International Journal of Scientific Research and Engineering Development-– Volume 8 Issue 2, Mar-Apr 2025
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Hybrid and Ensemble Machine Learning Models for Accurate Underwater Temperature Prediction in Coastal Islands of Santa Catarina

Omkar Singh[1], Femenca Noronha[2], Manik Borvadkar[3], Yash Yadav[4]
*(Data Science, Thakur College of Science & Commerce, Mumbai, India
Email: omkarsingh@tcsc.edu.in)
**(Data Science, Thakur College of Science & Commerce, Mumbai, India
Email: femenca1996@gmail.com)
***(Data Science, Thakur College of Science & Commerce, Mumbai, India
Email:manaikprabhu006@gmail.com)
****(Data Science, Thakur College of Science & Commerce, Mumbai, India
Email: yashyadav1211@gmail.com)

----------**********----------

Abstract:

Predictions of subsurface ocean temperatures are highly relevant to marine science since they contribute to ocean dynamics understanding and effects of climate change. This work offers an in-depth discussion of various machine learning methods used for predicting underwater temperatures from temporal data collection. A set of measurements of underwater temperatures was used, further complemented with supplementary variables: geographical coordinates and information on time. This dataset was used both to train and to test an Extra Trees Regressor, a LightGBM Regressor, and an ensemble consisting of a Stacking Regressor. Data Preprocessing Included feature creation, like extracting the temporal features from datetime variables and dropping columns by taking an assumption about columns being irrelevant to the model's training effectiveness. The methodology used in this research includes validating the model by using three appropriate indicators: Mean Absolute Error, RMSE, and R-squared (R^2). Based on these three indicators, the strength of each model was indicated in some applications but weakness in others regarding the predictive output of undersea temperature for each of the approaches used. The results show the ensemble model performs better than singular models; an R^2 of 0.976 is achieved, thus showing a possibility of using combinations of machine learning techniques to improve predictive capability. Further from this study, predictions realized using the model can be compared to actual available data in relation to underwater temperatures, therefore acting as a basis for further studies in this area. Beyond supplementing the body of knowledge already existing with regard to the prediction of subaqueous temperatures, it holds very important implications for marine scientists beyond this improvement. Results proved to highlight the importance of machine learning in enhancing the reliability and accuracy of environmental forecasts, hence propelling a deeper understanding of marine ecosystems.

Keywords — **Underwater Temperature, Environmental Forecasting, Machine Learning, Regression Models, Extra Trees Regressor, LightGBM, Stacking Regressor, Hybrid Model, Exploratory Data Analysis, Comparative Analysis, Model Performance, Marine Ecosystems, Southern Brazil**

----------**********----------

INTRODUCTION:

An introductory preface of a research manuscript is a preface that introduces the reader to the investigation along with contextual information necessary to understand the research issue placed in its proper framework. It synthesizes pertinent literature and presents the worth of the study in the broad academic context. This study aims to address

a set of concerns regarding the possibility to predict minimum temperature values close to Santa Catarina, an area that is greatly valued for its ecological significance in southern Brazil. Regions characterized by remarkable biological diversity are situated along coastal areas, resulting in the understanding that numerous species inhabit marine ecosystems. These ecosystems constitute intricate systems that experience various climatic influences, including temperature variations. The coastline of Santa Catarina is distinctly remarkable due to its exceptional diversity across coastal environments: it is ecologically important and represents significant variability in habitats, ranging from coral reefs to rocky substrates. Through this lens, understanding of intrinsic thermal dynamics is key to enhancing our ability to comprehend changes in temperature and the nature of marine biodiversity, as well as in climate as a whole. The first problem explored in this paper is one of uncertainty-inherent to marine condition forecasting: subsurface temperatures are but a small example from those areas of marine science where physical models and empirical correlations so frequently rely upon limited data points. In spite of this, complex interplays with many factors such as ocean currents, atmospheric conditions and specific geographical features greatly undermine the existence of different temperate conditions underwater. The growing demand for reliable temperature predictions to support resource mining ventures, conservation efforts, and research into local and global climate shifts have enhanced the difficulties along the coast of Santa Catarina. Data-driven approaches of the new generation in machine learning have transformed environmental forecasting practices by allowing the discovery of such a varied collection of practical nonlinear and multi-dimensional correlations embedded in large databases. This capability will be able to use any methodologies related to machine learning techniques like regression models including Extra Trees and LightGBM along with their hybrids. Underwater temperature data analysis with this will be fully useful for the discovery of latent patterns in data that will allow unstructured forecasting. Utilizing highly advanced computational software, this study will make predictions much more accurate. It will therefore allow for the acquiring of reliable measurements of sea temperatures within the coast of Santa Catarina in a demonstrably immediate necessity. This involves validating predictive models that predict the submerged temperatures at different depths of water, in addition to ecosystem diversity and different timescales. The methodology utilizes a synthesis of multiple datasets, temperature recordings, oceanographic parameters, and spatial variables to identify the dominant factors which influence temperature fluctuations in the region. Such a strategy holds high promise for providing critical information relevant to marine conservation and resource management activities, as well as adaptation efforts for climate change. This paper will therefore carry out a comparative analysis of results obtained from this study with alternative machine learning techniques, such as Extra Trees, LightGBM, and hybrid models, to determine the methodologies that produce the most promising ocean temperature forecasts. This analysis also serves in an effort to enlarge the understanding within marine science and environmental prediction circles.

Literature Review: This paper proposes an LSTM model in the forecasting of HWTs within coastal areas of Korea, making an emphasis on a scenario, particularly concerning aquaculture and disease incidents, with significant economic implications within the marine fishery sector. It underscores critical factors like the heat-dome phenomenon as well as the Tsushima warm current and simultaneously assesses its relative effectiveness in comparison with mainstream SST models, thus bearing testimony to the superiority of LSTM by utilizing multi-input datasets; however, it also underscores issues that would link it to prediction accuracy over long time-periods as well as possible apprehensions over data integrity. The implications reflect the current practical worth of the model to aquaculture practitioners and regulatory bodies in terms of bearing on economic vulnerabilities [1].

A system was designed for water temperature under winter conditions-Real-time forecasting for long-distance diversion projects. This paper introduces a real-time forecasting system for improving operational efficiency and solving icing problems in high-latitude regions. The goal of the current research is an important one: to allow for the prevention of ice jams at such points and thus the smooth conveyance of water. The model predicts changes in short-term water temperatures using air temperature predictions and hydraulic information with small errors within the range of ±0.3 to ±0.6°C for 1-7 days. It is on the long-term forecast that is 15 days that the model will be affected by the inaccurate data regarding air temperatures. The study uses latest monitoring technologies; these are GIS and imagery from satellites to avoid some of the disadvantages related to the traditional methods. The system's main limitations include reliance on air temperature data, varying performance across regions, and environmental factors affecting model accuracy. Future work suggests enhancing prediction algorithms and expanding the model's geographic applications to improve adaptability [2].
Paper "Water Temperature Prediction Using Improved Deep Learning Methods through Reptile Search Algorithm and Weighted Mean of Vectors Optimizer". A range of advanced deep learning techniques is applied to improve the accuracy of water temperature prediction by making full use of RSA along with INFO and CNN, LSTM. It forecasts the daily water temperature of the Bailong River in China, using parameters like air temperature and streamflow, where LSTM-INFO performs better. Results contain lesser predication errors in the form of lower values of RMSE, MAE, etc. Models might be difficult to adapt for other conditions. While further work goes on, the models could be tested under different settings with increased environmental factors to generally have broader applicability in terms of helping to control environmental management and resource planning [3].
A manuscript presents a deep learning-based model for SST near the Korean Peninsula, based on economic and ecological impacts of greater rapid growth in the SST due to climate change. Authors built an LSTM network from the time-series SST data pulled from the ECMWF ERA5 dataset. Much in terms of what it aimed to achieve-that high predictive capabilities were noted, particularly for short-term predictions (R^2 = 0.985 for 1-day). The model aims to reduce economic losses in aquaculture by anticipating high water temperature events that would cause extreme fish mortality. Despite the success of LSTM, the accuracy of long-term predictions was compromised, indicating further fine-tuning is needed. This study is valuable because it has presented a resilient SST prediction model and suggests the inclusion of ancillary environmental variables to improve prediction accuracy. The practical implications would be the assistance of policymakers in well-timed warnings, despite the problems of oscillating SSTs and observing large sea areas [4].
This study considers the use of machine learning algorithms in the forecasting of stream water temperature, which is a critical element for both ecological and socio-economic factors. The investigation considers six different models including sophisticated methodologies such as XGBoost and feed-forward neural networks (FNNs) over a sample of ten Austrian catchments, revealing a notable enhancement in predictive accuracy characterized by a mean RMSE of 0.55°C. It puts forward the requirement for hyperparameter tuning as well as input choice in achieving accuracy and makes an open-source R-package, wateRtemp, available to extend application. Successes are reported but challenges are identified in the availability of data and variability in model performance by the hyperparameter settings. Bettering input data and models for accommodating future climate-change scenarios are proposed as future work to extend usage [5].
The developed methodological framework offers an appropriate approach to predicting river water temperature (RWT) using advanced ML algorithms within a tropical river ecosystem in India, where emphasis is given on the importance of the ecological value of an accurate RWT forecast. More complex ML models that are used in the study are ridge regression, K-nearest neighbors, random forest, and support vector regression, which

are further integrated with Sobol' global sensitivity analysis for feature selection and Ensemble Kalman Filter data assimilation purposes. The investigation explores areas where the traditional models are lacking, often failing to integrate sensitivity analysis or scientific knowledge amalgamation. It has been shown that methods including support vector regression can well predict RWT, recognizing seasonal patterns, and improve the accuracy of the forecast. Extremes of thermal stress still cause problems in the handling of data and the performance of models. Further research should investigate more advanced data assimilation techniques and apply the framework to another river system to further test its robustness. The result presents realistic implications in environmental management as it supports strategies geared at ecosystem preservation and water quality monitoring [6].

The purpose of the paper is to increase the accuracy of weather forecasting using linear regression models by comparing two techniques namely normal equation methods and gradient descent. It mainly focuses on accurate weather forecasting being highly relevant for agricultural economies and underlines its crucial relevance to Indians. Even with developments in AI and machine learning, weather prediction does not yet meet the desired standards of accuracy. The derived conclusion is that normal equation works well compared to gradient descent, in the case of meteorological variables, such as temperature, humidity, and dew point for the period between the years 2014 to 2016 in the Vellore dataset with minimal error. Conclusions like these point toward its potential utility toward agricultural decision-making process. However, quality of the data and inefficiencies of the gradient descent showed signs of limitations. Another potential avenue for further research could be fine-tuning pre-processing and other algorithms to improve model robustness [7].

This paper introduces a novel approach to predicting sea surface temperature (SST) anomalies in the NINO region, critical for monitoring El Niño and La Niña events. The research employs a Long Short-Term Memory (LSTM) neural network, trained on historical SST data from 1854 to 2015, to forecast anomalies for the next 1 to 3 months. The LSTM model somehow outperformed the deficiencies of conventional statistical methods in the proper capturing of nonlinear SST patterns of change. The good results - correlation coefficients above 0.88, entail some reliability in the SST trends forecast. Still, it seems that performance in longer forecast periods is still not impressive, and therefore an enhancement and possible integration with other techniques might be necessary. The research has much practical implications for climate forecast and disaster preparation, although further studies should concentrate on extending time spans for forecasts and incorporating multiple sources of data [8].

Methodology: ***This chapter provides the systematic approach which will be taken in this study while predicting subaquatic temperatures using machine learning methodologies. It contains phases: data extraction and preprocessing, feature development, model choice, training, testing, and hybrid approach towards achieving accuracy which is discussed under the following subheadings.***

I. **Data Collection and Preprocessing:** The data collection procedure for this research was based on an acoustical record of submarine temperature, which was collected at every 20 minutes from December 2012 to July 2014. All collections were done along the southern coastline of Brazil, in Santa Catarina. In the raw data, the variable fields consisted of date, time, temperature (Temp), and Site.

II. **Data Preprocessing:** The preprocessed data awaits pre-training for preparing it to be used for training purposes. The next steps include:

- **Datetime Processing:** Date and time processing There are two separate columns, date and time, combined into

one column. Then, this column is converted into the format of datetime, and after the procedure, original Date and Time columns are deleted.

- **Feature Selection:** The site column was put only to avoid the explicit location information being added to the model.
- **Handling Missing Values:** The data set went through validation, finding missing values. These were filled or removed appropriately to ensure that the data consistent was generated.

III. **Feature Engineering:** Feature engineering was used to enrich the data set and yield more fit results for the model.

- **Temporal Features: These newly extracted attributes in the Datetime column are Year, Month, Day, Hour, etc.**
- **Final Feature Set: The engineered features (Year, Month, Day & Hour) were used as inputs (X), while the temperature readings (Temp) served as the target variable (y).**

IV. **Model Selection and Training: For predicting subsurface temperature, three top machine learning models were found to have the highest efficiency in tasks with regression.**

- **Extra Trees Regressor (ETR):** The Extra Trees Regressor is one such ensemble learning method based on tree-based approaches with randomized selection for optimal precision in predictions. It usually works better on datasets where the relationship among samples could not be linear in nature.
- **Light Gradient Boosting Machine (LightGBM):** The Light Gradient Boosting Machine is one of the gradient boosting frameworks known especially for its high efficiency in computation, and it can deal with large-scale and complex data. Here, an efficiency-optimal mode model has been used to discover complex patterns in the dataset.
- **Stacked Regressor:** An ensemble learning method which combines different regression models for better predictive accuracy is known as stacked regressor or stacking regression. Such a technique forms a hierarchy of models toward obtaining the "meta-model" better than any one model.

V. **Training and Evaluation:** There were two non-overlapping subsets in the database, with 80% of data covered by the training subset and the remaining 20% covered by the testing subset. To make sure it's reproducible, a random state is provided. For training each model, the training subset was used for the training process while the testing subset was used to evaluate each model in different aspects.

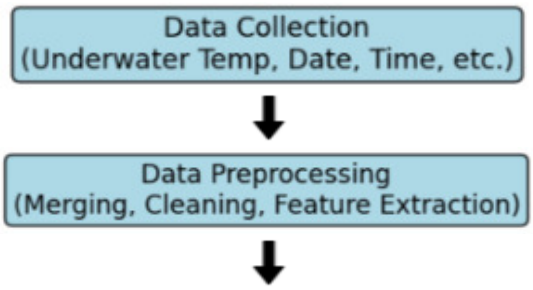
- **Mean Absolute Error (MAE):** Measures the average magnitude of errors in predictions.
- **Root Mean Square Error (RMSE):** Provides an indication of the standard deviation of prediction errors.
- **R-squared (R^2):** Indicates the proportion of variance in the target variable that is predictable from the input features.

VI. **Hybrid Model Approach:** Hybrid Methodology The hybrid methodology, which combines all the three models, is expected to deliver accurate predictions for verification purposes.

- **Stacking Regressor (Hybrid Model):** Hybrid Strategy The above models have constituted the basic framework, but the chosen model for prediction is a linear regression model that results from an application of the stacking ensemble technique. The hybrid approach is proposed to be a way of combining the strengths within each model simultaneously while addressing and mitigating corresponding drawbacks.

VII. **Visualization and Error Analysis: The final step involved visualizing the model results and analyzing prediction errors:**

- **Actual vs Predicted Plot:** A scatter plot was used to compare actual and predicted temperature values. It can be useful for graphically verifying model performance.
- **Error Distribution:** The distribution of prediction errors was plotted to understand the existence of any systematic biases in the predictions.



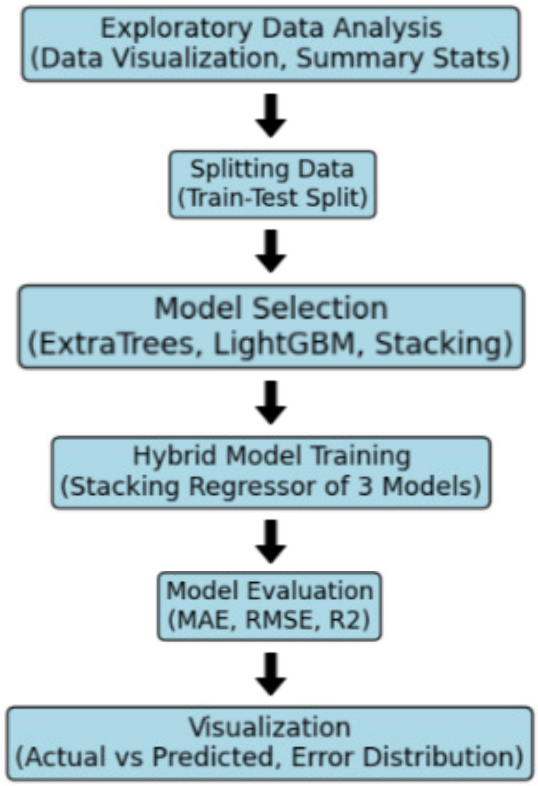


Figure 1: Methodology Flowchart

Results and Discussion:

1. Model Performance: Performance of the machine learning algorithms, in this case Extra Trees Regressor, LightGBM Regressor, and the ensemble known as Stacking Regressor, is evaluated using three main evaluation metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R^2). Results summary is presented in Table 1.

Model	MAE
Extra Trees Regressor (ETR)	0.26
Stacking Regressor (SR)	0.37
LightGBM Regressor (LGBMR)	0.80
Hybrid (ETR, SR, LGBMR)	0.209

Table 1: Model performance metrics

The High R^2 value was obtained from the Hybrid Model; the peak value reached was 0.976, which represents a good correlation between actual underwater temperature and predicted one. In

general, results show that aggregation conducts better than individual regressors at accuracy level.

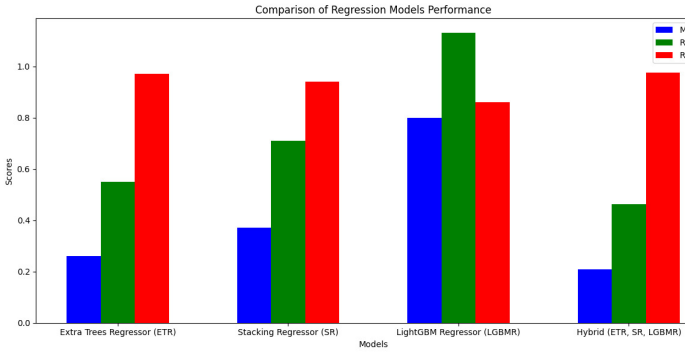


Figure 2: Model performance Comparison

3. Comparative Analysis: The given study is concluded to be consistent with the outcomes of many previous studies that tend to suggest that the machine learning method is a good way for environmental prediction. For example, it has been demonstrated that LSTM models enhance the performance of SST predictions (Choi et al., 2023). Nonetheless, this paper demonstrates that in the particular case of this work, the ensemble strategy but, specifically, the Stacking Regressor surpassed the LSTM approach presumably due to the diversity of the input variable's attributes along with their interrelationships. The precision achieved in this research, with an R^2 of 0.976, is either at par with, or above the results achieved in other studies that use highly complex models to simulate temperature in similar environmental conditions (Xu et al., 2024; Muhamm et al., 2023). This evidence, therefore, suggests that the hybrid model presented here can be one new and promising avenue for future research in environmental forecasting.

4. Implications for Marine Conservation: The results reported here have important implications for those conservation efforts specifically targeted to the marine biome in Santa Catarina and also spillover effects that spread beyond. Improved prediction of submarine temperatures can facilitate improved resource management, enable better predictions of ecological impacts of climate fluctuations, and build initiatives intended to reduce deleterious effects on marine biodiversity. The ability of such models to predict outcomes guides the researcher and policymakers in themarine sector to make judicious decisions that may prove useful in the conservation efforts for endangered species and ecosystems. Therefore, such models may prove to be a tool for further study on the effects of climate change on marine ecosystems.

5. Limitations and Future Work: While promising, there is a need to identify several drawbacks: it conditions highly on historical data, which leads to biases in some respects, especially if past temperature trends do not necessarily depict future changes in line with expected climate variability, and further validation is crucial to confirm relevance over multiple regions or conditions.

Smaller datasets could be integrated into larger ones with more environmental factors like salinity and currents and nutrient concentration for more exacting predictability models. Also, differences in the methodologies of machine learning, especially deep-learning frameworks, might lead to some interesting scientific understanding of the complex interactions involved in this type of marine temperature-prediction behavior.

CONCLUSIONS

In this research work, we created and validated machine learning algorithms that were targeted at precisely predicting underwater temperatures in the region of Santa Catarina, considering environmental and temporal variables. Further validating our results is the fact that ensemble techniques are powerful, as the Stacking Regressor achieved an R^2 value of 0.976, so improved predictability may be obtained by combining many algorithms. These results align well with recent scholarship addressing the appropriateness of machine learning for applications in environmental prediction and depict well the potential of ensemble approaches. Our results have great implications

for the conservation of marine ecosystems and the management of aquaculture resources, as we could provide stakeholders with accurate temperature forecasts that could inform the development of sustainable practices within marine environments. At the same time, however, we do recognize the limitations associated with an exclusive reliance on historical data and suggest further research to verify the applicability of our model in ecologically diversified scenarios. Further research should incorporate a higher range of environmental factors and examine more advanced machine learning techniques in order to more greatly improve prediction accuracy. This research deepens understanding of what the applications of machine learning have been towards the environment and the implication of having more innovative ideas so as to better face the challenges that the global warming aspect brings about into marine ecosystems.

ACKNOWLEDGMENT

The heading of the Acknowledgment section and the References section must not be numbered.

Causal Productions wishes to acknowledge Michael Shell and other contributors for developing and maintaining the IEEE LaTeX style files which have been used in the preparation of this template. To see the list of contributors, please refer to the top of file IEEETran.cls in the IEEE LaTeX distribution.

REFERENCES

1. Kim, M., Yang, H. (Corresponding Author), & Kim, J. (2020). Sea surface temperature and high-water temperature occurrence prediction using a long short-term memory model. https://doi.org/10.3390/rs12213654
2. Xu, Z., Liu, M., Huang, M., Wen, L., & Guo, X. (2024). Application of real-time water temperature prediction system in winter for long-distance water diversion projects. https://doi.org/10.2166/hydro.2024.064
3. Muhamm, R., Ikram, A., Mostafa, R. R., Chen, Z., Parmar, K. S., Kisi, Ö., & Zounemat-Kermani, M. (2023). Water temperature prediction using improved deep learning methods through reptile search algorithm and weighted mean of vectors optimizer. *Journal of Marine Science and Engineering*, 11(2), 259. https://doi.org/10.3390/jmse11020259
4. Choi, H.-M., Kim, M.-K., & Yang, H. (2023). Deep-learning model for sea surface temperature prediction near the Korean Peninsula. https://doi.org/10.1016/j.dsr2.2023.105262
5. Feigl, M., Lebiedzinski, K., Herrnegger, M., & Schulz, K. (2021). Machine-learning methods for stream water temperature prediction. https://doi.org/10.5194/hess-25-2951-2021
6. Rajesh, M., & Rehana, S. (n.d.). Prediction of river water temperature using machine learning algorithms: a tropical river system of India. https://doi.org/10.2166/hydro.2021.121
7. Gupta, S., & Singhal, G. (2016). Weather Prediction Using Normal Equation Method and Linear Regression Techniques. *International Journal of Computer Science and Information Technologies*, 7(3), 1490-1493.
8. Li, X. (2021). Sea surface temperature prediction model based on long and short-term memory neural network. *IOP Conference Series: Earth and Environmental Science*, 658, 012040. doi:10.1088/1755-1315/658/1/012040.

