Abstract-This paper presents an exploratory machine learning attack based on deep learning to infer the functionality of an arbitrary classifier by pooling it as a blackbox ,and using returned label to build a functionally equivalent machine.It is difficult and time consuming to build a classifier as it required a large training data,selecting a appropriate machine learning algorithm(through specific test and domain specific knowledge),and resolving the under lying hyperparameters (having a good knowledge of classifier’s structure).As all these information are trademarked so should be protected. With the proposed black-box attack approach ,an adversary can use this information by using deep learning with the help of labels previously obtained from the classifier under attack, and can build the same classifier with the same functionality without knowing the type, structure or underlying parameters of the original classifier. Results for text classification application demonstrate that deep learning can deduce Naïve Bayes and SVM classifier with accuracy and can take their functionalities. This new attack paradigm with deep learning introduces additional security challenges for online machine learning algorithms and raises the need for novel mitigation strategies to counteract the high fidelity inference capability of deep learning .

INTRODUCTION:

In the present day, information system are subject to frequent attacks and are utilized, and it is not possible to assume that any given program will be used as intended. This holds true not only for traditional programs where all behaviors are programmed manually,