1. 解决DLVD的数据不平衡的其他方法

**成本敏感分析：**

成本敏感的神经网络：在神经网络的训练中，通过调整损失函数来反映不同错误的成本

重新加权：调整每个类别的样本权重，使得模型在训练时对少数类给予更多关注。

**集成方法：**

集成方法（Ensemble Methods）是机器学习中的一种技术，它通过结合多个模型的预测来提高整体的预测性能。这种方法的基本思想是“众人拾柴火焰高”，即多个模型的集体智慧往往比单个模型更可靠。

1. DLVD是怎么学习漏洞的

通用：特征提取、模型训练、模型部署

**基于标记的方法：**

监督学习：这种方法依赖于有标记的数据集，即包含已知漏洞和无漏洞的代码样本。

特征提取：使用深度学习模型（如卷积神经网络CNN或循环神经网络RNN）从代

码中提取特征。

分类任务：将漏洞检测问题视为二分类或多分类任务，模型学习区分代码中是否存在漏洞。

**基于图的方法：**

把代码看成是图结构，比如数据流程图，程序流程图等，其中节点可以代表变量、函数等，边代表它们之间的关系。使用图神经网络来处理代码的图表示，捕捉代码元素之间的复杂关系。

1. 为什么ROS\_R最好

理论：

过采样优势：不像欠采样那样丢失信息

Sampling\_R优势：不像sampling\_l丢失信息

1. 小组分工

每个人都读了一遍论文，然后刘家骥、罗瑞、陈恩宝分别负责ppt前三部分，小组思考部分由我们五个人共同讨论获得，然后胡博浩负责整合ppt和准备讲稿，阿斯雅负责小组调度和展示部分。

1. 论文研究成果

Yes:

1：建议进行过采样，而不是欠采样。

2：建议对原始数据进行采样，而不是对潜在空间进行采样 。

3：建议对原始数据进行随机过采样来处理DLVD比较中的数据不平衡问题。

4：未来的研究建议开发新的数据增强，以提高DLVD方法学习真正脆弱模式的能力，因为仍有改进的空间。

Nays：

不建议使用OSS来处理DLVD 中的数据不平衡问题。(OSS并不能保证将多数阶级的规模缩小到少数阶级。因为它只会删除特定条件的数据样本)

1. 思考

**增强模型鲁棒性：**

从代码角度来看，原始论文方法中使用的三种数据集都是基于C语言漏洞的，缺乏其他编程语言的漏洞，而从漏洞的广泛性来看，有漏洞的不仅仅是代码，也有可能是文档或日志文件。

**增加新的数据采样方法：**

**寻找更合适的参数组合**：

原始论文并没有对DLVD中的实验设置进行修改，只是根据实验的目的进行了相应的微调。所以我们可以进一步实验寻找更为合适的超参数组合。

**引入数据增强技术：**

1. 应用场景

基于机器学习的漏洞检测

软件工程任务中的数据采样（缺陷预测，bug分类，软件质量预测和软件变化预测）

1. 信息安全四大顶会

**IEEE Security & Privacy (S&P):**

这是IEEE（电气和电子工程师协会）主办的一个会议，专注于安全和隐私领域的研究。它涵盖了广泛的主题，包括密码学、安全协议、系统安全、隐私保护技术等。

**ACM Conference on Computer and Communications Security (CCS):**

CCS是由ACM（美国计算机协会）主办的会议，是计算机安全和通信安全领域的顶级会议之一。它包括了从基础密码学到复杂的系统安全问题的广泛议题。

**USENIX Security Symposium:**

USENIX Security是USENIX协会主办的会议，它是信息安全研究者发布研究成果的首要平台之一。会议内容涵盖了系统安全、网络防御、隐私保护、安全政策等多个方面。

**NDSS (Network and Distributed System Security Symposium):**

NDSS是一个关注网络安全和分布式系统安全的会议，它通常由互联网协会（The Internet Society）主办。会议议题包括但不限于网络安全协议、安全架构、加密技术等。

1. 论文作者

加拿大马尼托巴大学: 徐杨、王少伟

美国新泽西理工学院：李毅、王少华

10、跟课程联系（细看）

基于标记的方法-程序切片技术

基于图的方法-程序依赖图，抽象语法树