DRAFT

MM-DGAT: Multi-Modal Dynamic Graph Attention Networks via disease dependency learning

Giovanni Officioso[*]

University of Milán-Bicocca, DEMS, Milán, Italy
[*]Corresponding author. Email: g.officioso@campus.unimib.it

Abstract

Accurate multi-label chest X-ray classification is essential for automated medical diagnosis, yet remains challenging due to complex inter-disease relationships that vary across patient presentations.

Existing approaches face two fundamental limitations: they either treat disease labels independently, ignoring clinical co-occurrence patterns or employ static disease correlation graphs that cannot adapt to individual cases. Moreover, current methods operate solely on visual features, disregarding the rich contextual information available in radiology reports.

We propose **MM-DGAT** (Multi-Modal Dynamic Graph Attention Networks), the first approach to generate patient-specific, multi-modal-conditioned disease correlation graphs for chest X-ray diagnosis. MM-DGAT addresses these limitations through three key innovations:

1. cross-modal fusion that integrates visual features from radiographs with textual embeddings provided by clinical reports via attention mechanisms;
2. dynamic graph generation that produces adaptive disease correlation structures conditioned on fused multi-modal representations;
3. graph attention networks that perform context-aware message passing over these patient- specific graphs.

Keywords: Graph attention networks, Multi-label classification, Multi-modal classifikcation framework

1. Introduction

Medical image diagnosis represents one of the most critical yet challenging applications of computer vision, where accurate identification of multiple co-occurring pathologies directly impacts patient care and clinical outcomes. Among medical imaging modalities, chest radiography (chest X-ray, CXR) stands as the most frequently performed diagnostic examination worldwide(Kh, Wong, and Fione 2017), serving as the first-line tool for detecting thoracic diseases ranging from pneumonia and tuberculosis to cardiovascular conditions and malignancies. The interpretation of chest X-rays demands not only the recognition of individual pathologies but also understanding the complex clinical relationships between diseases that frequently co-occur in predictable patterns.

1.1 Motivation and Clinical Context

While deep learning has achieved remarkable success in single-label image classification(He et al. 2015), multi-label medical diagnosis presents substantially greater challenges. Indeed this task requires simultaneous identification of multiple pathologies, each potentially presenting with varying severity and spatial distribution. More critically, diseases do not occur in isolation. For instance cardiomegaly frequently accompanies pulmonary edema due to shared cardiovascular pathophysiology, or pneumonia often presents alongside consolidation and infiltration or pleural effusion may indicate underlying malignancy, infection, or cardiac failure depending on the clinical context.

These disease co-occurrence patterns reflect fundamental aspects of human pathophysiology and represent crucial diagnostic knowledge that radiologists leverage during interpretation. However, the relationships between diseases are not uniform across all patients. Consider two patients presenting with pleural effusion: in the first case, it may co-occur with cardiomegaly in the context of congestive heart failure, exhibiting a strong positive correlation. In the second case, the same finding may appear alongside pneumonia in an infectious setting, demonstrating an entirely different correlation pattern. This *patient-specific* nature of disease relationships poses the following question: how can we build models that capture both the general statistical tendencies of disease co-occurrence and the specific manifestations in individual cases?

Moreover, radiologists do not interpret images in isolation. Indeed clinical reports containing patient history, symptoms, and prior findings, provide essential context that influences image interpretation. A radiograph showing mild cardiomegaly may be interpreted differently when accompanied by a report stating phrases like *"known history of congestive heart failure"* or *"acute chest pain with no prior cardiac history."* This integration of visual and textual information represents standard clinical practice, yet most automated systems operate solely on imaging features.

1.2 Limitations of Current Approaches

Current approaches to multi-label chest X-ray classification fall into three main categories, each with distinct limitations:

1. label-independent;
2. static graph-based;
3. multi-modal.

Label-independent methods employ standard deep learning architectures with independent binary classifiers for each disease(Rajpurkar et al. 2017). While achieving strong performance through sophisticated architectures and large-scale training, these methods fundamentally ignore disease co-occurrence patterns. Each pathology is predicted independently, precluding the model from leveraging the rich correlations between diseases that human experts routinely exploit.

On the other hand static graph-based methods address this limitation by explicitly modeling disease relationships through graph neural networks(G. Wang et al. 2023). These approaches construct a disease correlation graph, typically from training set co-occurrence statistics, and use graph convolutional or attention mechanisms to propagate information between related diseases. While demonstrating clear improvements over independent classifiers, they employ *fixed* graph structures: the same disease correlation matrix is applied to all patients, regardless of their specific presentation. This one-size-fits-all paradigm cannot capture the patient-specific variations in disease manifestations described above.

Last but not least, multi-modal methods leverage both imaging and textual information, primarily through vision-language models adapted to radiology(Zhang et al. 2025). These approaches typically focus on report generation or image-text retrieval tasks rather than structured classification. Crucially, they do not model disease relationships through explicit graph structures, treating diseases as independent despite their known clinical correlations.

Notably, no existing work combines all three capabilities: modeling disease relationships through graphs, adapting graph structure to individual patients and integrating multi-modal information. This represents a critical gap, as these three aspects are fundamentally complementary. Indeed disease relationships not only should be modeled (graphs), but also should adapt to specific cases (dynamics), and should be informed by all available information (multi-modality).

1.3 Our Approach: MM-DGAT

We propose **MM-DGAT** (Multi-Modal Dynamic Graph Attention Networks), a novel framework that addresses these limitations by introducing patient-specific multi-modal-conditioned disease correlation graphs, which is based on the key insight that disease relationship strengths should not be fixed, but rather generated dynamically on the back of the specific visual presentation and clinical context of each patient.

MM-DGAT operates through three synergistic components:

- cross-modal fusion module, which integrates visual features extracted from chest radiographs with textual embeddings from clinical reports through bidirectional cross-attention mechanisms. This enables the model to leverage complementary information: visual features capture spatial patterns and subtle imaging findings, while textual features provide clinical context and patient history;
- dynamic edge weights generator network which learns to produce patient-specific disease correlation graphs conditioned on the fused multi-modal representation. Rather than applying the same fixed adjacency matrix to all patients, this module generates adaptive edge weights that reflect the specific disease relationship patterns manifested in each case;
- graph attention network which performs message passing over dynamically generated disease correlation graph, where each disease node aggregates information from related diseases through attention-weighted message passing, where both the graph topology and attention weights adapt to the specific patient.

This design ideally enables MM-DGAT to capture several levels of adaptation, that is to say the base disease correlations learned from population statistics, the patient-specific modulation of these correlations based on visual and textual features and the attention-based weighting of information flow during message passing.

1.4 Contributions

This work is expected to achieve the following results:

1. **bi-modal dynamic graph neural network for medical imaging**. We propose MM-DGAT which generates patient-specific disease correlation graphs by fusing visual features from chest X-rays with textual features from radiology reports. To our knowledge, this is the first work to simultaneously address graph-based disease modeling, dynamic adaptation and multi-modal fusion;
2. **novel conditional graph generation architecture.** We design a neural architecture that learns to generate adjacency matrices conditioned on multi-modal input features, enabling disease correlation graphs to adapt to individual patient presentations preserving interpretability through explicit graph structures;
3. **cross-modal graph modulation mechanism.** We introduce a cross-attention fusion module that allows textual information from radiology reports to explicitly influence graph topology, enabling context-aware disease correlation modeling that reflects clinical practice.

2. Related Work

We review related work across four key areas: multi-label chest X-ray classification, graph neural networks for medical imaging, dynamic and adaptive graph learning, and multi-modal medical image analysis.

2.1 Multi-Label Chest X-Ray Classification

Automated chest X-ray diagnosis has witnessed remarkable progress through deep learning, with numerous works achieving radiologist-level performance on multi-label classification tasks. Recent large-scale challenges, such as CXR-LT 2024(Lin et al. 2025), have advanced the field with datasets containing $377,110$ images across 45 disease categories, tackling challenges in long-tailed distributions and zero-shot learning. State-of-the-art approaches predominantly employ convolutional neural networks (CNNs)(G. Huang et al. 2018) or Vision Transformers(Dosovitskiy et al. 2021), treating each disease label independently through binary classification heads.

However, this label-independent paradigm ignores critical clinical knowledge: diseases frequently co-occur in predictable patterns. For instance, cardiomegaly often presents with pulmonary edema due to shared pathophysiological mechanisms. Recognizing this limitation, graph-based approaches have emerged to model disease relationships explicitly.

2.2 Graph-based chest X-ray classification

Among the graph-based approaches, BB-GCN(G. Wang et al. 2023) pioneered the application of graph convolutional networks to chest X-ray diagnosis, constructing a disease correlation graph from training set co-occurrence statistics. Their bi-modal bridged architecture achieved mean AUC scores of 0.835 on ChestX-ray14 and 0.813 on CheXpert, demonstrating clear improvements over label-independent baselines. Building on this, CheXGAT(yan-wei, Huang, and Chang 2022) introduced graph attention networks (GATs) to dynamically weight disease relationships through self-attention mechanisms, further improving classification performance.

Despite these advances, existing graph-based methods share two fundamental limitations. First, they employ *static* graph structures: the disease correlation graph, computed once from training statistics, remains fixed across all test samples, which fails to capture patient-specific variations in disease presentations. For instance, the correlation between cardiomegaly and edema may vary substantially between acute and chronic cases. Moreover, these methods operate exclusively on *visual features*, ignoring the rich textual information present in radiology reports, which radiologists routinely use to contextualize imaging findings.

2.3 Graph Neural Networks for Medical Imaging

Graph neural networks have demonstrated utility across diverse medical imaging applications, from anatomical structure modeling to disease prediction(Ahmedt-Aristizabal et al. 2021). Moreover comprehensive surveys(J. Zhou et al. 2021) highlight the effectiveness of graph deep learning approaches in capturing non-Euclidean relationships that traditional CNNs cannot model naturally.

Also in medical field, graph deep learning approaches find way of applications. Particularly ImageGCN(Mao, Yao, and Luo 2022) proposed multi-relational GCNs for chest X-ray disease identification, constructing graphs at the *image level* where nodes represent individual radiographs and edges encode inter-image relationships. While demonstrating the value of relational modeling, this image-centric approach differs fundamentally from disease-level graphs. More recently, GANN-Med(Alanazi et al. 2025) applied graph attention networks to brain tumor segmentation and classification, combining GAT with wavelet-based multi-resolution features to achieve 93.2% accuracy on the dataset. Their work validates the effectiveness of attention mechanisms in medical imaging, though it operates on single-modality data with static graph structures.

2.4 Dynamic and Adaptive Graph Neural Networks

While traditional GNNs operate on static graphs where both topology and edge weights remain fixed, Dynamic GNNs relax this constraint, enabling graphs to evolve. Recent comprehensive survey(Zheng, Yi, and Wei 2024) categorize dynamic GNN approaches into temporal dynamics (graphs changing over time) and adaptive dynamics (graphs adapting to input characteristics).

In the field of dynamic GNN, the most of researches focuses on temporal evolution, where graph structure changes across timesteps. For instance, EvolveGCN(Pareja et al. 2019) framework uses RNNs to update GNN parameters between temporal snapshots, while TGL(H. Zhou et al. 2022) provides a general framework for temporal graph learning at billion-scale. However, even if temporal dynamics are powerful for social networks or traffic prediction, they do not directly address our goal of *input-conditional* graph generation.

More relevant to our work, several recent approaches generate graphs conditionally. While graph Transformers(Rajpurkar et al. 2017) implicitly learn graph structure through attention mechanisms, the cluster-wise graph transformer(S. Huang et al. 2024) introduces dual-granularity attention, attending to both node-level and cluster-level features. Even more pertinently, GTAT(Shen et al. 2025) proposed cross-attention between node features and graph topology representations, dynamically adjusting the influence of structural information. Their work demonstrates that adaptive weighting of topological information improves performance and reduces over-smoothing.

Indeed, since to our knowledge no prior works has explored conditional graph generation from multi-modal medical data, our work wants to extends these ideas to the medical domain, introducing *multi-modal conditional* graph generation: disease correlation graphs are patient-specific, generated from fused visual-textual representations. To our knowledge, no prior work has explored conditional graph generation from multi-modal medical data.

2.5 Multi-Modal Learning in Medical Imaging

From multi-modal point of view, this learning method has gained significant traction in medical imaging, particularly for vision-language tasks. Large-scale models like CLIP(Radford et al. 2021) have been adapted to radiology through domain-specific pre-training(Zhang et al. 2025). Recent works leverage vision-language models for radiology report generation(Zhanyu Wang et al. 2023), image-text retrieval(Zifeng Wang et al. 2022) and visual question answering(Li et al. 2023).

Existing multi-modal approaches typically employ early fusion (concatenating features), late fusion (combining predictions), or attention-based fusion(Jiang et al. 2021). For chest X-rays specifically, TieNet(X. Wang et al. 2018) pioneered text-image embedding networks proposing multi-level attention to combine visual features with report text. More recent works(Chen et al. 2022) employ transformer-based architectures for joint encoding.

However, not only these multi-modal approaches share a common limitation: they do not model *disease relationships* through explicit graph structures, but they typically focus on generation tasks (report writing) rather than structured classification. Our work, instead, introduces multi-modal learning to graph-based disease classification, where textual information explicitly influences the disease correlation graph.

References

Ahmedt-Aristizabal, David, Mohammad Ali Armin, Simon Denman, Clinton Fookes, and Lars Petersson. 2021. Graph-based deep learning for medical diagnosis and analysis: past, present and future. *Sensors* 21, no. 14 (July): 4758. ISSN: 1424-8220. https://doi.org/10.3390/s21144758. http://dx.doi.org/10.3390/s21144758.

Alanazi, Meshari D., Khaled Kaaniche, Mohammed Albekairi, Turki M. Alanazi, Munid Alanazi, and Ghulam Abbas. 2025. Graph attention neural network for advancing medical imaging by enhancing segmentation and classification. *Engineering Applications of Artificial Intelligence* 161:112372. ISSN: 0952-1976. https://doi.org/10.1016/j.engappai.2025.112372. https://www.sciencedirect.com/science/article/pii/S0952197625023802.

Chen, Zhihong, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. 2022. *Multi-modal masked autoencoders for medical vision-and-language pre-training.* arXiv: 2209.07098 [cs.CV]. https://arxiv.org/abs/2209.07098.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, et al. 2021. *An image is worth 16x16 words: transformers for image recognition at scale.* arXiv: 2010.11929 [cs.CV]. https://arxiv.org/abs/2010.11929.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. *Deep residual learning for image recognition.* arXiv: 1512.03385 [cs.CV]. https://arxiv.org/abs/1512.03385.

Huang, Gao, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2018. *Densely connected convolutional networks.* arXiv: 1608.06993 [cs.CV]. https://arxiv.org/abs/1608.06993.

Huang, Siyuan, Yunchong Song, Jiayue Zhou, and Zhouhan Lin. 2024. *Cluster-wise graph transformer with dual-granularity kernelized attention.* arXiv: 2410.06746 [cs.LG]. https://arxiv.org/abs/2410.06746.

Jiang, Cheng, Yihao Chen, Jianbo Chang, Ming Feng, Renzhi Wang, and Jianhua Yao. 2021. *Fusion of medical imaging and electronic health records with attention and multi-head machanisms.* arXiv: 2112.11710 [cs.CV]. https://arxiv.org/abs/2112.11710.

Kh, Ng, Jeannie Wong, and Tan Fione. 2017. *Technical specifications of medical imaging equipment.* https://doi.org/10.1007/978-981-10-1613-4_4.

Li, Chunyuan, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. *Llava-med: training a large language-and-vision assistant for biomedicine in one day.* arXiv: 2306.00890 [cs.CV]. https://arxiv.org/abs/2306.00890.

Lin, Mingquan, Gregory Holste, Song Wang, Yiliang Zhou, Yishu Wei, Imon Banerjee, Pengyi Chen, et al. 2025. Cxr-lt 2024: a miccai challenge on long-tailed, multi-label, and zero-shot disease classification from chest x-ray. *Medical Image Analysis* 106:103739. ISSN: 1361-8415. https://doi.org/https://doi.org/10.1016/j.media.2025.103739. https://www.sciencedirect.com/science/article/pii/S1361841525002865.

Mao, Chengsheng, Liang Yao, and Yuan Luo. 2022. Imagegcn: multi-relational image graph convolutional networks for disease identification with chest x-rays. *IEEE transactions on medical imaging* PP (February). https://doi.org/10.1109/TMI.2022.3153322.

Pareja, Aldo, Giacomo Domeniconi, Jie Chen, Tengfei Ma, Toyotaro Suzumura, Hiroki Kanezashi, Tim Kaler, Tao B. Schardl, and Charles E. Leiserson. 2019. *Evolvegcn: evolving graph convolutional networks for dynamic graphs.* arXiv: 1902.10191 [cs.LG]. https://arxiv.org/abs/1902.10191.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, et al. 2021. *Learning transferable visual models from natural language supervision.* arXiv: 2103.00020 [cs.CV]. https://arxiv.org/abs/2103.00020.

Rajpurkar, Pranav, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, et al. 2017. *Chexnet: radiologist-level pneumonia detection on chest x-rays with deep learning.* arXiv: 1711.05225 [cs.CV]. https://arxiv.org/abs/1711.05225.

Shen, Jiahao, Qura Ain, Yaohua Liu, Banqing Liang, Xiaoli Qiang, and Zheng Kou. 2025. Gtat: empowering graph neural networks with cross attention. *Scientific Reports* 15 (February). https://doi.org/10.1038/s41598-025-88993-3.

Wang, Guoli, Pingping Wang, Jinyu Cong, Kunmeng Liu, and Benzheng Wei. 2023. *Bb-gcn: a bi-modal bridged graph convolutional network for multi-label chest x-ray recognition.* arXiv: 2302.11082 [cs.CV]. https://arxiv.org/abs/2302.11082.

Wang, Xiaosong, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M. Summers. 2018. *Tienet: text-image embedding network for common thorax disease classification and reporting in chest x-rays.* arXiv: 1801.04334 [cs.CV]. https://arxiv.org/abs/1801.04334.

Wang, Zhanyu, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023. *R2gengpt: radiology report generation with frozen llms.* arXiv: 2309.09812 [cs.CV]. https://arxiv.org/abs/2309.09812.

Wang, Zifeng, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. *Medclip: contrastive learning from unpaired medical images and text.* arXiv: 2210.10163 [cs.CV]. https://arxiv.org/abs/2210.10163.

yan-wei, Lee, Sheng-Kai Huang, and Ruey-Feng Chang. 2022. Chexgat: a disease correlation-aware network for thorax disease diagnosis from chest x-ray images. *Artificial Intelligence in Medicine* 132 (August): 102382. https://doi.org/10.1016/j.artmed.2022.102382.

Zhang, Sheng, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, et al. 2025. *Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs.* arXiv: 2303.00915 [cs.CV]. https://arxiv.org/abs/2303.00915.

Zheng, Yanping, Lu Yi, and Zhewei Wei. 2024. *A survey of dynamic graph neural networks.* arXiv: 2404.18211 [cs.LG]. https://arxiv.org/abs/2404.18211.

Zhou, Hongkuan, Da Zheng, Israt Nisa, Vasileios Ioannidis, Xiang Song, and George Karypis. 2022. *Tgl: a general framework for temporal gnn training on billion-scale graphs.* arXiv: 2203.14883 [cs.LG]. https://arxiv.org/abs/2203.14883.

Zhou, Jie, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2021. *Graph neural networks: a review of methods and applications.* arXiv: 1812.08434 [cs.LG]. https://arxiv.org/abs/1812.08434.