Scorpions: Classification of poisonous species using shape features

J. Carlos Urteaga-Reyesvera
Telematics Engineering
Instituto Tecnologico Autonomo de Mexico
Email: jorge.urteaga@itam.mx

Andre Possani-Espinosa
Department of Digital Systems
Instituto Tecnologico Autonomo de Mexico
Email: andre.possani@itam.mx

***Abstract*—All around the world, poisonous scorpions are still considered as a public health issue. The scorpion's species can be determined by its physical characteristics. Different methods have been applied to differentiate among different insects, such as bugs, bees and moths. However, none have been applied to distinguish between different scorpion species. This paper presents a procedure to distinguish between two different species of scorpions (Centruroides limpidus and Centruroides noxius) using image processing techniques and three different machine-learning methods. First, the live scorpion is distinguished from the photograph image using a dynamic separation threshold obtaining its area and contour. A shape vector is obtained from both, area and contour, calculating the following features: aspect ratio, rectangularity, compactness, roundness, solidity and eccentricity. Finally, artificial neuronal network, classification and regression tree, and random forest classifiers are used to differentiate between both species. All three classifiers were evaluated by accuracy, sensitivity and specificity. Experimental results are reported and discussed. The best performance was obtained from the Random Forest algorithm with 82.5 percentage of accuracy.**

***Keywords*—Scorpions, species classification, shape feature, random forest.**

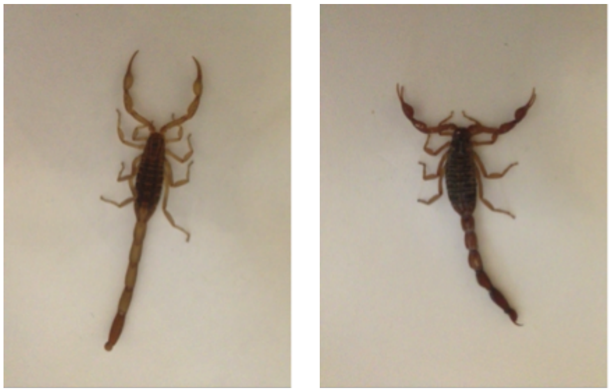
I. Introduction

Scorpions can be found almost everywhere in the world. They are classified in 18 different families and around 1500 species. Only the scorpions from Buthidae family are consider dangerous to people and only 12 species can produce serious envenomation or death [1].

Some of the most dangerous scorpions belong to the Centruroides genus. They can be found in most of the North American continent, especially in Mexico and in the south of the United States of America. In Mexico, the incidence of scorpion sting has reach around 600 stings per 100,000 inhabitants, around 93% of the sting take place in urban cities, however, in small communities the risk is nearly 12 times higher than in the cities [1]. In 2012, the American Association of Poison Control Center (AAPCC) counted around twenty thousand reports of scorpion sting in the USA [2]. In Mexico, the number of scorpion sting reported to the Epidemiology Department was over 300 thousand in 2014 [3]. However, the lethality of the venom differs between different species of scorpions. Thus identifying the scorpions species is important to know how dangerous the scorpion could be to humans. A system that can identify poisonous scorpion could help decrease the risk of fatalities in rural areas.

An identification of poisonous scorpion from a non poisonous can be performed knowing its morphology. For example, venomous scorpion has thick tails and thin pincers, and they are typically light colored (blond). In the other hand, non venomous scorpion have thin tails and broad pincers, and they are dark colored (black).

This paper aims to distinguishing between two different Centruroides species using three different machine-learning techniques. The two studied species are the C. limpidus shown in Figure 1(a) and the C. noxius in Figure 1(b). C. limpidus is typical located in the center and pacific coast of Mexico and C. noxius is located in north of Mexico. Both scorpion belong to the same genus and today the challenge of autonomous system is to identify species from genus [4].



(a) Centruroides limpidus. (b) Centruroides noxius.

Fig. 1. Photographs of the two studies scorpions.

To the best of our knowledge, there has not been any attempt to identify automatically the scorpions species using machine-learning algorithms. However, there are different systems that identify spiders, insects and other animal species.

In the present work, we address the problem of identify two different species of scorpions. Shape features from the contour and area images provide information to classify the scorpions species. This paper proposes a methodology to obtain the most relevant shape features. To validate the approach, experiments are carried out using a set of photographs of living scorpion.

The present manuscript is structured as follows: The initial section deals with a short revision about the related work. Next, an explanation of the implemented methodology. Third, description of the experimental method used for measuring

978-1-5090-0079-1/16/$31.00 ©2016 IEEE

living scorpions. Fourth, discussion of the results and its conclusion. Finally, a brief commentary on the future work is added.

II. Related Work

There are several approaches in the literature for identification of different insect using image processing [4] or using lasers [5], [6]. In the arthropods phylum, the following authors apply different machine learning techniques to classify the different animals. Huiyong Yang, et al. obtain 14 different shape features and propose a random tree algorithm to classify seven different insects [7]. These insects have completely different shapes between each other, so their classification method applies very well to the dataset. However, they specifically state that animals from the same species are harder to classify and that their method is not able to do it properly.

Jiangning Wang, et al. also classifies insects with clearly different shapes [8]. In this case, they use artificial neural networks (ANN) and support vector machine (SVM) [9] algorithms to classify them. Although their results are good, their algorithm is not fully automated and it can only be used for dissected insects. Tom Arbuckle, et al. developed an automated bee identification system (ABIS) [10]. The bees are captured in the wild to be cooled using an icebox, in order to photograph their main wings. From the images, areas between veins of the wing are identified to get particular key wing cells. Lengths angles and areas from cells are calculated. With these data the classification is performed using also SVM and kernel discriminant analysis (KDA). This system is also limited to dissected winged insects.

An identification of a moth can take place using a semantically related visual (SVR) attributes as Linan Feng, et al have shown [11]. The visual feature descriptor contains co-occurrence matrix and feature data from Scale-Invariant Feature Transform (SIFT) [12]. Probability theory is used to assign the best descriptor that has the largest posterior probability score. The specimen is identified by comparing the Euclidean distance for each known species. In this way, the closest distance defines the species. This method requires high computational cost for comparing each species.

ONeill proposed a digital automated identification system called DAISY [13] to identify various winged insect. Principal component analysis (PCA) is used to generate a series of Eigen images from the training set. The current variants of DAISY use a hybrid identification scheme based on a continuous n-tuple classifier (NNC) [14]]. The NNC simply compares an unknown image with a set of images from the training set. It also makes possible the addition of new species with only a small computational cost, nevertheless it requires adding enough instance of each new species. Other work using wings from owlflies (Ascalaphidae) was developed by H. Yang et al. [15], they propose a tool to identify subfamilies of owl flies using Elliptic Fourier Transform (EFT). The system requires images without background to extract coefficient of EFT. The tool use SVM with a radial basis kernel function to analyses the coefficients to identify the subfamilies of owl flies. This method can be extended to identify other types of flies using image without background.

In the same way, a classification of Tarantulas (Theraphosidae) is possible [16]. For each image, a color SIFT vector is extracted; the vector is treated as a document of visual words. Using k-means clustering, they created a codebook to map each feature vector into a code word by finding the closest cluster seed. They tested three methods of machine learning: naive Bayes classification, SVM and supervised Latent Dirichlet Allocation (sLDA). The sLDA results were better, comparing it to the other methods and it archived a good level of classification. The classification requires reducing their noise in the image and the SVM could be improved using well-known kernels.

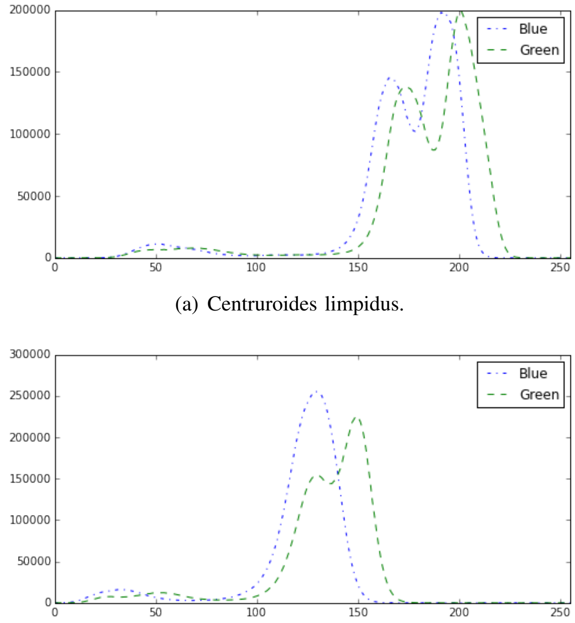
This work tackles the problem of scorpion specie identification using the typical reference of its geometry. As mentioned before, this approach has been proposed for different insect and arachnids, however, most of them use non-living animals and use a fixed position. The method in this paper does not require a specific position and uses a data set with living scorpions.

III. Methodology

In order to distinguish between two scorpion species, this paper presents a three steps implementation. First, image processing techniques are applied to remove noise and identify the scorpions contour and area. Then, six morphological parameters are extracted for each image. Finally, three different classification models are trained and compared to identify between the two studied species.

A. Image processing

The morphology of the scorpion is important to recognize each species. The body of the scorpion is distinguished from the background using a dynamic color threshold on the image. With an algorithm based on dilation and erosion process, the contour of the scorpion is obtained. Using an algorithm based



(a) Centruroides limpidus.

(b) Centruroides noxius.

Fig. 2. Color distribution of scorpions.

on dilation and erosion process, the contour of the scorpion is obtained. Figure 2 shows a histogram of the scorpion's green and blue color components. Pixels from the scorpion's body and pixels from the background are clearly separated by a single threshold value.

After processing the images with the threshold filter, the resulting image provides the scorpion's body with very little noise. Morphological operations are applied to eliminate that small noise. First an erode filter is applied with a box shaped, 5 pixels dimension filter. Then a dilation operation with the same structure is applied. The purpose is to recognize the body of the scorpion using a structural analysis [17]. Finally, the contour of the scorpion's body is obtained using a canny filter. Both the solid full body image and the contour image are used in the following process.

B. Feature extraction

Six features are extracted from each of the scorpion's contour and area images. Some of them are independent of translation, rotation and scale of the image. Others are noise resistant. A complete analysis of these features can be found in [18]. The six features are aspect ratio (1), rectangularity (2), compactness (3), roundness (4), solidity (5) and eccentricity (6).

$$A_R = \frac{d_{min}}{d_{max}} \tag{1}$$

where d_{min} and d_{max} are the minimum and maximum distance respectively.

$$PB = \frac{Area}{Area_{bb}} = \frac{Area}{major\ axis * minor\ axis} \tag{2}$$

where $Area_{bb}$ is the area of the minimal rectangle enclosing the component.

$$f_{circ} = \frac{4\pi Area}{P^2} \tag{3}$$

where P is the perimeters.

$$rd = \frac{Area}{d_{max}^2 \pi} \tag{4}$$

$$S = \frac{Area}{H} = \frac{Area}{convex\ hull\ area} \tag{5}$$

$$e = \frac{\sqrt{(\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2}}{\mu_{20} + \mu_{02}} \tag{6}$$

where the central moments are defined with (7).

$$\mu_{pq} = \sum_{x \in R} \sum_{y \in R} (x - \bar{x})^p (y - \bar{y})^q f(x,y) \tag{7}$$

To obtain the moments, the position of the centroid is calculated by (8).

$$\begin{aligned} \bar{x} &= \frac{1}{6A_c} \sum_{i=0}^{N-1} (x_i + x_{i+1})(x_i y_{i+1} - x_{i+1} y_i) \\ \bar{y} &= \frac{1}{6A_c} \sum_{i=0}^{N-1} (y_i + y_{i+1})(x_i y_{i+1} - x_{i+1} y_i) \end{aligned} \tag{8}$$

where N is the number of elements in the contour and A_c Is the contour's area defined by (9).

$$A_c = \frac{1}{2} \left| \sum_{i=0}^{N-1} (x_i y_{i+1} - x_{i+1} y_i) \right| \tag{9}$$

C. Classifier

Three different machine learning methods were used and their results were compared. These are: ANN, Classification and Regression Tree (CART) [19] and Random Forest (RF) [20]. For each model, a confusion matrix is calculated to compare their accuracy, sensitivity and specificity. Accuracy is determined with (10), which relates all the correctly classified result over the total tested [21].

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

where *TP* is the True Positive, correct classification of C. limpidus, *TN* is the True Negative, correct classification of C. noxius, *FP* is the False Positive, when C. limpidus is wrongly classified as C. noxius, and *FN* is the False Negative, when C. noxius is wrongly classified as C. limpidus.

Sensitivity is defined as the true positive rate (11) or is equivalent to know the rate of the successfully classification of the C. limpidus.

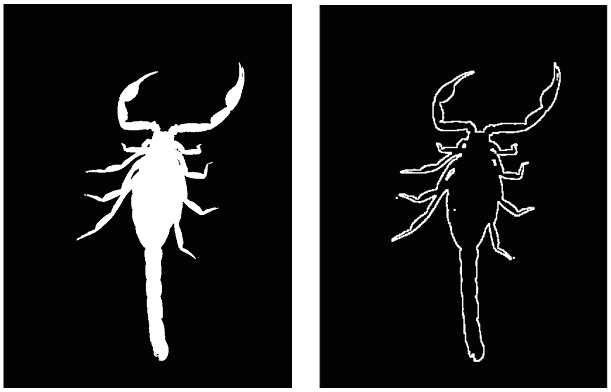
$$\text{Sensitivity} = \frac{TP}{TP + FP} \tag{11}$$

Successfully classification of the C. noxius is measured by the specificity is defined as true negative rate (12).

$$\text{Specificity} = \frac{TN}{TN + FN} \tag{12}$$

IV. Experiment

One hundred sixty photographs are obtained from living scorpions in a semi-controlled environment. The dataset contains eighty picture of each specie. The pictures are taken using an 8M pixels camera in autofocus mode. The scorpions are placed inside a crystal bowl with white background. The



(a) Area image. (b) Contour image.

Fig. 3. C. noxius results after image processing.

camera is placed facing down in order to obtain a clear image of the living scorpion. Images are stored and classified by species.

Using MATLAB 2014A, toolbox image-processing, the scorpion in each picture is distinguished from the background by analyzing the histogram as mentioned in section three. Then, the contour of the scorpion is obtained using erode and dilate methods. An example of the result is shown in Figure 3.

With the data of each feature, three data sets were obtained to know which feature is better to classify the scorpions. First, area features are calculated using the area image. An example of the area image is shown in Figure 3(a).The second group of features are extracted from the contour image. An example of the contour image is shown in Figure 3(b). The third and last data set is the collection of both previous data sets (A∪C). The machine learning methods were implemented in python language 2.7 with PyBrain [22] and SKlearn [23] library of Python. For the Artificial Neural Networks (ANN) and Random Forest (RF) algorithms, the number of neurons and the number of trees were obtained using the error rate. The third machine learning method used was a Classification and Regression Tree (CART). The criterions used for the CART were Gini index and the cross-entropy. The models were trained using cross validation to avoid over-fit.

V. Results

Each ANN was trained until the error converged, as shown in Figure 4. In order to have a constant error, the number of neurons was set to thirty. The average of the error in the training stage of each ANN was 14.19%. The result with the test sample is shown in the table I. The A∪C feature gave the best classification for the ANN method, but if the area feature is used, the ANN becomes over fitted.

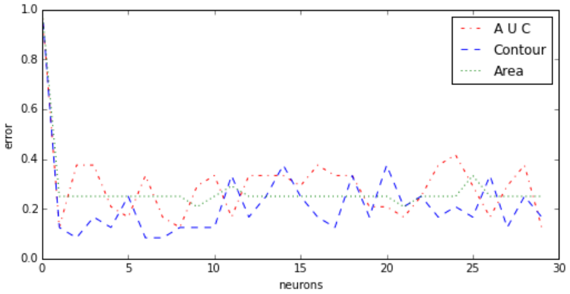


Fig. 4. Error rate between number of the hidden layers.

TABLE I. ANN results.

	Accuracy	sensitivity	specificity
ANN Area	50 %	0 %	100 %
ANN Contour	67.5 %	**70** %	65 %
ANN A∪C	**70** %	55 %	**85** %

The results of the different CART are shown in table II. Both methods, Gini and Cross-entropy, were over fitted using contour features. CART perform better when using the contour characteristics. The best accuracy was 77.5% with A∪C.

As shown in Figure 5, each RF was trained until the error converged. The error rate of A∪C and area feature was lower than using contour feature. The RF was considered to converge when using thirty trees. This number allows to have different splits for the RF, so that the result is uniform. The classification result is shown in table III. The classification was successfully with A∪C but was over fitted with contour feature.

TABLE II. CART results.

	Accuracy	sensitivity	specificity
CART gini area	60 %	55 %	65 %
CART gini contour	50 %	100 %	0 %
CART gini A∪C	65 %	70 %	60 %
CART cross-entropy area	65 %	70 %	60 %
CART cross-entropy contour	50 %	100 %	0 %
CART cross-entropy A∪C	**77.5 %**	**85 %**	**70 %**

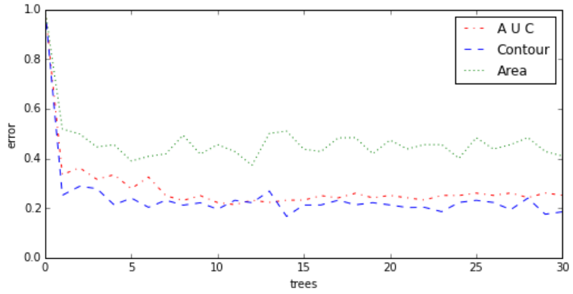


Fig. 5. Error rate between the number of tree in th RF.

TABLE III. RF results.

	Accuracy	sensitivity	specificity
RF Area	62.5%	55 %	70 %
RF Contour	50%	100%	0%
RF A∪C	**82.5%**	**85.0%**	**80%**

Another advantage of the RF is that it provides information about the importance for each variable. Using the Gini index, all properties are compared as shown in table IV. The first three variables with the highest importance came from the contour feature images. However, as table III indicates, when only contour is used, the results are not as good as when area and contour features are employed.

TABLE IV. Gini index.

Feature Type	Variable	Mean Decrease Gini
Contour	rectangularity	15.12
Contour	solidity	14.76
Contour	compactness	11.10
Area	compactness	8.25
Area	solidity	8.24
Contour	roundness	7.43
Area	roundness	6.87
Area	rectangularity	6.70
Area	aspectRatio	6.58
Contour	eccentricity	5.76
Contour	aspectRatio	5.15
Area	eccentricity	3.97

VI. Conclusion and future work

This work introduces the first method for classification of two different species from the Centruroides genus. Morphological features and machine learning techniques were used to identify different living scorpion species. These features provide enough information to classify them, if a color threshold allows distinguishing the scorpion from the background.

The highest accuracy from the ANN method occurs when using features from the collection of both area and contour images. This comes from a good C. limpidus classification. However, to recognize the C. noxius, the ANN with contour features provides the only not over-fitted solution.

The CART method is good to classify the scorpions species. The best accuracy was obtained from the entropy validation using both feature, although the Gini had a minor error rate for the specificity. The CART method presented a higher sensitivity than the ANN method.

The most important result is that the RF can be fitted to achieve the highest accuracy from the three implemented methods. The best case came from using both area and contour features. This is valid for the RF method as well as for the other two techniques but with lower score.

As future work, multiples image processing methods can be applied to extract other types of features using superpixeles and HSV color. Other feature extraction methods, such as Scale-Invariant Feature Transform (SIFT), can be used in order to obtain other scorpions features. Moreover, images of scorpions may be segmented in different body parts (claws, body, tail, etc.), in order to obtain separated values to classify in more detail other scorpions families. Finally, this method could be implement to classify poisonous or not poisonous scorpion using the same feature.

ACKNOWLEDGMENT

This work has been supported by Asociacion Mexicana de Cultura A. C. The authors would like to thank Dr. Lourival D. Possani and his staff for providing the living scorpions and allowing us to photograph them.

REFERENCES

[1] J.-P. Chippaux and M. Goyffon, "Epidemiology of scorpionism: a global appraisal," *Acta tropica*, vol. 107, no. 2, pp. 71–79, 2008.

[2] J. B. Mowry, D. A. Spyker, L. R. Cantilena Jr, J. E. Bailey, and M. Ford, "2012 annual report of the american association of poison control centers' national poison data system (npds): 30th annual report," *Clinical toxicology*, vol. 51, no. 10, pp. 949–1229, 2013.

[3] S. de Salud, "Direccion general de epidemiologia," *Boletin Epidemiologico*, vol. 32, no. 23, 2015.

[4] S. N. A. Hassan, A. Rahman, Z. Htike, and S. L. Win, "Advances in automatic insect classification," *Electrical and Electronics Engineering: An International Journal (EEELIJ)*, vol. 3, no. 2, pp. 51–63, 2014.

[5] D. F. Silva, V. M. Souza, D. P. Ellis, E. J. Keogh, and G. E. Batista, "Exploring low cost laser sensors to identify flying insect species," *Journal of Intelligent & Robotic Systems*, pp. 1–18, 2015.

[6] Y. Qi, G. T. Cinar, V. Souza, G. E. Batista, Y. Wang, and J. C. Principe, "Effective insect recognition using a stacked autoencoder with maximum correntropy criterion," in *Neural Networks (IJCNN), 2015 International Joint Conference on*. IEEE, 2015, pp. 1–7.

[7] H. Yang, W. Liu, K. Xing, J. Qia, X. Wang, L. Gao, and Z. Shen, "Research on insect identification based on pattern recognition technology," in *Natural Computation (ICNC), 2010 Sixth International Conference on*, vol. 2. IEEE, 2010, pp. 545–548.

[8] J. Wang, C. Lin, L. Ji, and A. Liang, "A new automatic identification system of insect images at the order level," *Knowledge-Based Systems*, vol. 33, pp. 102–110, 2012.

[9] V. N. Vapnik, "The nature of statistical learning theory. statistics for engineering and information science," *Springer-Verlag, New York*, 2000.

[10] T. Arbuckle, S. Schröder, V. Steinhage, and D. Wittmann, "Biodiversity informatics in action: identification and monitoring of bee species using abis," in *Proc. 15th Int. Symp. Informatics for Environmental Protection*, vol. 1, 2001, pp. 425–430.

[11] L. Feng and B. Bhanu, "Automated identification and retrieval of moth images with semantically related visual attributes on the wings," in *Image Processing (ICIP), 2013 20th IEEE International Conference on*. IEEE, 2013, pp. 2577–2581.

[12] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[13] M. O'Neill, I. Gauld, K. Gaston, and P. Weeks, "Daisy: an automated invertebrate identification system using holistic vision techniques," in *Proceedings of the Inaugural Meeting BioNET-INTERNATIONAL Group for Computer-Aided Taxonomy (BIGCAT)*, 1997, pp. 13–22.

[14] S. Lucas, "Continuous n-tuple classifier and its application to face recognition," *Electronics Letters*, vol. 33, no. 20, pp. 1676–1678, 1997.

[15] H.-P. Yang, C.-S. Ma, H. Wen, Q.-B. Zhan, and X.-L. Wang, "A tool for developing an automatic insect identification system based on wing outlines," *Scientific reports*, vol. 5, 2015.

[16] A. Utasi, "Local appearance feature based classification of the theraphosidae family," *Visual Observation and Analysis of Animal and Insect Behavior (VAIB), Tsukuba, Japan*, 2012.

[17] S. Suzuki *et al.*, "Topological structural analysis of digitized binary images by border following," *Computer Vision, Graphics, and Image Processing*, vol. 30, no. 1, pp. 32–46, 1985.

[18] M. Yang, K. Kpalma, and J. Ronsin, "A survey of shape feature extraction techniques," *Pattern recognition*, pp. 43–90, 2008.

[19] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[20] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[21] T. M. Mitchell, *Machine Learning*, 1st ed. New York, NY, USA: McGraw-Hill, Inc., 1997.

[22] T. Schaul, J. Bayer, D. Wierstra, Y. Sun, M. Felder, F. Sehnke, T. Rückstieß, and J. Schmidhuber, "Pybrain," *The Journal of Machine Learning Research*, vol. 11, pp. 743–746, 2010.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

978-1-5090-0079-1/16/$31.00 ©2016 IEEE