Mammography classification with multi-view deep learning techniques: Investigating graph and transformer-based architectures - Francesco Manigrasso, Rosario Milazzo et al.

**Application domain:**

The paper addresses an application in medical imaging: automated breast cancer detection from mammograms to lighten the workload for radiologists. The specific task involves classifying mammograms into cancerous and non-cancerous cases by analysing multiple views (cranio-caudal and medio-lateral-oblique views) of each breast. The aim is to relieve radiologists' burden by developing automated systems capable of interpreting these images efficiently. The paper investigates the ability of deep learning models to process such images in a multi-view setting, handling both ipsi-lateral and contra-lateral comparisons.

**The used (Machine Learning) methodology and evaluation metrics:**

The methodology involves three key architectures: CNNs, graph-based architectures, and transformer-based models.

1. CNN Baseline: The baseline uses a ResNet-22 CNN backbone, which downsizes the images for computational efficiency and combines features from four views using concatenation to classify the images. While CNNs are effective in general medical image analysis, they struggle to maintain high resolution, which is needed for detecting small lesions in mammography.
2. Graph-based Models: The Anatomy-aware Graph Convolutional Network (AGN4V) extends the CNN backbone with graph neural networks that simulate the relational reasoning used by radiologists when interpreting mammograms. This model captures spatial and geometric relationships between regions of the mammogram using pseudo-landmarks (anatomical reference points).
3. Transformer-based Models: Vision Transformers (ViTs) are included for their ability to model long-range dependencies in images. The transformer-based MaMVT architecture includes cross-view attention layers that enable it to integrate information from multiple views of each breast and between the breasts. Transformers are effective at focusing on the most important image regions and handling high-resolution images more efficiently than CNNs.

Evaluation Metrics:

The models are evaluated using the area under the receiver operating characteristic curve (AUC) to measure their performance in detecting cancer. Other metrics include the false positive rate at a sensitivity of 99% (FPR99) to assess the models' ability to rule out non-cancerous cases without missing any cancers. Final metrics include recall and precision. Visual explainability methods like Grad-CAM are used to generate heatmaps showing which regions of the image the model focuses on during classification.

**Strong and weak points:**

Strong Points:

1. Multi-view Integration: All architectures effectively combine information from multiple mammography views, which is crucial for accurate breast cancer detection.
2. Transformer-based Strength: The transformer model performs particularly well, benefiting from its ability to model long-range dependencies and focus computational resources on the most relevant parts of the image. This makes it more effective in handling high-resolution mammograms.
3. Graph-based Interpretability: The graph-based model mimics radiologists' reasoning by matching corresponding regions between views, offering better interpretability.
4. Ensemble Models: Combining CNNs, graph-based, and transformer models enhances overall performance and robustness since each architecture has its own strengths.

Weak Points:

1. Weak Supervision: All models are trained in a weakly supervised setting, using image-level labels rather than pixel-level annotations, limiting their ability to detect small lesions accurately.
2. High Data Requirements: Transformer-based models, while powerful, require large datasets and significant computational resources, posing challenges when dealing with limited medical datasets.
3. Overfitting Risk: Both CNN and graph-based models show signs of overfitting, especially with smaller datasets. While they perform well on validation sets, there is a risk of generalization issues on unseen data.
4. Explainability Gaps: Although Grad-CAM is used for interpretability, the attention maps produced by transformers tend to be broad and sometimes fail to pinpoint specific lesions accurately.

**Alternative methodology, evaluation metrics, and improvement ideas:**

Alternative Methodology

While transformers perform well, they could be further improved by integrating more domain-specific inductive biases, such as attention mechanisms explicitly designed for mammography or hybrid CNN-transformer architectures that combine the best of both worlds.

* Hybrid Models: Combining CNNs' feature extraction capabilities with transformers' long-range dependency modelling could mitigate some data inefficiency in transformers. For instance, using CNNs for initial feature extraction followed by transformer layers for relational reasoning could improve performance on small datasets.
* Self-supervised Learning: More extensive use of self-supervised learning approaches, like contrastive learning, could be beneficial, especially in the medical domain where labeled data is limited. Pretraining models on large unlabeled datasets could enhance their generalization abilities.

Evaluation Metrics

* Adding precision-recall curves (PRC) in addition to AUC could offer more insight into performance, especially for low-prevalence diseases. PRC would help evaluate how well the models perform in distinguishing between positive and negative cases in highly imbalanced datasets.
* Pixel-level Supervision: Incorporating pixel-wise annotations or using semi-supervised learning could help in detecting small lesions, improving the models' sensitivity for subtle cancer signs.

Data Augmentation and Synthesis

The paper includes data augmentation, but more advanced techniques such as generative adversarial networks (GANs) could be explored for synthetic data generation, helping address the challenge of small datasets in medical imaging.