Are Rogue Waves Predictable on Regional Datasets? A Comparative Analysis of LSTM, GRU, and Ensemble Models

Andrew Obermiller
University of California, San Diego
aobermiller@ucsd.edu

Mike Liu
University of California, San Diego
m7liu@ucsd.edu

Huabo Min
University of California, San Diego
9500 Gilman Dr, La Jolla, CA 92093
humin@ucsd.edu

Abstract

Rogue waves are waves characterized by sudden occurrence, large crest height, and unpredictable direction. Rogue wave detection is an important safety device to reduce danger to shipping and naval operations. Though detection methods have been produced based on the mathematical modeling of Benjamin-Feir instability and the training of neural networks using very large global datasets there remains a notable gap: the development of a model that requires less data for regional adaptation and testing and the testing of various network architectures.

This paper aims to fill this void through the investigation of rogue wave occurrences through generation of a novel region focused dataset derived from existing methodologies to train, test, and validate various network architectures under varied input data conditions. We will then evaluate models' performance to determine the most promising candidate for future applications in rogue wave prediction.

1. Introduction

Rogue waves are very large waves with significantly larger height than the other waves in the immediate area. When occurring in the vicinity of ships, offshore naval infrastructure, or coastal installations these waves present a serious risk for injury and equipment damage [2]. Especially prevalent where 40% of people live in coastal or near coastal region and virtually all levels of industry rely on naval shipping as core infrastructure. Research has demonstrated the predictability of this natural phenomenon through the the analysis of various models and development of various predictive models, but production of such a model in all existing cases requires a large existing data set spanning millions of hours of ocean data [2] [5] [7].

2. Background and Related Work

Before the advent of machine learning and computational models, predicting rogue waves was significantly challenging. Traditional approaches to understanding and predicting rogue waves largely relied on observational data, theoretical physics models, and statistical methods[4]. However, these approaches are highly specialized to specific conditions or require assumptions that limit their applicability, neural networks learn directly from data, continuously improving their predictions as more data become available. This adaptability is crucial for rogue wave precision, where conditions and contributing factors can vary widely across different regions and times. Current research focused on LSTM models and deep learning neglects evaluation of lightweight models which may provide an advantage in conducting short-term field research or set-up in areas without access to large computing power. An example would be a remote wave detection buoy would not have access to large memory banks or power capabilities to support large but powerful models like the LSTM.

3. Methodology

3.1. Data Acquisition

The dataset utilized in this study was sourced from the Coastal Data Information Program (CDIP) [1] at the Scripps Institution of Oceanography. We targeted a selection of 314 buoys from a total of 1504 deployments distributed across Pacific and Atlantic Oceans 1, accumulating a vast measurement of (0.8) million hours in set B and derived a smaller dataset focused on the East North American Coast in set A.

[1]CDIP website https://cdip.ucsd.edu/

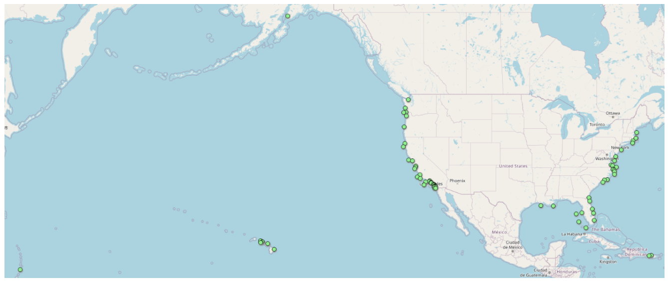

Figure 1. CDIP Buoy Distribution. Many buoys have multiple deployments, highest go up to 48 (Buoy 036)

Dataset	Hours (Million)
Breunung A	0.57
Breunung B	5.16
Breunung C	0.81
Häfner	6.13
Dataset A	0.3
Dataset B	0.8

Table 1. Dataset distribution from Breunung [2] and Häfner [5], Dataset A and B is collected by this group independently of other papers

From set B, we extracted a dataset consisting of (54237) samples in 30 minute time blocks, offering a valuable vertical ocean displacement data. In comparison to similar papers, we have significantly smaller dataset to work with.

Rogue wave occurrences are extremely rare in the vast measurement. To ensure the robustness of our analysis and to mitigate any bias, we balanced our dataset to include an equal number of rogue and non-rogue wave instances. This approach prevents model overfitting to the more prevalent class and allows for a fair assessment of the model's predictive capabilities across both outcomes. Rogue wave events were labeled using a threshold criterion where the wave height(H) exceeds twice the significant wave height(Hs)

$$\frac{H}{Hs} \geq 2$$

A thorough quality control process was implemented, which included:

- Adherence to manufacturer-quality flags
- Spike detection to identify and rectify abnormal data points
- Sensor malfunction checks, specifically for reading that surpassed the over-range threshold of 20.47 meters.
- Eliminating of repetitive data sequences to ensure the variability of the dataset

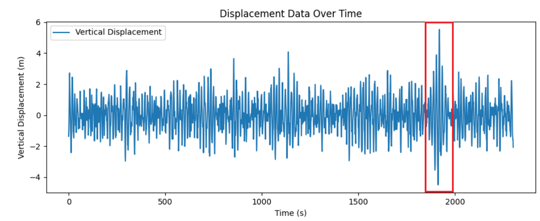

Figure 2. Rogue wave sample. The detected wave is boxed in red.

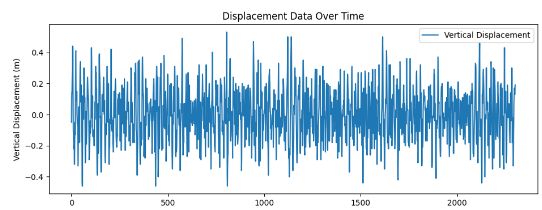

Figure 3. Non Rogue Wave sample. Note the absence of a significantly large wave at any point.

The quality control process that met. Consequently, we formulated two balanced datasets, each comprising 25-minute windows of data. In the rogue wave dataset, each window contains a verified rogue wave occurrence (Figure 2), whereas the non-rogue wave dataset does not contain the rogue wave occurrence in a 25-min window time frame (Figure 3). We collected a balanced dataset of rogue and non rogue windows because if we set rogue waves as true positive (T_p) and non rogue waves as true negatives (T_n), we would need a non-biased dataset in order to create an unbiased model.

$$\frac{T_p}{T_n} = 1$$

3.2. Model Implementation

In this paper, we are using three comparative machine learning methods for rogue wave prediction, which is a strategic approach that leverages the strengths and mitigates the weaknesses of individual models to enhance the accuracy and reliability of predictions. Besides this, a model that performs well across diverse conditions and datasets is more reliable for practical use. This comparative analysis helps identify which models are more versatile and robust against varying data patterns and noise, which are common in oceanographic data.

Here, we use two RNNs (Recurrent Neural Networks) methods LSTM (Long Short-Term Memory and GRU (Gated Recurrent Units) and one ensemble method with stacked CNN (Convolutional Neural Network). One of the advantages of RNNs is that they can model non-linear relationships within data through their layered structure and activation functions. Because of this structure, they are able to handle sequential data efficently. They can remember information over time and forget the information that is not needed for the output, making them particularly useful for time-series analysis.

3.3. LSTM

Recurrent neural networks are a common starting model for prediction tasks with minimal memory requirements. By augmenting an RNN with LSTM memory cells the ability of the network to process complex time series data is improved as the model may selectively retain information from each input vector [6]. RNNs with LSTM cells are a common starting point in the study of wave prediction models, where underlying trends are difficult to detect [2] [3] [1]. As LSTM models are the most common model for this task, it is assigned as the baseline model for later comparison.

The model is structured as a stacked architecture, consisting of in series: LSTM layer, batch normalization, dropout layer, fully connected (dense) layer. The LSTM layer acts to capture long-term dependencies, focusing on significant trends in the sequential data input. Batch normalization rescales data passing through the layer to zero mean and unit variance. A dropout layer follows to aid in the prevention of overfitting. A final fully connected layer processes input and makes a classification prediction. This process is repeated through for N_L stacked layers. This architectural design enables the model to discern subtle trends in the data without the extensive time and computational resources typically associated with more complex deep learning approaches.

Network weights are determined through use of the Adam optimizer. Initial hyperparameters were chosen based on a similar model using an advanced hyperparameter tuning algorithm [2]. The hyperparameters shown below were experimentally determined through incremental change until no improvement was observed.

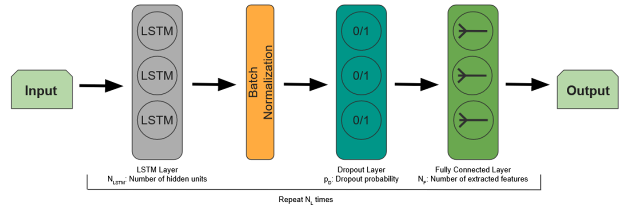

Figure 4. LSTM Network Architecture

Hyperparameter	Value
Optimizer	Adam
Learning Rate	0.001 (default)
Loss Function	Cat. Cross-Entropy
Batch Size	128
Number of Epochs	100
Hidden Units (N_{LSTM})	128
Activation Function in LSTM	tanh (default)
Activation Function in FCL	ReLU
Features in First FCL	50
Features in All Other FCL's	2
Number of Stacked Layers (N_L)	5
Dropout Rate	0.2

Table 2. Hyperparameter Selection for LSTM Model

3.4. GRU

One of our predictive models employs the Gated Recurrent Unit (GRU) architecture, known for its efficiency in processing sequential data, particularly in recognizing and preserving temporal dependencies within such data. This capability is crucial for accurately predicting rogue waves, phenomena whose occurrence is tied to patterns in sequential oceanographic data. The model begins with an initial GRU layer of 128 units, utilizing L2 regularization to prevent overfitting and ensure model robustness. Following this, batch normalization and dropout layers are applied to stabilize learning and mitigate overfitting further. A second GRU layer with 64 units refines temporal information, leading to an output architecture that includes a dense layer with 64 neurons (ReLU activation), a dropout layer, and a final dense layer with sigmoid activation for binary classification of wave data. This structure allows for nuanced feature extraction and processing, critical for distinguishing between rogue and non-rogue waves.

The GRU model is structured in s sequential layer configuration, with each layer engineered to progressively refine the input data. This design ensures the sequential extraction and refinement of features from the data, which is essential for the accurate classification of rogue and non-rogue wave occurrences.

3.4.1 Ensemble Method

The greatest advantage of the Ensemble method is the concert of a variety of models to achieve a higher accuracy than evaluated individually. For this project, we used an array of

Hyperparameter	Value
Optimizer	Adam
Initial Learning Rate	0.001
Loss Function	Binary Cross-Entropy
Batch Size	64
Number of Epochs	100
First GRU Layer Units	128
Second GRU Layer Units	64
Regularization	L2
Dropout Rate	0.3
Early Stopping Patience	10
Reduce LR on Plateau	Yes
LR Reduction Factor	0.5
Min Learning Rate	0.0001

Table 3. Hyperparameter Selection for GRU Model

CNN models with different initialized hyperparameters to boost performance. The basic architecture is an array of initialized CNN with different neurons in the RELU activation layer and a learning rate of 0.001 that will be increased and decreased as early stopping to prevent overfitting. The CNNs will be stacked on top of each other, training on the same dataset and output their guesses to a decision function. Then the CNN performance will be determined by their relative validation loss functions and sum each weighted prediction to receive the final prediction in the form of

$$Y_k = \sum_{i=1}^{N} w_i y_i$$

where Y_n is the prediction for the kth window, N is the number models, w_i is the weight determined by relative validation loss against each CNN model, and y_i is the prediction of each model visualized in Fig 5.

Hyperparameter	Value
Optimizer	Adam
Initial Learning Rate	0.001
Loss Function	Binary Cross-Entropy
Batch Size	100
Number of Epochs	100
Number of Models	3 (CNN)
First Dense Layer Units	64, 128, 256
Output Layer Units	1
Regularization	L2
Dropout Rate	0.2
Reduce LR on Plateau	Yes
LR Reduction Factor	0.5
Min Learning Rate	0.0001

Table 4. Hyperparameter Selection for LSTM Model

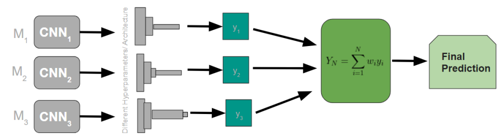

Figure 5. Representation of Ensemble Model

4. Experiments

4.1. Experiment Setup

The objective of the experiment is to explore whether a small dataset can generalize a rogue wave detection ML model. For this purpose, our wave data is compartmentalized as 30 minutes of recorded wave blocks and if there is a rogue wave it will be first detected at the 25th minute. The collected data is then separated into 80-10-10 (train-validation-test) sets standardized across each model for final comparisons in performance. Each model is standardized to train on 100 epochs with two warning time frames: a 20-minute and 5-minute wave recording window with a 1-minute warning time of rogue wave arrival.

4.2. Model Training and Validation

4.2.1 LSTM

Training of the LSTM model involved learning at a batch size of 128 over 100 epochs. Early-stopping was employed at epoch 50 with systematic slight hyperparameter alterations in order to observe positive or negative effect. Given the relatively small dataset size determining whether observed fluctuation in validation loss and accuracy was a result of model inefficiency or data sparsity proved challenging. Training of this model as a standard within the area of study provides a baseline for comparison to the other models. Though the LSTM model began to exhibit signs of overfitting as epochs increased, increases to the initial dropout rate improved the model's ability to generalize.

4.2.2 GRU

The GRU training involves early stopping and learning rate reduction to fine-tune the model's performance, using a batch size of 64 over 100 epochs. Validation on separate datasets ensures the model's reliability against the ground truth of the rogue wave formation.

4.2.3 Ensemble Method

Given the inherent difficulty in forecasting rogue waves without a comprehensive understanding of the regional wave dynamics, we were compelled to employ a black box model to predict incoming rogue waves within varying

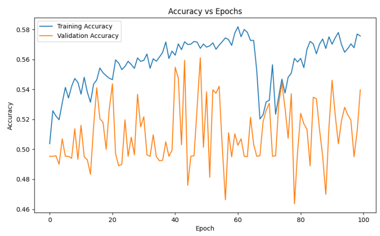

Figure 6. 5 minute training window for LSTM, dataset A. Unstable learning and near-random test accuracy indicate significant difficulty in training.

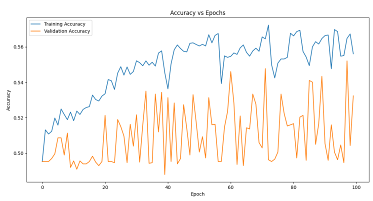

Figure 7. 20 minute training window for LSTM, dataset A. Unstable learning and near-random test accuracy indicate significant difficulty in training.

warning periods. Due to the ambiguity surrounding the intricacies of rogue wave prediction, we adopted an ensemble approach, utilizing three Convolutional Neural Networks (CNNs) with 64, 128, and 256 non-output neurons, respectively, within our Rectified Linear Unit (RELU) activation layer. The rationale behind employing a higher number of non-output neurons within the ensemble CNNs was to facilitate the exploration of the black box complexity in rogue wave prediction parameters. For validation, we scrutinized the performance of our models by comparing the accuracy and cross-entropy loss between the training and validation sets per epoch.

5. Results

5.1. LSTM

As a baseline model, the LSTM network performed reasonably given the difference in data set size in comparison to similar studies [2]. Training was unstable when finding results on the initial dataset, but was somewhat mitigated with the expanded dataset at higher epoch counts. Greater stability in training is observed in the 5 minute window with 1 minute warning period condition, suggesting that the immediate period before the rogue wave period is the easiest for the model to detect. With a smaller regional dataset (set A) this is invaluable knowledge as future training can be aimed at further reducing the training window to the most relevant period. An apparent plateau in training progress is visible for both the 20 minute and 5 minute training windows (and was present following different alterations in the hyperparameters), that, in comparison to similar models, suggests there is a lower bound on the dataset size required for a substantial increase in accuracy. Obtaining more data presents its own challenge because of the difficulty of distinguishing rogue wave-like sensor malfunctions and spikes from the waves themselves. Nonetheless, a final best accuracy of 0.591 suggests the potential of this model architecture and presents itself as a viable candidate for large global dataset training (set B).

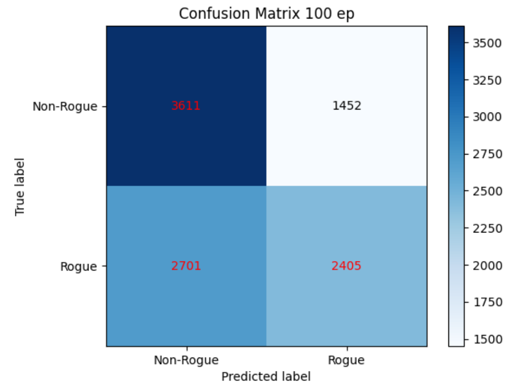

Figure 8. 5 minute training window for LSTM, dataset B. Note the increase in stability and improvement in convergence.

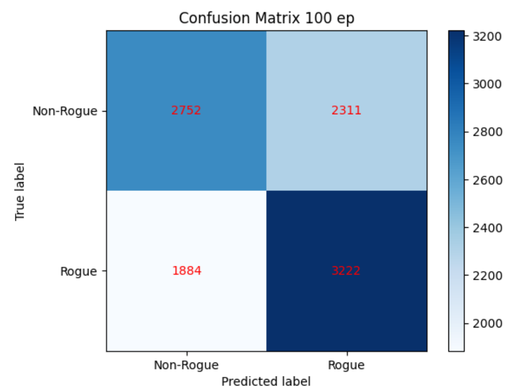

Figure 9. 5 minute training window for LSTM, dataset B.

Figure 6 and 7 are the accuracy vs epoch graphs trained for 100 epochs with either a 20 minute or 5 minute warning time from buoys located in the Gulf of Mexico and North American East Coast

Figure 8 and 10 are the accuracy vs epoch graphs and confusion matrices trained for 100 epochs with either a 20 minute or 5 minute warning time from an expanded buoy set.

5.2. GRU

The expansion from a smaller to a larger dataset has notably enhanced our GRU model's ability to predict rogue waves, as demonstrated by the accuracy increase from Figure 12 to Figure 14 for the 5-minute model and from Figure

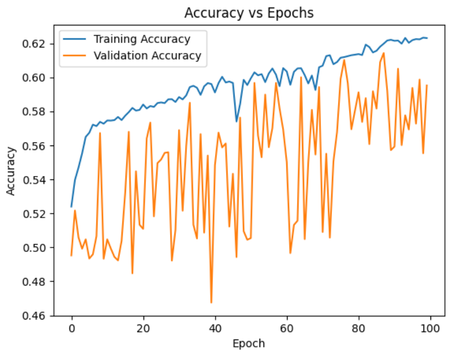

Figure 10. 20 minute training window for LSTM, dataset B. Similar end result suggests that the 5 minute window may contain trends more significant in learning the data.

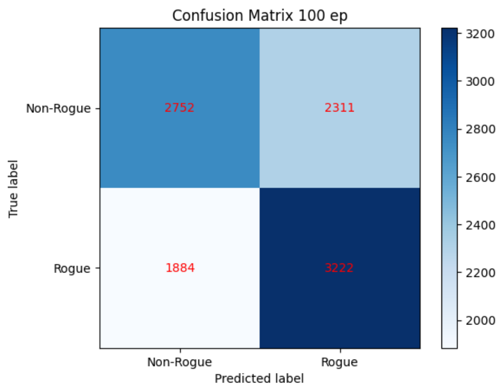

Figure 11. 20 minute training window for LSTM, dataset B

13 to Figure 16 for the 20-minute model. The more robust confusion matrices (Figures 15 and 17) for the larger dataset indicate improved classification precision, especially in the 5-minute model, which is critical for real-time maritime safety applications. Our GRU model was tested under two conditions for a larger dataset: a 5-minute and a 20-minute window, both with a 1-minute warning time before a potential rogue wave event. Figure 14, representing the 5-minute window model, shows a higher accuracy of 0.637 compared to Figure 16, the 20-minute window model, at 0.632 accuracy. The steadier trend in Figure 14 suggests that shorter window yield more reliable predictions by focusing on critical data and reducing noise. Figure's 16 spikes in accuracy indicate greater sensitivity to noise over longer periods, potentially leading to less stable predictions. The results favor shorter observation windows for timely and dependable rogue wave forecasts, ideal for real-world maritime application where efficiency is essential. The improved performance in the shorter window can be attributed to the model's enhanced ability to capture and utilize more immediate, relevant temporal features leading up to the event, reducing the noise and irrelevant information that might be present in longer window spans. This suggests that, for the specific task of rogue wave prediction, shorter observation windows preceding the event can provide more focused data, enabling the GRU model to make slightly more accurate prediction.

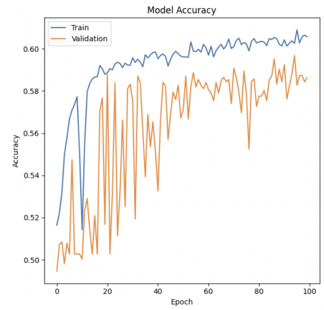

Figure 12. 5 minute training window for GRU, dataset A

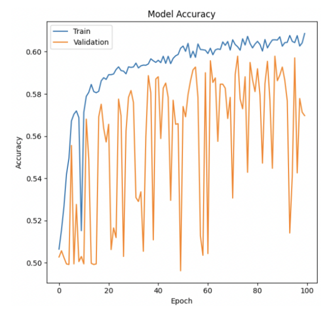

Figure 13. 20 minute training window for GRU, dataset A

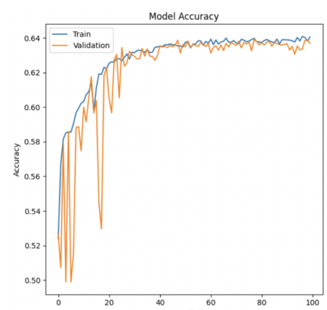

Figure 14. 5 minute training window for GRU, dataset B

The variations in model performance highlight the trade-off between capturing immediate, relevant features and integrating wider contextual information over longer observation periods, especially when working with smaller datasets from specific regions like the Gulf of Mexico and North America East. The model trained with the large dataset perform stabler and more accurate compared to trained with the small dataset.

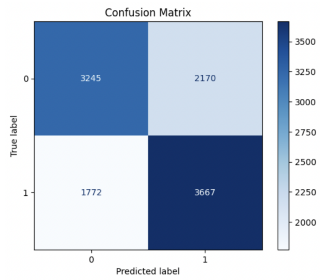

Figure 15. 5 minute training window Confusion Matrix, dataset B

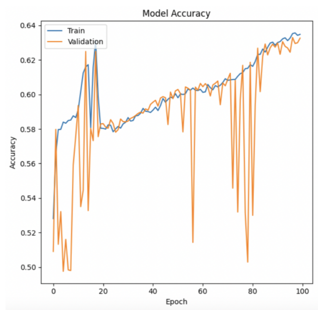

Figure 16. 5 minute training window for GRU, dataset B

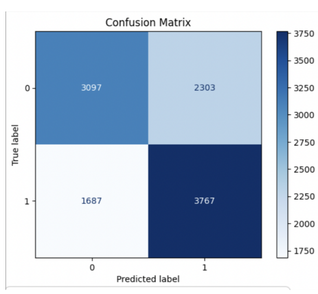

Figure 17. 20 minute training window Confusion Matrix,dataset B

5.3. Ensemble Method

Comparing the results to LSTM and GRU, the Ensemble CNNs failed horrendously in all marks except training speed. For all accuracy metrics on the validation sets, all CNNs hover around 50% accuracy no better than a coin flip. Their validation loss vs train loss curves clearly shows a lack of model convergence. Validation loss rises to several factors of training factors rapidly, while training loss converges close to zero nearly instantaneously. This demonstrates a problem in model selection, whereas CNNs have the necessary flexibility to predict patterns in a complex dataset. In this case, the CNNs learn more about the noise surrounding rogue wave detection rather than the rogue wave itself. Even with a larger dataset with the same hyperparameters, barely any performance can be identified, as a result, the Ensemble CNN is not a suitable model to predict rogue waves due to its extreme simplicity. The only metric that was better than LSTM or GRU is training time. At the same neuron or unit complexity, the ensemble is a far faster model trainer but the performance is left wanting.

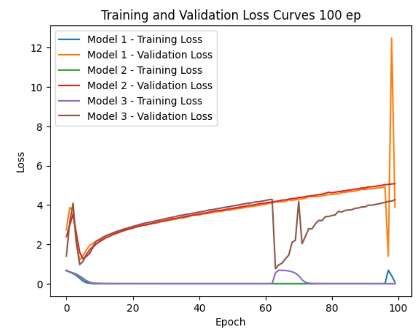

Figure 18. 5 minute training window for dataset A

Figure 18 is the 5 minute training window focused on the regional dataset A, demonstrate the immediate divergence of Validation and Training Loss for all three CNN models. Although Training and Validation loss function occasionally attempt to converge into each other at certain epochs, the general trend is very clear that training loss remains close to a value of 0 while validation loss has a logarithmic growth.

In hopes to address the clear effects of poor model selection and overtraining on the CNN ensemble, we reconstructed the model with additional data points with dataset B. We hoped by including a wider variance of data can mitigate the immediate divergence between Validation and Training loss, hence increase overall accuracy of the Ensemble CNNs. However, similar to results with dataset A, our training and validation loss never converged within the 100 epochs of training on the best performing data window of 5 minutes in Fig 19. And we were forced to conclude, using simple CNNs as an Ensemble is not a viable method due to the tendency to over fit on the training set regardless of hyperparamter tuning. At best, we were able to squeeze out 54% accuracy which is as reliable as a physical coin flip.

5.4. Discussion

As part of our exploration different data window size (20 and 5 minutes), the 5 minute window has a relatively apparent advantage over the 20 minute observation window. Highlighting that the most recent minutes leading up to a rogue wave has a significant weight to the prediction outcome, which carry across both datasets A and B. Interestingly, the effects of a regional dataset does not generalize rogue wave as well as a global rogue wave set. Our rationale was centered around wave prediction models commonly incorporate local bathymetry, in a similar vein, rogue waves should rely on regional bathymetry patterns. De-

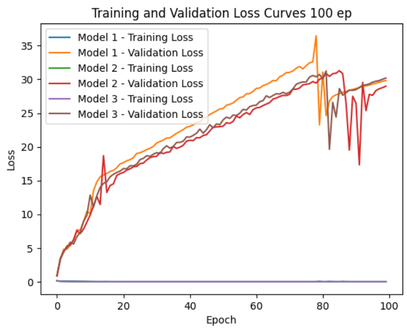

Figure 19. 5 minute training window loss, note the training vs validation loss diverge at epoch 0

Method	Accuracy (%)
LSTM 20 min	54%
LSTM 5 min	51%
GRU 20 min	57.0%
GRU 5 min	58.6%
Ensemble 20 min	48.2%
Ensemble 5 min	51.3%

Table 5. Accuracy Gulf of Mexico + N. America East Coast

Method	Accuracy (%)
LSTM 20 min	59.1%
LSTM 5 min	59.1%
GRU 20 min	63.2%
GRU 5 min	63.7%
Ensemble 20 min	51.5%
Ensemble 5 min	54.9%

Table 6. Accuracy all of N. America

spite this expectation, a global rogue wave set with dataset B shown a markedly higher accuracy than regional dataset A; we have a speculation that rogue waves may not have as strong as a tie to common wave prediction parameters such as bathymetry, surface temperature, etcetera. A markedly advantage ensemble has over LSTM and GRU is the fast training time and model size, LSTM and GRU takes upwards of 4 hours to train on dataset A while ensemble complete a model in half an hour.

6. Conclusion

In comparable papers, the dataset used to train machine learning models are far larger for the rogue wave definition of

$$\frac{H}{Hs} \geq 2$$

For example, in Breunung's [2] paper dataset B with an LSTM of similar epoch but far larger span of data (5.16 million hours vs 0.8 million hours) achieved a 0.76 accuracy while in our highest performing model achieved a 0.64 at most. Thus, even with a far smaller dataset and without stringent mathematical wave modeling, hints of rogue wave patterns can be discerned from the naturally noisy dataset.

In conclusion, our investigation into rogue wave prediction has highlighted the influence of data size on the performance of different neural network architectures. Despite the limited access to more free buoy data, our GRU model demonstrated an encouraging accuracy increase from 0.586 to 0.637 for 5-minute window time as the dataset was modestly expanded to the buoys over the North America. This improvement suggests that GRUs are particularly adept at learning from smaller datasets, possibly due to their streamlined architecture, which may require fewer data to effectively capture temporal dependencies.

In the goal of identifying a model adept at prediction using a smaller regional data set, we determined that while GRU models were found to perform best our results suggest that regional differences in wave trends may not supersede abundance of data in importance when training an accurate model.

Conversely, LSTM models, which started at a lower accuracy of 0.51, exhibited a marked improvement up to 0.59 when provided with a larger dataset. The significant gain implies that LSTMs, with their more complex architecture and additional parameters, have the potential for substantial improvements in performance as they are exposed to more data.

These findings suggest that while GRUs can offer a robust solution when data is scarce, LSTMs hold the promise of significant gains in accuracy with the acquisition of larger, more diverse datasets. Therefore, for applications where collecting extensive data is feasible, investing resources into expanding the dataset can be highly beneficial, particularly for LSTM-based models, which may ultimately lead to great predictive performance when it comes to a nonlinear predictive problems. [2]

References

[1] Cihan Bayindir. Early detection of the draupner wave using deep learning. *Physics Letters A*, 380:156–161, 2016.

[2] Thomas Breunung and Balakumar Balachandran. Are rogue waves predictable from field measurements? *International Conference on Offshore Mechanics and Arctic Engineering*, 2(42):1–9, 2023.

[3] Thomas Breunung and Balakumar Balachandran. Data-driven, high resolution ocean wave forecasting and extreme wave predictions. *Ocean Engineering*, 268, 2023.

[4] Alexander Cattrell, Meric Srokosz, Ben Moat, and Robert Marsh. Can rogue waves be predicted using characteristic

[2] Scripts and data used in this paper `https://github.com/rellimrebo/rogue-wave-prediction-model`

wave parameters? *Journal of Geophysical Rsearch: Oceans*, 123(8), 2018.

[5] Dion Häfner, Johannes Gemmrich, and Markus Jochum. Machine-guided discovery of a real-world rogue wave model. *Proceedings of the National Academy of Sciences*, 2023.

[6] Alex Sherstinksky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404, 2020.

[7] Alexey Slunyaev, Ira Didenkulova, and Efim Pelinovsky. Rogue waters. *Contemporary Physics*, 52(6):571–590, 2011.