在当今信息安全领域，基于AI的恶意PDF检测越发重要，基于文档的攻击通常具有针对性（targeted attack），加上其文件结构的多样性，攻击手段变得丰富且隐蔽，因此更容易成功。蓝盾着力于为工业界和学术界提供一个基于AI的恶意文档分类器原型。

现有方法约可分成两类：静态方法和动态方法；纵然这有一定程度上的重合。动态方法需要把文档放在某一个特定探针环境（instrumented environment）中打开运行；而静态方法则无需打开运行仅通过静态特征即可完成检测。

目前我们收集良性和恶意样本共20万个，经过对恶意样本静态解析，抽取具有显著分类能力的特征，训练生成AI模型。实验数据表明，模型的准确率达到99.82%，而误报率却只有0.01%，单个文件的平均检测时间仅需几毫秒。进一步地，我们研究了部分对抗性学习的方法，并用实验数据证明模型具有良好的抗逃逸能力和鲁棒性。最后，我们通过实际应用，表明此模型可广泛部署在终端安全产品，邮件服务器上等，这些均是非常有意义的应用场景。

关于对抗性学习主要有以下几个方面：

* 对于外置PDF解析器，现有攻击手段是解析器混淆攻击（Parser-Confusion Attacks）；
* 对于机器学习模型，现有的攻击手段是自动化分类器逃逸攻击（Automatic Classifier Evasion Attacks）；
* 对于假设性的“可检测的分辨力”(Detectable Discrepancy)，现有攻击手段为模拟和反向模拟攻击（Mimicry and Reverse Mimicry）。

这些攻击手段对于模型本身的安全提出了很大挑战，在我们的工作中不仅生成了一个准确度高的模型，并且在模型本身的安全也有所建树。且在实际运行时，CPU和内存的时空效能比（Time & Space Performance）比旧有基于规则的模型，有显著提升。

我们除了使用大量数据研究人工智能化应用安全，并且把人工智能本身的安全也放在了同等重要的位置上。我们通过大量的实验，模拟了（1）攻击者通过对恶意样本的增删改（如变更特征的值），以混淆分类器，达到逃逸的目的；（2）分类器经自身修正，通过重新训练模型，去除已被攻击者所利用的特征，以维持模型的健壮性。

基于AI的文档分类器是社会工程学、病毒分析等领域的重要研究课题。在未来，我们还会尝试解决以下研究问题：

基于深度学习的恶意PDF文档检测

动静态分析引擎的调优

对于Microsoft Office等其他文件格式的支持（如docx，pptx等）