15D.6 TROPICAL CYCLONE INTENSITY VIA DEEP LEARNING

Christopher C. Hennon * and Connor Algeri
Embry-Riddle Aeronautical University, Prescott, Arizona

1. INTRODUCTION

The Dvorak Technique (DT, Dvorak 1975, 1984) has long filled the void of reasonable tropical cyclone (TC) intensity (the maximum sustained wind speed) estimation. Despite its known shortcomings and subjectivity, it has been employed globally and has produced a reasonably consistent data record of TC intensity for several decades.

The "true" TC intensity is unknowable, even in storms with extensive reconnaissance coverage. Nevertheless, we can arrive at a reasonable estimate for technique skill by comparing Dvorak (or any technique) classifications to those storm times with recon-influenced data. For example, Knaff et al. (2010) determined a root mean square error (RMSE) of 10-15 kt when evaluating Dvorak classifications to best track fixes made within 2 hours of a recon flight.

Over the last two decades, significant progress has been made in both improving the technique and augmenting its estimates with other sources. For example, the Advanced Dvorak Technique (ADT, Orlander and Velden 2007) and its machine learning cousin (AiDT, Orlander et al. 2021) have eliminated the subjectivity of the DT and lowered RMSE errors.

The rise of artificial intelligence (AI) and specifically neural network (NN) applications in science has engendered new avenues of research in TC estimation. Several of these (e.g., Chen et al. 2019; Fu et al. 2024) use multiple satellite channels and report RMSE values on the order of 8-10 kt - notably lower than DT/ADT and comparable to AiDT.

However, there are shortcomings to approaches that use multiple channels and/or human identification of relevant TC features. First, multi-channel techniques may not be transferrable to an operational setting since all channels will not be available at any given analysis time. Second, any technique that relies on humans to extract relevant features of a TC image (e.g., ADT, AiDT) will not be able to replicate the efficiency and efficacy of a neural network.

Convolutional neural networks (CNNs) are the most popular means of image classification. CNNs identify relevant features in an image – an obvious example would be a TC eye – and over several layers optimizes ("convolves") the influence of those features on the classification.

Orlander et al. (2021) argue that CNNs are not practical to use in TC intensity estimation because: they are generally poor at generalization, require large amounts of training data, and have high computational costs. All these can be easily overcome; this has been demonstrated in recent work by Pradhan et al. (2018) and others. These studies showed that a CNN trained exclusively on IR images can yield skillful classifications with minimal investment. Once a CNN is trained, its application to new data is instantaneous.

Here we present a machine-learning algorithm to classify TC intensity that updates and improves upon the IR-exclusive work mentioned above. The algorithm employs both a CNN for image classification and a feed forward back propagation (FFBP) NN to include additional information into the system. The data and methods will be presented next, followed by the NN architectures and training. Results of the experiment are provided in Section 5.

2. DATA AND METHODS

The quantity and quality of the image data is critical to the success of the CNN classification. Generally, CNNs require large amounts of data to better "learn" features and increase generalization as well as classification skill. Furthermore, it is desirable to have a classifier with global scope, so that classifications can be made for any TC basin.

2.1 *HURSAT B1 Data*

The HURSAT B1 v07 dataset (Knapp and Kossin 2007) satisfies these criteria. HURSAT is a global collection of geostationary channels that corrects for intersatellite differences. Its long data record (1978-2023) and storm-centric geolocation are other features that make HURSAT an ideal dataset for this kind of study. One potentially challenging aspect is the relatively coarse (8 km) image resolution, which is necessary to include imagery from earlier eras but may likely degrade the ability of the CNN to identify relevant features.

Images were generated directly from the HURSAT "IRWIN" (IR) data for storms the met the following criteria: maximum sustained winds >= 23.5 kt, located over water, and located in the tropics (defined as equatorward of a

* *Corresponding author address:* Christopher C. Hennon; e-mail: chrishennon@yahoo.com

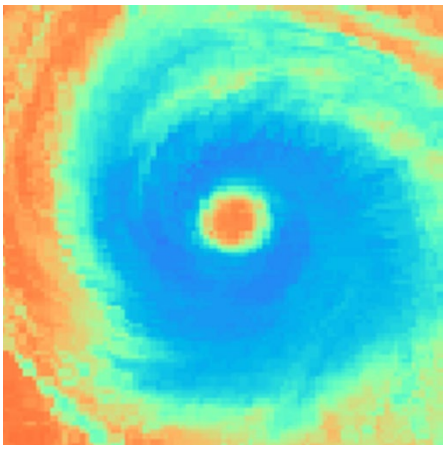

Figure 1. HURSAT B1 image of Hurricane Isabel (2003).

selected latitude which varies by basin). Outlines of land masses as well as latitude/longitude lines are excluded as they may confuse the classifier. The final images are color enhanced ("rainbow") with a size of 224x224 pixels, corresponding to a scene of approximately 700 km across. An example of an image is shown in Figure 1.

2.2 *Data Preparation*

Several image processing steps were executed in preparation for training. First, all southern hemisphere storms were flipped to mimic northern hemisphere circulations. Next, the images were quality controlled to remove bad renderings and storms that were obviously not in the center part of the image. Finally, the images were grouped into twenty categories, with each category representing a 5 kt bin of maximum sustained wind ("Windspd" variable in HURSAT). The categories ranged from 25 kt to 120+ kt. All storm times with winds higher than 120 kt were grouped into the last bin since there were not enough images to classify them into their own category. Figure 2 shows the distribution of images by

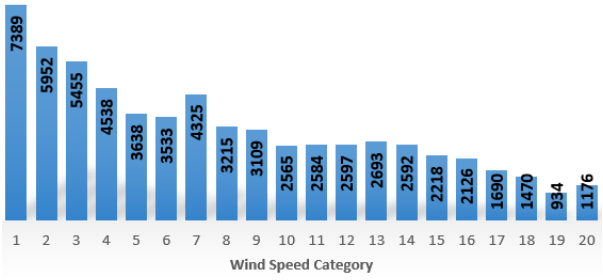

Figure 2. Number of HURSAT images by wind category.

category. As expected, lower wind speed categories contain significantly more images than higher wind speeds – and thus are theoretically easier to classify.

Finally, images were prepared for CNN training. To provide more classification and generalization power, each unique image was rotated 90°, 180°, and 270°.

The final dataset consists of 63,799 unique storm images from all TC basins. After rotation, 255,196 images were presented to the CNN. During CNN training, additional randomized rotations and zooms were employed to further enhance learning. The imagery was randomly divided into training (80%) and validation (20%) datasets prior to training.

2.3 *Independent Test Datasets*

Two groups of imagery were intentionally held out of the training process to be used as an independent test of classification skill: Extended Best Track (v1.0, 1,738 images) and the 2023 global TC year (4,461 images). The extended best track data includes all North Atlantic TC times that were influenced by reconnaissance surveillance, defined as ±12 hours from the image time.

3. NEURAL NETWORK ARCHITECTURES

3.1 *Convolutional Neural Network*

The CNN used in this experiment was designed from scratch but mimics other standard configurations for image classification. Figure 3 shows a diagram of the architecture, which contains 43 layers and nearly 20 million learnable parameters (weights and biases). Convolution layers are sometimes followed by "Pooling" layers, which reduce the dimensionality of the previous layer(s) while retaining the important features of the layer. Here we use "max" pooling, which selects the maximum value from the region surveyed by the pooling filter.

Following the convolution layers, the network finishes an iteration by employing fully connected, dropout, and softmax layers. The fully connected layer serves as a bridge between input neurons and the output layer by transforming all possible weights and biases from all input neurons to every neuron in the output layer. The dropout layer randomly "turns off" 15% of the neurons with the intention of reducing the tendency of CNNs to overfit to the training data. The softmax layer transforms the output of the CNN into probabilities of belonging to a particular category. Batch normalization and ReLU activation layers follow each convolution layer (not shown).

Typically, a CNN will run through many rounds ("epochs") of weight adjustments to reach a configuration with minimum classification error. We ran the CNN for 128 epochs.

3.2 *Feed Forward Back Propagation*

After the CNN classifications are complete, further refinement of the category predictions are performed by

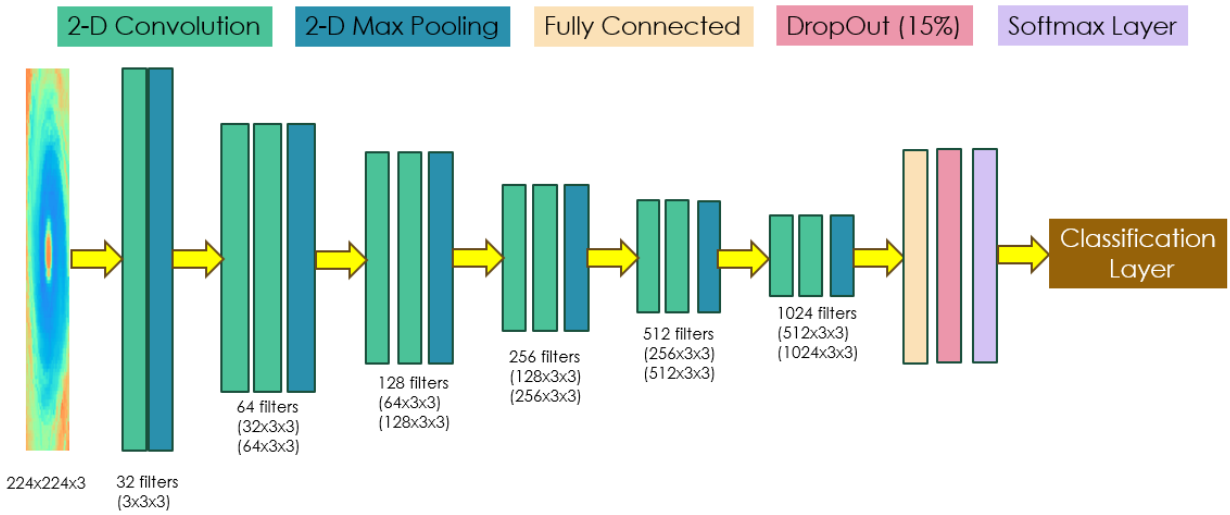

Figure 3. Layer architecture of the CNN. Batch normalization and ReLU layers follow each convolutional layer.

considering additional "feature data" (other metadata for each image). The feature data are: latitude, longitude, julian day, satellite viewing angle, and the CNN wind speed classification. A feed forward back propagation NN is used to classify the wind speed given this additional information. A simple 3-layer model with 10 hidden neurons is used with a ReLU activation function, softmax output activation, and quasi-Newton backpropagation.

4. NEURAL NETWORK TRAINING

The CNN was designed and executed using Matlab software. The network was tuned by running dozens of training sessions with varying hyperparameter configurations on a standard Linux workstation running a NVIDIA RTX A4500 graphics card with over 7,000 cores. A training session took approximately 23 hours to complete. A constant learning rate of 0.001 was used in the final configuration. The CNN achieved 87% training accuracy and 82% validation accuracy.

5. RESULTS

The CNN chooses the classification category with the highest output probability. Both test datasets (extended best track and the 2023 season, see Section 2.3) were classified and examined. Figure 4 shows the confusion matrix (predicted vs. actual category) of the extended best track (top) and 2023 season (bottom). The values on the figure represent the normalized selections of each category (i.e., 1 = 100% success). Note that the extended best track test produced a much lower bias (closer to the perfect line) and less spread than the 2023 season. This is likely due to the training dataset containing images from the same storms in the extended best track, just not within 12 hours of a reconnaissance survey. The 2023 season shows an overall low intensity bias, which appears consistent across all categories.

By using the probabilities of an image belonging to each category 'c' (P_C), we can determine the RMSE and bias for each test dataset by first calculating the wind speed error (WSE) prediction (kt):

$$\text{WSE (kt)} = \left(\sum_{c=1}^{c=20}(P_C * PWS_C) - \text{AWS}\right) * 5\text{kt} \quad (1)$$

where PWS = predicted wind speed category and AWS = actual wind speed category

Wind speed errors are summed (for the bias) and squared and summed (for the RMSE) for all images in the test dataset. The RMSE and bias are then calculated as follows:

$$\text{RMSE (kt)} = \sqrt{\frac{\sum_{i=1}^{N}(WSE)^2}{N}} \quad (2)$$

$$\text{Bias (kt)} = \sum_{i=1}^{N}(WSE) \quad (3)$$

Where N = number of images in the test dataset. The RMSE/Bias for the extended best track (2023 season) datasets are 10.68 kt/-2.6 kt (12.76 kt/-4.2 kt).

The further refinement of classifications in the feed forward neural network reduced the RMSE and bias significantly. Table 1 shows the improvement in RMSE (~15%) and bias (80-95%) from the addition of feature data to the CNN classifications.

6. DISCUSSION AND CONCLUSION

Other efforts to classify TC intensity based on satellite imagery have shown various levels of success. In general, RMSE values range from 8-15 kt, with multi-

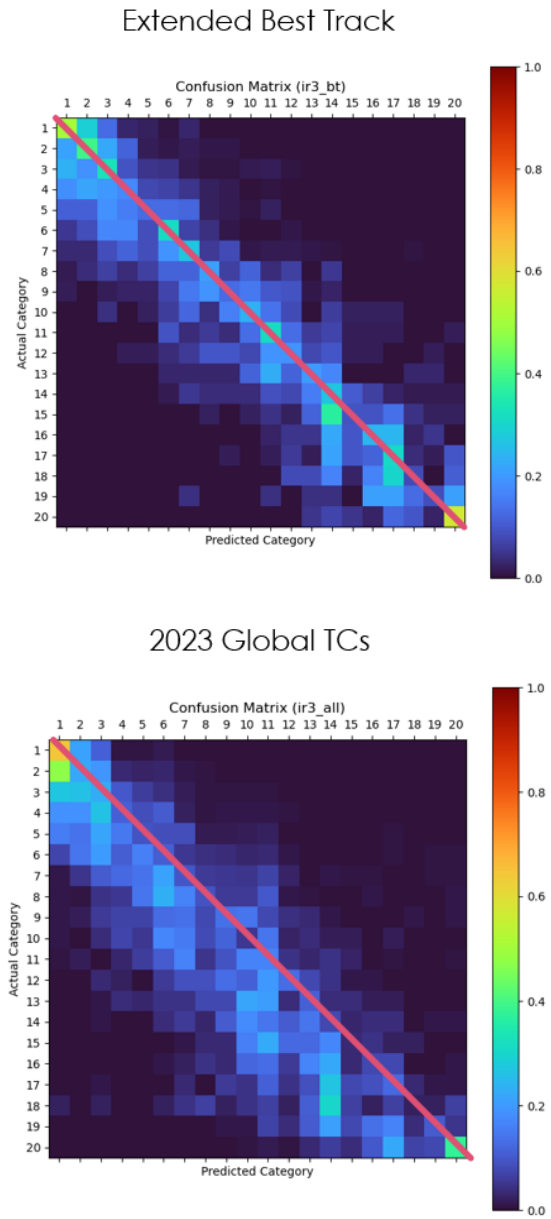

Figure 4. Confusion matrices for extended best track test (top) and the 2023 global TC season (bottom).

channel techniques (e.g., AiDT) and consensus methods (SATCON) exhibiting the lowest errors (8-9 kt). But these techniques may not always have the required information available to them to make an operational decision.

When compared to work using a singular IR channel for classifications, our RMSE (9-11 kt) and low bias are quite favorable. This is quite promising given that the IR imagery used here is very coarse in resolution (8 km) compared to the other methods. With nearly a dozen published techniques showing lower RMSE values than the original DT, one must wonder whether it is time to move on from the gray-scaled imagery of the past and embrace our AI tools. Ensemble TC estimation using this, and other established techniques already surpass the manual DT and the gap will continue to grow.

Extended BT	CNN Only	CNN + FFBP	Improvement
RMSE	10.68 kt	**9.15 kt**	14%
Bias	-2.60 kt	-0.14 kt	95%
2023 Season	**CNN Only**	**CNN + FFBP**	**Improvement**
RMSE	12.76 kt	**10.83 kt**	15%
Bias	-4.20 kt	0.78kt	81%

Table 1. RMSE and bias scores (kt) for CNN classifications and CNN+FFBP classifications.

7. REFERENCES

Chen, B-F, B. Chen and H-T Lin, and R.L. Elsberry, 2019: Estimating tropical cyclone intensity by satellite imagery utilizing convolutional neural networks. *Wea. Forecasting,* **34**, 447-465.

Dvorak, V., 1975: Tropical cyclone intensity analysis and forecasting from satellite imagery. *Mon. Wea. Rev.,* **103**, 420-430.

Fu, R., H. Hu, N. Wu, Z. Liu, and W. Jin, 2024: Spatiotemporal fusion convolutional neural network: Tropical cyclone intensity estimation from multisource remote sensing images. *J. Appl. Rem. Sens.,* **18**.

Knaff, J.A., D.P. Brown, J. Courtney, G.M. Gallina, and J.L. Beven, 2010: An evaluation of Dvorak Technique-based tropical cyclone intensity estimates. *Wea. Forecasting,* **25**, 1362-1379.

Knapp, K.R., and J.P. Kossin, 2007: New global tropical cyclone data from ISCCP B1 geostationary satellite observations. *J. Appl. Rem. Sens.,* **1**.

------, 1984: Tropical cyclone intensity analysis using satellite data. NOAA Tech. Rep. NESDIS 11, 47 pp.

Olander, T.L., and C.S. Velden, 2007: The Advanced Dvorak Technique: Continued development of an objective scheme to estimate tropical cyclone intensity using geostationary infrared satellite imagery. *Wea. Forecasting,* **22**, 287-298.

------, A. Wimmers, C. Velden, and J.P. Kossin, 2021: Investigation of machine learning using satellite-based Advanced Dvorak Technique analysis parameters to estimate tropical cyclone intensity. *Wea. Forecasting,* **36**, 2161-2186.

Pradhan, R., R.S. Aygun, M. Maskey, R. Ramachandran, and D.J. Cecil, 2018: Tropical cyclone intensity estimation using a deep convolutional neural network. *IEEE Trans. on Image Processing,* **27**, 692-702.