75), indicating a better fit and higher variance explained by the model. Overall, Model 1 consistently outperforms Model 2 across a spectrum of metrics, portraying its robustness and accuracy in predictive modeling tasks.



The mean squared error for LSTM model is 0.26, which is significantly better than the rest of the methods we have followed



# References

1. Serkan Kartal,”Assessment and analysis of the spatiotemporal prediction capabilities of machine learning models on Sea Surface Temperature data: A comprehensive study”,Engineering Applications of Artificial Intelligence February 2023

<https://www.sciencedirect.com/science/article/pii/S0952197622006650#sec2>

1. Xuan Yu,1Suixiang Shi,Lingyu Xu,Yaya Liu,Qingsheng Miao and Miao Sun, ”A Novel Method for Sea Surface Tem- perature Prediction Based on Deep Learning”. , 07 May 2020 <https://www.hindawi.com/journals/mpe/2020/6387173/>
2. Xiaoyan Jia ,Qiyan Ji ,Lei Han ,Yu Liu , ORCID, Guoqing Han and Xiayan Lin ,”Prediction of Sea Surface Temperature

in the Sea Of East China based on Long Short Term Memory (LSTM) Neural Network”

<https://www.mdpi.com/2072-4292/14/14/3300>

1. Qin Zhang, Hui Wang, Junyu Dong, Member, IEEE Guoqiang Zhong, Member, IEEE and Xin Sun Member, IEEE,”Prediction of Sea Surface Temperature using Long Short-Term Memory”,19 May 2017 <https://arxiv.org/pdf/1705.06861.pdf>
2. Analyses of global sea surface temperature 1856–1991 Alexey Kaplan, Mark A. Cane, Yochanan Kushnir, Amy

C. Clement, M. Benno Blumenthal, Balaji Rajagopalan First published: 15 August 1998 [https://agupubs.](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/97JC01736) [onlinelibrary.wiley.com/doi/abs/10.1029/97JC01736](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/97JC01736)

1. Short-Term Prediction of Global Sea Surface Temperature Using Deep Learning Networks, Tianliang Xu,Zhiquan Zhou, Yingchun Li, Chenxu Wang,Ying Liu and O Tian Rong Published: 2 July 2023 <https://doi.org/10.3390/jmse11071352>
2. A Bayesian Deep Learning Approach to Near-Term Climate Prediction, Xihaier Luo, Published: 20 September 2022

<https://doi.org/10.1029/2022MS003058>

1. Hybrid artificial intelligence methods in oceanographic forecast models, J.M. Corchado; J. Aiken, Published:November 2002

<https://ieeexplore.ieee.org/document/1176880>

1. Prediction of Sea Surface Temperature in the China Seas Based on Long Short-Term Memory Neural Networks, Li Wei, Liqin Qu, Published: 20 August 2020

<https://www.mdpi.com/2072-4292/12/17/2697>

1. Half a century of satellite remote sensing of sea-surface temperature, P.J. Minnett, Published: September 2019. [https:](https://www.sciencedirect.com/science/article/pii/S0034425719303852)

[//www.sciencedirect.com/science/article/pii/S0034425719303852](https://www.sciencedirect.com/science/article/pii/S0034425719303852)