Research Journal of Applied Sciences, Engineering and Technology 13(8): 638-651, 2016
DOI:10.19026/rjaset.13.3050
ISSN: 2040-7459; e-ISSN: 2040-7467
© 2016 Maxwell Scientific Publication Corp.
Submitted: April 2, 2016 Accepted: June 25, 2016 Published: October 15, 2016

Research Article

Applying Machine Learning Methods for Predicting Tropical Cyclone Rapid Intensification Events

[1]Hadil Shaiba and [2]Michael Hahsler
[1]Department of Computer Science and Engineering, Southern Methodist University, Dallas, TX 75275, United States
[2]Department of Engineering Management, Information and System, Southern Methodist University, Dallas, TX 75275, United States

Abstract: The aim of this study is to improve the intensity prediction of hurricanes by accounting for Rapid Intensification (RI) events. Modern machine learning methods offer much promise for predicting meteorological events. One application is providing timely and accurate predictions of Tropical Cyclone (TC) behavior, which is crucial for saving lives and reducing damage to property. Current TC track prediction models perform much better than intensity (wind speed) models. This is partially due to the existence of RI events. An RI event is defined as a sudden change in the maximum sustained wind speed of 30 knots or greater within 24 hours. Forecasting RI events is so important that it has been put on the National Hurricane Center top forecast priority list. The research published published on usingmachinelearning methods for RI prediction is currently very limited. In this study, we investigate the potential of popular machine learning methods to predict RI events. The evaluated models include support vector machines, logistic regression, naïve-Bayes classifiers, classification and regression trees and a wide range of ensemble methods including boosting and stacking. We also investigate dimensionality reduction and feature selection and we address class imbalance using the Synthetic Minority Over-sampling Technique (SMOTE). The evaluation shows that some of the investigated models improve over the current operational Rapid Intensification Index model finally; we use RI predictions to make improved storm intensity predictions.

Keywords: Ensemble learning, feature selection, feature extraction, machine learning, rapid intensification, SMOTE

INTRODUCTION

Tropical Cyclones (TCs) are a natural phenomenon which whichthat forms over large bodies of relatively warm water. TCs can intensify and reach hurricane strength, which causes which causes causing strong winds, heavy rain and flooding resultingthat result in loss of lifeves and significant damage whenonce it makes landfall. By employing systems for forecasting tracks and wind speeds (also called intensity) of TCs and establishing early warning systems, human lives can be saved and economic losses can be reduced. Despite the vast amount of research on climate observations and models, predicting the sustained wind speed (known as storm intensity) of a TC still remains a challenging (DeMaria *et al*., 2005). One issue is the existence of rapid intensification (RI) events, which are defined as sudden intensity increases of 30 knots or greater within 24 hours (Kaplan *et al*., 2010b). Predicting RI events is important and was added to the National Hurricane Center's (NHC) top forecast priority list in 2008 (Kaplan *et al*., 2010a). One of the main issues with RI events is that they are rare and the environmental conditions that favor their appearance are not clearly understood (Kaplan *et al*., 2010a). In this study, we introduce an easy to use data set and investigate how well popular machine learning methods, ensemble methods, feature selection and class balancing methods perform for predicting RI events.

The presented TC dataset is derived from the Statistical Hurricane Intensity Prediction Scheme (SHIPS) model (DeMaria *et al*., 2005; Kaplan *et al*., 2010a, 2010b) (version 2010) and includes storms from 1982 to 2011. The dataset includes information about the state of the ocean, clouds and atmosphere collected by satellites, ships, planes and buoys. Measurements are available in 6-hour time intervals (DeMaria, 2013). We will compare standard classifiers including support

Corresponding Author: Hadil Shaiba, Department of Computer Science and Engineering, Southern Methodist University, Dallas, TX 75275, United States, Tel.: +1(405)4105105

vector machines (SVM), logistic regression (LR), naïve-Bayes classifiers (NB) and Classification and Regression Trees (CART). To deal withaddress the large number of available features, we apply dimensionality reduction and feature selection. Class imbalance (rapid intensification events are relatively rare) is addressed by rebalancing the data with the Synthetic Minority Over-sampling Technique (SMOTE) and by using the area under the ROC curve (AUC) instead of accuracy. The results will be compared to the NHC operational Rapid Intensification Index (RII) model with a 30-knots RI threshold. The RII is based on a linear discriminate analysis function. We finally show that RI prediction can be used to improve intensity forecasts.

LITERATURE REVIEW

Rapid intensification: Kaplan *et al*. (2010b) define a Rapid Intensification (RI) event as a change in intensity (the maximum 1 min sustained wind speed at 10 meters above ground) of 30 knots or greater within 24 h. The number of RI events is small compared to the total number of observations. However, many major hurricanes in the Atlantic basin undergo RI events and their potential impact is large, representing a significant threat. They appear between June and November and their main causes are still not clear, which makes them hard to predict (Kaplan *et al*., 2010a). A study doneperformed by Wang and Zhou (2008) refines the RI definition as thea maximum intensity increase of at least 5 knots in 6 h, 10 knots in 12 h and 30 knots in 24 h and indicates that:

- Storms that undergo RI produce high forecast error rates.
- Seasonal, intra-seasonal and inter-annual environmental changes effect the potential of RI in different ways-; for example, it mentioned that there are more RI events in El Nino years than in La Nina years.
- The mean location of where the storms form differs from one month to another.
- It is more likely to have RI events when the storm's formation moves towards the south.
- The decrease in northerly vertical change in wind speed and the increase of in low-level westerly horizontal change in wind spinning increases the potential of RI.

Another study by Shay *et al*. (2000) shows that most RI events happen when the storm passes through a flow of deep warm water such as a Loop Current. Measurements of the water at different levels were derived from the radar altimeter placed on the TOPEX/Poseidon (T/P) satellite. The paper indicates that the measurements showed that h Hurricane Opal, which formed in 1995, underwent RI once it passed through a Loop Current. An article by Lippsett (2011) shows that Hurricanes Katrina and Rita are also examples of hurricanes that met came into contact witha Loop Current, that caused causing them to intensify. It shows that satellites could monitor these layers could be monitored during the winter through satellites while but that it is harddifficult to monitor them in summer. To solve the problem, they suggest employing a glider (a submarine vehicle), that which can be launched into the ocean and used to take measurements of the temperature of the water and send results that are then sent to a database. The glider can be controlled remotely and can stay in the ocean for months.

There have been several efforts to predict RI. We will discuss the most significant RI models next. The Rrapid intensification index (RII) (Kaplan *et al*., 2010b) is a discriminate analysis model used to forecast RI events based on large-scale weather, ocean and climate condition predictors, which are derived from the Statistical Hurricane Intensity Prediction Scheme (SHIPS) model. RII model is an operational model used by NHC that predicts RI and works in real time. Different RI thresholds can also be defined, such as e.g., 25, 35 and40. Different versions of the RII model werehave been implemented. The original version of RII (Kaplan and DeMaria, 2003) first calculates the average value (threshold) of each predictor for all RI events. Then, it compares the value of each predictor of the observed sample with its threshold and decides whether it satisfies the threshold or not. A total of five predictors were used to develop the model. A drawback to this step is that it does not show to whichthe level at which the features favor the appearance of RI (Kaplan *et al*., 2010a). Finally, the probability of RI is equal to the percentage of cases that satisfy the thresholds.

A revised version of the RII model was described by Kaplan *et al*. (2010a) for the Atlantic basin. Three predictors related to the inner-core were added to the five-predictor RII model. The first predictor is derived from GOES IR satellite images, the second predictor is derived from the microwave satellite images and the third is derived from the Global Forecast System (GFS) model and the Sea Surface Temperature (SST) predictor. GFS is global numerical weather prediction model run by the U.S. National Weather Service (NWS). This model provides predictions for aviation guidance several days ahead for the entire globe based on physical and mathematical equations that take several hours to run on supercomputers (Meisner, 2006). An improvement of in the new version is that predicted values of some quantities (SHRD, RHLO, POT, D200 and OHC) are averaged over the prediction interval, thus accounting for their influence on the TC. Another improvement is to rescale In addition, the values of the predictors are rescaled. The operational RII version set weights to the scaled predictors based on their importance in predicting RI. The Probability of Detection (POD) in the Atlantic basin for the years 2006 and 2007 was reported to be between 15 and 59% and the False Alarm Ratio (FAR) ranged between 71 and 85% (Kaplan *et al*., 2010b).

Kieper and Jiang (2012) show that a cyan ring appears in the 37 GHz microwave false color images around the TC eye when a RI event happens. It appears within the first 6 hours of the RI formation. This ring is used as an indicator of RI and is combined with the RII model to estimate the potential of RI. The model was evaluated on a dataset from 2003 to 2007 in the Atlantic basin. The result of the combined model shows a POD of 24% and a FAR of 26%. The RII model alone shows a POD of 77% and a FAR of 66%, whereas the ring alone shows a POD of 75% and a FAR of 9%. The combined method lessens the FAR, which is favorable but also reduces the POD. It is interesting to note that the ring performs much better alone than combined with RII.

Logistic regression and naïve-Bayes were investigated for RI prediction in Rozoff and Kossin (2011). The models apply feature selection on the set of predictors used by the RII model. For the naïve-Bayes model, forward feature selection and cross validation (in each round, 1 year is selected for testing and the remaining years for training) are used. For the logistic regression model, a sequential feature selection method called step wise regression is used. The results of the three classifiers, logistic regression, naïve-Bayes and the RII model, are then combined by simply averaging the RI probabilities predicted by the three models. The skill of the combined model was examined using the Brier skill score defined belowin Equation 8 on Page 5. Data from year 1995 to 2009 waswere used. The results show that compared to the RII model Brier skill score, the combination of the models improves by 33% in the North Atlantic Basin and by 52% in the North Pacific Basin. These results are encourageing for more furtherinvestigation of using statistical learning methods.

MATERIALS AND METHODS

Tropical cyclone dataset: For this study, we have gathered a historical dataset for TCs in the AtlanticBasin formfrom 1982 to 2011. The dataset includes the SHIPS model predictors (version 2010). SHIPS is an intensity prediction model that predicts a storm's maximum sustained wind several days ahead beforehand by applying multiple linear regression on the historical data described here. Feature weights are updated for every hurricane season based on their importance in the previous years. Therefore, an updated version of the SHIPS model is released to provide forecasts for each particular hurricane season (DeMaria *et al*., 2005). SHIPS was introduced in 1994 (National Hurricane Center, 2009) and is still one of the most skillful models used by the National Hurricane Center (NHC). The life-cycle of each TC is represented in 6-hour time intervals which and includes a set of ocean and climate observations. Environmental satellites, radars, buoys, ships and aircrafts are some a few of the methods of collectingused to collect climate data, which are then processed using complex models (Rhome, 2007). For comparison, RI probabilities produced by the RII model (Kaplan *et al*., 2010a) are also included toin the dataset with a 30, 35 and 40 knotsthreshold. Once the storm hits land, the intensity of the storm decaysddecreases rapidly (Kaplan and DeMaria, 1995). Therefore, overland cases were removed from the dataset for this study. A summary of the used SHIPS predictors and their description is shown in Table 1 (DeMaria, 2013; Kaplan *et al*., 2010a).

The cleaned and preprocessed dataset is made publicly available to researchers at: http://lyle.smu.edu/IDA/data/storms.

Comparing different machine learning methods for rapid intensification detection: We compare different well-known classification techniques to predict RI events in the Atlantic basin using the dataset described in this study. The investigated methods include:

- Support Vector machines (SVMs) (Cortes and Vapnik, 1995). The prediction in the SVM classifier is made based on a hyperplane that separates the classes.
- Logistic regression (Walker and Duncan, 1967), which models the log of the odds ratio (logit) of a class label as amultiple linear regression on a set of covariates.
- Naïve-Bayes (Russell and Norvig, 1995) is a simple model which thatpredicts the class probability conditional on given predictor values. It assumes that the predictors are conditionally independent.
- That CART (Breiman *et al*., 1984) classifier estimates the probability using adecision tree.

Compared to the methodology used by Rozoff and Kossin (2011), our study uses a larger dataset and examines more classification models and ensemble learning methods. We address the class imbalance problem using SMOTE and use for feature selection the ROC-based AUC score for feature selection instead of the normally employed accuracy, which is problematic for imbalanced data.

The training set used for the learning phase contains that years 1982 to 2008. Years 2009 and 2010 are held out for testing. At this stage, the best features for each supervised learning method are selected and extracted using 5-fold cross validation. Then, the selected and extracted features are used for each classification model.

Evaluating rapid intensification models: We add to the dataset a class attributes with the values "RI" and

Table 1: Tropical cyclone set of predictors and their description for SHIPS model (version 2010) and its derived predictors (DeMaria, 2013; Kaplan *et al*., 2010a). A dot (•) marks the predictors used by the RII model

Predictor	Description
SHIPS (version2010)	
ID	The first two characters "AL" represent the Atlantic basin, the second two characters represent the sequence number of a TC in a certain year, and the remaining four characters represent the year when the TCh appended.
DATE	A 6-hour time interval represented as follows: "yymmdd". The first two digits represent the year, the second twodigits represent the month, and the last two digits represent the day of aTC.The current UTC time interval is related to a given record in hours. Once a TC takes place, data points are produced.
TIME	Every 6 h starting from time t, which means that this predictor will have one of the following values: t = 0, 6, 12 or 18 hours. 0, 6, 12, and 18 mean 12:00 AM, 6:00 AM, 12:00 PM, and 6:00 PM, respectively.
LAT,LON	Latitude in 1/10 degrees North of the center of a TC, longitude in 1/10 degrees West of the center of the TC.
VMAX	Presents TC's maximum intensity in knots. It measures the strength of a TC, and its values are a factor of five and range between 15 and 160. The Saffir-Simpson scale (Schott *et al*., 2012) is used to categorize TCs, where the weakest TC is called a tropical depression, and once its maximum intensity exceeds 64 knots, it turns into a hurricane.
PER$^{\bullet}$	Previous 12-h change in VMAX (persistence).
ADAY	A Gaussian functions of the day of the year when a current TC took place (jdate) relative to the peak day of the Atlantic hurricane season (day 253). Transformation: exp ((jdate−253)^2/900).
SPDX	Zonal component of initial storm motion, which refers to the horizontal (East-West) direction of the persistence over the sea surface. An SPDX positive value represents the horizontal wind movement from West to East, and its value represents its wind speed in knots.
PSLV	The pressure at a level where the water flow is similar to the storm's motions (the leading layer).
VPER	A derived predictor of V MAX ×PER.
PC20	A channel four GOES predictor that returns the percentage of area covered by cold-cloud top brightness temperature ≤ −20°C.
GSTD	A derived predictor of VMAX times the standard deviation of PC 20.
POT	The estimated Maximum Potential Intensity (MPI) that is averaged over the TC's track minus VMAX at time t = 0.
SHDC	850 millibar (pressure level) shear magnitude with vortex removed and averaged from 0-500 km of storm center.
T200, T250P	200,250 millibar temperature averaged from 0-1000 km of storm center.
EPOS	Average thata difference between a parcel lifted from the surface and its environment averaged from 200-800 km.
RHMD	700-500 millibar relative humidity (%) averaged from 200-800km.
TWAT	GFS model mean tangential wind speed.
Z850	850 millibar absolute vortices.
D200	200 millibar divergence.
LSHDC	A derived predictor of SHDC × sin (LAT).
VSHDC	A derived predictor of SHDC × VMAX.
POT2	A derived predictor ofPOT^2.
RHCN	Ocean heat content from satellite altimeter data (KJ/cm2).
SDIR	Reference direction for shear direction predictor (sdp).
SHGC	850-200 millibarshear magnitude (kt × 10) with vortex removed and averaged from 0-500 km.
X_AVG	Where X = POT •, SHDC, T200, T250P, EPOS, RHMD, TWAT, Z850, D200 •, LSHDC, VSHDC, POT2, RHCN •, SDIR, or SHGC with values averaged from time t = 0 to t = 24h.
RHLORI$^{\bullet}$	850-700mb relative humidity percentage from 200-800 km radius averaged from time t = 0 to t = 24h.
SBTRI$^{\bullet}$	Standard deviation of cold cloud-top from 50-200 radius with a GOES channel 4 brightness temperature.
PCRI30$^{\bullet}$	Percentage area of cold cloud-top from 50-200 radius with a GOES channel 4 brightness temperature ≤−30°C.
RHCRI$^{\bullet}$	Reynold heat content averaged from time t = 0 to t = 24 h.
Derived Predictors	
PERX	Previous change in VMAX where X = 6, 18, or 24 h.
VPERX	VMAX×PERX with X = 6,18, or 24 h.
YDAYS	Day of the year.
VPC20	PC20×V MAX.
RII Model Outcome	
RIIX	RII model estimate of RI with X = 30, 35, or 40knots threshold.

"NRI" which is derived from the VMAX variable. RI marks the duration of a 30 knots RI event initial point until 6 h before its completion (6, 12, 18 and 24 h prior to a 30 knots change in intensity). In the case that ifan RI event overlaps with another RI event, they will be considered asa single RI event that lasts for more than 24 h. Figure 1 and Table 2 demonstrates an example of an RI event that lasted for 54 h. The TC dataset includes 421 different storms that include 171 RI events. Table 3 summarizes the total number of RI and NRI instances in the dataset. The table shows a strong class imbalance with roughly 9 times more NRI events than RI events.

Performance metrics: Model predictions for a test set are compared with the known correct class labels (RI/NRI) to obtain the number of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) as shown in the confusion matrix in Table 4. TP is the number of thecorrect forecasts of RI events, whereas FP is the number of the incorrect forecasts. The values in the confusion table areused

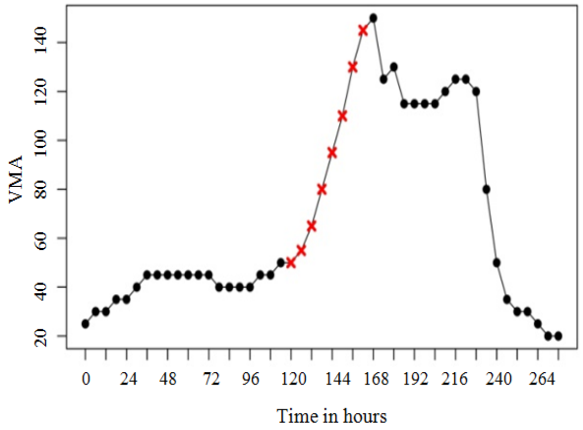

Fig. 1: An example of a storm that underwent RI. The red Xs represent an RI event that lasted for 54 hours.Theplot represents the maximum intensities of the storm in 6-h time intervals

Table 2: A summary of the RI event shown in Fig. 1 from time t = 102 to t =186. The table shows the time in 6-hour intervals and the actual intensity (VMAX). The numbers in bold represent the 24-hour intensity change during the RI event

Time in hours	VMAX	ΔVMAX24	Class
102	40	-5	NRI
108	45	5	NRI
114	45	5	NRI
120	50	10	RI
126	50	10	RI
132	55	10	RI
138	65	20	RI
144	80	30	RI
150	95	45	RI
156	110	55	RI
162	130	65	RI
168	145	65	RI
174	150	55	NRI
180	125	15	NRI
186	130	0	NRI

Table 3: Summary of the number of RI and NRI instances and the imbalance ratio

Storms' Instances	
NRI	7901
RI	865
Imbalance ratio	9.134

Table 4: Confusion matrix for a binary RI/NRI classifier

Predicted/Actual	RI	NRI	Total
RI	TP	FP	P'
NRI	FN	TN	N'
Total	P	N	

to calculate sensitivity, specificity and other performance measures.

Accuracy (ACC) is used to measure the overall performance of a binary classifier and is measured as:

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

However, strong class imbalance is a problem with accuracy, since as always predicting the majority class will already results in a very high accuracy. For binary classification models which the provide not just a class label but probabilities of RI, the Receiver Operating Characteristic (ROC) curve can be used. The ROC plots sensitivity versus specificity. Different thresholds on the predicted RI probability are chosen and then the TPs, TNs, FPs and FNs are calculated. Sensitivity is equal to the True Positive Rate (TPR), whereas specificity is one minus the False Positive Rate (FPR). TPRs and FPRs are defined as:

$$TPR = \frac{TP}{TP + FN} \qquad (2)$$

$$FPR = \frac{FP}{FP + TN} \qquad (3)$$

In this study, we use the Area under the Curve (AUC), which is a performance measure that computes the area under the ROC curve to evaluate the performance of each model. In comparison to accuracy, AUC is based on rates (TPR and FPR) and therefore is not affected by class imbalance.

Kaplan *et al.* (2010b), the performance of the RII model is measured using the Probability of Detection (POD) and false alarm ratio (FAR). POD is equivalent to the TPR and is the ratio of the correct forecasts of RI occurrences to the actual number of RI occurrences, while FAR is the number of incorrect forecasts of RI divided by the total number of RI forecasts. The greater the value of POD and the lower the value of FAR, the better the model performs. FAR is calculated as

$$FAR = \frac{FP}{FP + TP} \qquad (4)$$

The Brier Skill Score (BSS) is another performance measure used by Kaplan *et al.* (2010a) to evaluate the skill of a model relative to climatology. A positive BSS is an indication of a skillful model, whereas a negative BSS is an indication of a model that is not. The climatologically probability of RI is simply defined as the unconditional probability of RI estimated by

$$PRI_{climatology} = \frac{RI}{RI + NRI} \qquad (5)$$

The Brier Score (BS) of a prediction is calculated as the squared difference of the predicted probability from the observed value (0 or 1)(is used)

$$BS = \begin{cases} (1 - PRI)^2, if RI; \\ (0 - PRI)^2, otherwise. \end{cases} \qquad (6)$$

The Brier Scores for all observations are summed upfor each model, resulting in BSCs for climatological probability and BSMfor each model's forecasts (BSM). Finally, the BSS is calculated as:

$$BSS = \left[1 - \frac{BSM}{BSC}\right] \times 100 \qquad (7)$$

There are different ways to test how well a predictive model performs on given data. Cross validation is a standard evaluation method in data mining (Tan *et al*., 2005). The data is split into k random folds. Then, the classifier is tested on each fold while being trained on the remaining folds. This guarantees that all data is are used at least once for testing. The results are then averaged. Usually, observations are individually assigned to folds. However, our data is are organized by hurricanes and observations in the same hurricane are related with each other. To produce useful results, we needit is necessary to avoid testing observations against models which that were able to learn from the future behavior of the same hurricane. Therefore, we use a custom assignment strategy whichthat assigns complete hurricanes to folds. In our experiments, we use 5-fold cross validation, where each fold contains approximately the same number of storms.

The holdout method is another standard evaluation method in data mining, where a holdout sample is used for testing and the remaining samples are used for training (Tan *et al*., 2005). In our study, the holdout method is also used for evaluating the predictions of the different RI models using the selected and extracted features. The test set includes an independent year (2009 or 2010), while the training set includes years from 1982 up-to the test year to have all the previous year's included in the learning process and to simulate a real-time RI forecast. The test results offor 2009 and 2010 are combined and the AUC is measured for each RI model. This type of evaluation is used to emulatethe a s real application case where all past data is are available at the beginning of a hurricane season to predict the RIs of that season.

Synthetic minority over-sampling technique: SMOTE (Torgo, 2010) is a sampling method that over-samples the minority class by introducing synthetic observations based on k-nearest neighbors and at the same time under-samples the majority class based on the user's preference and the number of generated samples. The SMOTE instance generated is a random number between the selected sample and its neighbor. We use SMOTE with SVM, LR, NB and CART to address the problem of class imbalance for RI events.

Table 5 summarizes the results of the holdout experiments for the years 2009 and 2010 and shows the model performed for each experiment on the original data and two settings for SMOTE. SMOTE 1 increases the imbalance to about 1 to 4 and SMOTE 2 balances the dataset. By looking at tBased on the table, we can see that sampling does not improve LR and does not have an influence on NB. However, having an identical number of RI and NRI instances by decreasing the size of the NRI instances and increasing the size of the RI instances increases the performance of SVM and CART.

Table 5: AUC results for basic classifiers with the SMOTE sampling technique. The number of RI and NRI instances is shown at the top of the table

Class	Original	SMOTE 1	SMOTE 2
RI	530	1060	1060
NRI	4456	4240	1060
Classifier	Original	SMOTE 1	SMOTE 2
SVM	0.836	0.861	0.870
LR	0.905	0.902	0.889
NB	0.838	0.841	0.839
CART	0.852	0.770	0.861

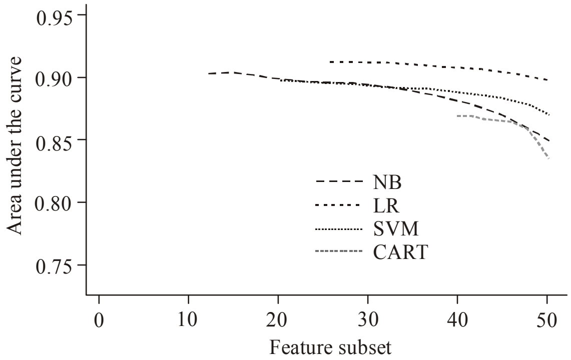

Fig. 2: AUC improvement using backward feature se- lection algorithm (from right to left)

Feature selection methods: We test the well-known feature selection techniques of hill climbing (Greiner, 1992) and backward and forward feature selection (Tan *et al*., 2005). Twelve experiments using 5-fold cross validation were conducted to select the best predictor subset out of the 53 available features. We make two important changes to the way standard feature selection is performed. First, we create folds by assigning complete storms rather than individual observations to avoid overestimating accuracy by temporal correlations of observations in the same storm. Second, instead of accuracy, we use the area under the ROC curve (AUC) as the selection criterion for feature selection to account for the class imbalance in the data set.

The Backward Feature Selection (BFS) algorithm starts with the entire set of features. At each round, all models with a single feature removed are tested. The feature subset creating the model with the highest AUC remains for the next round. The search proceeds until model quality worsens and the subset of the round with the highest AUC is finally chosen. Figure 2 represents the improvement of AUC in each BFS iteration for the different classifiers. Note that the selection process starts at the right-hand-side of the plot (includes all features) and proceeds towards the left.

To perform forward feature selection (FFS), the algorithm starts with the empty feature set. One feature is chosen at a time and the feature that produces the highest AUC is selected for the next round. In each remaining round, the feature which that improves the AUC the most is added. Figure 3 represents the improvement of in AUC in each FFS iteration.

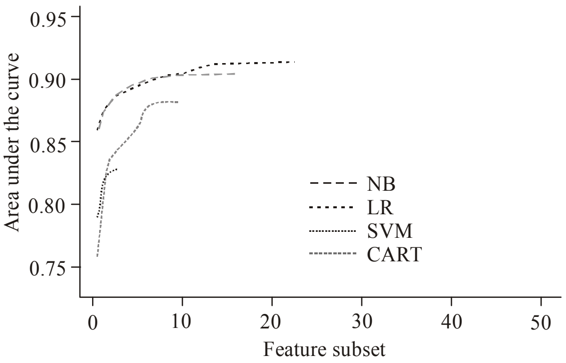

Fig. 3: AUC improvement using forward feature selection algorithm

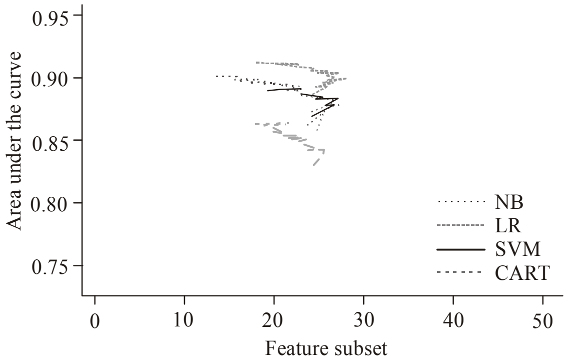

Fig. 4: AUC improvement using hill climbing feature selection algorithm

Hill climbing employs forward and backward selection. It starts with a random set of features it, keeps adding and removing features (one at a time) and selects the subset with the highest AUC. This process is iteratively repeated until the AUC stops improving. The result of the hill climbing algorithm is shown in Fig. 4.

The best result achieved by each feature selection method is shown in Table 6. As we can see from the table, the Logistic Regression (LR) models outperform other models in spite of the feature selection algorithm used. The p value of the t test of every pair of the feature selection models is shown in Table 7.

The p-value of the t-test of every pair of the feature selection models is shown in Table 7. The reported p-values are adjusted for multiple comparisons using Bonferroni correction. The results show that LR models are significantly better than the other models (p-values <0.05).

The left half of Table 8 summarizes which the feature that is selected (marked by an asterisk) by each model and feature selection method. Forward selection selects the smallest number of features, which is not surprising given the selection strategy. It is interesting to evaluate the importance of individual features. The purpose of Feature selection has the aim is to include the most important features into the model. Therefore, one way to measure feature importance is by counting how many models and selection methods select each particular feature. Ten out of twelve feature selection modelsselected the predictor PER18 (persistence over 18 hours), while seven choose PER, which is also chosen by the RII model. PER24 was the least chosen predictor amongPER, PER6, PER18 and PER24. POT AVG, which is also an RII predictor, was chosen by six RI models. Table also shows that the previous changes in intensities are chosen frequently by different models. Two RII predictors called RHCN AVG and PCRI30 have only been chosen by three and one models, respectively, while SBTRI has never been chosen.

Table 6: Best AUC (5-fold cross validation) of each feature selection method. The standard deviation is shown in parentheses. The best three models are highlighted

Classifier	Hill Climbing	Backward	Forward
SVM	0.894 (0.013)	0.899 (0.021)	0.826 (0.016)
LR	0.913 (0.011)	0.912 (0.013)	0.912 (0.012)
NB	0.903 (0.014)	0.903(0.015)	0.903 (0.013)
CART	0.864 (0.022)	0.868(0.024)	0.880 (0.027)

Table 7: P-values of every pair of feature selection models adjusted using Bonferroni. The values that are < 0.05 are highlighted. P-values < 0.004 are rounded to 0

		SVM	SVM	LR	LR	LR		NB	NB	CART	CART
Classifier	SVM Hill	BFS	FFS	Hill	BFS	FFS	NB Hill	BFS	FFS	Hill	BFS
SVM BFS	0										
SVM FFS	0	0									
LR Hill	1	0.23	0								
LR BFS	1	0.11	0	1							
LR FFS	1	0.78	0	1	0.11						
NB Hill	0.02	1	0	0	0	0					
NB BFS	0.01	1	0	0	0	0	0				
NB FFS	0.01	1	0	0	0	0	1	0.04			
CART Hill	0	0	0	0	0	0	0	0	0		
CART BFS	0	0	0	0	0	0	0	0	0	1	
CART FFS	0	0	0	0	0	0	0	0	0	0.07	1

Counting the number of models a feature is chosen for ignores the importance of the feature in each model. The right half of the table includes the rank of each feature based on feature importance reported by SVM, LR and CART using each feature selection method. If a feature is not chosen, then we use the highest rank of 44. We exclude the naïve-Bayes model, since because feature importance is not available. The average rank is reported in the last column. The rank-based importance ordering is similar to the ordering by the number ofmodels using each feature. However, there are some exceptions. While PC20 is chosen by nine models, its

Table 8: Features selected by different models and feature selection methods (left half of table) and feature rank (lower is better) by model (right half of table). The dot (•) marks predictors used by the RII model

	Forward				Backward				Hill Climbing			
Predictor	SVM	LR	NB	CART	SVM	LR	NB	CART	SVM	LR	NB	CART
PER18		*	*	*		*	*	*	*	*	*	*
PER6	*	*	*		*	*	*	*		*	*	
GSTD		*	*		*	*	*	*	*		*	*
D200 AVG•		*	*		*	*	*	*	*	*	*	
PC20		*	*	*		*	*	*		*	*	
POT		*	*		*		*	*	*	*	*	
SDIR AVG		*	*			*	*	*		*	*	*
VSHDC			*		*		*	*	*	*	*	*
PER•	*	*	*			*	*			*	*	
TWAT		*	*				*	*	*		*	*
VPER6		*		*		*		*	*	*		*
SDIR			*		*		*	*	*		*	*
EPOS AVG		*				*		*	*	*	*	*
TWAT AVG		*			*	*		*	*	*		*
POT AVG•			*	*		*	*	*			*	
RHCRI•			*		*	*	*	*			*	
VSHDC AVG	*	*				*		*				*
VPER18		*		*	*	*				*		
TIME		*			*	*		*		*		
VPER				*		*		*		*		*
T250P		*				*		*		*		*
EPOS			*			*	*	*				*
VMAX				*	*				*			*
POT2 AVG		*				*		*		*		
RHLORI•			*		*			*	*			
VPER24		*				*		*		*		
VPC20				*	*			*	*			
YDAYS		*				*		*		*		
LSHDC		*				*		*				
SHDC AVG		*				*		*				
T200 AVG		*				*		*				
PER24				*				*				*
SPDX					*				*			*
PSLV					*				*			*
SHDC						*		*				*
T200					*				*	*		
Z850					*			*				*
D200					*			*	*			
POT2						*		*				*
RHMD AVG								*	*			*
RHCN AVG•						*		*	*			
LON		*						*				
SHGC			*								*	
T250P AVG				*				*				
LAT					*			*				
ADAY								*	*			
RHMD					*							*
Z850 AVG					*				*			
LSHDC AVG						*		*				
RHCN								*				
SHGC AVG								*				
PCRI30•								*				
SBTRI•												
# of	3	23	16	10	21	28	14	43	21	19	15	22

Table 8: Continue

		Forward Rank			Backward Rank			Hill Climbing Rank			
Predictor	# Selected	SVM	NB	CART	SVM	NB	CART	SVM	NB	CART	Average Rank
PER18	10	44	2	10	44	2	43	20	1	22	20.89
PER6	9	2	20	44	20	22	42	21	13	44	25.33
GSTD	9	44	8	44	15	14	24	17	44	11	24.56
D200 AVG$^{\bullet}$	9	44	18	44	16	24	36	18	14	44	28.67
PC20	8	44	23	4	44	27	35	44	19	44	31.56
POT	8	44	22	44	3	44	28	9	18	44	28.44
SDIR AVG	8	44	17	44	44	21	33	44	12	7	29.56
VSHDC	8	44	44	44	17	44	20	19	17	14	29.22
PER•	7	3	15	44	44	5	44	44	4	44	27.44
TWAT	7	44	1	44	44	44	3	5	44	2	25.67
VPER6	7	44	12	8	44	15	41	44	10	21	26.56
SDIR	7	44	44	44	6	44	19	7	44	3	28.33
EPOS AVG	7	44	19	44	44	19	26	15	15	13	26.56
TWAT AVG	7	44	7	44	8	13	31	1	8	5	17.89
POT AVG$^{\bullet}$	6	44	44	5	44	28	32	44	44	44	36.56
RHCRI$^{\bullet}$	6	44	44	44	7	6	16	44	44	44	32.56
VSHDC AVG	5	1	16	44	44	20	15	44	44	15	27.00
VPER18	5	44	10	6	21	10	44	44	7	44	25.56
TIME	5	44	6	44	1	4	6	44	6	44	22.11
VPER	5	44	44	9	44	3	40	44	2	20	27.78
T250P	5	44	4	44	44	1	2	44	3	1	20.78
EPOS	5	44	44	44	44	12	9	44	44	9	32.67
VMAX	4	44	44	2	19	44	44	4	44	8	28.11
POT2 AVG	4	44	21	44	44	23	27	44	16	44	34.11
RHLOR	4	44	44	44	9	44	8	11	44	44	32.44
VPER24	4	44	9	44	44	16	21	44	9	44	30.56
VPC20	4	44	44	3	18	44	23	10	44	44	30.44
YDAYS	4	44	5	44	44	9	12	44	5	44	27.89
LSHDC	3	44	13	44	44	25	34	44	44	44	37.33
SHDC AVG	3	44	14	44	44	11	14	44	44	44	33.67
T200 AVG	3	44	11	44	44	17	29	44	44	44	35.67
PER24	3	44	44	7	44	44	39	44	44	19	36.56
SPDX	3	44	44	44	10	44	44	3	44	17	32.67
PSLV	3	44	44	44	14	44	44	16	44	10	33.78
SHDC	3	44	44	44	44	18	7	44	44	18	34.11
T200	3	44	44	44	4	44	44	13	11	44	32.44
Z850	3	44	44	44	12	44	5	44	44	4	31.67
D200	3	44	44	44	2	44	25	8	44	44	33.22
POT2	3	44	44	44	44	7	17	44	44	16	33.78
RHMD AVG	3	44	44	44	44	44	11	12	44	6	32.56
RHCN AVG$^{\bullet}$	3	44	44	44	44	8	4	6	44	44	31.33
LON	2	44	3	44	44	44	18	44	44	44	36.56
SHGC	2	44	44	44	44	44	44	44	44	44	44
T250P AVG	2	44	44	1	44	44	1	44	44	44	34.44
LAT	2	44	44	44	11	44	38	44	44	44	39.67
ADAY	2	44	44	44	44	44	13	2	44	44	35.89
RHMD	2	44	44	44	5	44	44	44	44	12	36.11
Z850 AVG	2	44	44	44	13	44	44	14	44	44	37.22
LSHDC AVG	2	44	44	44	44	26	37	44	44	44	41.22
RHCN	1	44	44	44	44	44	10	44	44	44	40.22
SHGC AVG	1	44	44	44	44	44	22	44	44	44	41.56
PCRI30$^{\bullet}$	1	44	44	44	44	44	30	44	44	44	42
SBTRI$^{\bullet}$	0	44	44	44	44	44	44	44	44	44	44.00
# of											

average rank is higher than the other features, which even were chosen less often. On the other hand, TWAT AVG has the best average rank, but was only chosen by 7 models. Some of these discrepancies are due to the fact that variables might contain redundant information and different models choose only one of the redundant variables.

Feature extraction methods: Principle Components Analysis (PCA) (Dunteman, 1989) converts a dataset into a new set of un-correlated features called Principle Components (PCs). Depending on the rank of the original matrix, the number of PCs is less or equal to the number of the original features. Principal components are a linear combination of original features calculated using the eigenvectors of the covariance matrix. The principle components are ordered such that the first principle component explains the highest degree of variability in the data and the last principle component explains the least amount. PCA can be used for

Table 9: Best model AUC (and standard deviation for cross validation) and the selected number of selected principal components. Highlighted are the three best models

Classifier	AUC	#PCs	AUC(normalized)	#PCs
SVM	0.865 (0.017)	25	0.875(0.030)	23
LR	0.901(0.018)	38	0.891(0.022)	25
NB	0.832(0.021)	5	0.866(0.017)	8
CART	0.834(0.028)	6	0.842(0.043)	27

Table 10: Best model AUC (and standard deviation for cross validation) and the selected number of selected principal components. Highlighted are the three best models

Classifier	Nor. SVM	Nor. LR	Nor. NB	Nor. CART	Non-nor. SVM	Non-nor. LR	Non-nor. NB
Nor. LR	0						
Nor. NB	0	0					
Nor. CART	0	0	0				
Non-nor. SVM	1	0	0	0			
Non-nor. LR	0	0	0	0	0.11		
Non-nor. NB	0	0	0	0	0	0	
Non-nor. CART	0	0	0	1	0	0	0.02

Table 11: P-values of every pair of the PCA models adjusted using the Bonferroni correction. Significant differences (p-values<0.05) are highlighted

Classifier	Nor. SVM	Nor. LR	Nor. NB	Nor. CART	Non-nor. SVM	Non-nor. LR	Non-nor. NB
Nor. LR	0						
Nor. NB	0	0					
Nor. CART	0	0	0				
Non-nor. SVM	1	0	0	0			
Non-nor. LR	0	0	0	0	0.11		
Non-nor. NB	0	0	0	0	0	0	
Non-nor. CART	0	0	0	1	0	0	0.02

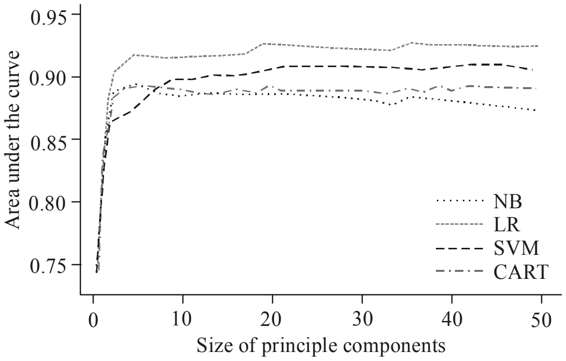

Fig. 5: AUC by number of used principal components

dimensionality reduction by only using a certain number of the top PCs.

We ran LR, NB, SVM and CART using the extracted PCs. To decide ondetermine the number of PCs to use, we apply k-fold cross validation while increasing the set of PCs. The set of PCs that produces the highest AUC is chosen. For our first experiment, we normalized the values of the features to z-scores before extracting the PCs (Fig. 6). For the second experiment, we extracted the PCs without normalization (Fig. 5). The best number of PCs for each model is presented in Table 9 and 10. It shows that the normalization has improvesd the performance forall but logistic regression (LR). It also shows that the number of selected PCs varies significantly from one model to another and that, similar to the feature selection experiments, LR outperforms the other classifiers.

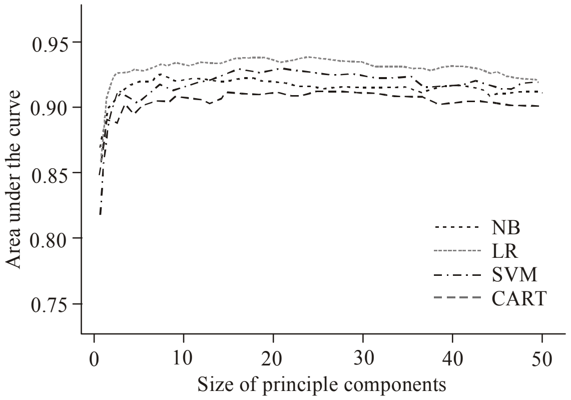

Fig. 6: AUC by number of used principal components (normalized features)

The p-value of every pair of the feature selection models is shown in Table 11. The results show almost all differences are significant, indicating that normalized LR is better than all other models with the exception of non-normalized LR.

RESULTS AND DISCUSSION

For model comparison, we use observations from 2009 and 2010 for testingto test the performance of RI models after applying feature selection and extraction on data from 1982 to 2008. We will also evaluate different ensemble learning techniques whichthat combine the results of the base classifiers investigatedso thus far.

Table 12: AUC scores of SVM, LR, N B, and CAR Tusing different feature subsets (by selection method) for the 2009 and 2010 hurricane seasons. The RII model achieves a value of 0.743

Classifier	Hill C.	BFS	FFS	Nor. PCs	Non-nor. PCs
SVM	0.801	0.812	0.854	0.845	0.847
LR	0.896	0.902	0.897	0.900	0.895
NB	0.868	0.864	0.868	0.843	0.770
CART	0.820	0.850	0.815	0.780	0.852

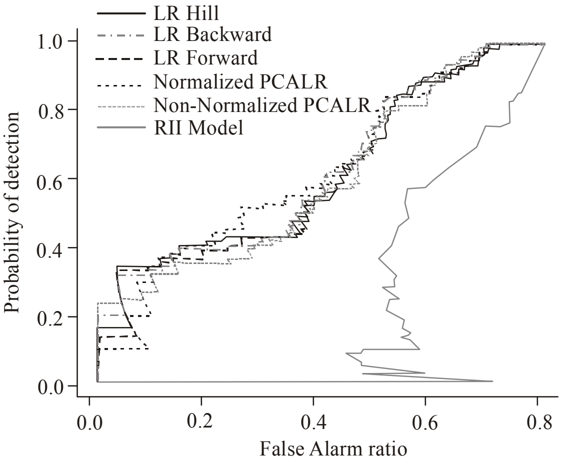

Fig. 7: Comparison of POD and FAR for different models

Test results for classification models: We evaluate SVM, LR, NB and CART to predict RI events in the Atlantic Basin using the subset of predictors produced by each feature selection and extraction method discussed above. The years 2009 and 2010 are used for testing and the years 1982 to 2008 are used for training. This testing method is used sincebecause it emulates the real application case where only data up to the current year isare available for training. Table 12 summarizes the test results for the 2008 and 2009 hurricane seasons. The table shows that LR reaches high AUC values compared to the other models despite of the selected subset used in building the model. The POD and FAR results of the LR models are shown in Fig. 7. The results closer to the top left corner with high POT and low FAR are better. It They also show that all models improve over the RI model with a 30-knots threshold.

Test results for ensemble learning: A method to incorporate the results of multiple models is Ensemble Learning (EL) (Maimon and Rokach, 2005). EL is a technique that combines the prediction of different models aiming to produce a more stable and often even better performing aggregate model. The idea is to build a two-stage model. In the first stage, each classifier, which is called a weak learner, produces a decision that is more efficient than a random guess. In the second stage, the, results are aggregated to produce a single decision. Maimon and Rokach (2005) mentioned that there is amentionedthe disagreement about the rightcorrect number of classifiers used for ensemble learning. They pointed outnoted a paper that states that ten classifiers are enough to enhance the results of an ensemble and another paper that proves that AdaBoost in particular requires as large manyas 25 classifiers to improve its performance. We used a value between these of 20 classifiers for our customized ensemble averaging and stacking models.

There existare several EL aggregation techniques. Here we examine the following techniques: averaging, weighted averaging, Bayesian Model Averaging (BMA), bagging, Random Forest (RF), boosting and stacking. The first stage of ensemble learning for averaging, weighted averaging and BMA is identical, where the different classifiers are run for the years 2009 and 2010. The second stage of the ensemble learning combines the probabilities produced by the classifiers built in the first stage as follows:

- Averaging (Maimon and Rokach, 2005; Galar *et al.*, 2012) assumes that all models carry the same importance and simply averages the RI probabilities produced by the individual classifiers as:

$$P_{average} = \frac{P_1+P_2+\cdots+P_n}{n} \tag{8}$$

where, P$_i$, with i = 1, 2, ... , n, is the probability of RI of model i and n is the number of classifiers.

- Weighted averaging (Maimon and Rokach, 2005) is similar to simple ensemble averaging except that each classifier has a relative weight w_i associated with it. The weight represents the importance of the classifier, so a classifier with a higher weight has more influence on the decision of the ensemble averaging. It combines the models as:

$$P_{weightedavg} = \frac{w_1 \times P_1+w_2 \times P_2+\cdots+w_n \times P_n}{n} \tag{9}$$

where, w_i represents the weight and $\sum_{i=1}^{n} w_i = 1$.Rescaled accuracy measures of the models are usually used as weights. We used the AUC measure instead of accuracy since because it is not influenced by class imbalance.

The AUC ranges between 0.77 and 0.90, which gives all models nearly equally high weights. To increase the influence of the highly performing models and decrease the less performing ones, we scale the weights between zero and one before averaging the models.

$$wi = \frac{AUCi-min(AUC)}{max(AUC)-min(AUC)} \tag{10}$$

where, *AUC = (AUC$_1$, AUC$_1$, ... , AUC$_n$)* and w_i is the *i*th scaled weight.

- Like the weighted averaging technique, BMA (Maimon and Rokach, 2005) applies weights for each model. The used weight used is the posterior probability of the model given the training data instead of the AUC values and is calculated using the Bayes' theorem formula:

$$w_i = P(M_i|X) = \frac{P(X|M_i)P(M_i)}{P(X)} \tag{11}$$

where, $M_i$represents model *i* and *X* represents the training data. *P(X|M$_i$)* is the conditional probability of the training data given the model, *P(M$_i$)* is the prior probability and *P(X)* is new evidence that has a constant value.

Bagging and random forest build a set of decision trees to predict RI in the first stage of the ensemble learning method. Bagging (Maimon and Rokach, 2005) draws m random subsets with replacement (bootstrap samples) from the training set. Each subset is of size *n*, which is equal to the size of the original dataset and is used to train *m* different classifiers. RF (Breiman, 2001) creates *m* random subsets that have the same probability distributions of size *n*. The major difference between RF and bagging is that the RF algorithm randomly draws a different set of features for each decision tree, creating more variety between the trees.In the second stage, the trees are aggregated by simply averaging the probabilities of the RI obtained from each tree.

Adaptive boosting (AdaBoost) (Maimon and Rokach, 2005) starts with a weak learner and then improves in subsequent training rounds, the performance of the model in subsequent training rounds by iteratively oversampling misclassified training instances. This procedure creates a series of models where each model concentrates more and more on hard-to-classify instances. In our experiment, we performed 30 iterations and therefore built 30 models.

The first stage of stacking builds a set of weak learners. However, in contrast to other EL methods, a classification model is used for aggregation. The basic idea is that this model will learn the best waymethod of aggregation. To perform the first stage of our stacking experiment, we divided the training set from 1982 to 2008 into 5-folds and in each round, one fold was used for testing and the remaining folds were used for training. The probabilities of RI for the five folds are predicted by the used base classifiers SVM, LR, NB, CART, RF, bagging and adaBoost. These predictions represent the training set that is used in the second stage to learn the aggregation model. SVM, LR, NB and CART were used as aggregation models and the performance was evaluated using the holdout samples of the years 2009 and 2010.

Table 13: AUC measures for performing ensemble learning methods. Highlighted are the best models are highlighted

Ensemble Learning Method	AUC
Averaging	0.879
Weighted Average (WA)	0.88
Scaled WA	0.885
BMS WA	0.89
Random Forest	0.88
Bagging	0.856
AdaBoost	0.866
Stacking SVM	0.762
Stacking LR	0.879
Stacking CART	0.862
Stacking CART	0.857

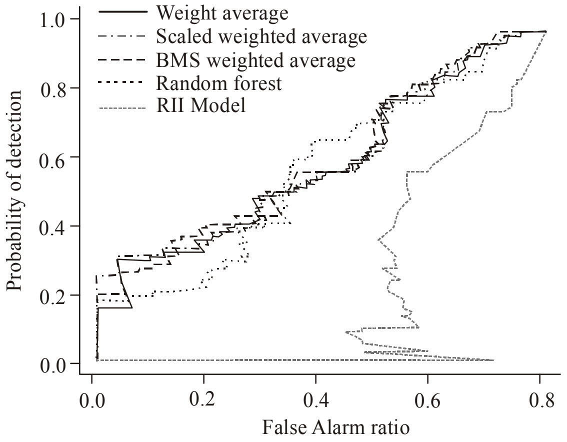

Fig. 8: POD and FAR for the best performing ensemble learning techniques

The results of the different ensemble techniques are shown in Table 13. The results show that simple averaging, weighted averaging, scaled weighted averaging, BMS weighted averaging and random forest are the best performing models. Figure 8 is a representation of the results of POD and FAR for the best performing ensemble learning methods along with the RII model.

Weighted intensity forecast: An important application of using of RI predictions is to improvethe intensity forecasts. An effort to improve the intensity prediction of a hurricane was done undertaken by Kaplan *et al.* (2015) by incorporating the change in intensity forecast of the 24-h RII model of 25, 30, 35 and 40 knot thresholds with the variable intensity five-member consensus (IVCN) forest model. The combined model of IVCN and RII is called RAPID. IVCN (National Hurricane Center, 2009) is an ensemble of five different official NHC models that averages their available forecasts. At least two forecasts have to be present for the model to work. The five models are Decay-SHIPS (DSHP), Logistic Growth Equation Model (LGEM), interpolated NWS/Geophysical Fluid Dynamics Laboratory (GHMI), interpolated NWS/Hurricane Weather Research and Forecasting Model (HWFI) and the navy version of the interpolated NWS/Geophysical Fluid Dynamics Laboratory (GFNI). If RII models predict an RI probability of greater than 40%, then the IVCN intensity prediction is increased by averaging it with the largest threshold value of the RII models. This method showed a slight improvement to the change in intensity forecast of the 24-h rapidly intensifying events (Kaplan *et al*., 2015).

For simplicity, we use here instead of an ensemble, only we use a Multiple Linear regression (LM) model using all the…..For simplicity, instead of an ensemble, we use a multiple linear regression (LM) model using all the features introduced above. This model is very similar to the operational intensity model SHIPS. First, weusea method very similar to RAPID, but we average the highest threshold among RI30, RI35 and RI40 of RII that exceeds 40% with the linear regression (LM) model and call it LM RII.

Next, we present a new way of how to use using the probabilities of RI learned from the weighted average. The EL model (with SVM, LR, NB, CART, RF, adaBoost and bagging) can be used to improve intensity forecasts. We first learn two LM intensity forecast models. One only usesd NRI observations from the training set and the other only uses the RI observations. This method produces the two following models: LMRI and LMNRI. For each test instance, we predict the probability of RI using the EL model and the increase in intensity$\Delta VMAX$using LMRI and LMNRI. We combine LMRI and LMNRI predictions using the probabilities of RI and α that range from 0 to 10. α is used to increase or decrease the effect of the intensity models. SoThus, if we want to give the LMRI model highergreater influence, we decrease α. The prediction of ΔVMAX is defined as follows:

$$\Delta VMAX = {P_{RI}}^{\alpha} \times \Delta VMAX_{LMRI}$$
$$+(1 - {P_{RI}}^{\alpha}) \times \Delta VMAX_{LMNRI} (12)$$

Figure 9 presents the predicted changes in intensities Mean Absolute Error (MAE) of the intensities. The MAE on the predictions' RI observations is plotted against MAE on the NRI observations. This method illustrates how the model improves the intensity forecast on RI observations and how it affects the forecast of NRI observations. The square point in each figure represents the LM model learned on all (RI and NRI) instances. It has a high RI observation mean absolute error compared to the NRI observations. This results from the fact that NRI cases are overrepresented in the dataset. The circle point in Fig. 9 represents the LM_RII model. It has lower error when forecasting RI instances compared to pure LM but slightly higher error when forecasting NRI instances. The points connected by the line represent our proposed model (LM_alpha_RI). Each point on the line represents a different α value. It shows improvement over the other models with lower RI and NRI forecast error.

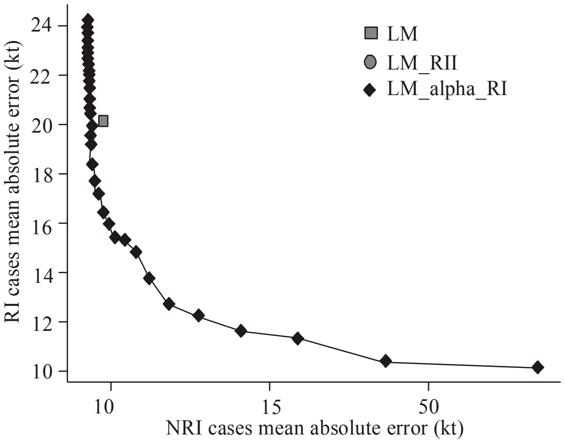

Fig. 9: Mean absolute error of RI cases vs. NRI cases intensity forecast

CONCLUSION

The investigation in this study indicates that machine learning methods have the potential to significantly improve rapid intensification event prediction. It is interesting to note that each of the feature selection methods (forward, backward and hill climbing) selected very different subsets of features, but they all selected at least one of the previous changes in intensity predictors. However, not all predictors deemed important by the official RII model were selected by the models.

The simple Ensemble Learning (EL) technique based on the Bayesian Model Averaging (BMA) showed a relatively high performance with an AUC of 0.89, which is very similar to the performance of the logistic regression model alone. Balancing the data using SMOTE had positive effects for SVM and CART but did not affect the performance of LR and NB.

The results of this investigation show that using machine learning methods have the potential to improve the status quo for predicting rapid intensification events. In addition, we introduced a new method to improve storm intensity prediction by combining several intensity models with a rapid intensification model. An advantage of this method over RAPID is that it can learn to which degree the degree to which intensity will increase in RI instances without having multiple thresholds for a certain time interval. As part of this study, we are releasing the useddata set used in this study with the hope to of encouragingspark more research in this important application area.

REFERENCES

Breiman, L., 2001. Random forests. Mach. Learn., 45(1): 5-32.

Breiman, L., J. Friedman, C.J. Stone and R.A. Olshen, 1984. Classification and Regression Trees. Wadsworth, Belmont, Calif.

Cortes, C. and V. Vapnik, 1995. Support-vector networks. Mach. Learn., 20(3): 273-297.

DeMaria, M., 2013. Ships predictor file. RAMMB Technical Report, Colorado State University, Regional and Mesoscale Meteorology Branch (RAMMB) of NOAA/NESDIS. 2013. Retrieved form: http://rammb.cira.colostate.edu/research/tropical_cyclones/ships/docs/SHIPS_predictor_file_2013.doc.

DeMaria, M., M. Mainelli, L.K. Shay, J.A. Knaff and J. Kaplan, 2005. Further improvements to the statistical hurricane intensity prediction scheme (SHIPS). Weather Forecast., 20(4): 531-543.

Dunteman, G.H., 1989. Principal Components Analysis. Sage University papers Series, Newbury Park, Calif.

Galar, M., A. Fernandez, E. Barrenechea, H. Bustince and F. Herrera, 2012. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. IEEE T. Syst. Man Cy. C, 42(4): 463-484.

Greiner, R., 1992. Probabilistic hill-climbing: Theory and applications. Proceeding of the 9th Biennial Conference of the Canadian Society for Computational Studies of Intelligence (CSCSI-92), pp: 60.

Kaplan, J. and M. DeMaria, 1995. A simple empirical model for predicting the decay of tropical cyclone winds after landfall. J. Appl. Meteorol., 34(11): 2499-2512.

Kaplan, J. and M. DeMaria, 2003. Large-scale characteristics of rapidly intensifying tropical cyclones in the north Atlantic basin. Weather Forecast., 18(6): 1093-1108.

Kaplan, J., J.J. Cione, M. DeMaria, J. Knaff, J. Dunion *et al*., 2010a. 9C.4 Enhancements to the operational ships rapid intensification index. Proceeding of the 29th Conference on Hurricanes and Tropical Meteorology, Rickenbacker Causeway.

Kaplan, J., M. DeMaria and J.A. Knaff, 2010b. A revised tropical cyclone rapid intensification index for the Atlantic and eastern north pacific basins. Weather Forecast., 25(1): 220-241.

Kaplan, J., C.M. Rozoff, M. DeMaria, C.R. Sampson, J.P. Kossin, C.S. Velden, J.J. Cione, J.P. Dunion, J.A. Knaff, J.A. Zhang *et al*., 2015. Evaluating environmental impacts on tropical cyclone rapid intensification predictability utilizing statistical models. Weather Forecast., 30(5): 1374-1396.

Kieper, M.E. and H. Jiang, 2012. Predicting tropical cyclone rapid intensification using the 37 GHz ring pattern identified from passive microwave measurements. Geophys. Res. Lett., 39(13).

Lippsett, L., 2011. Gliders tracked potential for oil to reach the east coast. OCEANUS Mag., 48(3).

Maimon, O. and L. Rokach, 2005. Data Mining and Knowledge Discovery Handbook. Springer-Verlag, New York, Inc., Secaucus, NJ, USA.

Meisner, B.N., 2006. An overview of NHC prediction model. NOAA Technical Attachment SR/SSD 95-36.

National Hurricane Center, 2009. NHC Track and Intensity Models. pp: 18. Retrieved form: http://www.nhc.noaa.gov/modelsummary.shtml. (Accessed on: September 17, 2013)

Rhome, J.R., 2007. Technical summary of the national hurricane center track and intensity models. Technical Report, National Hurricane Center, 2007. Retrieved form: http://www.nhc.noaa.gov/pdf/model_summary_20070912.pdf.

Rozoff, C.M. and J.P. Kossin, 2011. New probabilistic forecast models for the prediction of tropical cyclone rapid intensification. Weather Forecast., 26(5): 677-689.

Russell, S. and P. Norvig, 1995. Artificial Intelligence: A Modern Approach. Prentice-Hall, Egnlewood Cliffs, NJ.

Schott, T., C. Landsea, G. Hafele, J. Lorens, A. Taylor, H. Thurm, B. Ward, M. Willis and W. Zaleski, 2012. The Saffir-Simpson Hurricane Wind Scale. National Weather Services, National Hurricane Centre, National Oceanic and Atmospheric Administration (NOAA) Factsheet. Retrieved form: http://www.nhc.noaa.gov/pdf/sshws.pdf.

Shay, L.K., G.J. Goni and P.G. Black, 2000. Effects of a warm oceanic feature on hurricane opal. Mon. Weather Rev., 128(5): 1366-1383.

Tan, P.N., M. Steinbach and V. Kumar, 2005. Introduction to Data Mining. 1st Edn., Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Torgo, L., 2010. Data Mining with R: Learning with case studies. Chapman and Hall/CRC, Taylor & Francis Group, Boca Raton, London, New York. Retrieved form: http://www.dainf.ct.utfpr.edu.br/~kaestner/Mineracao/RDataMining/Data%20Mining%20with%20R-Kumar.pdf.

Walker, S.H. and D.B. Duncan, 1967. Estimation of the probability of an event as a function of several independent variables. Biometrika, 54(1-2): 167-179.

Wang, B. and X. Zhou, 2008. Climate variation and prediction of rapid intensification in tropical cyclones in the western north pacific. Meteorol. Atmos. Phys., 99(1): 1-16.