# Malicious Document Detection and Robust ML Model Construction恶意文档检测与强鲁棒性ML模型的构建

**摘要：**随着信息安全技术的迅速发展，对恶意文档检测变得越来越重要，但是，由于文档结构的多样性，使攻击者可以获得更多的攻击机空间，在本文中，我们旨在工业界和学术界中构建一个人工智能（AI）文档分类器。我们一共收集了200,000个样本用于训练AI模型与优化。我们的实验结果表明，经过多次训练与调优，我们的模型准确度高达99.82%，并且误报率仅低于0.01%。此外，通过对抗性学习的研究，该模型有一定的抗逃逸性与良好的鲁棒性。最后，我们展示了模型可以广泛的部署在实际的应用场景中，例如安全产品和终端安全中。

实验部分

## 构建基于机器学习的文档分类器

在这个章节中，我们主要集中在对设计分类器，一种将数据分类成两种预定义的学习系统。分类器的预测是通过计算一些数学或者概率的分数或者计算特定的阈值，我们主要关注于分类器的以下几个方面：

A 通用机器学习分类框架

B 数据集

C 特征工程

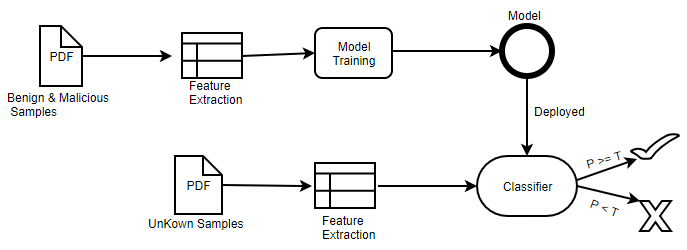
D 分类算法选择

E 模型优化

### 通用机器学习分类框架

通用机器学习的分类器框架如图1所示，我们的目标是训练一个鲁棒性的模型用于恶意文档分类，首先，在数据收集阶段，我们需要收集到大量有标签的恶意文件和良性文件；第二，在特征工程阶段，我们通过对比恶意文件与良性文件的差异，手动提取上百个文件的典型特征，这是一个将文档从原始向量转换为特征向量；最后，我们通过机器学习来训练模型，使模型可以很好的拟合训练数据分布。训练通常是离线的，并且需要很多的时间和内存占用，当预测是就可以打包到在线服务进行预测。在这时，我们的模型已经准备好预测了，当新的样本提交到了模型中，它就可以返回出自信的预测分数，来判定这个样本是恶意的还是良性的。如图1 很好的展示了一个基于机器学习的分类器框架。

Figure 1. Basic machine learning classification framework.



### Dataset

我们的数据集一共有201,368 个PDF文件，可以分为两类：正常样本28332个，恶意样本173,036个。

关于这些文件，其中有156,035个恶意样本是从VirusShare 下载下来的，还有9000的恶意样本是从Contagio 的公开数据集下载下来的。其余的则是来自两个常用的搜索引擎爬取下来的。除此之外，我们还有一个开源数据集的样本特征集，