Training Data Heterogeneity Effects on Code Model Vulnerability Detection Performance

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does training data heterogeneity across C, C++, and Python affect the F1 score of 7B-parameter code models compared to 70B-parameter models in CWE vulnerability detection. Abstract Deep learning (DL) is one of the fastest-growing topics in materials data science, with rapidly emerging applications spanning atomistic, image-based, spectral, and textual data modalities. DL allows analysis of unstructured data and automated identification of features. 13 claims were extracted from source literature; 13 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Recent advances and applications of deep learning methods in materials science. Research question: How does training data heterogeneity across C, C++, and Python affect the F1 score of 7B-parameter code models compared to 70B-parameter models in CWE vulnerability detection?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.0/10.

3 Results

13 papers retrieved. 13 claims extracted; 13 independently verified. Quality review score: 8.0/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Deep learning (DL) is one of the fastest-growing topics in materials data science.	✓	0.31
Deep learning applications in materials science span atomistic, image-based, spectral, and textual data modalities.	✓	0.30
Deep learning allows analysis of unstructured data.	✓	0.21
Deep learning allows automated identification of features.	✓	0.18
The recent development of large materials databases has fueled the application of DL methods in atomistic prediction.	✓	0.34
Advances in image and spectral data have largely leveraged synthetic data enabled by high-quality forward models.	✓	0.32
Advances in image and spectral data have largely leveraged synthetic data enabled by generative unsupervised DL methods.	✓	0.33
The article presents a detailed discussion of recent developments of deep learning in atomistic simulation, materials im	✓	0.36
The article discusses applications involving both theoretical and experimental data for each modality.	✓	0.17
The article discusses typical modeling approaches with their strengths and limitations for each modality.	✓	0.18
The article discusses relevant publicly available software and datasets for each modality.	✓	0.15
The review concludes with a discussion of recent cross-cutting work related to uncertainty quantification in this field.	✓	0.25
The review concludes with a perspective on limitations, challenges, and potential growth areas for DL methods in materia	✓	0.32

References

- https://doi.org/10.48550/arxiv.2307.06435

- https://doi.org/10.1038/s41524-022-00734-6
- https://doi.org/10.1186/s40537-023-00727-2