

# **Comparative Analysis of Analytical Methods for Predicting NDVI Using Multi-Dimensional Environmental Data**

**Abstract ID: 5260**

Yahya Al Bustanji  
Texas A&M University-Kingsville

Hua Li  
Texas A&M University-Kingsville

Jianhong Ren  
Texas A&M University-Kingsville

Tushar Sinha  
Texas A&M University-Kingsville

Jong-Won Choi  
Texas A&M University-Kingsville

Kai Jin  
Texas A&M University-Kingsville

## **Abstract**

In the past years, remote sensing has been used by scientists to estimate vegetation greenness due to advancements that have reduced accessibility and cost constraints. The Normalized Difference Vegetation Index (NDVI), a popular metric derived from satellite imagery's reflectance in the red and near-infrared spectral bands, has been widely used in the estimation of the vegetation greenness. However, the accuracy of NDVI can be affected by various environmental factors, including wind speed, wind direction, precipitation, humidity, sea level pressure, and cloud cover. To address these influences, analytical techniques are essential for predicting NDVI based on multi-dimensional environmental data, which enhances forecast precision and provides a deeper understanding of vegetation health. The objective of this study is to compare the accuracy in predicting NDVI using various approaches with multi-dimensional data, including multiple linear regression, support vector regression, random forest, and long short-term memory. A dataset spanning eight years and seven months (January 2016 to July 2024) of NDVI satellite data with high spatial resolution was used. This research provides valuable insights into NDVI estimation, with findings revealing that long short-term memory models incorporating time-lag analysis on NDVI data significantly outperform traditional regression methods. The use of time-lag, particularly a 1-month delay in NDVI data, proved critical in capturing temporal dependencies and long-term patterns in greenness of areas. These insights offer valuable guidance for researchers and practitioners in coastal ecosystem management, emphasizing the role of time-lag in improving decision-making and enabling more effective conservation strategies.

## **Keywords**

Machine Learning, Normalized Difference Vegetation Index (NDVI), Comparative Analytical Methods.

## **1. Introduction**

The Normalized Difference Vegetation Index (NDVI) is an essential metric in remote sensing, it provides critical insights into vegetation cover changes over time and across different regions, influenced by surrounding environmental factors [1]. NDVI has broad applications in fields such as coastal studies, forestry, ecology, and climate science. Accurate NDVI predictions are vital for informed decision-making in these areas. NDVI accuracy is

influenced by a range of environmental factors, and these factors introduce complexities in modeling and forecasting NDVI, requiring the development of robust analytical methods. Predicting NDVI using climate variables is particularly challenging due to the dynamic interactions between environmental factors and vegetation health [2]. While various analytical methods, such as Multiple Linear Regression (MLR), have been used in previous studies, no single approach has proven universally accurate across different climatic conditions.

By leveraging a high-resolution dataset sourced from the National Solar Radiation Database (NSRDB), spanning eight years and seven months (January 2016 to July 2024), with daily data collected and monthly averages calculated for each climatic factor, the project analyzes monthly NDVI and climate variable recordings totaling 103 data points, this project studies the prediction of NDVI values at the Padre Island which is the biggest barrier island not just in Texas but also globally. The Island is situated in close proximity to the southern coastline of Texas. It is geographically isolated from the mainland due to the presence of the Laguna Madre water body [3]. Between the months of July and October 2020, South Padre Island witnessed storm surges resulting from the impact of two hurricanes, namely Hurricane Hanna and Hurricane Delta, as well as a tropical storm known as Tropical Storm Beta [4]. This research contributes to the body of knowledge on machine learning usage in vegetation health monitoring. The findings provide practical insights for researchers and practitioners involved in machine learning modeling, vegetation health management, and climate-driven land use planning.

## 2. Research Methods

Previous studies have demonstrated NDVI utility in tracking environmental changes and monitoring deforestation. However, NDVI values can be affected by seasonal variations, atmospheric conditions, and environmental factors, making accurate prediction a complex challenge. According to Hitzfelder [5], the range of output values for NDVI spans from -1 to 1. Specifically, values closer to -1 are associated with clouds, water, or snow, while values closer to 0 indicate barren soils. Conversely, values closer to 1 indicate the presence of healthy vegetation. NDVI was calculated using the following formula:

$$NDVI = \frac{\rho_{NIR} - \rho_{red}}{\rho_{NIR} + \rho_{red}} \quad (1)$$

In this study, four analytical methods - Multiple Linear Regression (MLR), Random Forest (RF), Support Vector Regression (SVR), and Long Short-Term Memory (LSTM) - were implemented and compared for predicting NDVI values on Padre Island. Environmental factors such as temperature, wind speed, humidity, sea level pressure, cloud cover, and precipitation were collected from the National Solar Radiation Database (NSRDB) at a 4 km x 4 km resolution, with daily data averaged monthly to align with NDVI measurements. These variables were selected based on their known influence on vegetation health.

MLR is a fundamental statistical technique used to model relationships between multiple independent variables and a dependent variable. In NDVI prediction, MLR has been employed to quantify the effects of climatic factors like temperature, humidity, and precipitation. While MLR offers interpretability, its linear assumptions often limit its predictive accuracy, especially in capturing complex, nonlinear relationships between NDVI and environmental variables. RF, an ensemble machine learning technique, constructs multiple decision trees to enhance predictive accuracy and reduce overfitting, capturing nonlinear interactions among environmental factors and outperforming traditional regression methods in prior NDVI studies [6]. However, RF demands significant computational resources and may lack interpretability. SVR, based on Support Vector Machines (SVM), is designed to handle nonlinear relationships and manage high-dimensional input spaces, proving effective in NDVI modeling, though its hyperparameter tuning is complex and generalization across conditions can vary. Lastly, LSTM is a specialized recurrent neural network (RNN) that excels at handling sequential data with long-term dependencies [7], leveraging historical and time-lagged NDVI and climatic data to capture temporal patterns, often surpassing conventional regression methods in time-series forecasting for environmental applications.

An in-depth exploration of the collected data was conducted to identify patterns, trends, and outliers, alongside an analysis of temporal variations in NDVI values to understand vegetation dynamics in coastal regions. Statistical and visual tools were utilized, with MLR providing insights into relationships between NDVI and environmental variables.

The predictive performance of each model was assessed using multiple metrics - Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-Squared ( $R^2$ ) - computed across training (70% of data, 72 points), testing (15%, 15 points), and validation (15%, 16 points) subsets, split chronologically to preserve temporal order. For reproducibility, RF used 100 trees with a maximum depth of 10, SVR employed a radial basis function kernel with  $C=1.0$  and  $\epsilon=0.1$ , and LSTM was configured with 50 units, a 0.2 dropout rate, and trained for 100 epochs with a batch size of 32, optimized using the Adam algorithm. Time-lagged NDVI data, incorporating values from previous months (e.g., 1-month lag uses NDVI from  $t-1$  to predict  $t$ ), was included to capture temporal dependencies, accounting for delayed vegetation responses to environmental changes, such as precipitation impacting growth weeks later, and tested with lags of 1, 2, and 3 months to determine the optimal delay.

### 3. Results

To evaluate the influence of environmental factors and time-lagged NDVI data on model performance, we conducted regression analysis under various conditions. The analysis compared the effects of excluding one environmental factor at a time while assessing the impact of different time-lagged NDVI inputs on prediction accuracy. Table 1 presents a summary of regression analysis results, showing the impact of time lag, NDVI data type, and environmental factors on model performance. Here, Data Type refers to the NDVI input used: NDVI indicates unchanged NDVI values as directly obtained from the dataset, while Average NDVI represents the average of the current month's NDVI and the previous month's NDVI. The Key Factors column denotes the climatic variables included in each analysis, where All Factors encompasses all available climatic factors (temperature, wind speed, humidity, sea level pressure, cloud cover, precipitation, etc.), and exclusions like Excludes Cloud Type mean all factors are retained except the specified one (e.g., cloud type), with similar logic applied to exclusions such as wind direction, humidity, or precipitation.

Table 1: Summary of regression analysis results using MLR

| Time-Lag (months) | Data Type    | Key Factors             | $R^2$ Value |
|-------------------|--------------|-------------------------|-------------|
| No Time-Lag       | NDVI         | All Factors             | 0.468       |
| 1 month           | NDVI         | All Factors             | 0.572       |
| 2 Months          | Average NDVI | All Factors             | 0.705       |
| 3 Months          | Average NDVI | All Factors             | 0.591       |
| 2 Months          | Average NDVI | Excludes Cloud Type     | 0.692       |
| 2 Months          | Average NDVI | Excludes Wind Direction | 0.686       |
| 2 Months          | Average NDVI | Excludes Humidity       | 0.659       |
| 2 Months          | Average NDVI | Excludes Precipitation  | 0.649       |

The results indicate that incorporating a time lag improves NDVI prediction accuracy, with the highest  $R^2$  value (0.705) achieved using a 2-month time-lagged Average NDVI dataset including all factors. When key environmental factors were excluded one at a time from this configuration,  $R^2$  values declined, confirming their collective importance in the model. Specifically, excluding cloud type and wind direction resulted in a slight drop in accuracy ( $R^2 = 0.692$  and  $0.686$ , respectively), suggesting these factors have a modest influence, whereas removing humidity and precipitation led to more substantial decreases ( $R^2 = 0.659$  and  $0.649$ , respectively), highlighting their critical roles. Additionally, the best  $R^2$  value among 1-month lag trials was observed when using unchanged NDVI data with all environmental factors ( $R^2 = 0.572$ ). This suggests that a shorter time lag benefits from retaining localized NDVI variations without averaging, while longer time lags, such as 2 months, require averaging across months to enhance predictive stability.

To evaluate the predictive capabilities of different analytical methods for NDVI estimation, we assessed the performance of LSTM, RF, and SVR using multiple metrics, including  $R^2$ , MAE, and RMSE. The results were obtained across training, testing, and validation phases and summarized in Table 2. The findings in Table 2 reveal that the LSTM model consistently outperformed RF and SVR in NDVI prediction, particularly during validation. LSTM achieved the highest  $R^2$  score of 0.670 during validation, indicating its superior ability to capture temporal

dependencies in NDVI fluctuations. In contrast, RF exhibited overfitting characteristics, as evidenced by its high training  $R^2$  (0.876) but lower validation  $R^2$  (0.498). Meanwhile, SVR demonstrated the weakest predictive power, with the lowest validation  $R^2$  (0.201), suggesting difficulties in handling the complex relationships within the dataset. Figure 1 illustrates the  $R^2$  values for LSTM, RF, and SVR models across training, testing, and validation phases.

One possible explanation for the observed differences in model performance is the relatively small dataset used in this study. With 103 data points available, the models had limited opportunities to learn robust patterns, particularly those requiring complex feature extraction and generalization. Deep learning models such as LSTM typically require large amounts of data to perform optimally; however, despite this limitation, LSTM still outperformed the other models, likely due to its inherent ability to learn sequential patterns and capture temporal dependencies in NDVI trends. The ability of LSTM to recognize sequential relationships in NDVI and environmental variables, especially when using time-lagged inputs, likely contributed to its relatively strong validation performance.

Table 2: Performance metrics comparison of machine learning models

| Model | Metric | Training | Testing | Validation |
|-------|--------|----------|---------|------------|
| LSTM  | $R^2$  | 0.878    | 0.544   | 0.670      |
| RF    | $R^2$  | 0.876    | 0.629   | 0.498      |
| SVR   | $R^2$  | 0.421    | 0.498   | 0.201      |
| LSTM  | MAE    | 0.056    | 0.120   | 0.080      |
| RF    | MAE    | 0.018    | 0.041   | 0.056      |
| SVR   | MAE    | 0.037    | 0.042   | 0.050      |
| LSTM  | RMSE   | 0.071    | 0.178   | 0.128      |
| RF    | RMSE   | 0.024    | 0.052   | 0.063      |
| SVR   | RMSE   | 0.055    | 0.050   | 0.067      |

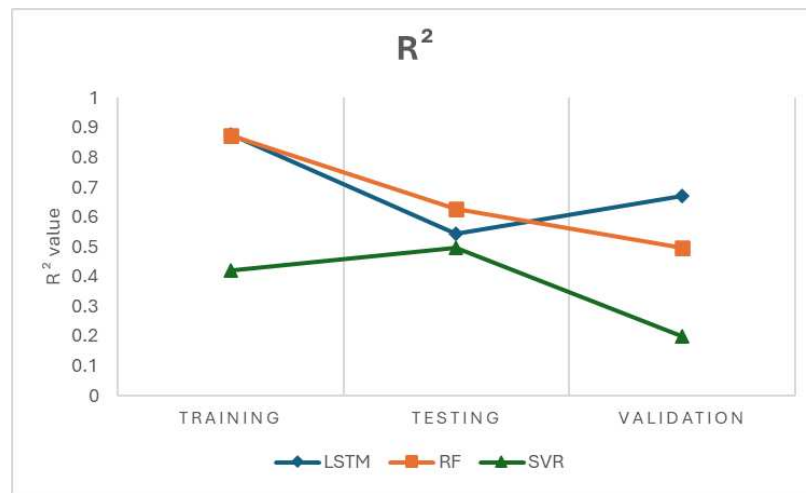


Figure 1:  $R^2$  Comparison Across Training, Testing, and Validation Phases

In contrast, RF, while demonstrating strong performance in training, struggled during validation, suggesting that it may have learned noise within the dataset rather than generalizable patterns. The ensemble nature of RF makes it highly effective in complex pattern recognition, but with a small dataset, it becomes more susceptible to overfitting. The relatively high difference between RF's training and validation performance further indicates that it memorized training data rather than generalizing well to unseen data. SVR, on the other hand, exhibited the lowest  $R^2$  scores

across all phases, highlighting its struggles with high-dimensional and nonlinearly correlated datasets. The difficulty in selecting an optimal kernel function and tuning hyperparameters could be one of the reasons for SVR's subpar performance. Furthermore, SVR generally requires a large amount of data to define clear support vectors, and in this study, the limited dataset likely restricted its effectiveness, making it less suitable for capturing the intricate relationships between NDVI and environmental factors.

Figure 2 presents the MAE values for LSTM, RF, and SVR models across different phases. The LSTM model exhibits low MAE during training but experiences a significant increase in error during testing, highlighting possible overfitting. In contrast, the RF model maintains a relatively low MAE throughout all phases, indicating a more stable generalization. The SVR model shows a moderate increase in MAE over testing and validation phases, maintaining a more balanced performance. These results suggest that while LSTM achieves better fitting on training data, RF provides more reliable generalization across unseen data.

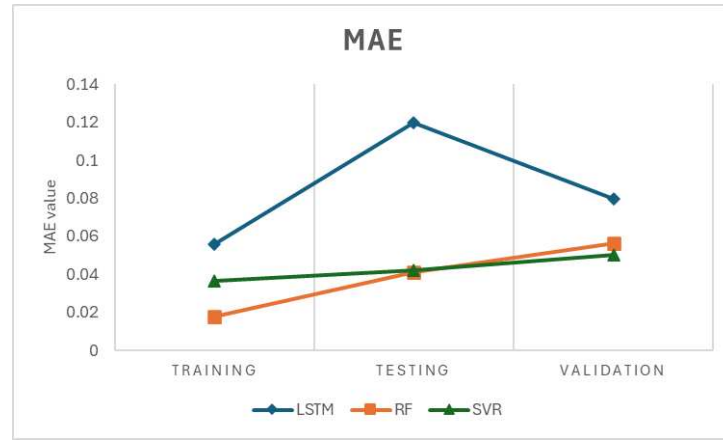


Figure 2: MAE Comparison Across Training, Testing, and Validation Phases

The RMSE results in Figure 3 indicate that the RF model achieved the lowest error across all data splits, with values of 0.024 for training, 0.052 for testing, and 0.063 for validation. The SVR model had slightly higher RMSE values of 0.055 for training, 0.050 for testing, and 0.067 for validation, showing competitive performance in testing but slightly weaker generalization. The LSTM model exhibited the highest RMSE, particularly in the testing phase (0.178), suggesting that it struggled to generalize compared to RF and SVR.

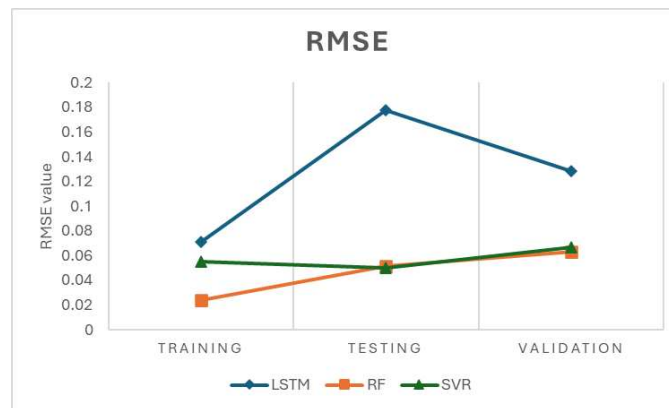


Figure 3: RMSE Comparison Across Training, Testing, and Validation Phases

The comparative analysis highlights the advantage of LSTM in leveraging sequential data for NDVI forecasting. While RF demonstrated competitive performance, its tendency to overfit suggests that further regularization

techniques could improve its generalization ability. The poor performance of SVR indicates that it may not be well-suited for highly dynamic and nonlinear NDVI prediction tasks.

Overall, these results suggest that deep learning methods, particularly LSTM, hold significant potential for improving the accuracy of NDVI forecasts. However, to maximize their effectiveness, larger datasets and enhanced feature engineering are essential. Future research should focus on integrating additional environmental factors, increasing temporal resolution, and applying hybrid models that combine deep learning and ensemble techniques to mitigate the limitations observed in this study.

#### 4. Conclusions

This study conducted a comparative analysis of various machine learning techniques for NDVI prediction using multi-dimensional environmental data. The results demonstrated that LSTM-based models incorporating time-lagged NDVI data significantly outperformed traditional regression methods, particularly in capturing long-term temporal dependencies. A 2-month time lag was found to be the most effective in improving model accuracy, as evidenced by the highest  $R^2$  value obtained when considering both machine learning techniques and regression analysis.

Among the models tested, LSTM exhibited superior predictive capability across training, testing, and validation phases, achieving the highest validation  $R^2$  of 0.670. In contrast, RF showed strong training performance but suffered from overfitting, while SVR struggled to generalize effectively. The findings underscore the necessity of using deep learning models for NDVI forecasting, as traditional machine learning methods failed to capture the complexity of NDVI-environment interactions.

Despite the promising results, limitations remain, including a relatively small dataset and the exclusion of additional remote sensing features such as soil moisture. Future work should focus on expanding the dataset, incorporating more environmental variables, and leveraging hybrid models that integrate deep learning with ensemble learning techniques.

This study's comparative analysis demonstrates that incorporating time-lagged NDVI data, particularly with LSTM, enhances the ability to model vegetation dynamics under varying environmental conditions. These findings offer targeted guidance for optimizing NDVI predictions in coastal ecosystem monitoring, supporting improved land-use planning and conservation strategies by capturing temporal influences on vegetation health and climate resilience.

#### Acknowledgements

The authors are thankful for the support from Texas General Land Office (TGLO) Coastal Management Program (Contract No. 22-045-008-D105) and the U.S. National Science Foundation (award 2244523). Any opinions, findings, or recommendations expressed were created by the authors and not reviewed by nor necessarily reflect the view of the TGLO and NSF.

#### References

- [1] Y. Everingham, J. Sexton, D. Skocaj, and G. Inman-Bamber, "Accurate prediction of sugarcane yield using a random forest algorithm," *Agronomy for sustainable development*, vol. 36, pp. 1–9, 2016.
- [2] M. Abdipour, M. Younessi-Hmazekhanlu, S. H. R. Ramazani, and A. Hassan Omid, "Artificial neural networks and multiple linear regression as potential methods for modeling seed yield of safflower (*Carthamus tinctorius* L.)," *Industrial Crops and Products*, 2019.
- [3] E. J. Farrell, I. Delgado-Fernandez, T. Smyth, B. Li, and C. Swann, "Contemporary research in coastal dunes and aeolian processes," *Earth Surface Processes and Landforms*, 2023, doi: 10.1002/esp.5597.
- [4] D. P. Brown, R. Berg, and B. Reinhart, *National Hurricane Center Tropical Cyclone Report: Hurricane Hanna (AL082020)*. National Oceanic and Atmospheric Administration (NOAA), 2020.
- [5] E. Hitzfelder, *Utilizing Normalized Difference Vegetation Index to Assess the Activity of the Maverick Badlands, Big Bend National Park, Texas: 1999–2018*, 2019.
- [6] J. H. Jeong *et al.*, "Random forests for global and regional crop yield predictions," *PLoS ONE*, vol. 11, no. 6, 2016.
- [7] N. Minallah and W. Khan, "Comparison of neural networks and support vector machines for the mass balance ablation observation of glaciers in Baltoro region," *J. Inf. Commun. Technol. Robot. Appl.*, pp. 37–45, 2019.