Hurricane Rapid Intensification Forecasting and Historical Trends. (CS 7641 Proposal - Group 1)

Rick H Nguyen, Austin O'Connell, Inhyeong Park, Brandon Lee Parker, Achintya Raya Polavarapu

***Abstract*—This is the abstract.**

***Index Terms*—AI, Machine Learning, Supervised Learning, Unsupervised Learning, Weather Radar Prediction**

I. Introduction / Background

RAPID intensification (RI) is a phenomenon in which a tropical system's maximum wind speed increases by at least 30 knots within a 24-hour period, often in an environment of low wind shear, higher relative humidity, and higher sea surface temperatures [1]. Numerous applications of machine learning have been used to forecast hurricanes and RI. For example, support vector machines (SVMs) have been applied to classify tropical cyclone formations, mapping meteorological parameters with storm development [2]. Similarly, neural networks such as multilayer perceptrons (MLPs) have been used to model non-linear relationships between such parameters to identify atmospheric thresholds that lead to RI in storms [3].

This project involves timeseries data on recorded Atlantic tropical storms. The first dataset is HURDAT2 track data containing 6-hour storm records, including geographic position, maximum sustained wind speed, and central pressure. The second dataset is ERA5 global reanalysis data since 1940, which we will confine to the Atlantic Ocean. The relevant features utilized include sea surface temperatures, vertical wind shear, mean sea level pressure, and relative mid-level humidity, which are features known to contribute to hurricane intensity [4].

II. Problem Definition

RI in hurricanes is an increasingly common and unpredictable phenomenon that is worsened due to rising sea temperatures from climate change. RI remains very difficult to forecast in terms of timing and magnitude [5], which can make proper large-scale preparation difficult. Current physics-based and numerical weather models struggle to capture subtle inner-core dynamics in storms [6]. Minor parameterization errors can result in major changes, which is where we hope a machine learning model can bridge this gap.

III. Methods

A. Data Preprocessing Methods

We begin with the provided HURDAT2 (6-hour) storm-track records and the accompanying ERA5 atmospheric variables. Our preprocessing focuses on making these sources consistent and model-ready. First, we align ERA5 predictors to each HURDAT2 timestamp and storm location (e.g., using the nearest ERA5 grid point or bilinear interpolation), so that each 6-hour storm observation has a matched set of environmental features. When appropriate, we optionally summarize the local environment by averaging ERA5 variables over a small neighborhood around the storm center to reduce sensitivity to single-grid-cell noise.

We then clean the data by flagging invalid entries and imputing missing values (within-storm time-neighbor fill when possible, otherwise training-set medians). We build a compact feature set (storm state plus recent 6–12h changes and translation speed) and standardize continuous predictors for scale-sensitive models.

B. ML Methods

We employ two complementary approaches to study rapid intensification (RI) in Atlantic hurricanes. The *unsupervised* component clusters storm observations by their atmospheric/environmental conditions to identify recurring regimes and relate them to RI behavior. The *supervised* component trains a predictive model to forecast whether RI will occur over a specified future horizon. We will tune clustering and classifier hyperparameters on validation data (silhouette/BIC for clustering; trees/depth/class weights for random forest) and evaluate with PR-AUC and F1/recall to handle class imbalance [7]. Details of each approach, including model choices, hyperparameter selection, and evaluation under class imbalance, are provided in the sections below.

IV. Unsupervised Learning Model

The unsupervised learning portion of this project will involve clustering hurricanes by leveraging observations on their atmospheric conditions to classify their progression. Further, whether differing magnitudes of RI is represented within specific atmospheric conditions.

The first candidate learning method is K-Means clustering, which will partition storm observations into k clusters by iteratively assigning each observation to the nearest centroid. A limitation is that K-Means produces hard assignments and assumes equal-sized clusters, raising concerns since the smaller population of RI prone storm environments may be absorbed into the conditions associated with that of a larger neighboring cluster.

To address this, the second candidate learning method is Gaussian Mixture Models (GMM), which assigns each datapoint a probability of belonging to each cluster and, most importantly, allows clusters to vary in size and density, making

it better suited to present differing scales of small distinct RI prone storms. This soft assignment works to reduce the chance of cluster absorption. For both methods, the hyperparameter will be determined empirically using the elbow method and silhouette scoring for K-Means, and BIC for GMM.

V. Supervised Learning Model

The supervised learning portion of this project will involve building a model to predict if a storm will experience rapid intensification within a designated time period in the future (for example, within 24 hours). The model will be a binary classifier, determining whether or not rapid intensification will occur.

Alongside building a classifier, one of the project's hypotheses is that rapid intensification is occurring more frequently now compared to 20-30 years ago. Therefore, this binary classifier will be trained on data from 20-30 years ago, and tested against more recent data. This separation of the dataset will provide dual use: testing the accuracy of our binary classifier, and testing our hypothesis that storms more frequently experience rapid intensification.

A potential candidate for the binary classifier model is random forest. This model is ideal for the problem due to the high-dimensional dataset that has been selected [8]. A random forest model can be used to easily perform feature selection and reduction, and thus will help understand which metrics affect storms undergoing rapid intensification, which may yield insightful analysis as to (if proven true) why rapid intensification is increasing over time.

One challenge in building a proper supervised learning model for this topic is that a lower percentage of storms experience rapid intensification, thus leading to a potentially misleading level of accuracy. For example, it is reported that only 20%-30% of tropical storms in the Atlantic experience rapid intensification, which means a classifier could answer "No" and be correct 70%-80% of the time. Thus, our model will account for the uneven amount of labeled data for RI storms vs No-RI storms [7].

VI. Results: Metrics & Project Goals

The performance of the RI classifier will be evaluated using robust metrics to class imbalance: Recall, Precision, F1-score, and PR-AUC. The primary goal is to outperform logistic regression in F1-score and PR-AUC while maintaining high Recall to minimize RI events.

Regarding sustainability, we will minimize our computational footprint by selectively downloading ERA5 variables and regions. Ethically, we will explicitly report the trade-off between false negatives and false positives, including a disclaimer that this research-oriented model is not intended for direct real-world evacuation decisions. We expect tree-based models, such as Random Forest, to demonstrate superior predictive power. Furthermore, we anticipate that physically significant features-including wind shear, relative humidity, and sea surface temperature-will emerge as the top predictors, aligning with established meteorological theory.

VII. Gantt Chart - Team Responsibilities

The project timeline and team responsibilities are available in our shared Gantt chart spreadsheet: `CS7641 Group 1 - Gantt Chart (Google Sheet)`.

VIII. Contribution Table

Team Member	Contributions
Rick H. Nguyen	Intro/background/problem definition; GitHub repo; M1 design + feature reduction; M2 coding; M3 design/cleaning.
Austin O'Connell	Supervised learning section; video; M1 design + visualization; M2 design + visualization.
Inhyeong Park	Results/discussion; Gantt chart; M1 implementation; midterm report; M3 design/feature reduction/cleaning; final report.
Brandon L Parker	Unsupervised learning section; GitHub page; M1 design + data cleaning; M2 design + data cleaning; M3 implementation.
Achintya Polavarapu	Methods/ML algorithms; contribution table; M2 design + feature reduction; M3 design/visualization/cleaning.
All: M1/M2 results evaluation; M3 results evaluation; M1–M3 comparison.	

References

[1] J. Kaplan, M. DeMaria, and J. A. Knaff, "A revised tropical cyclone rapid intensification index for the atlantic and eastern north pacific basins," *Weather and Forecasting*, 2010.

[2] M. Wei, G. Fang, and Y. Ge, "Tropical cyclone genesis prediction based on support vector machine considering effects of multiple meteorological parameters," *Journal of Wind Engineering and Industrial Aerodynamics*, 2023.

[3] W. Xu, K. Balaguru, A. August, N. Lalo, N. Hodas, M. DeMaria, and D. Judi, "Deep learning experiments for tropical cyclone intensity forecasts," *Weather and Forecasting*, 2021.

[4] S. S. Ganesh, F. I.-H. Tam, M. S. Gomez, M. McGraw, M. DeMaria, K. Musgrave, J. Runge, and T. Beucler, "Multidata causal discovery for statistical hurricane intensity forecasting," 2025. [Online]. Available: https://arxiv.org/abs/2510.02050

[5] F. Judt and S. Chen, "Predictability and dynamics of tropical cyclone rapid intensification deduced from high-resolution stochastic ensembles," 2016.

[6] M. DeMaria, J. Franklin, M. Onderlinde, and J. Kaplan, "Operational forecasting of tropical cyclone rapid intensification at the national hurricane center," *Atmosphere*, 2021.

[7] L. Yi, T. Youmin, T. Ralf, and W. Shuai, "Revisiting the definition of rapid intensification of tropical cyclones by clustering the initial intensity and inner-core size," *Journal of Geophysical Research: Atmospheres*, 2022.

[8] O. NS, J. BC, B. GS, and S. JL, "A comparison of random forest variable selection methods for regression modeling of continuous outcomes," *Briefings in Bioinformatics*, 2025.