# LLM-CBC: Confidence-Based Code Classification With Large Language Models

宋锐科*，杨帆*，董洪伟，司凌宇
通讯作者邮箱：lingyu@iscas.ac.cn

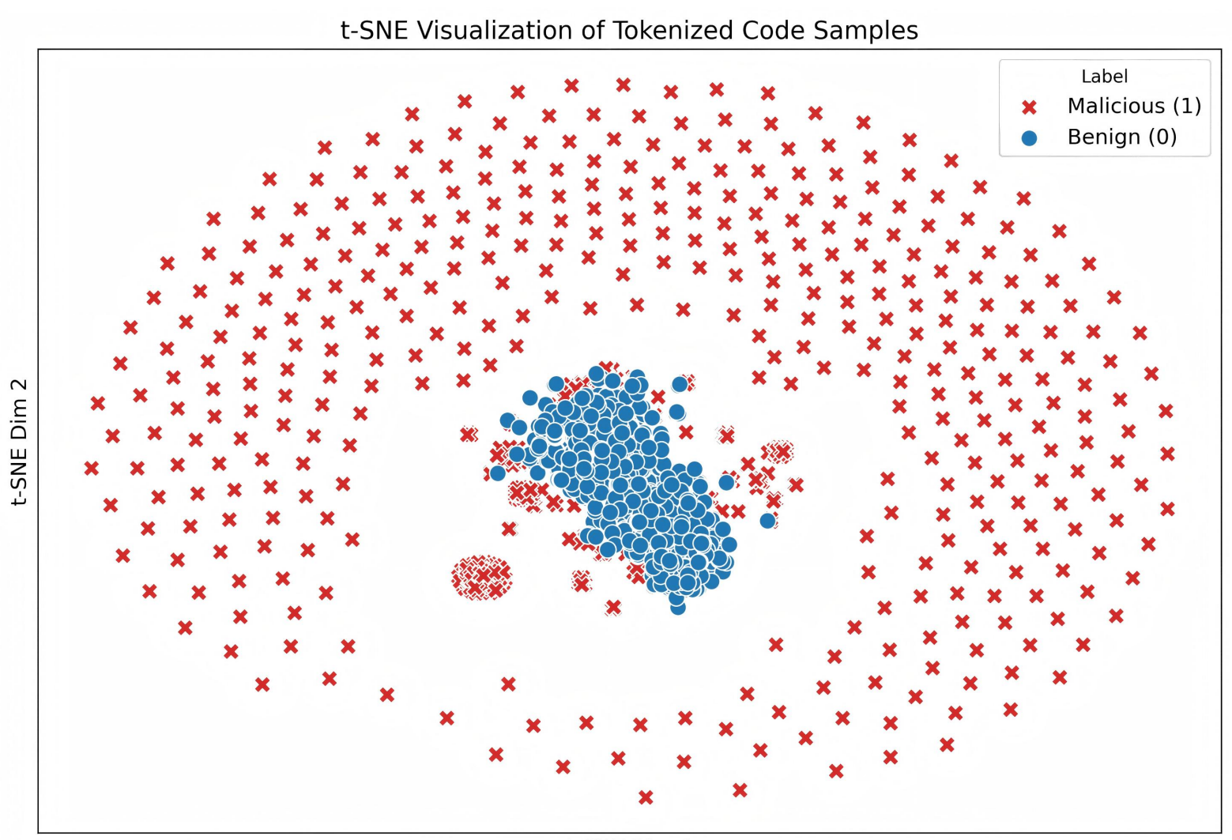
IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING

## 概述

本文提出了一种新颖的跨语言恶意代码检测框架LLM-CBC，结合大语言模型与双VAE架构，分别建模良性与恶意代码的语义分布特征。通过生成样本与提取潜在变量，构建高维判别向量并输入分类模块C-model，实现基于置信度的精细分类。实验表明，该方法在多个跨语言数据集上显著优于现有检测工具，具备出色的准确率和泛化能力，为语义建模驱动的代码安全分析提供了有效路径。

## 动机与分析

现有恶意代码检测方法多依赖固定规则或语法特征，难以应对跨语言环境和语义混淆攻击。为探究代码类型的本质差异，本文通过BERT编码与t-SNE可视化发现：良性代码在语义空间中分布更紧凑，而恶意代码分布分散、方差更大。该发现表明不同类型代码在语义分布上存在显著差异，为构建基于分布建模的跨语言检测方法提供了理论支持，也激发了对代码潜在语义建模与精细分类策略的进一步研究。



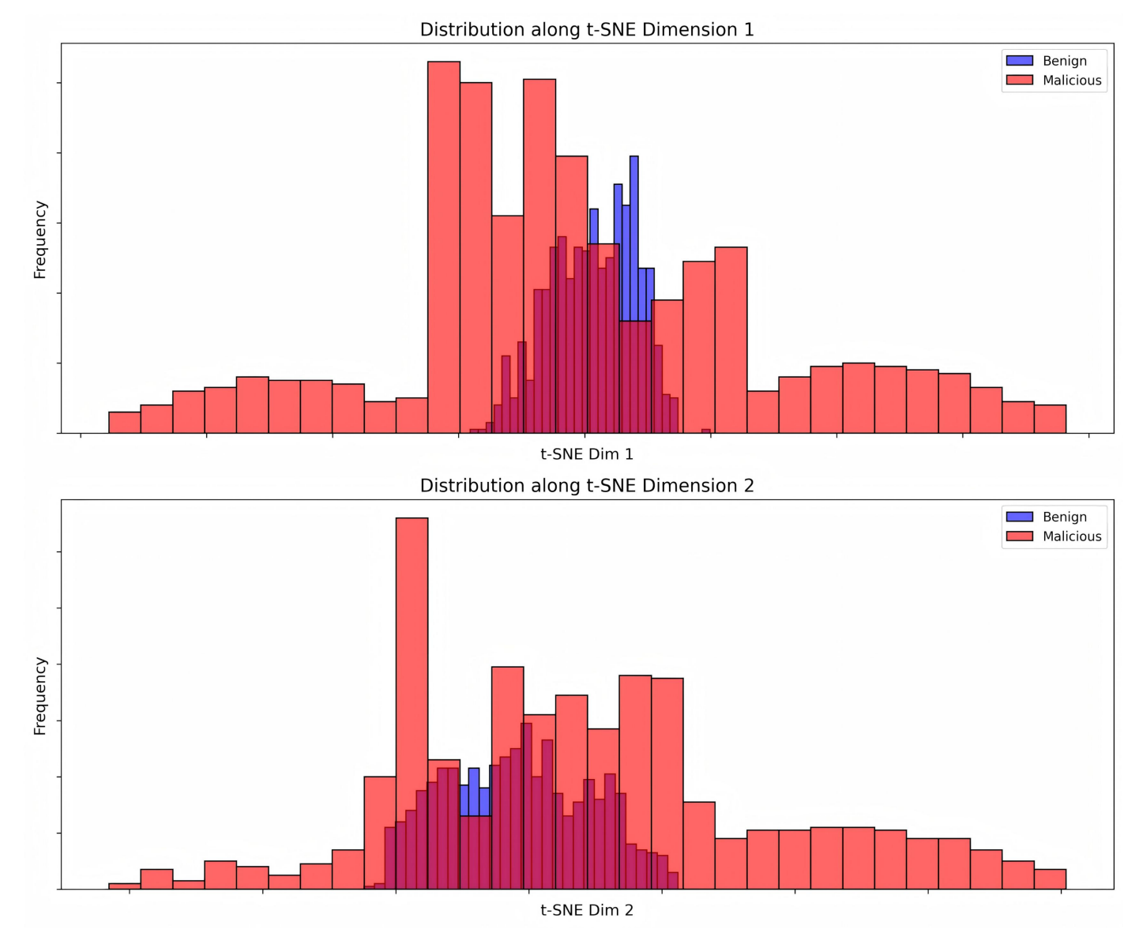


图1展示经过t-SNE降维后的代码嵌入分布，红色叉号为恶意代码，蓝色圆点为良性代码，显示出两者在语义空间中分布显著不同。图2分别展示在两个t-SNE维度上的分布直方图，进一步突出恶意代码分布更分散、方差更大。

## 方法

本文提出LLM-CBC方法，包含分布建模与分类两阶段。首先构建两个独立VAE模型，分别学习良性与恶意代码的语义分布，通过BERT提取特征并生成样本，获取重构距离和潜在变量统计量。随后将上述信息构成高维特征向量输入三层MLP分类器C-model，输出0~1之间的置信分数以判别样本类型。该方法实现了跨语言的细粒度恶意代码检测，兼顾语义建模与分类精度。

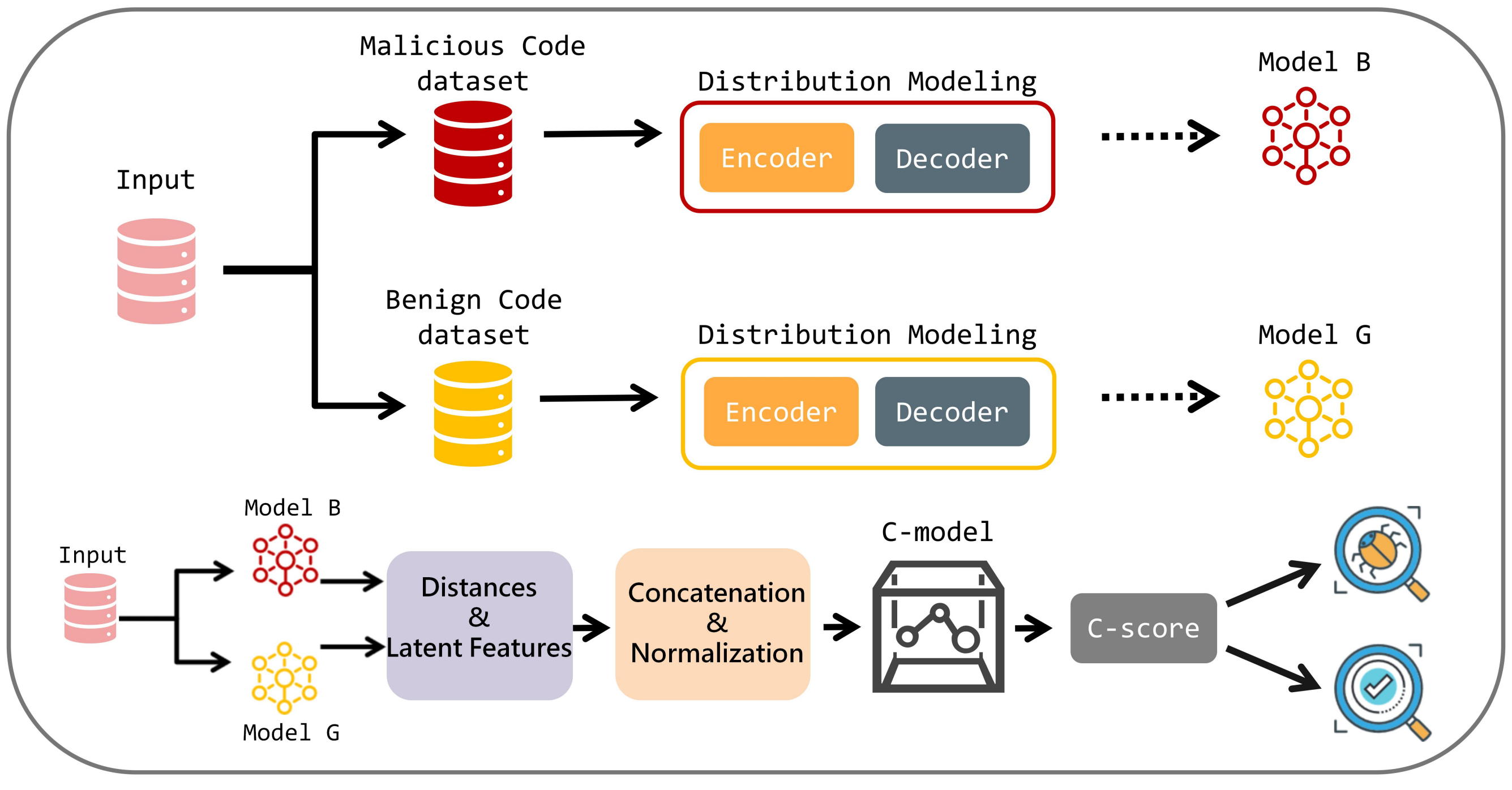


图3展示方法架构：通过双VAE建模良性与恶意代码分布，提取特征后由C-model输出置信分数，实现细粒度分类。

## 实验

实验在多种跨语言恶意代码数据集上进行，LLM-CBC在准确率、召回率和F1值等指标上均显著优于现有方法。对比与消融实验表明，双VAE建模和C-model分类器对性能提升均有关键作用，验证了方法的有效性与鲁棒性。

| Tool | Malicious Code Classification | | | |
|---|---|---|---|---|
| | Acc. | Prec. | Rec. | F1. |
| *LLM-CBC* | **98.42** | 98.67 | **98.67** | **98.67** |
| *Grep* | 21.31 | 47.04 | 20.63 | 28.68 |
| *Clamav* | 38.07 | **100.00** | 1.32 | 2.61 |

(a) Various programming languages

| Tool | Phishing Dataset | | | |
|---|---|---|---|---|
| | Acc. | Prec. | Rec. | F1. |
| *LLM-CBC* | **95.02** | **95.83** | 90.58 | **93.13** |
| *Sanitize* | 49.35 | 49.41 | **99.35** | 66.00 |
| *Semgrep* | 66.57 | 47.37 | 5.14 | 9.28 |

(b) HTML

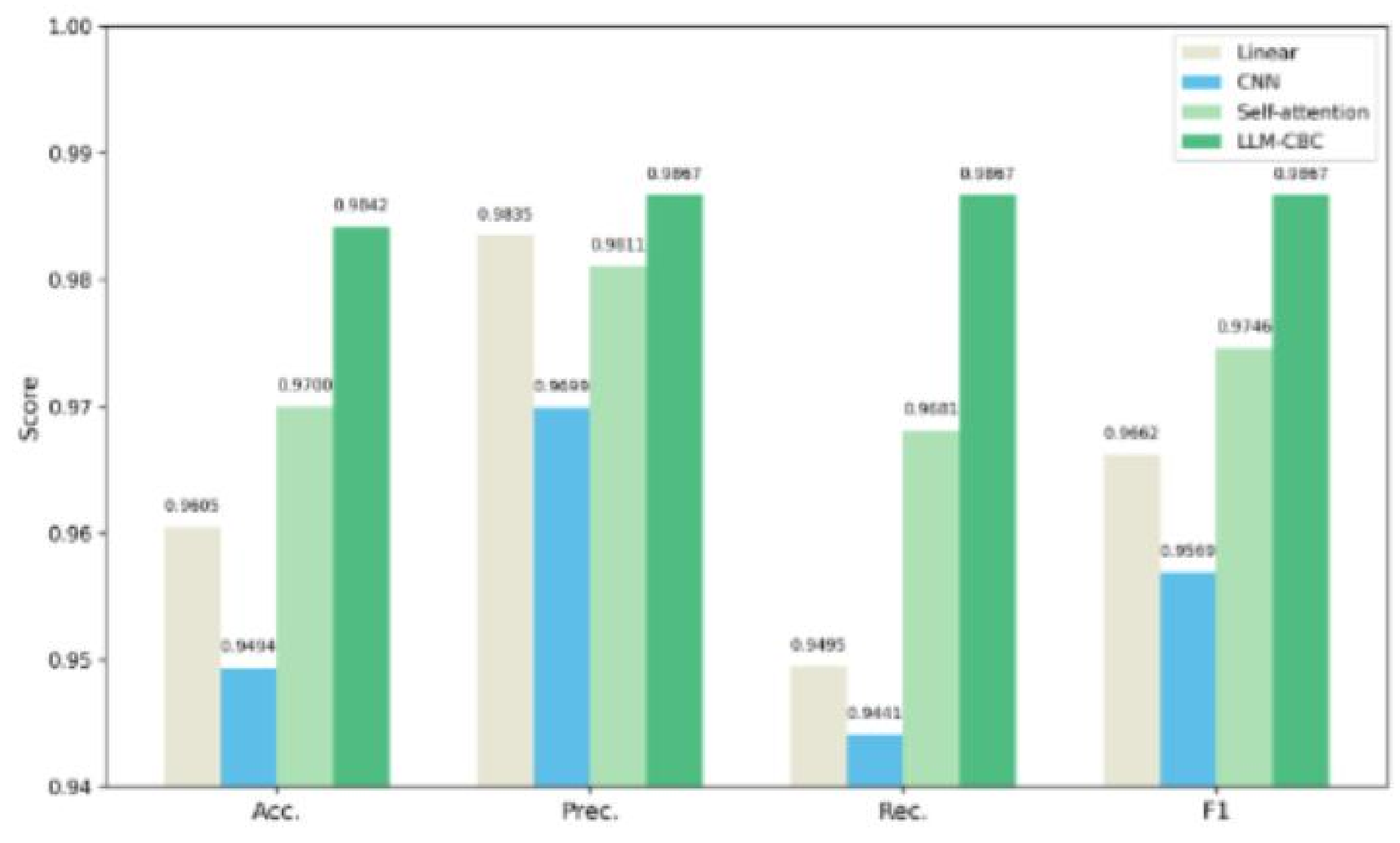
| Tool | Malicious Software Packages Dataset | | | | PyPI Malregistry | | | |
|---|---|---|---|---|---|---|---|---|
| | Acc. | Prec. | Rec. | F1. | Acc. | Prec. | Rec. | F1. |
| *LLM-CBC* | **96.03** | **96.09** | **95.96** | **96.03** | **97.27** | **96.28** | **98.34** | **97.30** |
| *Bandit* | 59.68 | 61.74 | 50.89 | 55.79 | 83.79 | 76.83 | 96.75 | 85.65 |
| *Pyre* | 91.30 | 91.20 | 91.41 | 91.31 | 95.17 | 92.99 | 97.66 | 95.27 |

(c) Python

| Tool | Malware Bench | | | | JavaScript Malware | | | |
|---|---|---|---|---|---|---|---|---|
| | Acc. | Prec. | Rec. | F1. | Acc. | Prec. | Rec. | F1. |
| *LLM-CBC* | **93.49** | **95.62** | **94.43** | **95.02** | **98.14** | **97.14** | **99.20** | **98.16** |
| *ESLint* | 46.27 | 66.94 | 30.14 | 41.56 | 82.73 | 83.47 | 89.00 | 86.14 |
| *Maltracker* | 87.82 | 88.29 | 90.50 | 89.38 | 85.59 | 88.60 | 85.50 | 87.02 |

(d) JavaScript

表III比较LLM-CBC与多种现有检测工具在不同语言数据集上的性能，LLM-CBC在各项指标上均表现最佳。



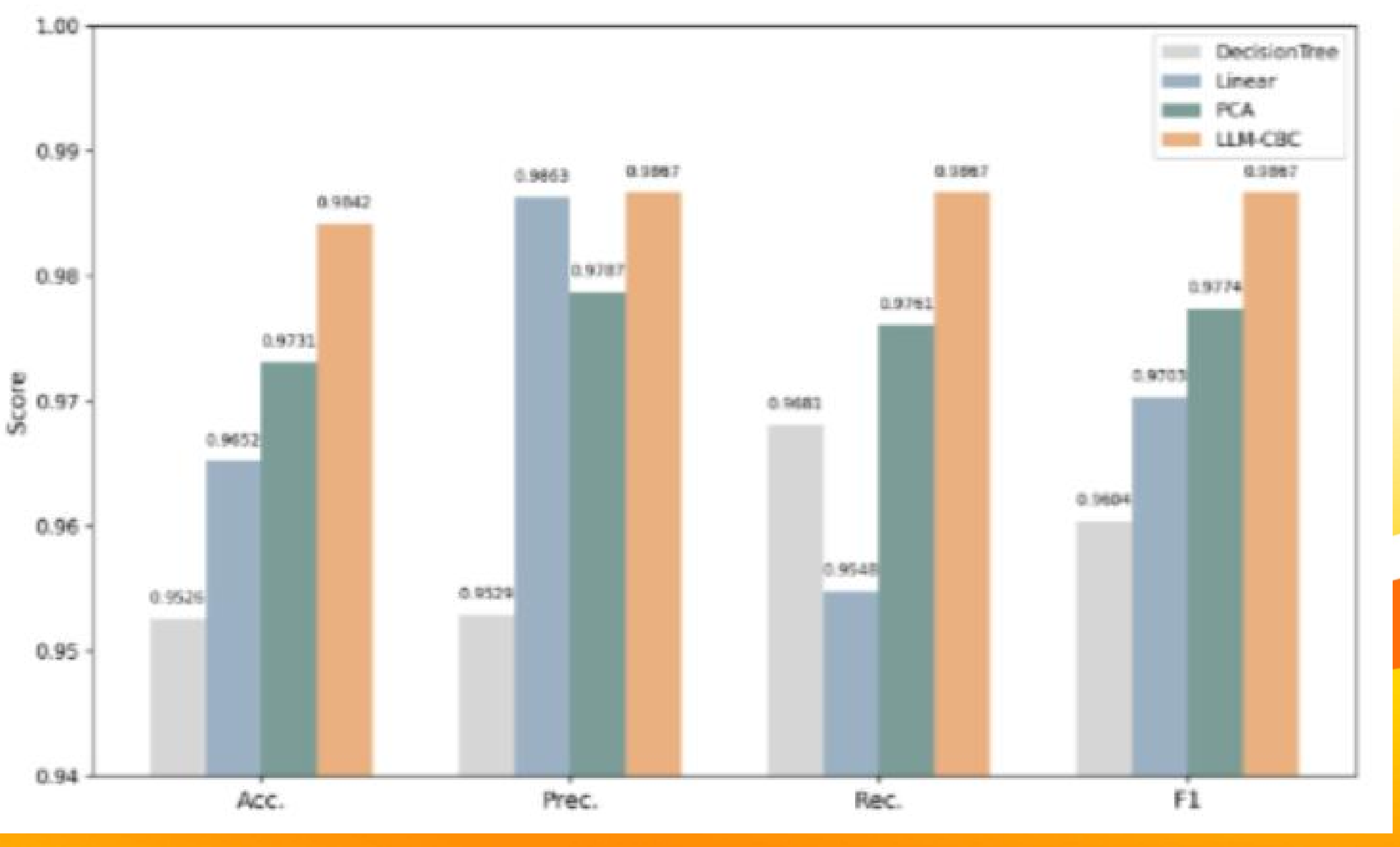


图6和图7展示在替换分布建模组件与分类器组件后的性能对比，结果表明原始双VAE与C-model设计在准确率和F1值等方面明显优于替代方案。