Identifying the Machine Learning Family from Black-Box Models*

*Note: The full contents of this paper have been published in the volume *Lecture Notes in Artificial Intelligence 11160* (LNAI 11160)

Raül Fabra-Boluda, Cèsar Ferri, José Hernández-Orallo,
Fernando Martínez-Plumed, M. José Ramírez-Quintana
DSIC
Universitat Politècnica de València
Valencia, Spain
{rafabbo,cferri,jorallo,fmartinez,mramirez}@dsic.upv.es

***Abstract*—We address the novel question of determining which *kind* of machine learning model is behind the predictions when we interact with a black-box model. This may allow us to identify families of techniques whose models exhibit similar vulnerabilities and strengths. In our method, we first consider how an adversary can systematically query a given black-box model (oracle) to label an artificially-generated dataset. This labelled dataset is then used for training different surrogate models (each one trying to imitate the oracle's behaviour). The method has two different approaches. First, we assume that the family of the surrogate model that achieves the maximum Kappa metric against the oracle labels corresponds to the family of the oracle model. The other approach, based on machine learning, consists in learning a meta-model that is able to predict the model family of a new black-box model. We compare these two approaches experimentally, giving us insight about how explanatory and predictable our concept of family is.**

***Index Terms*—machine learning families, black-box model, dissimilarity measures, adversarial machine learning**