Daphne Tuncer
Final Project
AOS 204

Introduction

Coastal erosion presents a significant threat to communities, infrastructure, and ecosystems along the coastline. Accurately predicting erosion patterns is essential for mitigating these risks and planning effective coastal defenses. This project aims to leverage machine learning to develop a model capable of forecasting coastal erosion, using wave period, tidal data, wind speed, and sediment transport.

The data for this study is sourced from La Jolla station (Station ID: 9410230), located in La Jolla, California, which is affiliated with Scripps Institution of Oceanography at UC San Diego. This station provides a robust and reliable dataset, making it ideal for this type of analysis. Coastal erosion has been a persistent issue in La Jolla and along the broader California coastline, exacerbated by rising sea levels driven by climate change. As sea levels continue to rise, the need for accurate predictive models becomes increasingly urgent.

Currently, few cohesive coastal erosion models are publicly available. The project seeks to address this gap by estimating wave heights through machine learning, utilizing historical measurements of water levels, wave heights, wind speed, dominant wave period (DPD), and average wave period (APD). Additionally, the project incorporates time shifting techniques to assess the model's ability to predict future wave heights, with the ultimate goal of enhancing erosion forecasting capabilities over time.

Data

The input features selected for this project include wave height, wind speed, water level, dominant wave period, and average wave period. The water level data for this project was obtained from the La Jolla Station, maintained by NOAA. Data for wind speed, dominant wave period (DPD), and average wave period (APD) were sourced from the National Data Buoy Center (NDBC). All datasets cover the period from January 2023 to December 2023. Quantifying coastal erosion posed a significant challenge, as readily available erosion models for comparison were scarce. A review of existing models revealed that most coastal erosion models developed by agencies such as the USGS are accessible only in GIS or shapefile formats, limiting their usability for direct comparison in this study.

Consequently, machine learning was employed to predict wave heights, which play a crucial role in coastal erosion dynamics. As noted by the National Park Service, "Waves are the dominant force driving the nature of a beach. The energy carried through waves moves beach sediment and transforms beach shape. The more energy, the greater the extent of change" (Coastal Processes – Waves).

Figure 1 below illustrates the dataset collected from the La Jolla Buoy and La Jolla station. The most frequent water levels range between 2 and 4 feet, while wave heights predominantly fall between 1 and 3 feet. Wind speeds commonly range from 0 to 10 feet per second. The dominant wave period exhibits a well-distributed pattern, peaking at 12 seconds, while the average wave period shows a significant concentration around 100. Water level data represents the ocean's surface height relative to the Mean Lower Low Water (MLLW) datum. High wind speeds are closely linked to high-energy waves, which in turn, significantly influence the extent of coastal erosion.

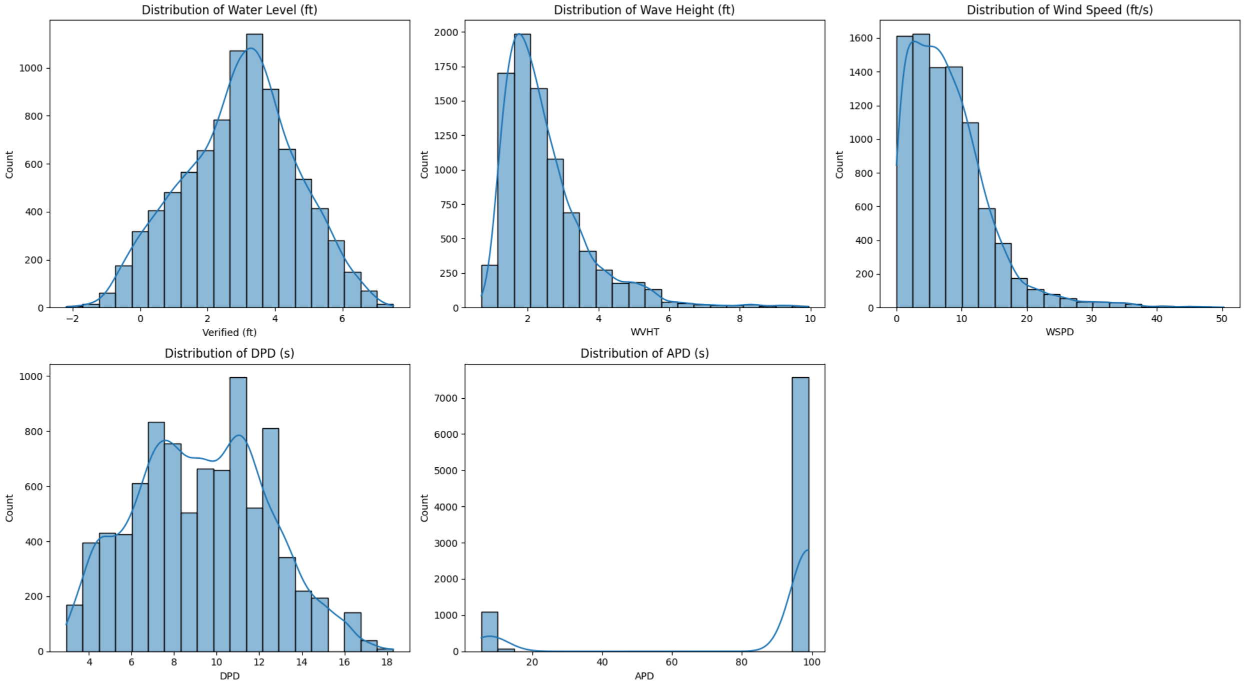


Figure 1: Data available from La Jolla Bouy and La Jolla Station

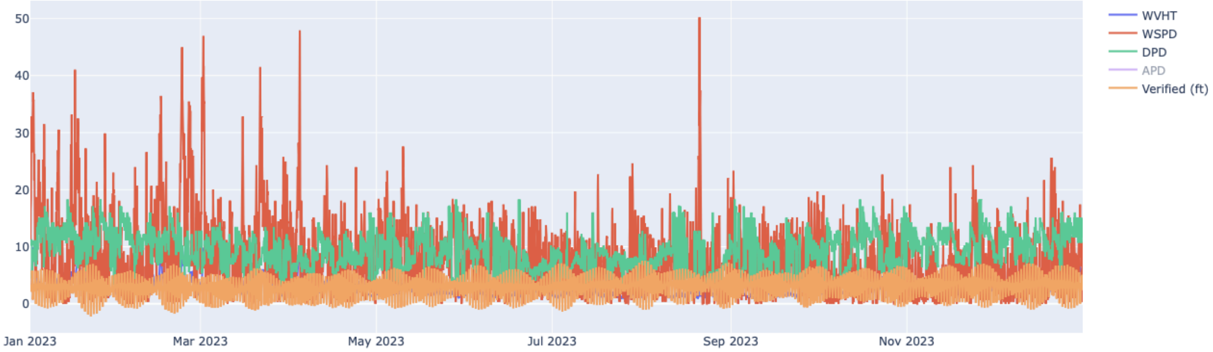


Figure 2: Combined graph of wave height, windspeed, verified water levels, and dominant wave period

Modeling

For this project I compared the random forest regressor, the linear regression, and the gradient boosting regressor as the predictive models. The random forest regressor is a good model for this project due to its ability to effectively handle non-linear relationships. One key advantage of random forest is its minimal requirement for data preprocessing, which streamlines the workflow. The linear regression model serves as a solid baseline for comparison, as it assumes a linear relationship between the features, even though the actual relationships may not be linear. On the other hand, the gradient boosting model excels at capturing non-linear patterns in the data and is robust at handling missing values effectively.

Before model training, some preprocessing was necessary. I worked with two separate datasets that recorded data hourly from January 2023 to December 2023. These datasets were merged based on their timestamps to align the features accurately. For the random forest regressor . To evaluate each model's performance, I utilized the Mean Squared Error (MSE) and R^2 score metrics from the sklearn.metrics module. I used a forest consisting of 200 trees for the random forest model and the gradient boosting regressor has a n estimators of 200.

Results

Random Forest Model

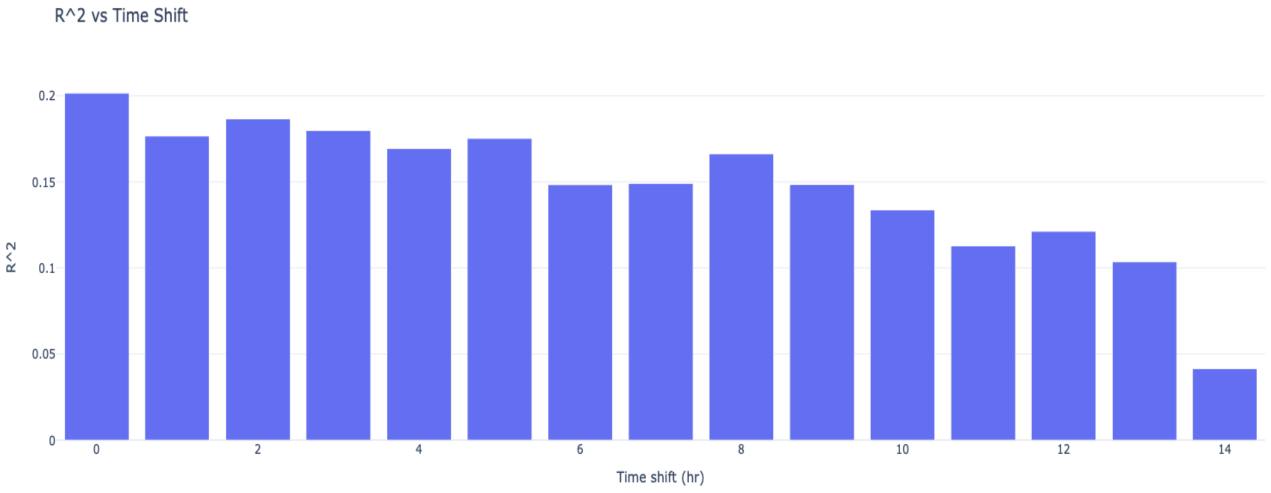
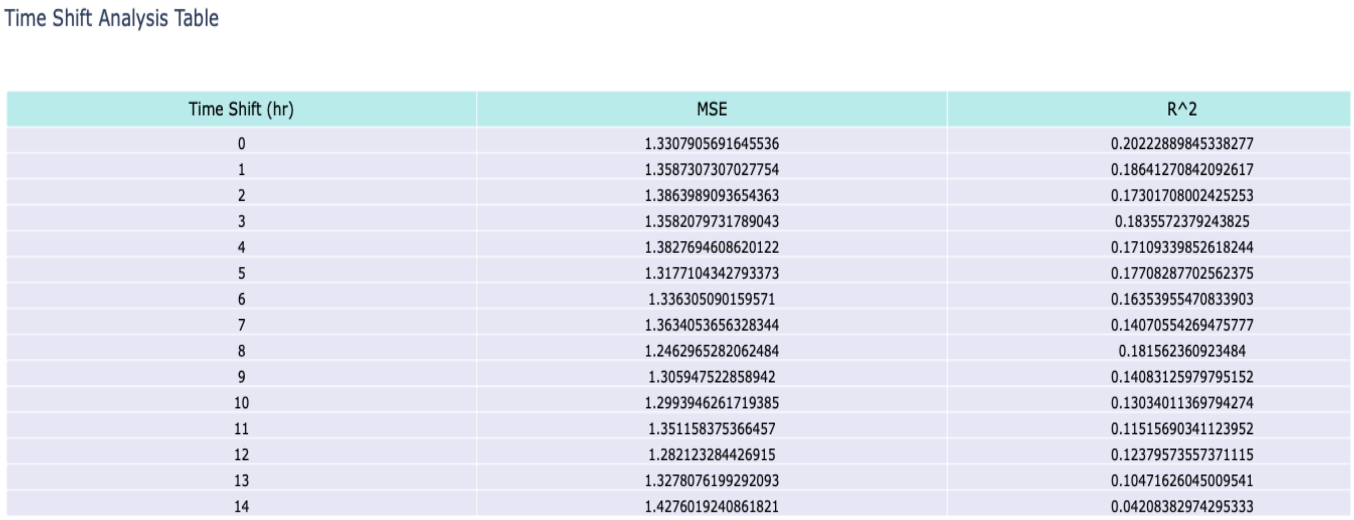


Figure 3: Graph of time shift vs R^2 value for the random forest regressor



Time Shift Analysis Table

Time Shift (hr)	MSE	R^2
0	1.3307905691645536	0.20222889845338277
1	1.3587307307027754	0.18641270842092617
2	1.3863989093654363	0.17301708002425253
3	1.3582079731789043	0.1835572379243825
4	1.3827694608620122	0.17109339852618244
5	1.3177104342793373	0.17708287702562375
6	1.336305090159571	0.16353955470833903
7	1.3634053656328344	0.14070554269475777
8	1.2462965282062484	0.181562360923484
9	1.305947522858942	0.14083125979795152
10	1.2993946261719385	0.13034011369794274
11	1.351158375366457	0.11515690341123952
12	1.282123284426915	0.12379573557371115
13	1.3278076199292093	0.10471626045009541
14	1.4276019240861821	0.04208382974295333

Figure 4: Table of time shift vs R^2 value for the random forest regressor

Linear Regression

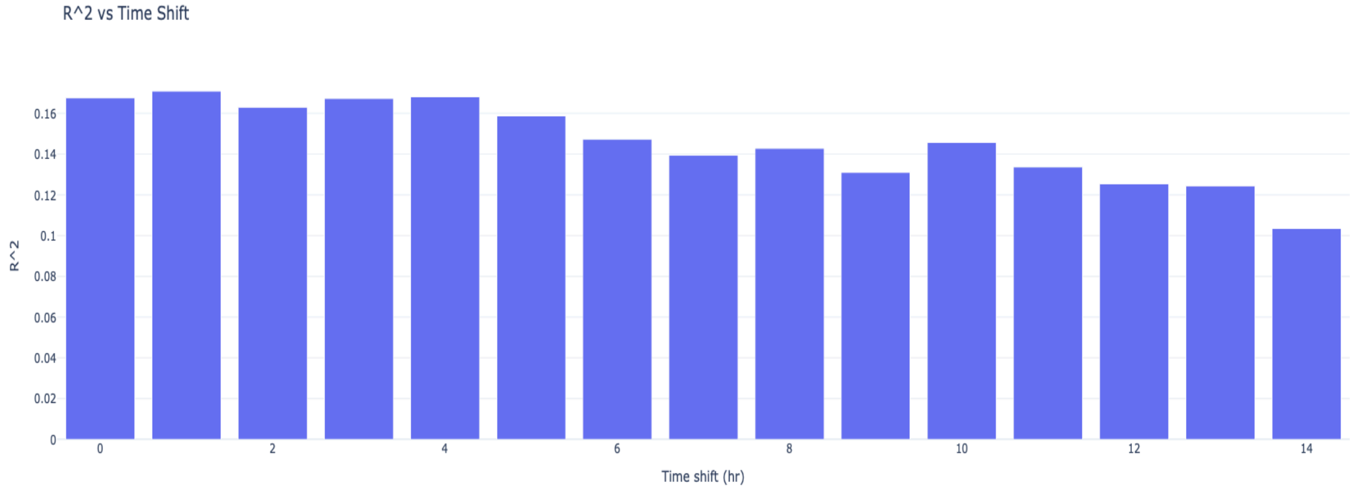
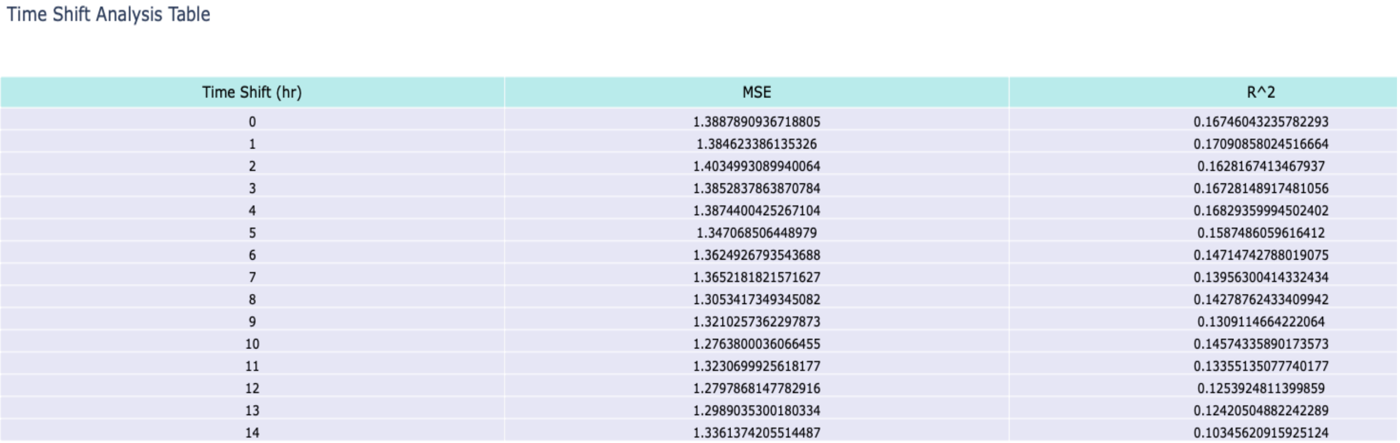


Figure 5: Graph of time shift vs R^2 value for the linear regression



Time Shift Analysis Table

Time Shift (hr)	MSE	R^2
0	1.3887890936718805	0.16746043235782293
1	1.384623386135326	0.17090858024516664
2	1.4034993089940064	0.1628167413467937
3	1.3852837863870784	0.16728148917481056
4	1.3874400425267104	0.16829359994502402
5	1.347068506448979	0.1587486059616412
6	1.3624926793543688	0.14714742788019075
7	1.3652181821571627	0.13956300414332434
8	1.3053417349345082	0.14278762433409942
9	1.3210257362297873	0.1309114664222064
10	1.2763800036066455	0.14574335890173573
11	1.3230699925618177	0.13355135077740177
12	1.2797868147782916	0.1253924811399859
13	1.2989035300180334	0.12420504882242289
14	1.3361374205514487	0.10345620915925124

Figure 6: Table of time shift vs R^2 value for the linear regression

Gradient Regressor

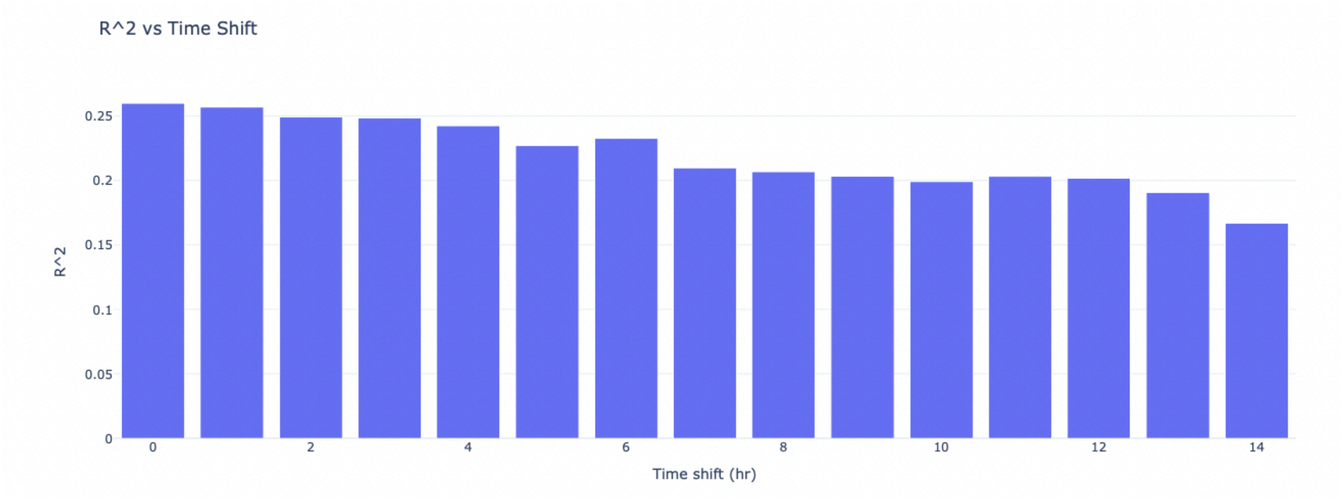
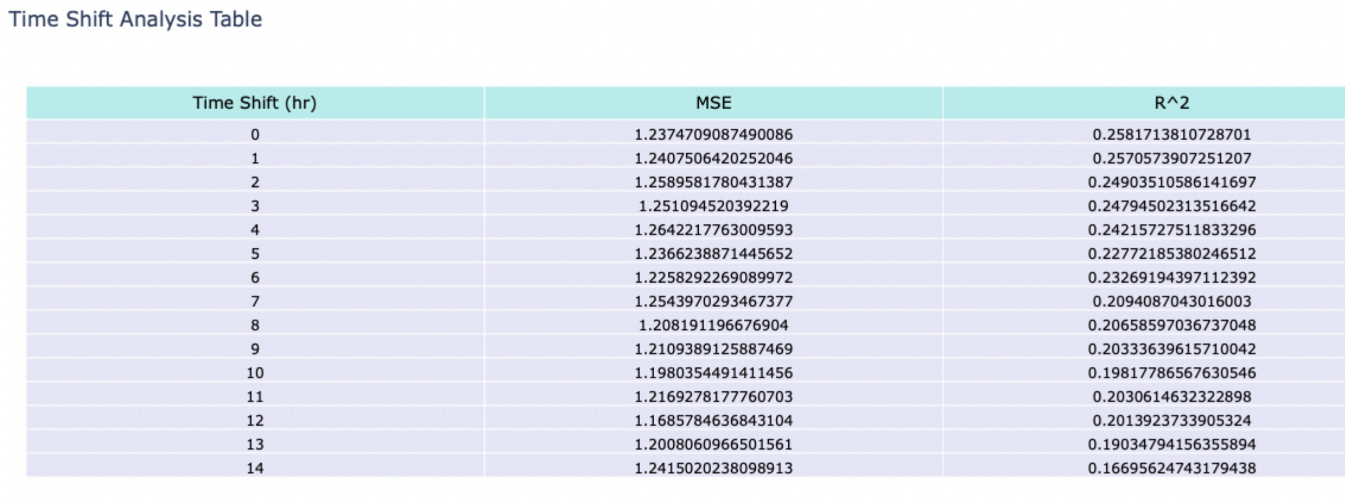


Figure 7: Graph of time shift vs R^2 value for the gradient regressor

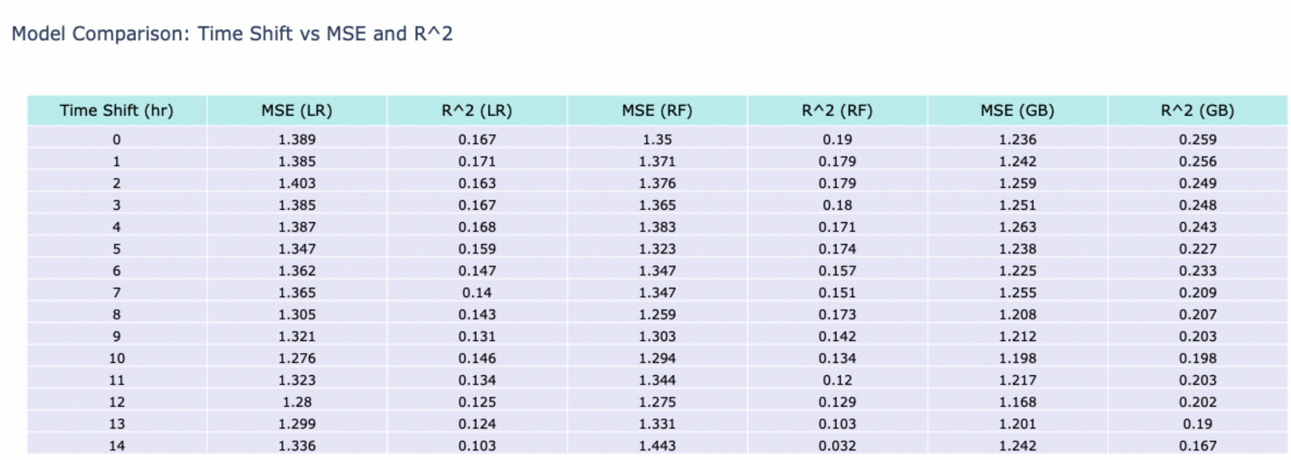


Time Shift Analysis Table

Time Shift (hr)	MSE	R^2
0	1.2374709087490086	0.2581713810728701
1	1.2407506420252046	0.2570573907251207
2	1.2589581780431387	0.24903510586141697
3	1.251094520392219	0.24794502313516642
4	1.2642217763009593	0.24215727511833296
5	1.2366238871445652	0.22772185380246512
6	1.2258292269089972	0.23269194397112392
7	1.2543970293467377	0.2094087043016003
8	1.208191196676904	0.20658597036737048
9	1.2109389125887469	0.20333639615710042
10	1.1980354491411456	0.19817786567630546
11	1.2169278177760703	0.2030614632322898
12	1.1685784636843104	0.2013923733905324
13	1.2008060966501561	0.19034794156355894
14	1.2415020238098913	0.16695624743179438

Figure 8: Table of time shift vs R^2 value for the gradient regressor

Discussion



Model Comparison: Time Shift vs MSE and R^2

Time Shift (hr)	MSE (LR)	R^2 (LR)	MSE (RF)	R^2 (RF)	MSE (GB)	R^2 (GB)
0	1.389	0.167	1.35	0.19	1.236	0.259
1	1.385	0.171	1.371	0.179	1.242	0.256
2	1.403	0.163	1.376	0.179	1.259	0.249
3	1.385	0.167	1.365	0.18	1.251	0.248
4	1.387	0.168	1.383	0.171	1.263	0.243
5	1.347	0.159	1.323	0.174	1.238	0.227
6	1.362	0.147	1.347	0.157	1.225	0.233
7	1.365	0.14	1.347	0.151	1.255	0.209
8	1.305	0.143	1.259	0.173	1.208	0.207
9	1.321	0.131	1.303	0.142	1.212	0.203
10	1.276	0.146	1.294	0.134	1.198	0.198
11	1.323	0.134	1.344	0.12	1.217	0.203
12	1.28	0.125	1.275	0.129	1.168	0.202
13	1.299	0.124	1.331	0.103	1.201	0.19
14	1.336	0.103	1.443	0.032	1.242	0.167

Model	Combined R^2
Linear Regression	2.188
Random Forest	2.236
Gradient Boosting	3.259

When comparing the combined R² values, gradient boosting emerges as the top-performing model, yielding the highest R² and thus providing the best predictions for time-shifted wave heights. As shown in Figure 7, the R² value for gradient boosting decreases with increasing time shift, indicating that the model's accuracy diminishes as the time shift grows. The random forest model, on the other hand, does not offer as accurate predictions as gradient boosting. Additionally, the R² value at the 14-hour time shift, shown in Figure 3, appears to be an outlier when compared to the other values. Across various time shifts, the R² for random forest fluctuates, leading to some uncertainty about its overall accuracy. The linear regression model, while following a similar trend to gradient boosting (with R² decreasing over time), consistently provides a much lower fit, suggesting it is less effective. It would be interesting to explore whether gradient boosting can maintain its strong performance as the time shift extends to weeks, months, or even years. In reality, the R² values for all the models are relatively low, suggesting that the problem may be too complex or that additional data is required to make more accurate predictions of wave heights.

Conclusion

This report detailed the development and training of various machine learning models to estimate wave heights at La Jolla, California. Among these, the gradient boosting model emerged as the most effective model for predicting wave heights using water levels, wind speed, dominant wave period (DPD), and average wave period (APD). The model demonstrated fine performance in predicting wave heights at multiple future time intervals, including 0 through 14 hours. The ability to predict wave heights accurately is crucial for understanding and anticipating coastal erosion patterns. By providing reliable short-term forecasts, this model contributes valuable data for coastal management and erosion mitigation efforts. It enables planners to better assess erosion risks, which are vital for protecting infrastructure, ecosystems, and communities along the coastline.

Expanding the time horizons of wave height predictions to weeks, months, or even years would offer profound insights into long-term coastal erosion trends. This would be especially beneficial for regions like La Jolla, where rising sea levels and dynamic wave activity continuously reshape the coastline. This work highlights the potential of machine learning in advancing coastal management practices. Continued efforts to refine predictive models and incorporate additional environmental factors, such as sediment transport and sea-level rise, will further enhance their utility in addressing the challenges posed by coastal erosion.

References

U.S. Geological Survey. (n.d.). *Shoreline change projections for California: CoSMoS*. Retrieved from https://www.sciencebase.gov/catalog/item/57f1d519e4b0bc0bebfee13d.

Wang, H., Zhang, D., & Xu, Z. (2023). Investigating global coastal wind dynamics. *Scientific Reports, 13*, 38729. https://doi.org/10.1038/s41598-023-38729-y.

Li, X., & Cheng, Y. (2024). Wind erosion modeling and climate analysis. *Scientific Reports, 14*, 74714. https://doi.org/10.1038/s41598-024-74714-9.

Walkden, M., & Hall, J. (2024). Modeling erosion on rocky coastlines. *Geoscientific Model Development, 17*, 3433–3447. https://doi.org/10.5194/gmd-17-3433-2024.

National Park Service. (n.d.). *Coastal Processes: Waves*. Retrieved from https://www.nps.gov/articles/coastal-processes-waves.htm.

National Data Buoy Center. (n.d.). *La Jolla Station (LJPC1) station page*. National Oceanic and Atmospheric Administration. Retrieved from https://www.ndbc.noaa.gov/station_page.php?station=ljpc1

NOAA Tides & Currents. (n.d.). *Station 9410230 - La Jolla, CA*. National Oceanic and Atmospheric Administration. Retrieved from https://tidesandcurrents.noaa.gov/stationhome.html?id=9410230