**成绩：**

**厦门大学**

**课程项目报告**

**课程名称： 网络攻防技术基础**

**项目题目：加密恶意流量分析**

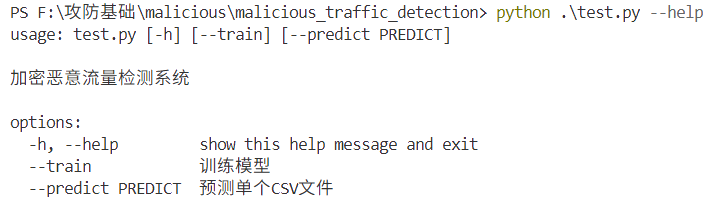
**任课教师： 冯超、张翼**

**组员： 庞轶轲 陆璟杰**

**年级： 2022级**

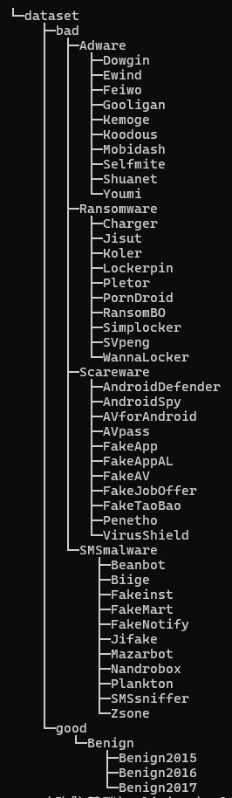
**专业： 网络空间安全**

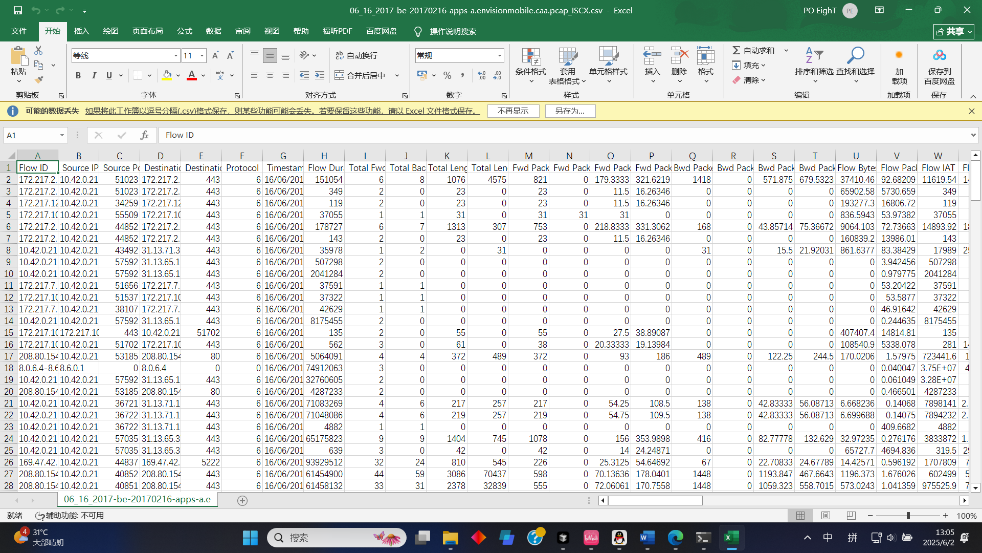
**一、源代码使用说明**

本项目具有模型训练和预测两个功能，可以通过命令行中的—train和—predict两个参数选择是进行训练还是预测。

**二、项目所用数据集**

本项目使用的数据集为课程项目指导中给出的CIC-AndMal2017数据集。CIC-AndMal2017数据集由加拿大网络安全研究所（CIC）于2017年创建，旨在为移动设备上的恶意软件检测提供一个全面且标准化的基准。该数据集的核心研究问题是如何在复杂的移动环境中准确识别和分类恶意软件，它的构建基于对Android平台上的恶意软件进行深度分析。研究团队通过模拟真实环境中的网络流量，捕获了大量恶意软件的行为数据。这些数据经过严格的筛选和标注，确保了数据集的高质量和代表性。构建过程中，采用了多种先进的检测技术和算法，以识别和分类不同类型的恶意软件。样本分布为良性样本121万左右，恶意样本140万左右，总共有260万左右的样本





样本内容

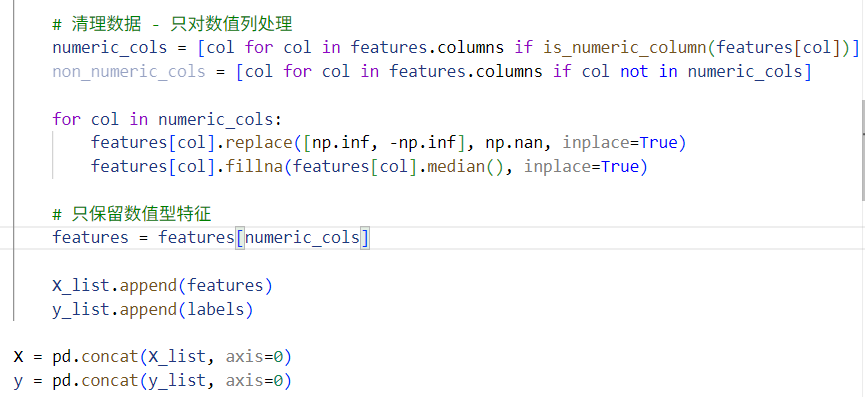
数据集结构

**三、算法流程思路**

1、特征处理

本项目的主题是加密恶意流量的分析，数据集是来自软件传输的数据包，而由软件传输的数据包基本上都是加密的流量，比如TLS/SSL加密，所以我们无法像普通的恶意流量分析一样，直接让模型通过读取数据包的内容判断其中是否包含恶意代码，而只能让模型对数据包的各种特征进行学习，观察数据包是否具有恶意流量的某些特征来分辨恶意流量。

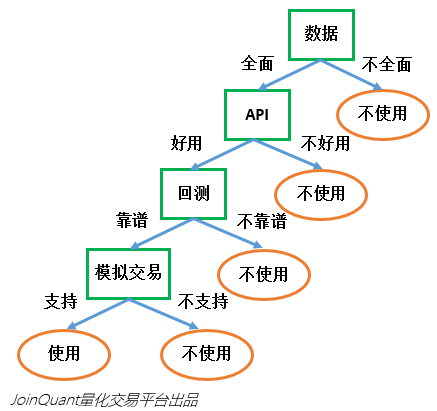
由于CIC-AndMal2017数据集本身已经完成了特征的处理，csv文件的每一列分别对应一个数据包的特征，包括源端口、目标端口、使用的协议等等，总共有84个特征。而在这些特征中，有一些对恶意流量的检测的作用不是很大，比如源IP和目标IP，因为恶意流量可以从任何一个公网IP发出，攻击者也可以伪造任意IP发送流量，同样的被攻击者也可以是任意一个IP；还有像Flow ID这样的序号类特征，和判断恶意流量没有直接的关系；最后还有时间戳特征，我们一开始考虑的是这个可能可以根据这个信息提取出攻击持续时间或是速率这种特征，但是之后观察数据集发现时间戳只精确到秒，如果想要分析出上述特征至少也应该精确到毫秒，所以最后我们还是清理了这一特征。正好我们选择清理的特征都是数值型，所以我们只需要提取数值型特征进行训练即可。提取特征后使用scaler进行标准化，并保存scaler数据便于之后预测时进行调用。



2、算法模型

我们在查阅过几篇恶意流量检测相关的文章，浏览过几个流量检测的项目，并尝试过几个不同的模型后，最终选择在本项目中使用随机森林算法来训练模型。

介绍随机森林之前需要先介绍一下决策树。决策树是一种基本的分类器，主要工作就是选取特征对数据集进行划分，最后把数据贴上两类不同的标签，构建好的决策树呈树形结构。具体流程类似下图。



而随机森林就是一个包含多个决策树的分类器，并且其输出的类别是由个别树输出的类别的众数而定，该算法的具体流程如下。

（1）用N来表示训练用例（样本）的个数，M表示特征数目。

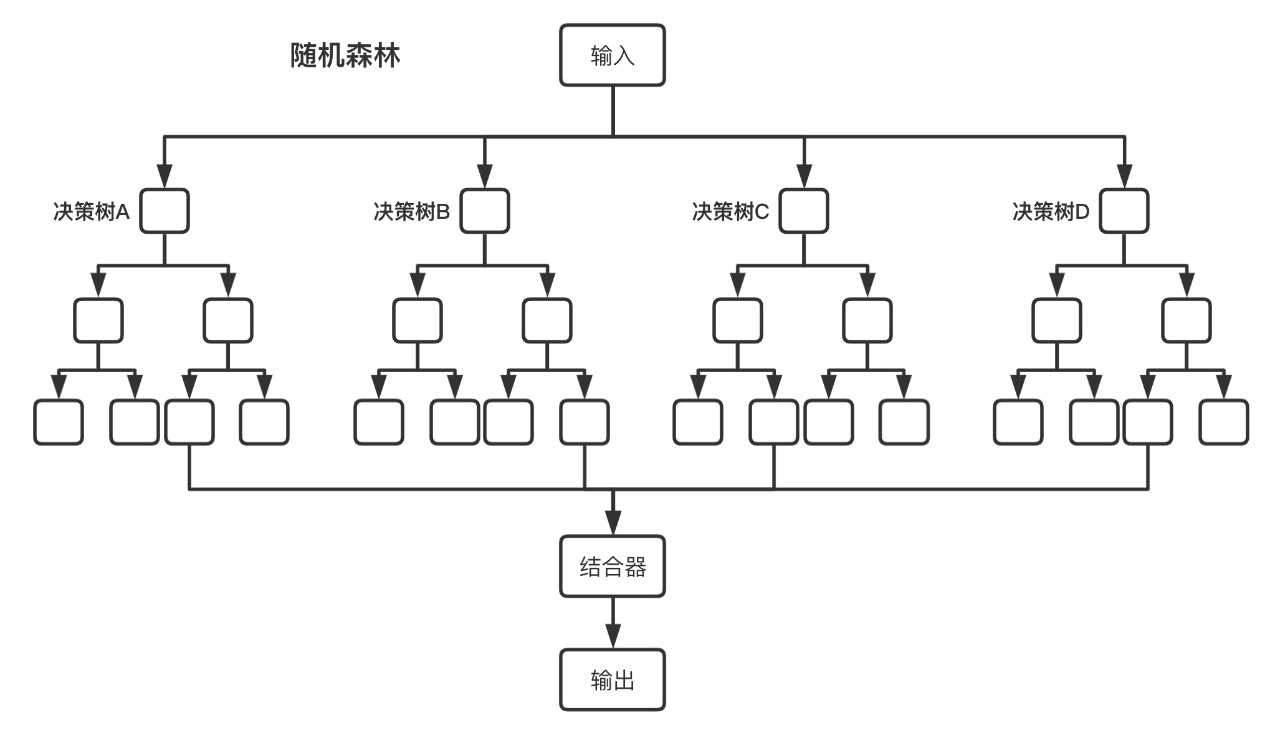
（2）输入特征数目m，用于确定决策树上一个节点的决策结果；其中m应远小于M。

（3）从N个训练用例（样本）中以有放回抽样的方式，取样N次，形成一个训练集（即bootstrap取样），并用未抽到的用例（样本）作预测，评估其误差。

（4）对于每一个节点，随机选择m个特征，决策树上每个节点的决定都是基于这些特征确定的。根据这m个特征，计算其最佳的分裂方式。

（5）每棵树都会完整成长而不会剪枝，这有可能在建完一棵正常树状分类器后会被采用）。

整个流程可以概括为下图。



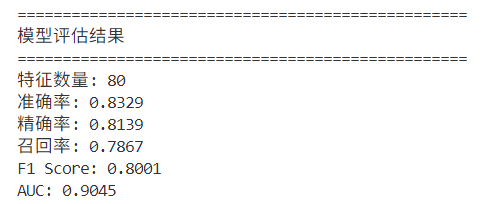
在代码中，我们训练模型的参数如下图。其中n\_estimators是抽样生成决策树的个数，由于数据集比较庞大，所以需要设置为较大的值，本项目中设置为200；max\_depth是决策树的最大深度，由于样本量多，特征也多，如果不限制最大深度会导致模型训练十分缓慢，且经过我们尝试不会对模型有很大的提升，还可能导致过拟合，所以我们设置最大深度为100；min\_samples\_split是节点可分的最小样本数，2是默认值；min\_samples\_leaf是叶子节点最小的样本权重和，叶子节点如果小于这个值，则会和兄弟节点一起被剪枝，只保留该叶子节点的父节点。默认是0，即不考虑样本权重问题，但是在本项目的数据集中，有很多数据包其实是有很多特征为空的，即有较多样本的缺失值，所以这里需要设置该参数值。



**四、性能展示与分析**

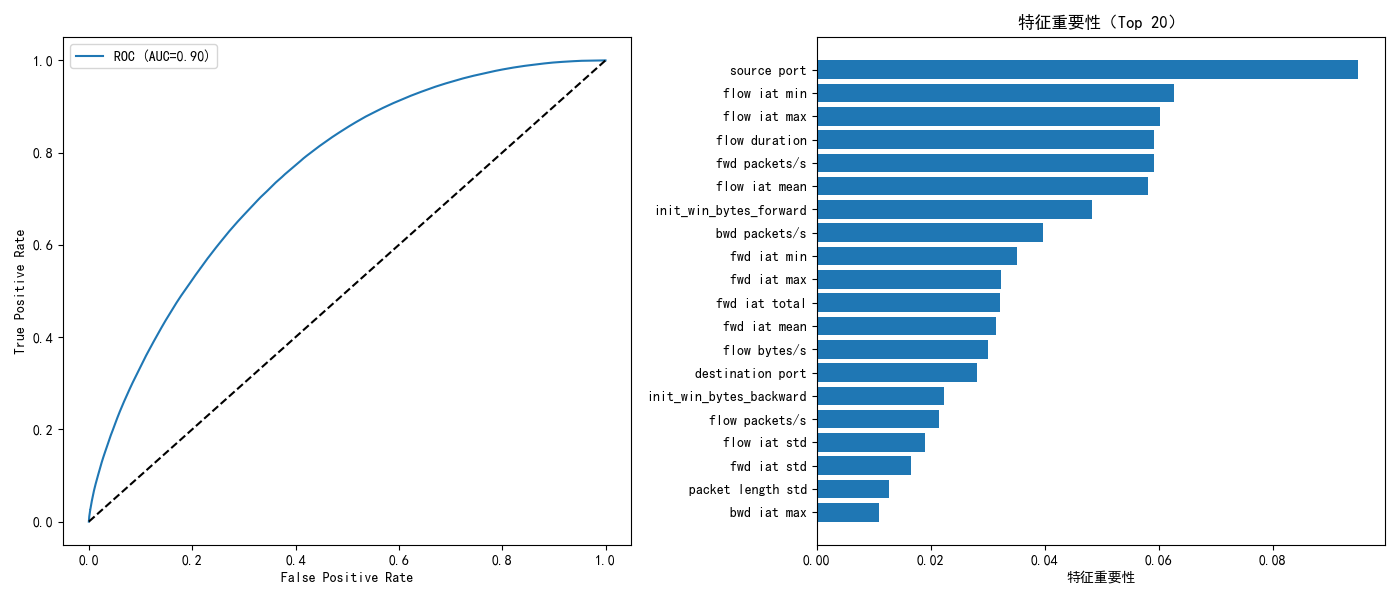
1、模型训练结果

经过几天的训练，设置过很多不同的参数，也尝试了很多从各种文章中了解到的处理算法，我们将模型的正确率提高到了百分之83左右，最终还是没有能达到90以上，猜测可能还是样本数据比较复杂的原因。



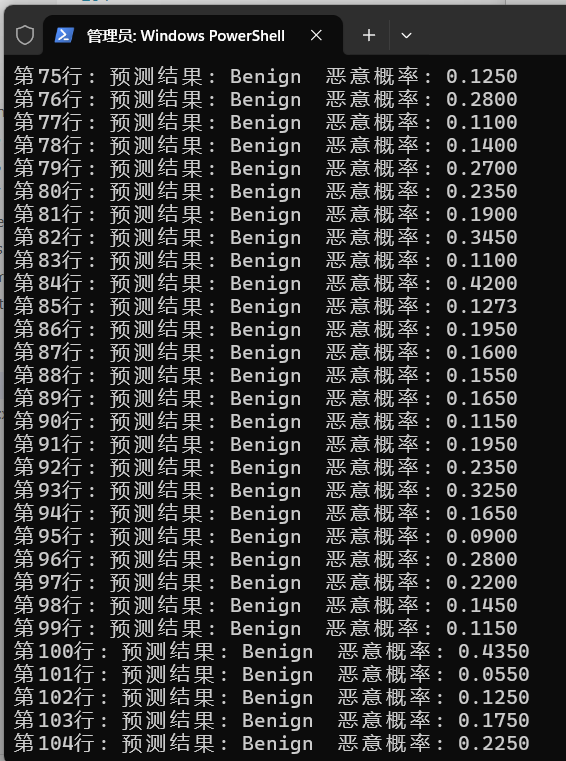
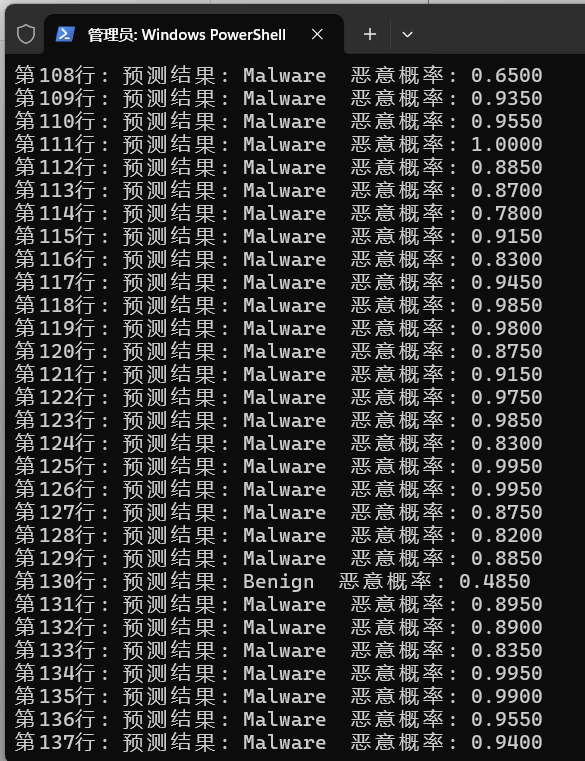
根据模型训练结果绘出ROC曲线，并通过随机森林模型的feature\_importances\_函数绘出特征重要性图如下。可以看到源端口和流时间特征重要性较高。





2、使用模型进行预测

使用训练后保存的模型文件对数据集中的csv文件进行预测，需要注意因为训练时使用了scaler标准化，所以预测时也要使用保存的scaler数据对目标文件进行特征标准化。对良性文件进行预测，可以看到对数据包的预测基本都为良性Benign；而对恶意文件进行预测，也能看到预测结果绝大部分为恶意malicious。

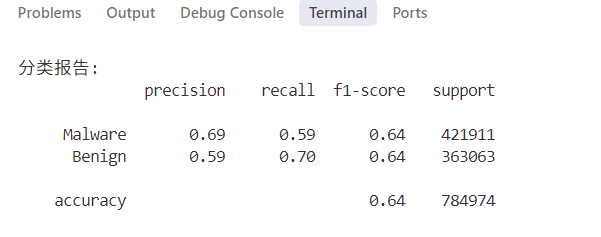


**五、改进方向与总结**

1、最主要的改进方向就在提高模型准确率方面，虽然我们花费了很多时间进行训练，但是模型的准确率还是没有达到一个很高的水平。为了提高模型准确率我们也尝试过很多办法。

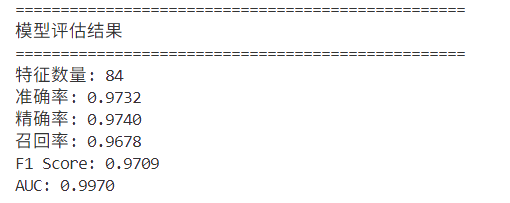
首先因为特征工程的好坏基本决定了模型的质量上限，所以我们尝试对数据集中已经提取好的特征再进行特征工程，我们在查阅到的文章中了解到了relief这个特征权重算法，它可以根据特征和类别的相关性赋予不同权重，当权重小于某个阈值时移除特征，这种选取策略比我们直接清除非数值特征的策略更为科学合理，所以我们尝试使用这个算法进行特征工程，但效果并不好，模型的准确率甚至更低了，我们研究认为原因可能是数据集的特征过多过于复杂，有可能每个特征对判断恶意流量的作用相差不大，使用relief算法又需要选取保留的特征数，如果保留得过少会丢失过多的信息，如果保留得过多使用这个算法意义就不大了，所以最后我们还是放弃了这个算法。

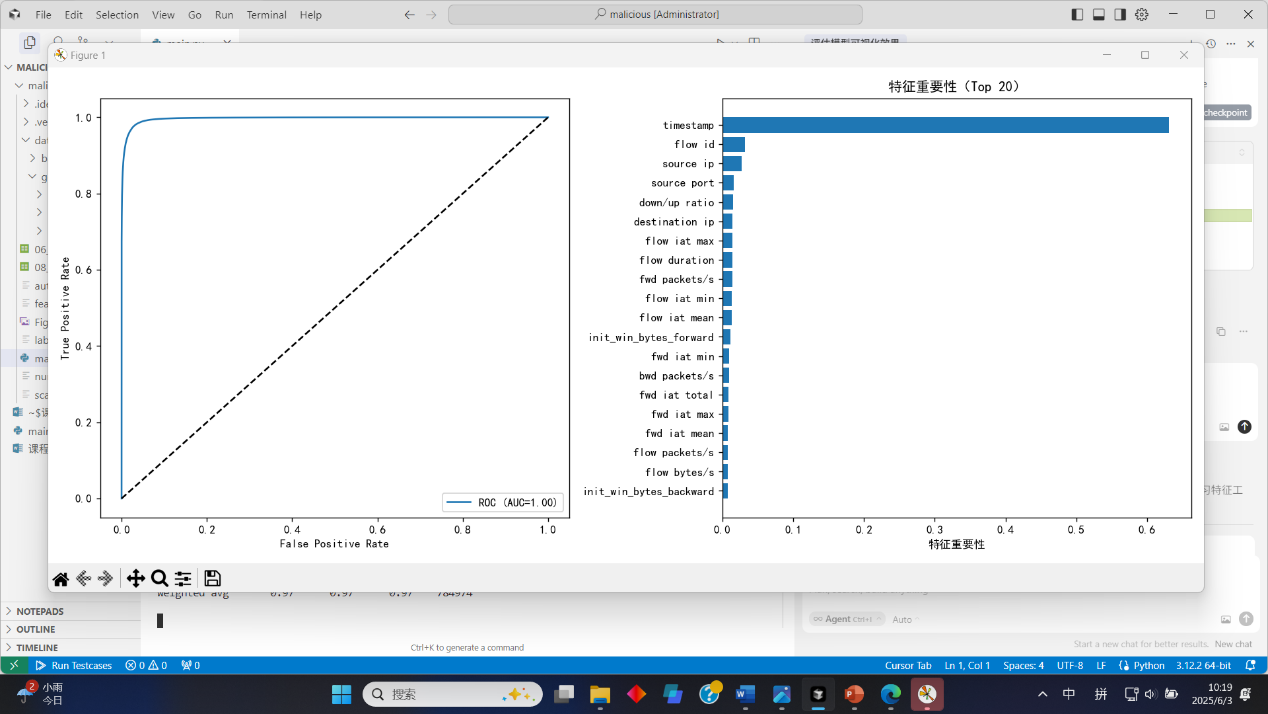
除了特征工程，我们还尝试过几个其他的算法模型进行训练。比如XGBoost极端梯度提升算法，SVM支持向量机算法等，但是效果都不如随机森林算法好。



XGBoost算法性能

2、在选择特征时我们最后选择了清除了非数值型的特征，其实我们也尝试过将所有的特征都用于训练，要达成这个目标就需要对非数值型特征进行编码，并且只能使用lebelencode编码而不能使用onehot编码，因为我们的数据集存在很多的样本缺失值，使用onehot编码会导致特征爆炸，导致内存爆满。将特征进行编码后进行训练，得到的模型确实准确率有显著的提高，但这并不意味着模型是正确的。从图片上可以看出，首先ROC曲线的上升非常不合理，其次时间戳和流ID特征的重要性断层式的高，这说明模型仅仅依据时间戳或者流ID这种序号类的特征就判断流量是否恶意，而我们知道这种特征实际上是与恶意流量没有直接关系的，这就导致模型只能对训练集中的数据包进行判断，对训练集以外的数据包将失去效果，也就是我们所说的过拟合，这样的模型一定是错误的。





**六、参考文献**

1、基于机器学习的TLS恶意加密流量检测方案<https://www.secrss.com/articles/18679>

2、基于随机森林的网络入侵检测方法[基于机器学习的TLS恶意加密流量检测方案 - 安全内参 | 决策者的网络安全知识库](https://www.secrss.com/articles/18679)

3、随机森林算法参数解释及调优[【实践】随机森林算法参数解释及调优（含Python代码）\_随机森林参数调优-CSDN博客](https://blog.csdn.net/wzk4869/article/details/126256780)

4、基于Relief F-PSO特征提取的加密恶意流量检测研究[基于Relief F-PSO特征提取的加密恶意流量检测研究.docx - 人人文库](https://www.renrendoc.com/paper/421096359.html)

5、结合多特征识别的恶意加密流量检测方法[结合多特征识别的恶意加密流量检测方法-Robust Malicious Encrypted Traffic Detection based with Multiple Features](https://jcs.iie.ac.cn/xxaqxb/ch/reader/view_abstract.aspx?file_no=20210209&flag=1)