Text-enriched graph embeddings for HVAC-related maintenance triage

Sajith Wettewa[a,b], Lei Hou[a,b*], Kevin Zhang[a,b]

[a] School of Engineering, RMIT University, Melbourne, Victoria, Australia
[b] Centre for Future Construction, RMIT University, Melbourne, Victoria, Australia
[*]Corresponding author

sajith.wettewa@rmit.edu.au, lei.hou@rmit.edu.au, kevin.zhang@rmit.edu.au

Abstract-
Facilities management teams encounter complex HVAC maintenance scenarios as faults propagate across interconnected subsystems. This study presents HVACONNECT, a text-enriched heterogeneous graph learning framework for HVAC-related maintenance triage. Using six years of CMMS work orders and BAS signals from a 19-building campus, a heterogeneous graph capturing building hierarchies, HVAC systems, and work orders was constructed. A large language model (LLM) pipeline was used to extract semantic data from raw work order text. A relation-aware heterogeneous GATv2 was then used to generate node embeddings via masked feature reconstruction. Feature ablation showed that removing text features reduced clustering quality by 44%, highlighting the critical value of semantic context. Downstream priority prediction achieved an operationally useful 83% precision despite severe class imbalance, and the learned component embeddings enabled meaningful maintenance-zone clustering.

Keywords-
Graph Deep Learning, GNNs, Priority Prediction, Maintenance Triage

1 Introduction

Heating, ventilation, and air conditioning (HVAC) systems dominate commercial building operational loads, requiring proactive maintenance to prevent comfort complaints and energy waste [1]. Traditionally, Fault Detection and Diagnosis (FDD) relies on Building Automation System (BAS) sensor streams. However, scaling BAS analytics across diverse portfolios is hindered by cross-building variability, inconsistent instrumentation, and high operational overhead [2].

Computerised Maintenance Management Systems (CMMS) offer a highly accessible alternative. Work orders routinely capture symptoms, actions, and priority signals without needing extra sensors. While recent Natural Language Processing (NLP) based methods can mine this text for triage, they typically treat records independently, ignoring the interconnected nature of building subsystems [3,4]. Graph learning naturally models these interdependencies. Graph Neural Networks (GNNs) couple equipment topology, operational context, and text-based maintenance events into a unified formulation. Heterogeneous graph architectures, in particular, allow the integration of mixed node and edge types while preserving narrative semantics [5,6].

To bridge this gap, this study introduces HVACONNECT (Heating Ventilation and Air Conditioning Operations Network for Node Embedding, Clustering, and Triage). By unifying BAS signals and CMMS work orders into a heterogeneous graph, our framework learns robust node representations. We demonstrate its operational utility through a three-level priority prediction task and maintenance zone clustering. Therefore, Section 2 reviews related work, followed by the proposed methodology in Section 3. Section 4 presents experimental results, and Section 5 concludes the paper.

2 Related work

BAS-driven fault detection and diagnosis has been widely studied using both domain-informed modelling and data-driven learning. However, persistent organisational and technical barriers exist, including inconsistent data quality, portfolio variability, and weak alignment between algorithm outputs and actionable maintenance workflows, which are explored below.

2.1 CMMS work order analytics

Maintenance logs were repeatedly shown to contain predictive signals when analysed using text mining and machine learning. Work order narratives support component-level fault frequency analysis, while portfolio benchmarking indicates that these could be used to distinguish maintenance scenarios by building user populations [3,7] More recent work that applied NLP to automate labour assignment used neural networks for priority prediction, while clustering and taxonomy generation supported structured knowledge discovery from unstructured records [4,8]. However, these approaches rarely modelled explicit relationships between assets, symptoms, and interventions as a unified relational system.

2.2 Graph learning for maintenance representation

GNNs were progressively used to encode relational structure for diagnostics and prognostics [5]. In HVAC-specific settings, semi-supervised Graph Convolutional Networks (GCNs) improved fault diagnosis under limited labels, which indicated that topology and relational neighbourhood signals were beneficial when data were scarce [6]. Broader graph learning methods, including neighbourhood aggregation, attention-based message passing, relational convolutions, and heterogeneous graph transformers, enabled scalable modelling of multi-typed nodes and edges [9]. Text-based graph learning further demonstrate the capabilities of integrating text narratives into embeddings [10]. Together, these methods motivate our approach of treating maintenance planning as learning over a heterogeneous, text-bearing building graph rather than as isolated prediction over sensors or tickets.

3 Methodology

3.1 Data set

The data set covered BAS variables with CMMS work orders from a North American university complex as summarised in Table 1.Work orders were available in text format as exported tabular records. Each record contained a text description field alongside structured metadata fields such as dates, component identifiers, and cost values, which were directly ingested into the processing pipeline.

The scope focused on centralised HVAC assets aligned with Uniformat II[11] and Whitestone specifications [12]. Work orders were organised by system, subsystem, and component using unified identifiers for cross-building consistency, and each work order was originally linked to a HVAC component. Only reactive maintenance events were analysed.

Table 1.Work order dataset characteristics

Dataset attribute	Description
Data collection timeline	2014-2019 (6 years)
Institutional distribution	19 Buildings
Dataset volume	2999 distinct work order records
HVAC classification architecture	System, Subsystem, Component
Performance metrics	Labour hours, Material costs, Total maintenance costs
BAS Data	Temperature, Humidity, Atmospheric pressure, Wind Conditions

3.2 Analysis of unstructured work order descriptions

CMMS descriptions used inconsistent terminology and documentation styles. A data extraction step was applied to recover problem and action narratives, floor designations, and room markers, as illustrated in Figure 1. A 300-row gold standard was constructed by two domain experts with over five years of facilities management experience, who independently extracted problem and action labels from raw CMMS descriptions. Disagreements were resolved by a third expert. Four extraction sources were scored against this gold standard. LLM variants comprised three Gemini 2.5 Pro [13] prompt configurations (zero-shot, one-shot, and few-shot, with an increasing number of examples) and an open-source model Qwen3-30B-A3B-Instruct [14]. Each row was scored as correct when all fields matched the gold standard. The few-shot Gemini prompt achieved 99.7% accuracy, with similar performance from Qwen. The few-shot Gemini was retained as the production configuration. Across the full dataset, only 9 rows were flagged for manual checks, confirming that the pipeline operates reliably at scale with minimal human intervention.

Two candidate embedding models, OpenAI text-embedding-3-large [15] and Gemini embedding-001 [15], were evaluated across five metrics to be used as the text embedding model. While Gemini achieved higher thematic cohesion (0.730 vs 0.586), OpenAI demonstrated stronger inter-theme separation (0.646 vs 0.840, lower is better) and higher negation contrast (0.309 vs 0.198). OpenAI was selected as separation and negation sensitivity are more critical for distinguishing opposite system

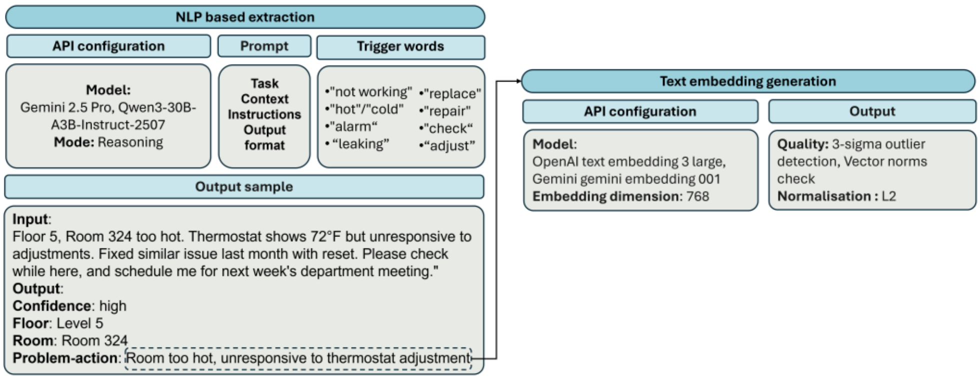

Figure 1. NLP extraction and text embedding generation workflow with other leading models used for comparison of extraction and embedding quality

states, such as "no airflow" versus "adequate airflow" in HVAC maintenance records. 768 embedding dimension is chosen as a standard widely used sentence transformer dimension [16]

3.3 Heterogeneous graph and fault propagation modelling

A heterogeneous HVAC operations graph was constructed to represent potential fault propagation through spatial, functional, and temporal relationships. The design followed a tripartite framing that aligned the representation with standards for system classification, maintenance documentation, and control mapping. The physical infrastructure layer encoded six containment levels, linking buildings to floors, floors to rooms, rooms to HVAC systems, systems to components, and components to work orders. Beyond containment, system-to-system edges captured control interactions and functional flow, and work order-to-work order edges captured temporal adjacency. Edge types and semantics are listed in Table 2.

Table 2. Graph edges and relationships

Source node	Target node	Relationship type
Building	Floor	contains floor
Floor	Room	contains room
Room	System	has system
System	Component	has component
Component	Work Order	has workorder
System	System	controls system
System	System	flows to system (flow type)
Work Order	Work Order	temporal (time difference)

An HVAC-specific flow topology was encoded using system flow edges to reflect service delivery pathways. Cooling generation supplied distribution networks that served downstream terminals, and heat generation followed a similar structure, while control edges represented feedback relationships. Temporal edges linked work orders within a seven-day window to reflect operational interdependence. The median inter-event time between consecutive maintenance events on related components is 4.2 days, with 62.4% of events falling within seven days. Temporal edge weights follow a linear decay from 1.0 at same-day to near-zero at the window boundary. A sensitivity analysis across 3, 7, 14, and 30-day windows confirmed stable results, with chain counts ranging from 225 to 357 and the functional similarity threshold having negligible effect, confirming that the seven-day window is a balanced operating point rather than an arbitrary choice.

3.4 Heterogeneous graph feature engineering

After graph construction, feature sets were assigned to each node type using BAS and CMMS data, as summarised in Table 3. Work order nodes combined text embeddings with numeric attributes representing problem diversity, thermal context, cost signals, and priority information. Features were normalised to stabilise message passing using z-score standardisation for scalar features and L2 normalisation for text embeddings.

Table 3. Summary of feature engineering

Node type	Feature groups	Representation
Building	Size, age, condition and cost indicators	Normalised numerics
Floor	Floor index, room count, subsystem counts	Normalised counts
Room	Floor index, subsystem presence	Indices, binary flags

System	System category, component count	One hot, normalised counts
Component	System family, component class	One hot, categorical codes
Work order	Text semantics, environment, priority, cost, problem diversity	Embeddings, engineered numerics, frequency features

3.5 Graph representation learning and downstream evaluation

Standard graph attention on homogeneous graphs does not preserve relation-specific semantics, therefore a customised heterogeneous GATv2 was implemented to learn embeddings under edge-type attention and multimodal feature integration. The building mask (BM) module given in Eq. (1), (2) reduces cross-building information leakage for within-building analyses. The overall architecture is shown in Figure 2.Figure 2. HVACONNECT model architecture

3.6 Heterogeneous graph attention architecture

Each message passing layer applied relation-specific GATv2 operators and aggregated relation-wise messages using a mean operator to avoid domination by high-degree relations. Four edge categories, building hierarchy, HVAC system structure, HVAC flow, and temporal linkage, were handled by separate operators. Each operator used four attention heads with 32 dimensions per head, producing 128-dimensional hidden representations after concatenation. Work order nodes used a two-stage projection with increasing dropout, while other node types used a single projection to 128 dimensions.

3.7 Domain-informed learning objective

The model was trained transductively using masked feature reconstruction. During training, subsets of input features were masked and the model reconstructed only masked positions, which encouraged reliance on neighbourhood information and reduced identity mapping. Masking rates were 30% for 768-dimensional language features and 10% for engineered scalar features. Reconstruction loss in Eq. (3) used mean squared error on masked indices only. A diversity regulariser Eq. (4), adapted from redundancy reduction principles in Barlow Twins [17] and VICReg [18], maximised mean pairwise Euclidean distance across all component embeddings to counteract global embedding collapse. A separation loss Eq. (6), grounded in supervised contrastive learning [19], operated on building-masked embeddings using cosine similarity to increase within-building alignment while reducing across-building similarity, preserving building-level identity distinct from the global geometric spread enforced by the diversity term. Total loss is given in Eq. (7).

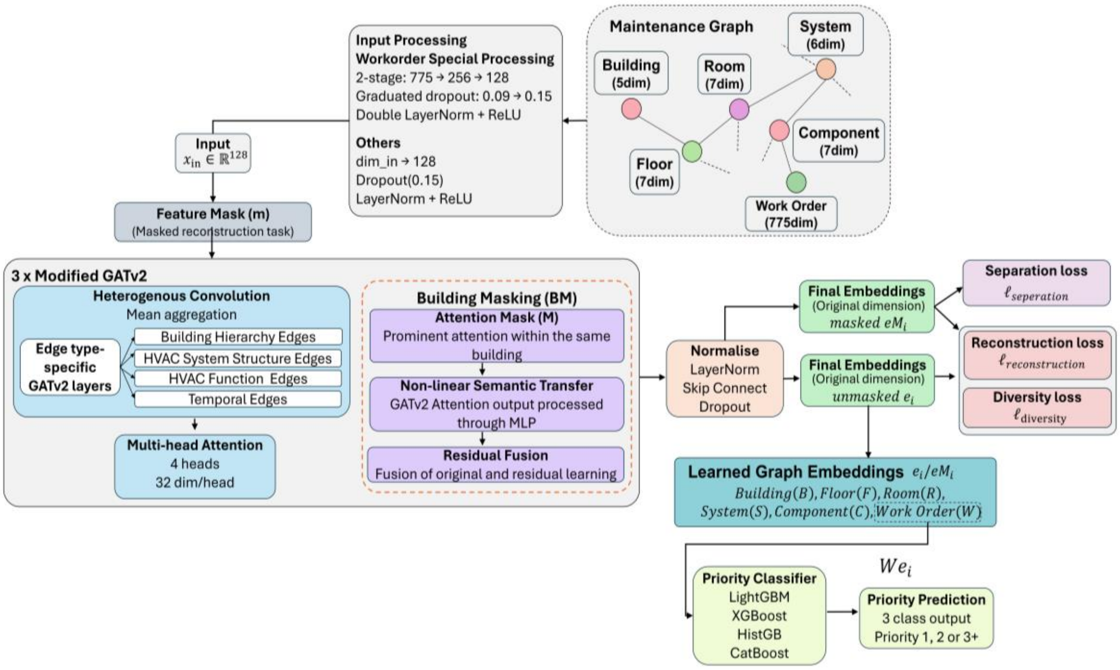

Figure 2. HVACONNECT model architecture

$$M_{ij} = \begin{cases} 0, & b(i) = b(j) \\ -\infty, & b(i) \neq b(j) \end{cases} \quad (1)$$

$$\alpha_{ij}^{BM} = \mathrm{softmax}_j(e_{ij} + M_{ij}) \quad (2)$$

$$\ell_{\text{reconstruction}} = \frac{1}{|m|} \sum_{i \in m} \| D_t(e_i) - x_i^{in} \|^2 \quad (3)$$

$$\ell_{diversity} = -\frac{1}{N(N-1)} \sum_{i \neq j} \| e_i - e_j \|_2 \quad (4)$$

$$\bar{e}_i^M = \frac{e_i^M}{\|e_i^M\|_2}, S_{ij} = (\bar{e}_i^M)^\top \quad (5)$$

$$\ell_{\text{separation}} = -\mathrm{mean}(S_{ij} \mid b(i) = b(j), i \neq j) + \mathrm{mean}(S_{ij} \mid b(i) \neq b(j)) \quad (6)$$

$$\ell_{total} = \sum_t w_t \; \ell_{reconstruction}^{(t)} + \lambda + \gamma \, \ell_{separation} \quad (7)$$

Here, $b(i)$denotes the building membership of node i, $e_{ij}$the raw GATv2 attention score, and $\alpha_{ij}^{BM}$the attention weight after building masking. The set $m$contains the masked feature positions, $x_i^{in}$is the original input feature of node i, $e_i$and $e_i^M$are the node embeddings before and after masking, $D_t$is the decoder associated with node type t, $N$is the number of nodes, $S_{ij}$is the cosine similarity between L2-normalised masked embeddings, and $w_t$is the reconstruction weight for node type t. In total loss, $w_t$is the reconstruction weight for node type t, and λand γare the weights for the diversity and separation losses, respectively.

4 Experiments and results

4.1 Priority triage

The applicability of GNN-derived embeddings was first analysed using a priority prediction task using downstream classifiers. Raw CMMS priority values were grouped into three classes to reduce sparsity while retaining operational distinctions of urgent, high, and all remaining priorities.

Models were evaluated using two splits as given in Table 4. Priority labels were severely imbalanced, with 1,186 urgent work orders, 1,731 high-priority work orders, and 82 work orders in the remaining group. Under the stratified split, gradient boosting models achieved strong discrimination. With XGBoost, macro precision reached 0.832 and macro AUROC reached 0.837, indicating that the embedding representation separated classes despite macro averaging penalising minority performance. Under the temporally ordered split, performance decreased modestly. A text-only baseline trained on problem action embeddings underperformed the full HVACONNECT representation, which indicated that triage benefited from combining text semantics with relational and contextual signals learned through heterogeneous message passing.

4.2 Embedding validation and ablation evidence

Ablation analysis (Figure 3) reveals that removing problem and action text features causes the largest degradation in clustering quality, followed by thermal context feature Figure 3. Feature ablation. All changes are sign-normalised (negative:degradation). Baseline: Silhouette:0.896, DB:0.352, IBC:0.221, IBD:0.189. Overall is the unweighted mean of all four metrics. Removing priority deviation has minimal impact, confirming the model does not rely on it as an implicit shortcut. Clustering metrics, namely Intra-Building Cohesion (IBC), Inter-Building Distinction (IBD), Silhouette, and Davies-Bouldin (DB), confirm the embeddings successfully capture building context while maintaining coherent, well-separated structure.

Table 4. Priority prediction metrics

Model	Precision (weighted)	Recall (weighted)	F1 (weighted)	Precision (macro)	Recall (macro)	F1 (macro)	AUROC (macro)
80/20 Split							
LightGBM	0.7652	0.7650	0.7622	0.7958	0.6454	0.6925	0.8349
XGBoost	0.7615	0.7583	0.7563	0.8316	0.6424	0.6961	0.8374
HistGB	0.7612	0.7583	0.7558	0.8325	0.6415	0.6959	0.8340
Temporal Split							
CatBoost	0.7189	0.7200	0.7192	0.5622	0.5381	0.5473	0.7779
XGBoost	0.7349	0.7480	0.7409	0.4958	0.5002	0.4976	0.7950
LightGBM	0.7314	0.7427	0.7362	0.4935	0.4957	0.4940	0.7694
Text embedding only (CatBoost)	0.6252	0.6275	0.6239	0.4412	0.4399	0.4384	0.7050

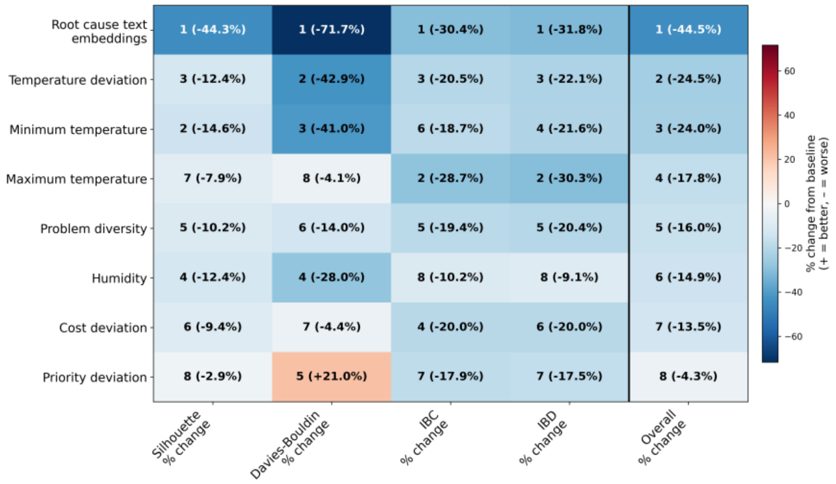

Figure 3. Feature ablation . All changes are sign-normalised (negative = degradation). Baseline: Silhouette:0.896, DB:0.352, IBC:0.221, IBD:0.189. Overall is the unweighted mean of all four metrics

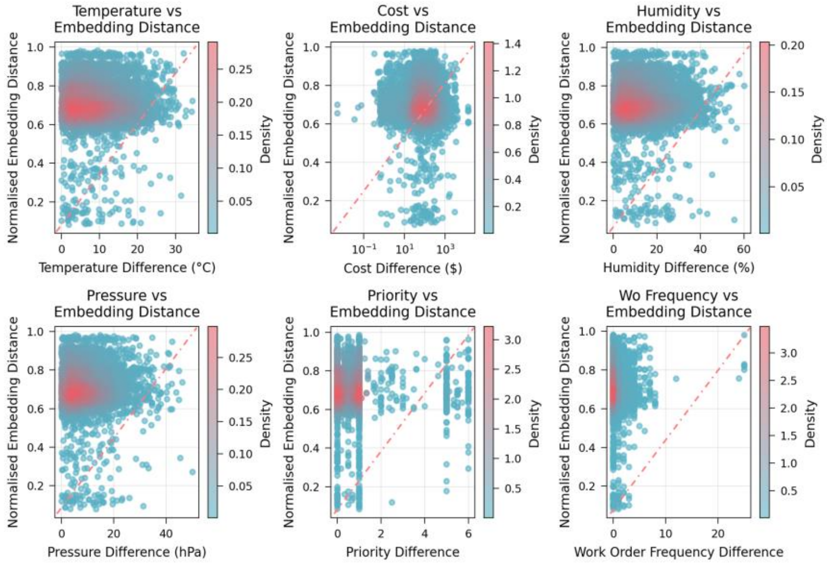

Figure 4. Embedding distance visualisation for feature reconstruction using HVACONNECT

Figure 4 maps pairwise operational attribute differences against normalised embedding distances. Most points sit above the hypothetical y = x line (where a single attribute would explain separation), proving HVACONNECT integrates multiple features simultaneously. Dense clusters with near-zero attribute differences but high embedding distances confirm that contextual factors, such as spatial location, component type, and message-passing dependencies, heavily drive the representations. Additionally, the log-scaled cost subplot shows components with drastically different maintenance costs can remain close in embedding space. This demonstrates that the attention mechanism effectively balances environmental, operational, and economic attributes without allowing any single feature to dominate.

4.3 Maintenance zone identification

A further application of the learned HVACONNECT component embeddings focused on optimising maintenance zones. Traditional zone-based strategies often relied on static, generalised boundaries that were not sensitive to operational conditions or functional interdependencies, which could create inefficiencies and overlapping responsibilities [20]. HVACONNECT addressed this gap by deriving data-driven maintenance zones directly from learned embeddings.

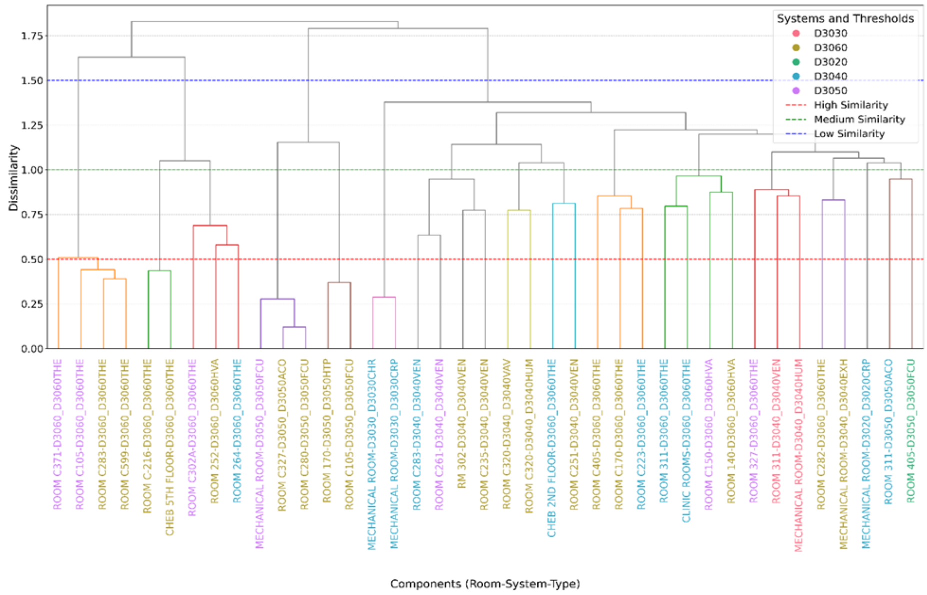

Figure 5. Hierarchical component relationships in building F280

4.4 Portfolio-scale zoning

Portfolio-scale zones were formed by applying DBSCAN to L2-normalised component embeddings using cosine-based dissimilarity. A semantic cut-off of $\theta \geq 0.80$ was used, corresponding to a neighbourhood radius of $\varepsilon = 1 - \theta$, which yielded $\varepsilon = 0.20$ in cosine distance. This configuration produced five dense and well-separated zones that broadly aligned with HVAC subsystem groupings, including heating, cooling, distribution, terminal units, and controls. Location attributes and work order history were attached after clustering to characterise each zone for scheduling and resourcing, while ε remained adjustable to match the operational setting.

The derived zones offered a compact view of resource demand and historical patterns. For example, a controls-centric zone captured a large share of thermostats and panels and exhibited substantial historical labour and cost, while a critical cooling zone isolated a small number of chillers and cooling towers that required specialist coverage despite a comparatively low cost profile.

4.5 Granular building-level relationships

To support building-specific planning, hierarchical clustering was applied to building-masked component embeddings using Ward's method to generate per-building dendrograms. This analysis provided a more granular view of how components clustered based on maintenance histories, and it supported three operational tiers defined by dendrogram distance thresholds (Figure 5). Tier A captured near-identical relationships for $0 \leq d \leq 0.50$, Tier B captured moderate coupling for $0.50 < d \leq 1.00$, and Tier C captured weak coupling for $d > 1.00$. The dendrogram visualisation was reported per building in a Room-System-Component format, which supported rapid inspection of within-system structure and cross-system relationships.

4.6 Limitations and future work

The framework's reliability depends on the quality of historical CMMS data, as inconsistent documentation can degrade the extracted semantics and embedding quality. Validation is limited to a single campus. While operational deployment requires pilot testing with facility management teams, the framework's outputs support several potential workflows. For reactive maintenance, priority predictions and nearest-neighbour searches could enable dispatcher triage and ticket bundling for co-located or similar components. In terms of planned maintenance, aggregated priority predictions by zone could inform labour and inventory forecasting. Additionally, for strategic maintenance, annual summaries of zone workloads could support threshold reviews and long-term replacement planning. Future work should evaluate model transferability across diverse portfolios, integrate real-time sensor attributes, and apply temporal graph learning to capture evolving operational states.

5 Conclusion

This study presented HVACONNECT, a text-enriched heterogeneous graph representation of

HVAC maintenance that unified physical hierarchy, operational relations, and work order history. The learned work order embeddings supported triage through priority prediction under severe class imbalance, while ablation evidence indicated that problem-action text contributed strongly to meaningful embedding structure. Beyond triage, the same embedding space supported data-driven maintenance zoning at portfolio scale and more granular building-level relationship inspection through dendrograms. Together, these outputs suggested a single representation learning backbone that could support reactive routing, planned scheduling, and strategic review of maintenance activities when deployed with human-in-the-loop processes.

6 References

[1] J.H. Lai, Maintenance performance: Examination of the computer-aided maintenance data of a large commercial building, Journal of Performance of Constructed Facilities 29 (2015) 04014118.

[2] C. Fan, F. Xiao, C. Yan, A framework for knowledge discovery in massive building automation data and its application in building diagnostics, Automation in Construction 50 (2015) 81–90.

[3] H.B. Gunay, W. Shen, C. Yang, Text-mining building maintenance work orders for component fault frequency, Building Research & Information 47 (2019) 518–533.

[4] Y. Li, Y. Liu, J. Zhang, L. Cao, Q. Wang, Automated analysis and assignment of maintenance work orders using natural language processing, Automation in Construction 165 (2024) 105501.

[5] T. Li, Z. Zhou, S. Li, C. Sun, R. Yan, X. Chen, The emerging graph neural networks for intelligent fault diagnostics and prognostics: A guideline and a benchmark study, Mechanical Systems and Signal Processing 168 (2022) 108653.

[6] C. Fan, Y. Lin, M.S. Piscitelli, R. Chiosa, H. Wang, A. Capozzoli, Y. Ma, Leveraging graph convolutional networks for semi-supervised fault diagnosis of HVAC systems in data-scarce contexts, Build. Simul. 16 (2023) 1499–1517.

[7] P. Nojedehi, W. O'brien, H.B. Gunay, Benchmarking and visualization of building portfolios by applying text analytics to maintenance work order logs, Science and Technology for the Built Environment (2021).

[8] M. Ensafi, W. Thabet, X. Gao, Work order prioritization using neural networks to improve building operation, Journal of Information Technology in Construction (ITcon) 29 (2024) 324–346.

[9] S. Wettewa, L. Hou, G. Zhang, Graph Neural Networks for building and civil infrastructure operation and maintenance enhancement, Advanced Engineering Informatics 62 (2024) 102868.

[10] Z. Yu, D. Jin, Z. Liu, D. He, X. Wang, H. Tong, J. Han, Embedding text-rich graph neural networks with sequence and topical semantic structures, Knowl Inf Syst 65 (2023) 613–640.

[11] R.P. Charette, H.E. Marshall, UNIFORMAT II elemental classification for building specifications, cost estimating, and cost analysis, US Department of Commerce, Technology Administration, National Institute of …, 1999.

[12] D. Abate, M. Towers, R. Dotz, L. Romani, The Whitestone facility maintenance and repair cost reference 2009-2010, Whitestone Research, California (2009).

[13] Gemini Pro - Google DeepMind, On-line:https://deepmind.google/technologies/gemini/pro.

[14] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, Qwen3 technical report, arXiv Preprint arXiv:2505.09388 (2025).

[15] New embedding models and API updates, On-line:https://openai.com/index/new-embedding-models-and-api-updates.

[16] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019: pp. 3982–3992.

[17] J. Zbontar, L. Jing, I. Misra, Y. LeCun, S. Deny, Barlow twins: Self-supervised learning via redundancy reduction, in: International Conference on Machine Learning, PMLR, 2021: pp. 12310–12320.

[18] A. Bardes, J. Ponce, Y. LeCun, Vicreg: Variance-invariance-covariance regularization for self-supervised learning, arXiv Preprint arXiv:2105.04906 (2021).

[19] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised Contrastive Learning, in: Advances in Neural Information Processing Systems, Curran Associates, Inc., 2020: pp. 18661–18673.

[20] B. Atkin, A. Brooks, Total facility management, (2021).