## Title

When A Tree Falls Using Diversity In Ensemble Classifiers To Identify Evasion In Malware Detectors

## Authors

Charles Sumutz & Angelos Stavrou from GMU

## Abstract

Machine learning classifiers are a vital component of modern malware and intrusion detection systems. However, past studies have shown that classifier based detection systems are **susceptible to evasion attacks** in practice. **Improving the evasion resistance of learning based systems is an open problem.** To address this, we introduce **a novel method for identifying the observations on which an ensemble classifier performs poorly.** During detection, when a sufficient number of votes from individual classifiers disagree, the ensemble classifier prediction is shown to be unreliable. The proposed method, **ensemble classifier mutual agreement analysis, allows the detection of many forms of classifier evasion without additional external ground truth.**

**We evaluate our approach using PDFrate, a PDF malware detector.** Applying our method to data taken from a real network, we show that **the vast majority of predictions can be made with high ensemble classifier agreement.** However, most classifier evasion attempts, including nine targeted mimicry scenarios from two recent studies, **are given an outcome of uncertain indicating that these observations cannot be given a reliable prediction by the classifier.** To show the **general applicability of our approach,** we tested it **against the Drebin Android malware detector where an uncertain prediction was correctly given to the majority of novel attacks.** Our evaluation includes **over 100,000 PDF documents and 100,000 Android applications.** Furthermore, we show that our approach can be generalized to **weaken the effectiveness of the Gradient Descent and Kernel Density Estimation attacks against Support Vector Machines.** We discovered that **feature bagging** is the most important property for enabling ensemble classifier diversity based evasion detection.

## I. INTRODUCTION

The use of machine learning has emerged as one of the primary techniques employed to address a wide range of malfeasance and malicious activities. Applications of machine learning include **clustering of malware families** [7], [20], **detection of malicious downloads** [12], [34], **detection of account misuse in social networks** [14], [44], **and detection of commonly exploited file formats such as Java archives** [36] and **documents** [24], [25], [39]. Moreover, statistical or machine learning techniques have been used successfully **for years to identify SPAM** [11], [21], [35].

One of the main weaknesses of systems that **employ machine learning classification in adversarial environments is their susceptibility to evasion attacks**. With evasion attacks, we refer to the classes of attacks that take advantage of knowledge of how the machine learning system operates, and in many cases utilize access to the training set and features, to **evade detection passively or actively** [8], [9], [15], [33], [45].

A common technique used in **evasion attacks** against machine learners is **mimicry**. Mimicry attacks thwart detection by making the **attack data appear benign** according to the model used by **the intrusion detection system**. Often this is achieved by **hiding overtly malicious content through encoding or encryption** [28], [42] or minimizing the footprint of malicious content **through data misuse or code re-use attacks** [17], [37]. For instance, content aligning with a benign observation is added to cover up or drown out **the malicious content**. **Many detection systems are evaded by exploiting differences in the detection system and the systems being protected** [16], [19]. **Even if operational details of defense systems are kept secret, enough knowledge to conduct evasion can often be obtained solely from external testing** [18]. With all of these potential **evasion vectors**, **preventing detection evasion remains an open problem**.

Our approach is not to prevent all possible evasion attacks, but to introduce a mechanism that **provides detection of poor classifier performance**. We analyze the use of introspection in an ensemble classifier to detect **when the classifier provides unreliable results at classification time. The use of ensemble classifier mutual agreement analysis relies on the intuition that when individual classifiers in an ensemble vote for the same prediction, the prediction is likely to be accurate.** When a sufficient number of the votes are in opposition, then the classifier prediction is not trustworthy. In this state of internal classifier disagreement, the detector returns the outcome of uncertain **instead of a prediction of benign or malicious**. In operation, confidence in the predictions of the classifier is improved at the cost of a small portion of the samples being labeled as uncertain, indicating that the classifier is not fit to provide an accurate response. This separation of accurate predictions from uncertain predictions is possible because the majority of the misclassifications, including evasion attempts, have a classifier voting score distribution distinct from the accurate predictions.

To evaluate our technique, we applied **mutual agreement analysis** to **two well-studied malware detection systems**: PDFrate [40] and Drebin [4]. PDFrate uses features derived from document structure and metadata fed into a Random Forest classifier to detect Trojan PDFs. **PDFrate is used in real world intrusion detection systems and can be evaluated by the public through submissions to pdfrate.com**. PDFrate was selected because it is **publicly accessible**, well documented, uses an ensemble classifier which returns the raw voting score, and **has been subjected to（饱受） multiple recently published mimicry attacks** [26], [27], [43]. Our evaluation includes over 100,000 documents sourced from an operational environment and hundreds of malicious documents in **nine unique evasion scenarios** from two independent evasion studies. To demonstrate the general applicability of our approach, we apply **mutual agreement analysis** to the Drebin Android malware detector using over 100,000 applications, including over 5,000 malicious applications in over 20 labeled malware families. We find that mutual agreement analysis enables the identification of novel malware that would otherwise not be detected reliably.

In building **an evasion resistant ensemble using Support Vector Machines (SVM) as base classifiers**, we find that feature bagging, or constructing many individual classifiers with randomized subsets of the whole feature set, is crucial to providing this discriminatory power. Using this method, we counter the Gradient Descent and Kernel Density Estimation (GD-KDE) attack, which is highly successful against a traditional SVM classifier.

## II. RELATED WORK