Predicting Significant Wave Height Using Embedded Sensors on Surfboards

Purisa Jasmine Simmons, Jennifer Chien, Adrian Salguero, Martha Gahl
University of California San Diego

December 3, 2019

0.1 Introduction

Oceanographers are interested in obtaining accurate predictions of sea state conditions; however, access to accurate ocean wave height observations are costly to obtain in part due to a lack of robust autonomous ocean sensors [1]. Therefore, our project focuses on the prediction of significant wave height from the Smartfin's Inertial Measurement Unit (IMU) data using machine learning techniques. The Smartfin is a surfboard fin with an embedded IMU, which collects data related to linear acceleration, angular velocity, and compass heading. Published literature in oceanography fields indicate that it may be possible to predict wave characteristics such as significant wave height, wave period, and wave direction from the data collected by IMU sensors [7]. The inputs to our machine learning model will consist of information that is feature engineered from the Smartfin IMU data, and the outputs will be the "ground truth" labels of significant wave height, which are provided by the nearest CDIP buoy that collects these values.

0.2 Literature Review

Numerical and spectral analysis techniques have long been used in ocean sciences fields on IMU data to determine sea state characteristics such as significant wave height, wave period, and wave direction [2], [3], [4]. In particular, these techniques focus on directly calculating sea state characteristics from stationary buoys floating on the ocean surface. The main issue with these types of analyses are that errors (accelerometer bias and gyroscope misalignments) can easily compound and propagate through these systems over longer period of time, such as when data is collected over many hours [5]. Additionally, buoys and other autonomous sensors are often placed in offshore regions because of the lack of high energy wave dynamics, which make it particularly difficult for oceanographers to collect data in nearshore regions. The primary differences between our work and previous work are that we: (a) use IMU data from non-stationary sensors placed on surfers in order to collect wave data in nearshore regions, and (b) use machine learning techniques rather than numerical techniques to determine ocean characteristics in order to minimize error propagation.

The scope of this project aligns more closely with recent oceanographic research, which has started to use machine learning techniques in order to predict sea state characteristics from input features, most commonly using wind as an input feature to predict significant wave height [6]. We found only one instance of a published research paper using machine learning techniques directly on IMU data to determine sea state characteristics; in it, an IMU is placed on a sea vessel to determine significant wave height, wave period, and wave direction using machine learning on statistical properties from the IMU such as, "the standard deviation of acceleration and the rate of rotation around the primary axis" as input features

[7]. This paper serves as the main source of motivation for our work and has led us to hypothesize that feature engineering statistical properties of the IMU data would allow us to be successful in computing these sea state parameters as well. The dataset that we used was collected by Smartfin sensors created and distributed by a researcher at Scripps Institute of Oceanography (SIO), Phil Bresnahan. We scraped both the Smartfin IMU data as well as the publicly available CDIP wave height data ourselves in order to create a model which could predict significant wave height from Smartfin IMU data using machine learning techniques. The aforementioned research papers have used similar datasets taken from the National Oceanic and Atmospheric Administration (NOAA) for predicting wave height using numerical methods [2], or buoy IMU data collected by Khazar Exploration and Production Company (KEPCO) in the Caspian Sea to predict wave height using machine learning on wind models [6]. However, to our knowledge, this was the first project which scraped Smartfin data in conjunction with CDIP wave parameter data in order to train a machine learning model to predict significant wave height directly from IMU data. In addition to taking a novel approach and constructing our own dataset, our conclusions produce similar results to existing work (see Results).

0.3 Exploratory Data Analysis

Our dataset consists of Smartfin IMU data from 135 surf sessions along the San Diego coastline at La Jolla Shores, Scripps, and Black's beaches. The Smartfin's IMU contains an accelerometer, gyroscope, and magnetometer, which collect linear acceleration, angular velocity, and surfboard heading data in the x, y, and z directions respectively at a rate of 4 Hz. We scraped the Smartfin IMU data from the Smartfin Lostbird website using a headless browser and exponential backoff, then matched each datapoint with corresponding output labels from CDIP using Coordinated Universal Time (UTC) timestamps. For the input labels, the IMU A1, A2, A3, G1, G2, G3, M1, M2, and M3 features correspond to the raw accelerometer, gyroscope, and magnetometer values in the x, y, and z directions respectively. For example, IMU A1 corresponds to the accelerometer reading in the x-direction. Overall, our data from the 135 surf sessions had a total of 7,124,311 observations; each surf session ranges in duration from 45-130 minutes over a time period ranging from September 2016 to July 2019.

From our exploratory data analysis, we sought to verify literature indicating that statistical measures of our raw features were significant features to include in training our models. We saw moderate linear correlations between an input feature's mean, standard deviation and the corresponding output label, which indicated that including these statistical measures as features might help strengthen our model's predictive ability. In Figures 1 and 2 the overall distribution of acceleration, gyroscope, and magnetometer data are visualized to determine if assumptions about error distributions were relevant in model building. As expected, the distribution of the means of A1, G1 and M1 per ride session are normally distributed, whereas the standard deviations were right skewed. Additionally, in Figures 3 and 4 we see possible linear correlations between the input feature (IMU G2 and IMU M3 means) and the significant wave height using the line of best fit. Additionally, in the bottom center graph of Figure 2, we see that the standard deviation of IMU M2 has a negative correlation with significant wave height. Our exploratory data analysis also showed that the significant wave heights range from 0.89 ft. to 6.4 ft with a mean of 2.4 ft. and a standard deviation of 0.92 ft. (In the following figures, the plots are shown with data sampled from 30 surf sessions for visualization clarity).

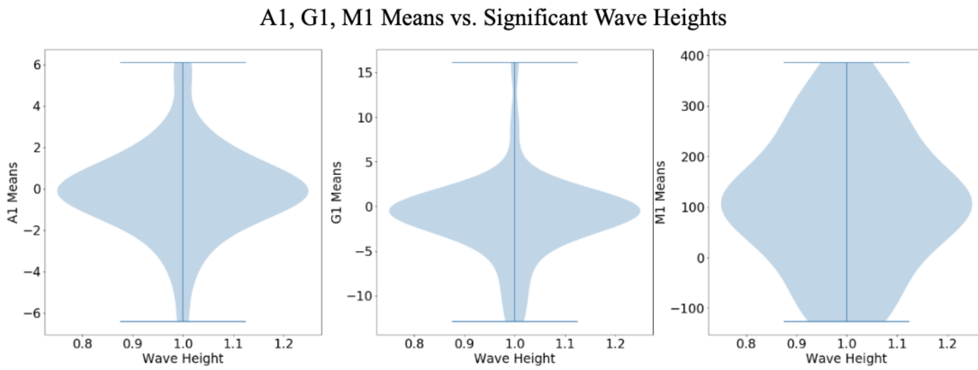

Figure 1: Distribution of A1, G1, and M1 means according to wave height.

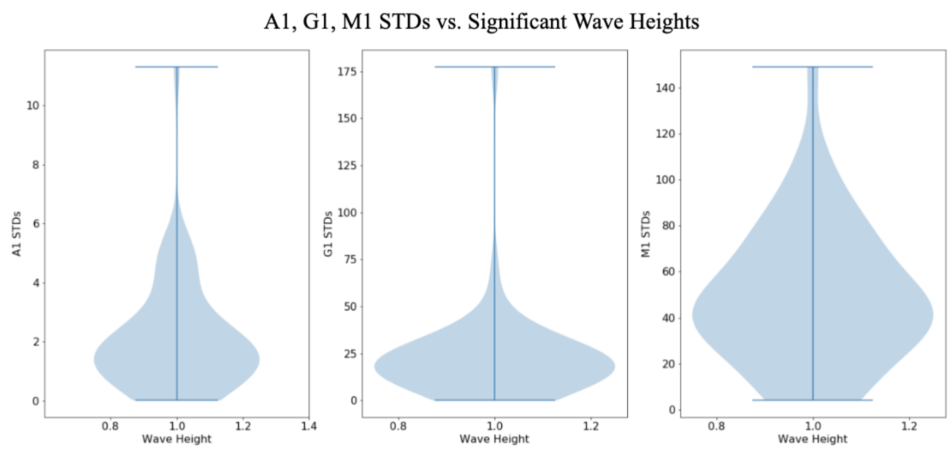

Figure 2: Distribution of A1, G1, and M1 standard deviations according to wave height.

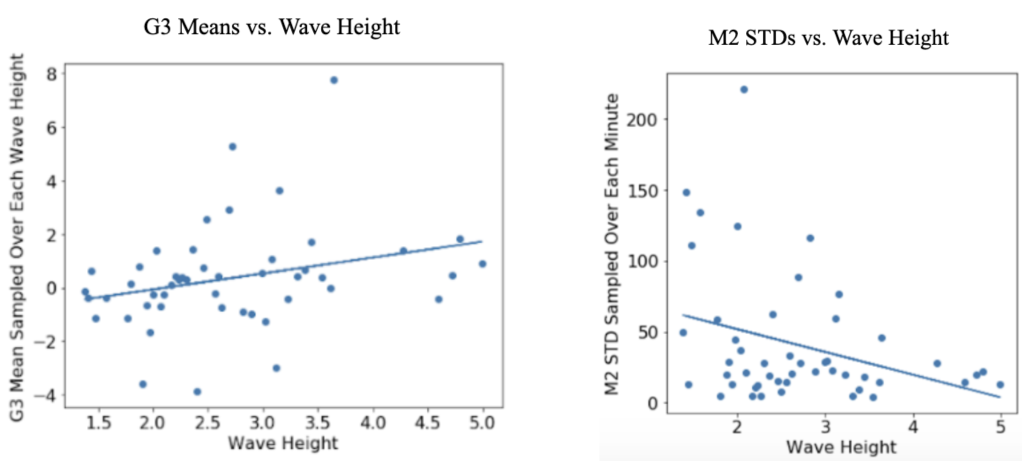

Figure 3: Possible linear corrleations between IMU means, standard deviations and wave height

0.4 Predictive Task

For this dataset, we wanted to be able to make predictions about wave characteristics from input IMU data. There are a variety of different wave characteristics, such as wave periodicity or wave direction, which we could have chosen to predict. We ultimately decided to explore prediction of significant wave height, which can vary between different times of the year and even across the time of day. The chosen baseline for our dataset was a model that always guesses the mean significant wave height, as it is simple to implement and performs reasonably well as a simple benchmark. We chose to evaluate our models based on the MSE as it allows us to determine the distance between our model's predicted label and the actual label. Our models' task is to minimize that distance in order to optimize our model to the task at hand. We considered the MSE as our measure of model success because we felt the underlying assumption that errors were normally distributed about zero seemed appropriate. MSE is also a standard error metric when working with regression models.

We created additional features using the mean and standard deviation of the IMU data over minute long and ride long intervals. This was motivated by a paper in our literature review, which determined statistical properties of IMU data such as mean and standard deviation to be good predictors of significant wave height [6]. This meant that we had to compress our 4 Hz data into 0.016 Hz data in order to have consistent input features for each observation in our dataset. The motivation for this step was that the Smartfin data was sampled at a higher frequency than the data collected by CDIP's buoys; we believed that having IMU information for every 0.25 seconds of data would bias the model to overfit to the training data. Therefore, in each of these three test scenarios, the input features were either: (1) the 9 IMU values (IMU A1, A2, A3, M1, M2, M3, G1, G2, G3), (2) the 9 IMU values and the mean and standard deviation of each of the 9 IMU values over a minute long interval, which corresponds to 27 input features, or (3) the 9 IMU values and the mean and standard deviation of each of the 9 IMU values over the entire surf session, which again corresponds to 27 input features.

We also considered a number of different models for this predictive task and ended up deciding that linear regression and decision tree regression were the most logical models to pursue. We considered using a latent factor model, but preliminary models performed poorly because our data doesn't have the user-item pair relationship for the latent factor model to discover. Logistic regression models were also considered, but they too are unsuitable for our data as they predict binary outputs and we wanted a model to predict real valued variables. Decision tree regression was a simple and relevant model because it uses observations about the dataset's features to make conclusions about the corresponding output label. Ultimately, the linear regression approach was deemed appropriate because it maps the relationship between real valued variables. This is an intuitive choice because we observed potential linear relationships between certain IMU input features and their corresponding output labels during our exploratory data analysis.

0.5 Models

For each of our models, we used 70% of our data in the training set, 20% in the validation set, and 10% in the testing set. Because data points within each surf session were highly correlated, we grouped data by surf session and then shuffled the data to avoid any bias due to the initial ride session order. Grouping data points in this way before splitting into training, validation, and test sets prevented testing on data points from the same surf session, which would have led to an inaccurate measure of model performance. A scalability issue we ran into whilst constructing additional features for our models was that of our limited computational resources since our original complete dataset had over 7,000,000+ observations. As a result, we decided to look at a percentage of all observations from each surf session. Each model below is tested using 5% of each ride's total IMU observations.

0.6 Decision Tree Regression

Decision tree regression is a form of supervised learning that serves as a more rudimentary approach to prediction that recursively partitions the dataset into smaller subsets as the decision tree grows. Each branch in the decision tree is formed by a split on a feature that maximizes the information gain based on some defined impurity measure (each child node has a larger information gain and a lower impurity measure than the parent node). For this model, we defined the impurity measure as the weighted MSE of the children nodes and no maximum depth. The primary benefit of using decision regression is that it is fairly simple to train and retains a low error rate on the training set. Limitations of this approach include overfitting to the data due to their increasing complexity with each branch and difficulty in concisely and reasonably interpreting model prediction decisions. Overall, this approach was unsuccessful as it did not produce models that performed better than or similar to the baseline. However, these MSE values serve as an informative point of comparison for later models. Additionally, similar to results in literature [7], we found that we had a lower MSE for datasets that included mean and standard deviation features for all of the raw IMU features (Figure 5), but overall MSE results were still slightly worse than the baseline.

0.7 Linear Regression

Linear regression is a supervised learning technique that produces real-valued predictions based on given features of the dataset. Based on our exploratory data analysis, we observed possible linear correlations between IMU features and their corresponding wave height labels. In terms of feature representation, our models appear to support already established literature [7].

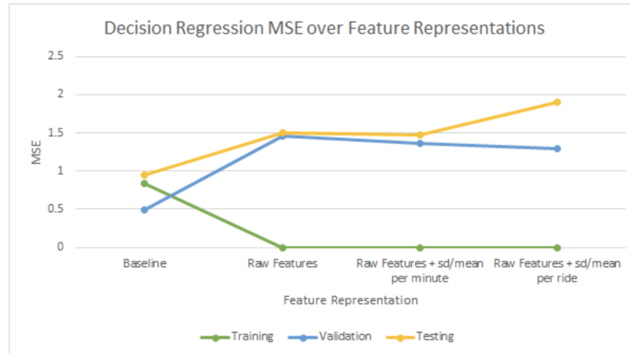

Figure 4: Decision Regression MSE Performance

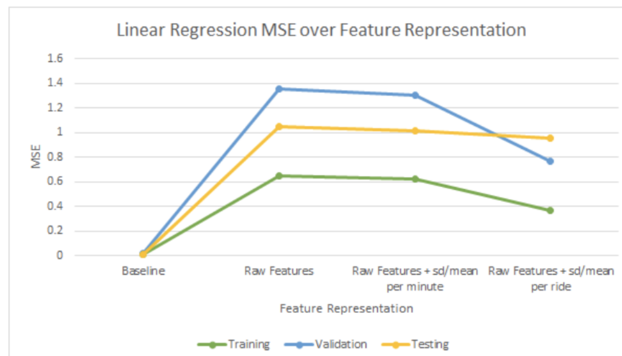

Figure 5: Linear Regression MSE Performance

When only the raw features are used for prediction, the MSEs for our validation and testing sets are fairly large, which indicates that our predictions are over 1 ft. incorrect on average (Figure 6). By adding the standard deviations and means of the raw IMU features, our predictions became more accurate as our MSE was minimized. While the MSEs of the three sets only improve marginally when the mean and standard deviation of the IMU features are included over a minute long interval; the MSE of the validation set improves dramatically when we include these statistical features computed over the entire ride session. This result suggests that using sample per minute may not be the best approach when trying to make predictions.

Overall, our linear regression model sup-

Table 1. Summary of Models' MSE Results on Feature Sets

Features	Training 1 MSE	Training 2 MSE	Validation 1 MSE	Validation 2 MSE	Testing 1 MSE	Testing 2 MSE
Baseline	0.837	0.008	0.492	0.015	0.950	0.009
A	1.89E-30	0.648	1.455	1.358	1.502	1.045
B	4.50E-28	0.627	1.36	1.302	1.472	1.015
C	8.07E-28	0.37	1.298	0.768	1.905	0.955

ports published literature which stated that the addition of the IMU's statistical properties would be good features for predicting wave height. We optimized the predictive strength of our model by taking two different versions of mean and standard deviation, per minute and per ride. Based on the MSE values reported, the addition of mean and standard deviation per ride appeared to be most successful in this optimization. In terms of scalability, our model suffers from the fact that it needs to calculate many means and standard deviations across many data points in order to work, making it difficult to run on our entire dataset of over 7,000,000 data points. The fact that we randomly sample 5% of data points per ride may be an additional source of bias, as there exists the possibility that the random samples are not a good representation of the ride as a whole. Despite this setback of scalability, using linear regression has strengths. The theta values we obtain from running the model are easily interpretable and clearly highlight which features have a stronger influence in prediction. Furthermore, the MSE is easy to interpret, giving us a good idea of how well our model is performing. We found the most successful model overall for our predictive task to be a linear regression model as it out performed our decision tree regression model on every feature representation. By examining the coefficients of each feature in our linear regression model, we determined the most important parameter to be A2, which is the acceleration in the y direction. This is reasonable as the acceleration in the y direction is correlated to the height of a wave, or the displacement in the y direction. Linear regression and decision tree regression models were more successful than any other models we considered because we wanted to predict real valued variables, making them more appropriate models to use over logistic regression or latent factor models.

0.8 Results and Conclusions

We compared the results of different combinations of features within each of our models to find those that were most successful (Table 1). For our first feature representation, we used the nine raw features taken directly from the data, with only preprocessing and unit conversion done as modification. Using this representation with our linear regression model, we obtained a validation MSE of 1.35810 and a testing MSE of 1.04543. When we added the standard deviation and mean of the features per minute, this resulted in a validation MSE of 1.30160, and a testing MSE of 1.01482, out performing the model with only raw features. The standard deviation gives a measure of how the wave changes over a minute, which gives insight to the changing wave height. The mean gives insight to the significant wave height during one minute compared to others. This was helpful in determining whether the significant wave height is above or below average. The final feature combination we tested included the raw features with standard deviation and mean, but calculated over the course of a whole ride. The validation MSE was 0.76806 and the testing MSE was 0.95465, making it the best performing model. This is likely because there is a correlation between standard deviation and mean for periods of time within one ride, but considering each ride as an individual example introduces more variation into the data, on which the model can better train.

Overall, the baseline that we created for this task, which always predicted the mean significant wave height, outperformed our testing values. With the linear regression model, the baseline MSE values were 0.01512 for validation and 0.00855 for testing. With the decision tree regression model the baseline MSE values were 0.49177 for validation and 0.94958 for testing. The baselines likely performed better than the predicted values on the task because of the characteristics of the data used; our labeled data from CDIP computed significant wave height as a statistical measure of the top third largest waves observed over a 30 minute period. In order to get the significant wave height, wave heights that were very small, occurred between surges, or were not perceivable, were excluded. The data we used to make predictions were recorded at 4 Hz, and no values were discarded. This means our model was using characteristics of the waves over a period of time that the labeled data set had thrown out. This discrepancy in the format of the data collection is likely the cause of predictions having a poorer performance than the baselines.

Our models confirmed our findings in the literature, which stated that the addition of statistical properties of IMU data, such as standard deviation and mean, as features would improve model performance in predicting wave characteristics, such as significant wave height [7]. Not only was our model able to verify these trends seen in data, it was able to perform similarly, and subject to the full data variance, even outperform previous models on similar prediction tasks [7]. In Bailey et al. RMSE is calculated for a wave height prediction task, which is found to be between 0.8 and 1.0. When converted to MSE for the sake of comparison to our model, the range of their MSE values is 0.64 to 1.0 [7]. Our MSE values are comparable to this range and, with certain feature representations and the full variance of the full dataset, can reach lower MSE values.

References

[1] Bresnahan et al. *Engineering a Smartfin for surf-zone oceanography.*. IEEE Oceans 2017, Sept. 2017.

[2] Marshall D. Earle *Nondirectional and directional wave data analysis procedures.* Stennis Space Center, Jan. 1996.

[3] L.C. Bender III, N.L. Guinasso Jr., and J.N. Walpert. *A Comparison of Methods for Determining Significant Wave Heights.* Journal of Atmospheric and Oceanic Technologies, June 2010.

[4] *Datawell Waverider Reference Manual.* Datawell BV Oceanographic Instruments. 10 Oct. 2009.

[5] J. Coyte, D.A. Stirling, M. Ros, H. Du & A. Gray. *Displacement Profile Estimation Using Low Cost Inertial Motion Sensors with Applications to Sporting and Rehabilitation Exercises* IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM), 2013, pp. 1290-1295.

[6] A. Zamani, D. Solomatine, A. Azimian, A. Heemink. *Learning from data for wind-wave forecasting.* Ocean Engineering, July 2008.

[7] J. Bailey, I. Milne, M. Van Someren. *Seastate Detection from Vessels at Motion using Learning Algorithms.* CEED Seminar Proceedings, 2019.