Literature Review for Synthetic Clone Pair Dataset Generation

Lero Technical Report No. 2024_TR_01_Synthetic_Clone_Pair_Dataset_Generation

James Vincent Patten[*], Gul Aftab Ahmed[†], Yuanhua Han[‡], Guoxian Lu[§],
David Gregg[¶], Jim Buckley[∥] and Muslim Chochlov[**]
Email: [*]james.patten@lero.ie, [†]ahmedga@tcd.ie, [‡]hanyuanhua2@huawei.com,
[§]luguoxian@huawei.com, [¶]david.gregg@tcd.ie, [∥]jim.buckley@ul.ie, [**]muslim.chochlov@ul.ie

10 June 2024

T<small>ABLE OF</small> C<small>ONTENTS</small>

I. INTRODUCTION

This report outlines the design, execution and results of a literature review conducted to find research articles related to the application of methods/techniques to automatically generate synthetic clone pair datasets. We have proposed an approach that uses grammatical evolution (GE) [1], and more specifically context-sensitive grammatical evolution (CSGE) [2], in combination with a large language model (LLM) to generate such datasets. We conducted this review to capture details of any existing state-of-the-art (SOTA) techniques that already support this, and additionally to gather any useful information that may support and inform the development of an approach that uses CSGE and LLMs. The search was conducted using the Elsevier database, Scopus, with several key search terms being defined in the form of a string that is compatible with the Scopus search interface. This search are discussed in more detail in Section II.

II. SEARCH OF AVAILABLE LITERATURE

An initial search returned seventy-five articles, but within these results were preprints, non-English language articles, and a range of articles from subject areas outside of software engineering, computer science, and areas such as artificial intelligence and machine learning. The initial search string was therefore updated to exclude such articles and the updated search string, as detailed in Listing 1, was used to search Scopus once again.

Listing 1. Outline of the search string used to find research articles on Scopus.

```
TITLE-ABS-KEY (
( "synthetic data" OR "data synth*" OR "data generation" OR
  "program synth*" OR "program generation" )
  AND ( "software clone*" OR "clone pair*" OR "clone*" ) )
  AND ( EXCLUDE ( SUBJAREA, "CHEM" ) OR EXCLUDE ( SUBJAREA, "MEDI" ) OR
        EXCLUDE ( SUBJAREA, "AGRI" ) OR EXCLUDE ( SUBJAREA, "PHYS" ) OR
        EXCLUDE ( SUBJAREA, "MULT" ) OR EXCLUDE ( SUBJAREA, "MATE" ) OR
        EXCLUDE ( SUBJAREA, "HEAL" ) OR EXCLUDE ( SUBJAREA, "IMMU" ) OR
        EXCLUDE ( SUBJAREA, "ENVI" ) OR EXCLUDE ( SUBJAREA, "EART" ) OR
        EXCLUDE ( SUBJAREA, "DECI" ) OR EXCLUDE ( SUBJAREA, "BIOC" ) )
  AND ( LIMIT-TO ( LANGUAGE, "English" ) )
```

This updated search string returned thirty-three articles. The title and abstract of each article were subsequently reviewed and based on this a decision was made as to the relevance of each article to the proposed synthetic clone pair dataset generation study. Details of these thirty-three articles, as well as the inclusion/exclusion decision made, can be found in Section III.

III. RESULTS OF LITERATURE SEARCH

The titles of the articles found, as well as a reference to their bibliography entry in References, can be found in Table I. After reviewing the titles and abstracts of these thirty-three articles, a total of three were deemed to be directly relevant to research into automatic generation of synthetic clone pair databases, with an additional four being considered as possible relevant and needing a full text review to determine this. This left twenty-six articles, all of which were deemed not to be relevant, or useful, to this research study.

Additional discussions were held by the Lero TREES group members to decide on the inclusion of articles which may not have been found as part of the Scopus search. This resulted in two additional articles being chosen for inclusion, namely:

1) Detection and analysis of near-miss software clones [3]
2) GPTCloneBench: A comprehensive benchmark of semantic clones and cross-language clones using GPT-3 model and SemanticCloneBench [4]

TABLE I

LIST OF TITLES AND ASSOCIATED BIBLIOGRAPHY ENTRIES FOR THE 33 ARTICLES FOUND ON SCOPUS.

	Title	Reference
1	Assessing Similarity-Based Grammar-Guided Genetic Programming Approaches for Program Synthesis	[5]
2	Unleashing the Power of Compiler Intermediate Representation to Enhance Neural Program Embeddings	[6]
3	Unified Pre-training for Program Understanding and Generation	[7]
4	Maze: Data-free model stealing attack using zeroth-order gradient estimation	[8]
5	XML version detection	[9]
6	Evolving assembly programs: How games help microprocessor validation	[10]
7	Problems Creating Task-relevant Clone Detection Reference Data	[11]
8	Loop Rerolling for Hardware Decompilation	[12]
9	A Novel Framework to Generate Synthetic Video for Foreground Detection in Highway Surveillance Scenarios	[13]
10	Finding Substitutable Binary Code by Synthesizing Adapters	[14]
11	NeuralPot: An Industrial Honeypot Implementation Based on Deep Neural Networks	[15]
12	29th International Symposium on Logic-Based Program Synthesis and Transformation, LOPSTR 2019	[16]
13	Output-oriented software testing data generation based on artificial immune algorithm	[17]
14	BENEVOL 2019 - Proceedings of the 18th Belgium-Netherlands Software Evolution Workshop	[18]
15	Discrimination between genuine and cloned gait silhouette videos via autoencoder-based training data generation	[19]
16	Implementation of threshold comparator using cartesian genetic programming on embryonic fabric	[20]
17	Synergy of two mutations based immune multi-objective automatic fuzzy clustering algorithm	[21]
18	Scrible: Ultra-accurate error-correction of pooled sequenced reads	[22]
19	A model for cancer tissue heterogeneity	[23]
20	Comparison of Hadoop multiple node cluster performance over physical and virtual nodes using inverted index data structure for search over Wikipedia data set	[24]
21	A point symmetry-based clonal selection clustering algorithm and its application in image compression	[25]
22	On the modeling of heterogeneity in cancer tissue	[26]
23	Logic-Based Program Synthesis and Transformation - 22nd International Symposium, LOPSTR 2012, Revised Selected Papers	[27]
24	Test data generation for object-oriented methodology using clonal selection algorithm	[28]
25	Logic-Based Program Synthesis and Transformation - 21st International Symposium, LOPSTR 2011, Revised Selected Papers	[29]
26	Immunodomaince based Clonal Selection Clustering Algorithm	[30]
27	A clonal selection clustering algorithm using pointed symmetry-based distance measure	[31]
28	Gene transposon based clonal selection algorithm for clustering	[32]
29	AN SS-SVM approach to generate synthetic network delays	[33]
30	Applying mesh conformation on shape analysis with missing data	[34]
31	Composition and Cloning in Modeling and Meta-Modeling	[35]
32	Problems creating task-relevant clone detection reference data	[36]
33	Proceedings - 16th annual international conference on automated software engineering, ASE 2001	[37]

IV. Conclusion

TABLE II
LIST OF ARTICLES DEEMED AS RELEVANT TO THE PROPOSED STUDY.

Title	Include Decision	Reference
Assessing Similarity-Based Grammar-Guided Genetic Programming Approaches for Program Synthesis	Yes	[5]
Unleashing the Power of Compiler Intermediate Representation to Enhance Neural Program Embeddings	Yes	[6]
Unified Pre-training for Program Understanding and Generation	Yes	[7]
Detection and analysis of near-miss software clones	Yes	[3]
GPTCloneBench: A comprehensive benchmark of semantic clones and cross-language clones using GPT-3 model and SemanticCloneBench	Yes	[4]
Maze: Data-free model stealing attack using zeroth-order gradient estimation	Maybe	[8]
XML version detection	Maybe	[9]
Evolving assembly programs: How games help microprocessor validation	Maybe	[10]
Problems Creating Task-relevant Clone Detection Reference Data	Maybe	[36]

[5] was included as it directly deals with program synthesis using a grammar guided evolutionary technique, which is very much in line with what CSGE will be used to accomplish. [6] was included as it presents research on how evolved embeddings can be used to improve code clone detection. [7] was included as it details research on a system, PLBART, designed to support program synthesis for tasks such as clone detection. [3] was included as it is fundamental related to the detection and analysis of near-miss software clones. Finally, [4] was included as it outlines research in which a LLM, namely GPT-3, was used to produce a range of different clone pair types. The full text of the "maybe" articles, [8], [9], [10], and [36], will be reviewed and subsequently included/excluded in the study based on their review.

REFERENCES

[1] M. O'Neill and C. Ryan, "Grammatical evolution," *IEEE Transactions on Evolutionary Computation*, vol. 5, no. 4, pp. 349–358, 2001.

[2] J. V. Patten, "Context sensitive grammatical evolution: A novel attribute grammar based approach to the integration of semantics in grammatical evolution," Ph.D. dissertation, University of Limerick, 2022.

[3] C. K. Roy, "Detection and analysis of near-miss software clones," in *2009 IEEE International Conference on Software Maintenance*, 2009, pp. 447–450.

[4] A. I. Alam, P. R. Roy, F. Al-Omari, C. K. Roy, B. Roy, and K. A. Schneider, "Gptclonebench: A comprehensive benchmark of semantic clones and cross-language clones using gpt-3 model and semanticclonebench," in *2023 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE, 2023, pp. 1–13.

[5] N. Tao, A. Ventresque, and T. Saber, "Assessing similarity-based grammar-guided genetic programming approaches for program synthesis," *Communications in Computer and Information Science*, vol. 1684 CCIS, pp. 240–252, 2022, cited By 0. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85145251993&doi=10.1007%2f978-3-031-22039-5_19&partnerID=40&md5=937d7ba95291cbdffe231b2bbb85417f

[6] Z. Li, P. Ma, H. Wang, S. Wang, Q. Tang, S. Nie, and S. Wu, "Unleashing the power of compiler intermediate representation to enhance neural program embeddings," vol. 2022-May, 2022, pp. 2253–2265, cited By 9. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85133516020&doi=10.1145%2f3510003.3510217&partnerID=40&md5=7e8c2a23f25cdeb7ecc4bd4aa83fc7e0

[7] W. Ahmad, S. Chakraborty, B. Ray, and K.-W. Chang, "Unified pre-training for program understanding and generation," 2021, pp. 2655–2668, cited By 188. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85137738691&partnerID=40&md5=f9ac41e96cce90ac9d53338225cca064

[8] S. Kariyappa, A. Prakash, and M. Qureshi, "Maze: Data-free model stealing attack using zeroth-order gradient estimation," 2021, pp. 13 809–13 818, cited By 44. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85117666011&doi=10.1109%2fCVPR46437.2021.01360&partnerID=40&md5=f4315f702730045c40746bf5fe637c34

[9] D. De Brum Saccol, N. Edelweiss, R. De Matos Galante, and C. Zaniolo, "Xml version detection," 2007, pp. 79–88, cited By 5. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-37849018673&doi=10.1145%2f1284420.1284441&partnerID=40&md5=37f8f1788f0b020d6c8960a1b62a9baf

[10] F. Corno, E. Sánchez, and G. Squillero, "Evolving assembly programs: How games help microprocessor validation," *IEEE Transactions on Evolutionary Computation*, vol. 9, no. 6, pp. 695–706, 2005, cited By 39. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-29244453993&doi=10.1109%2fTEVC.2005.856207&partnerID=40&md5=1531351db2d178399b9746509341dba7

[11] A. Walenstein, N. Jyoti, J. Li, Y. Yang, and A. Lakhotia, "Problems creating task-relevant clone detection reference data," vol. 2003-January, 2003, pp. 285–294, cited By 27. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84944630622&doi=10.1109%2fWCRE.2003.1287259&partnerID=40&md5=c63d1ca47b2906009be2624688aeed35

[12] Z. Sisco, J. Balkind, T. Sherwood, and B. Hardekopf, "Loop rerolling for hardware decompilation," *Proceedings of the ACM on Programming Languages*, vol. 7, 2023, cited By 1. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85162050431&doi=10.1145%2f3591237&partnerID=40&md5=b9948b19a8a905aa605994beb6c85eb0

[13] X. Li, H. Duan, B. Liu, X. Wang, and F.-Y. Wang, "A novel framework to generate synthetic video for foreground detection in highway surveillance scenarios," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 6, pp. 5958–5970, 2023, cited By 3. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85151496537&doi=10.1109%2fTITS.2023.3253919&partnerID=40&md5=d475e7501753b2267102e796fffb4ecb

[14] V. Sharma, K. Hietala, and S. McCamant, "Finding substitutable binary code by synthesizing adapters," *IEEE Transactions on Software Engineering*, vol. 47, no. 8, pp. 1626–1643, 2021, cited By 1. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85071562859&doi=10.1109%2fTSE.2019.2931000&partnerID=40&md5=62b0d09a053c425756d55cdbdfd4e3dc

[15] I. Siniosoglou, G. Efstathopoulos, D. Pliatsios, I. Moscholios, A. Sarigiannidis, G. Sakellari, G. Loukas, and P. Sarigiannidis, "Neuralpot: An industrial honeypot implementation based on deep neural networks," vol. 2020-July, 2020, cited By 13. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85094130058&doi=10.1109%2fISCC50000.2020.9219712&partnerID=40&md5=c4c648a52bb8b81080165307362dc0e3

[16] "29th international symposium on logic-based program synthesis and transformation, lopstr 2019," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12042 LNCS, 2020, cited By 0. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85084218142&partnerID=40&md5=8227dcfe55c86fdf5d3cc52a3b8dd4f1

[17] W. Zhang, Y. Qi, Z. Dou, W. Liu, B. Wei, and M. Zhang, "Output-oriented software testing data generation based on artificial immune algorithm," 2019, pp. 298–303, cited By 0. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85076421044&doi=10.1109%2fDDCLS.2019.8909069&partnerID=40&md5=fe510fe4171fe03bbfcadb370593cf28

[18] "Benevol 2019 - proceedings of the 18th belgium-netherlands software evolution workshop," vol. 2605, 2019, cited By 0. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85089913118&partnerID=40&md5=eeb2e926a303945d8153fc141b715295

[19] Y. Hirose, K. Nakamura, N. Nitta, and N. Babaguchi, "Discrimination between genuine and cloned gait silhouette videos via autoencoder-based training data generation," *IEICE Transactions on Information and Systems*, vol. E102D, no. 12, pp. 2535–2546, 2019, cited By 0. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85076432798&doi=10.1587%2ftransinf.2019EDP7042&partnerID=40&md5=860aa2134d2f3da2837217b3efb7790f

[20] G. Malhotra, V. Lekshmi, S. Sudhakar, and S. Udupa, "Implementation of threshold comparator using cartesian genetic programming on embryonic fabric," *Advances in Intelligent Systems and Computing*, vol. 939, pp. 93–102, 2019, cited By 6. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85066052765&doi=10.1007%2f978-3-030-16681-6_10&partnerID=40&md5=8f0a370440ae2a7534224e0d0f131e35

[21] R. Liu, L. Zhang, B. Li, Y. Ma, and L. Jiao, "Synergy of two mutations based immune multi-objective automatic fuzzy clustering algorithm," *Knowledge and Information Systems*, vol. 45, no. 1, pp. 133–157, 2015, cited By 4. [Online].

Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84942295469&doi=10.1007%2fs10115-014-0805-4&partnerID=40&md5=dfd4de6b71367adc7f1d83b003fc76e1

[22] D. Duma, F. Cordero, M. Beccuti, G. Ciardo, T. Close, and S. Lonardi, "Scrible: Ultra-accurate error-correction of pooled sequenced reads," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9289, pp. 162–174, 2015, cited By 1. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2. 0-84947810099&doi=10.1007%2f978-3-662-48221-6_12&partnerID=40&md5=95bea4cd5cccbcfac5e5f0575cff78df

[23] A. Mohanty, A. Datta, and V. Venkatraj, "A model for cancer tissue heterogeneity," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 3, pp. 966–974, 2014, cited By 12. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2. 0-84896843907&doi=10.1109%2fTBME.2013.2294469&partnerID=40&md5=92597056afc6b2a58616a65149130315

[24] D. Venkitaramanan, N. Vijayaraju, K. Velusamy, G. Suresh, and M. Divya, "Comparison of hadoop multiple node cluster performance over physical and virtual nodes using inverted index data structure for search over wikipedia data set," *International Journal of Applied Engineering Research*, vol. 9, no. 16, pp. 3515–3532, 2014, cited By 0. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84904353694&partnerID=40&md5=5d3b02b7c28783b149fe28a8a8db624d

[25] R. Liu, F. He, J. Liu, W. Ma, and Y. Li, "A point symmetry-based clonal selection clustering algorithm and its application in image compression," *Pattern Analysis and Applications*, vol. 17, no. 3, pp. 633–654, 2014, cited By 0. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84903983359&doi=10.1007%2fs10044-013-0344-8&partnerID=40&md5=89cf1d4139903ca04da920d87d0c9384

[26] A. Mohanty, A. Datta, and J. Venkatraj, "On the modeling of heterogeneity in cancer tissue," 2013, p. 93, cited By 0. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84897705778&doi=10.1109%2fGlobalSIP.2013.6736821&partnerID= 40&md5=617ca1047bae2b3165fdc549dfb2b83a

[27] "Logic-based program synthesis and transformation - 22nd international symposium, lopstr 2012, revised selected papers," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7844 LNCS, 2013, cited By 1. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84887276304&partnerID= 40&md5=c3b0af3e3d761adc8a3e35b61618e55f

[28] Y. Suresh, M. Sharma, S. Islam, and S. Rath, "Test data generation for object-oriented methodology using clonal selection algorithm," 2013, pp. 27–33, cited By 0. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84898836253&partnerID= 40&md5=88f11a8982c6d1899976e58f260d58dd

[29] "Logic-based program synthesis and transformation - 21st international symposium, lopstr 2011, revised selected papers," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7225 LNCS, 2012, cited By 0. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84865113074&partnerID= 40&md5=997ea4a05c93448469526a600fe86bc3

[30] R. Liu, X. Zhang, N. Yang, Q. Lei, and L. Jiao, "Immunodomaince based clonal selection clustering algorithm," *Applied Soft Computing Journal*, vol. 12, no. 1, pp. 302–312, 2012, cited By 14. [Online]. Available: https://www.scopus.com/inward/record.uri? eid=2-s2.0-81155155632&doi=10.1016%2fj.asoc.2011.08.042&partnerID=40&md5=049b956fc33b814d1fdfc30a550b7d5f

[31] R. Liu, H. Ning, W. Zhang, and L. Jiao, "A clonal selection clustering algorithm using pointed symmetry-based distance measure," 2010, pp. 1063–1064, cited By 0. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-77955894026&doi=10. 1145%2f1830483.1830676&partnerID=40&md5=e7a025f0f2a9cc2cb95c0ed4d8ca021f

[32] R. Liu, Z. Sheng, and L. Jiao, "Gene transposon based clonal selection algorithm for clustering," 2009, pp. 1251–1258, cited By 8. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-72749088683&doi=10.1145%2f1569901.1570069& partnerID=40&md5=1dca402c0649711bc0263183c102130b

[33] J. Hernández, I. Phillips, and J. Moguerza, "An ss-svm approach to generate synthetic network delays," 2005, pp. 125–131, cited By 1. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84856595356&partnerID=40&md5= 453c8f8f6ac7bdbc653321ee46ab2ebb

[34] X. Ju, Z. Mao, J. Siebert, N. McFarlane, J. Wu, and R. Tillett, "Applying mesh conformation on shape analysis with missing data," 2004, pp. 696–702, cited By 5. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-16244390248&partnerID= 40&md5=e6fae30567a0c2943a07f6b481ae53f6

[35] G. Karsai, M. Maroti, A. Ledeczi, J. Gray, and J. Sztipanovits, "Composition and cloning in modeling and meta-modeling," *IEEE Transactions on Control Systems Technology*, vol. 12, no. 2, pp. 263–278, 2004, cited By 59. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-1942536180&doi=10.1109%2fTCST.2004.824311& partnerID=40&md5=adafc2a0e5617fe45755b50eebb4abd8

[36] A. Walenstein, N. Jyoti, J. Li, Y. Yang, and A. Lakhotia, "Problems creating task-relevant clone detection reference data," 2003, pp. 285–294, cited By 24. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0346151254&doi=10.1109% 2fWCRE.2003.1287259&partnerID=40&md5=5c3afaee0248099ac6f92fa6152db742

[37] "Proceedings - 16th annual international conference on automated software engineering, ase 2001," 2001, cited By 0. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85116890332&partnerID=40&md5=14dc353cf7d6d25d216833dcd3a73486