Final Year Project: Preliminary Report

Development of an Explainable Deep Learning Model for
Early Breast Cancer Detection in Singaporean Women with Dense Tissue

Justin Lim

May 23, 2025

Contents

1 Introduction

1.1 Project Template

3.2 Project Idea 2: Deep Learning Breast Cancer Detection

1.2 Project Background Information

Breast cancer remains the most common cancer among women in Singapore [10]. While national screening programs such as BreastScreen Singapore aim to improve early detection, multiple barriers—such as dense breast tissue, lower participation among certain ethnic groups, and limitations in clinical capacity—still affect screening efficacy. As shown in the Singapore Breast Cancer Cohort (SGBCC) and recent public health literature, under-screening and delayed diagnosis are especially prominent among older women and specific demographic groups.

This project addresses these challenges by developing a localized, explainable AI system that enhances early detection for high-risk cases. The value lies in improving diagnostic accuracy in dense breast tissue scenarios, increasing clinician trust through explainability, and supporting targeted interventions in under-screened populations.

1.3 Project Concept

This project proposes an explainable deep learning system that classifies histopathology or mammography images to support breast cancer detection. The system is optimized for diagnostic workflows in Singapore, particularly for patients with dense breast tissue or late-stage presentation. Convolutional Neural Networks (CNNs), combined with interpretability tools such as Grad-CAM, will be used to generate heatmaps that highlight model attention areas, allowing clinicians to visually validate predictions.

1.4 Project Objectives

The main objectives of this project are:

- To design and train a CNN-based model that accurately classifies breast cancer from mammography or histopathology image datasets.
- To evaluate model generalization performance on profiles relevant to Singaporean women, particularly those with dense breast tissue and older age.
- To integrate explainability features such as Grad-CAM that generate visual outputs aligned with known diagnostic markers, thereby increasing clinician trust and interpretability.

1.5 Deliverables

- A trained and tested CNN-based image classification model (e.g., ResNet or EfficientNet)
- Grad-CAM visualizations for model interpretability
- Evaluation metrics (e.g., accuracy, sensitivity, specificity, F1-score) measured across typical and edge-case scenarios

2 Literature Review

This chapter reviews the limitations of traditional breast cancer risk models, the role of deep learning in diagnostic imaging, the importance of explainability, and challenges in generalizing models across diverse populations. Each section provides justification for the design choices and objectives outlined in this project—emphasizing the need for a localized, explainable deep learning system in Singapore's screening context.

2.1 Limitations of Traditional Risk Models

Risk prediction models form the foundation of clinical decision-making in early breast cancer detection. Conventional tools, such as the Gail and Tyrer-Cuzick models [1], have been widely adopted in clinical settings for population-level risk stratification. These models use structured demographic and clinical inputs to estimate cancer risk but are limited by their rule-based design and assumptions of feature independence.

For this project, understanding the limitations of these traditional models is crucial—they underscore the need for more individualized and image-informed approaches. These models are unable to capture spatial or visual features present in medical images, which are essential when diagnosing cancer in women with dense breast tissue.

A key study [1] compared the Tyrer-Cuzick model with CNN-based deep learning models trained directly on mammograms. The deep learning models significantly outperformed the clinical models, with AUC values ranging from 0.68 to 0.70 compared to 0.62. These results were especially compelling for dense breast tissue subgroups—directly aligning with this project's target population of Singaporean women, who often present with dense tissue at younger ages [6].

Appendix Figure 7 further illustrates these findings. It shows that the deep learning model identified over 31.2% of future cancers within the highest-risk decile—almost double the detection rate of the Tyrer-Cuzick model, which captured only 18.2%. This performance gap highlights not just improved sensitivity, but also the model's utility in stratifying high-risk cases early—an essential capability for any national screening program.

Beyond performance metrics, traditional models are inherently limited in adaptability and interpretability. They cannot be fine-tuned for local demographics or provide spatial reasoning about predictions. For instance, local analysis [6] found that the Tyrer-Cuzick model underestimates risk in Singaporean women, leading to reduced screening recall rates and delayed diagnoses. This directly supports this project's motivation to build a CNN-based pipeline that can be trained and evaluated using regional risk profiles.

By shifting from predefined questionnaire-based inputs to deep, image-based learning, this project enables the model to extract and prioritize risk signals that are visually embedded in mammograms. Furthermore, with the integration of Grad-CAM, the proposed system will offer transparency in predictions—something traditional risk models fundamentally lack.

In summary, the weaknesses of traditional models in both predictive power and adaptability strongly support the development of an explainable, localized CNN-based model tailored for dense tissue scenarios in Singapore.

2.2 Role of Deep Learning in Breast Cancer Imaging

To justify the deep learning approach used in this project, it is essential to examine how convolutional neural networks (CNNs) have enhanced breast cancer diagnostics. Unlike traditional models that rely on structured variables such as age and family history, CNNs operate directly on full-field mammograms. This allows them to detect complex spatial features—such as tissue asymmetry, mass shape, and density patterns—that often precede clinical symptoms. Such characteristics are especially critical in screening contexts involving dense breast tissue, a trait common among women in Singapore [6].

Evidence from Yala et al. [1] provides a compelling foundation for this project's approach. Their study compared three models: the Tyrer-Cuzick clinical risk model, a CNN-based image-only model, and a hybrid model combining clinical and image data. In all cases, the deep learning models outperformed the traditional baseline, particularly in high-density subgroups. Appendix Figure 8 illustrates how hybrid DL models captured significantly more cancers in high-risk, dense-tissue populations—underscoring CNNs' ability to capture nuanced malignancy signals that traditional methods miss.

Appendix Figure 9 further strengthens this argument by visualizing ROC curves across the three models. The CNN-based models demonstrated superior true-positive rates at nearly every threshold. For this project, these findings justify the selection of a CNN backbone (e.g., ResNet or EfficientNet) for risk prediction. They also support the hybrid design direction, where image features can be supplemented with clinical data for marginal gains.

Beyond statistical performance, deep learning models offer clinical interpretability when paired with tools like Grad-CAM. This is exemplified in Shen et al. [7], where CNNs produced heatmaps overlayed on mammograms to highlight regions of diagnostic interest. Appendix Figure 10 shows how the model accurately localized high-risk regions while minimizing false positives. This interpretability aligns directly with the project's goal of supporting radiologist decision-making—not replacing it—by providing transparent visual justifications for model outputs.

From a local systems perspective, these capabilities are highly relevant to Singapore's healthcare landscape. DL models offer scalable, automated triage solutions for national screening programs that face resource constraints. By fine-tuning these models on localized data or using targeted augmentation strategies, this project aims to build a system that is both technically robust and demographically sensitive.

In summary, deep learning offers a powerful and adaptable foundation for early breast cancer detection. The literature reviewed here justifies the core technical choices of this project: adopting CNN-based models, integrating explainability, and designing for dense tissue scenarios within the Singaporean context.

2.3 Explainability in Medical AI

While deep learning (DL) models have achieved strong performance in medical imaging, their "black box" nature remains a major barrier to clinical adoption. Unlike traditional models with clearly defined features, CNNs often produce predictions without transparent reasoning. In high-stakes domains like cancer diagnosis, this lack of interpretability poses ethical, legal, and practical concerns—especially when errors may result in misdiagnosis or delayed treatment.

This issue is particularly relevant to this project, which targets breast cancer screening in Singapore. Clinicians need to understand how and why a model makes a prediction in order to trust and act on it. As highlighted by Ching et al. [3], explainability is not just a technical add-on but a prerequisite for the responsible use of AI in healthcare. Without it, medical professionals are less likely to adopt such tools, regardless of performance gains.

To overcome this limitation, this project incorporates Gradient-weighted Class Activation Mapping (Grad-CAM) into its CNN architecture. Grad-CAM generates a heatmap that shows which regions of the input mammogram contributed most strongly to a model's prediction. These visual explanations serve as interpretive cues for radiologists, helping them validate whether the model is focusing on clinically relevant structures like calcifications or asymmetries.

For instance, Selvaraju et al. [5] demonstrated how Grad-CAM could be used to highlight malignancy-related areas in medical images. Appendix Figure 11 shows a representative output where red activation zones correspond with tumor-like regions. Such interpretability directly supports this project's goal of clinician-aligned AI: when a model's attention visibly aligns with known diagnostic features, trust in its predictions increases. Conversely, if the heatmap focuses on irrelevant regions, it can signal model failure or the need for retraining—adding a valuable layer of quality control.

This capability is especially important in the Singaporean healthcare context, where patients and providers come from diverse ethnic backgrounds, and diagnostic baselines may differ. Explainable models help bridge anatomical and cultural variability by providing a shared visual rationale for risk scores. They also reduce automation bias by keeping radiologists in the loop.

Moreover, integrating Grad-CAM into the system architecture allows for routine auditing and retrospective case review. This aligns with the Ministry of Health's emphasis on accountability and traceability in AI-driven diagnostics. Visual heatmaps can be logged, reviewed, and compared over time to refine both human and machine performance.

In summary, explainability is not peripheral—it is central to the viability of AI in breast cancer screening. By embedding Grad-CAM into the model pipeline, this project ensures that predictions are not only accurate but also interpretable, auditable, and clinically actionable.

2.4 Dataset Diversity and Generalization

A major consideration in designing AI systems for breast cancer screening is the extent to which models trained on one population can generalize to another. Many state-of-the-art deep learning models in this domain have been trained on datasets collected from Western populations [1]. These datasets—while large and well-annotated—reflect clinical practices, imaging modalities, and demographic profiles that differ from those in Southeast Asia.

This limitation is particularly relevant in the Singaporean context. For example, Chau et al. [6] found that Singaporean women tend to have denser breast tissue and are diagnosed at younger ages compared to their Western counterparts. Dense tissue not only increases cancer risk but also reduces the effectiveness of mammography—making detection more difficult and increasing the likelihood of false negatives. Moreover, socio-cultural factors such as screening hesitancy and lower awareness compound the challenge. For this project, these differences emphasize the importance of designing a model that accounts for regional characteristics.

Yala et al.'s [1] hybrid model—which combines image and clinical features—performs well on U.S. data, but its effectiveness in Singaporean populations is uncertain. Appendix Figure 8 shows how cancer incidence varies significantly across tissue density and risk score categories. The stratification observed in U.S. data suggests potential for image-driven models to capture visual malignancy signals; however, the model must be adapted or fine-tuned for local demographics to ensure meaningful predictions.

This concern is further reinforced by Raghu et al. [2], who investigated the value of transfer learning from ImageNet in medical imaging. Their findings, shown in Appendix Figure 12, indicate that pretrained CNNs did not significantly outperform models trained from scratch. In fact, most pretrained features were overridden during fine-tuning. This supports this project's strategy to adopt a lightweight CNN architecture trained specifically on mammographic data, rather than relying on generalized features learned from non-medical domains.

Finally, the problem of dataset scarcity must also be addressed. In regions like Singapore, access to annotated medical data is limited due to privacy regulations and small population sizes. Cheplygina et al. [4] highlight the utility of alternative learning paradigms such as semi-supervised learning (SSL), multi-instance learning (MIL), and weak supervision. These approaches reduce dependence on manual labeling and are more viable in data-constrained settings.

This project incorporates several of these ideas: synthetic augmentation is used to simulate dense tissue variability; focal loss is employed to prioritize underrepresented malignancy cases; and the architecture is designed to accommodate future integration of local data for fine-tuning. Together, these strategies ensure that the model is not only technically sound but also aligned with the demographic and infrastructural realities of Singapore.

In summary, model generalization cannot be assumed across geographic or cultural boundaries. By prioritizing localization, training flexibility, and data-efficient learning, this project builds an AI system that is clinically relevant, demographically sensitive, and adaptable to future expansions in real-world Singaporean screening programs.

2.5 Summary of Gaps and Relevance to This Project

The literature reviewed across clinical models, deep learning strategies, explainability, and dataset generalization reveals critical gaps that inform this project's design decisions. Rather than offering a generic summary, this section presents an analysis of how each limitation directly influences the features, architecture, and deployment considerations of the proposed system.

- **Gap 1: Limited Predictive Power of Clinical Risk Models**
  Existing tools like the Gail and Tyrer-Cuzick models rely on structured questionnaire inputs and ignore visual indicators present in mammograms. These models struggle particularly in women with dense breast tissue—a common trait among Singaporean women [6]. Deep learning models [1, 7] have shown stronger performance in identifying malignancy through image features. *Therefore, this project adopts a CNN-based model trained on full-field mammograms to provide higher-resolution risk detection that accounts for local anatomical traits.*

- **Gap 2: Lack of Explainability Hinders Clinical Trust**
  CNNs are often criticized for their opacity [3, 5]. Without interpretability, clinical

uptake is limited. Grad-CAM offers a partial solution by generating saliency maps aligned with diagnostic regions [5]. *This project integrates Grad-CAM to ensure radiologists can visually verify AI predictions and maintain control over diagnostic decisions, addressing trust and accountability concerns in real-world deployments.*

- **Gap 3: Western-Centric Data Limits Generalization**
  Models trained on Western cohorts often fail to generalize to Southeast Asian populations [6]. High-density breast tissue, screening hesitancy, and sociocultural differences alter risk profiles. *This project uses stratified evaluation and region-specific augmentation strategies to simulate local data characteristics and ensure performance relevance in Singaporean women.*

- **Gap 4: Scarcity of Annotated Data in Southeast Asia**
  Annotated medical imaging data is often inaccessible due to privacy laws and cohort limitations. Alternative training paradigms such as semi-supervised and weakly supervised learning have been proposed [4]. *This project incorporates data augmentation and simplified CNNs with lower parameter counts to reduce the need for extensive labeled data while maintaining robustness.*

- **Gap 5: Overreliance on Transfer Learning from Natural Images**
  While transfer learning is common, recent studies show that pretrained features from ImageNet are frequently overwritten during fine-tuning in medical domains [2]. *This project evaluates both pretrained and scratch-trained CNN variants to determine the most effective approach under low-data, domain-specific conditions.*

Collectively, these gaps justify the core architecture of this project:

- CNN-based image models are prioritized over questionnaire-based tools for precision screening;
- Grad-CAM explainability ensures clinical interpretability and user trust;
- Training pipelines are adapted to local demographic factors;
- Lightweight architectures and weak supervision accommodate Singapore's data limitations;
- Transfer learning is validated critically rather than assumed.

These strategies form the foundation of a culturally responsive, explainable AI pipeline optimized for breast cancer screening in Singapore.

2.6 Analysis of Similar Projects and Tools

To guide the design of this project's prototype, several real-world and academic AI systems in breast cancer screening were reviewed. These systems provide critical insights into what works—and what must be improved—to ensure technical success and clinical adoption in real-world healthcare settings like Singapore.

- **Google Health / DeepMind – Mammogram AI System**
  A 2020 study evaluated an AI system trained on over 90,000 mammograms from the US and UK, showing that it could outperform radiologists in controlled settings [11]. However, its performance declined when applied to unseen populations. *This project incorporates stratified testing and local demographic simulation to ensure model generalizability to Singapore's multi-ethnic population.*

- **Zebra Medical Vision – Scalable Cancer Detection Tools**
  Zebra developed FDA-cleared AI tools for breast cancer detection and deployed them at scale via cloud-based infrastructure [12]. Their success came from integration into existing workflows. *This project uses lightweight models and explainable visual outputs (via Grad-CAM) to enable similar compatibility with systems like BreastScreen Singapore.*

- **Shen et al. (2019) – CNN-Based Mammogram Classifier**
  This academic study demonstrated that CNNs can significantly boost mammographic sensitivity and specificity [7], but also emphasized the importance of high-quality training data. *This project uses data augmentation, synthetic sampling, and reduced-parameter CNNs to mitigate local data scarcity.*

- **Salim et al. (2024) – AI in Population-Based Screening (Sweden)**
  A recent real-world deployment across 58,000 mammograms found that AI-assisted screening improved workflow efficiency while maintaining diagnostic accuracy [13]. *This project follows a similar assistive philosophy—offering clinicians interpretable risk assessments, not automated diagnoses.*

These examples reinforce key design choices in this project: use of explainable CNNs, integration-readiness for local infrastructure, data-efficient training strategies, and a clinician-in-the-loop approach tailored for Singapore's screening ecosystem.

3 Project Design

3.1 User and Domain Context

This project is designed for use by clinical radiologists and public health administrators within Singapore's national breast cancer screening infrastructure. The domain focuses on AI-assisted diagnostic tools for mammogram analysis, especially among women aged 40–60 with dense breast tissue. This demographic is underrepresented in most Western training datasets, necessitating a locally contextualized system.

3.2 System Architecture

The system architecture is designed to support accurate and interpretable breast cancer risk prediction using mammographic images, with a focus on dense tissue scenarios commonly observed in Singaporean women. Each component of the architecture is included for its role in fulfilling clinical performance requirements while supporting transparency in decision-making.

Input Preprocessing Layer. Mammographic images from different sources often vary in resolution, lighting, and embedded metadata. To ensure model robustness, this stage standardizes image inputs for consistent downstream processing. Based on best practices in mammogram preprocessing [7, 15], and observations from the DDSM dataset [17], the following steps are included:

- *Grayscale conversion:* Images are converted to 8-bit grayscale to reduce computational complexity while preserving critical tissue structures.
- *Histogram normalization:* Applied to minimize contrast variability across acquisition devices, aiding generalization.
- *Resizing:* Images are resized to 224x224 pixels to match CNN input requirements [1], enabling model reuse and compatibility.
- *ROI extraction:* Regions of interest are isolated by removing annotations and borders, preventing irrelevant artifacts from biasing the model. This step was adapted manually after inspecting DDSM-specific patterns [17].

These preprocessing steps are crucial to achieving consistent input quality, especially when working with heterogeneous public datasets and simulating local screening conditions.

CNN Backbone. Feature extraction is performed using a convolutional backbone—either ResNet-50 or EfficientNet-B0—chosen based on prior success in medical imaging [1, 7]. ResNet-50 provides a balance of accuracy and depth, while EfficientNet's compound scaling offers computational efficiency. Both are capable of learning fine-grained features such as mass margins, asymmetries, and microcalcifications—relevant for identifying early-stage cancers in dense tissue.

This component supports the project goal of developing a model that can capture diagnostic subtleties missed by rule-based systems and generalize across patient profiles.

Prediction Head. Following feature extraction, the output is flattened and passed through fully connected layers, ending in a sigmoid neuron that outputs a cancer risk probability score. Binary cross-entropy is used as the default loss function. However, due to the class imbalance typical in screening datasets, focal loss is optionally applied [2] to improve sensitivity to rare malignancy cases—supporting early detection without overfitting on benign cases.

Explainability Module. Explainability is a design priority, not an afterthought. To address clinical transparency, Grad-CAM is integrated into the architecture. It visualizes attention maps by highlighting regions of the input image that contributed most to a model's decision [5]. This enhances radiologists' trust and provides a layer of clinical validation—particularly important in high-density or ambiguous cases where interpretability directly impacts diagnostic decisions.

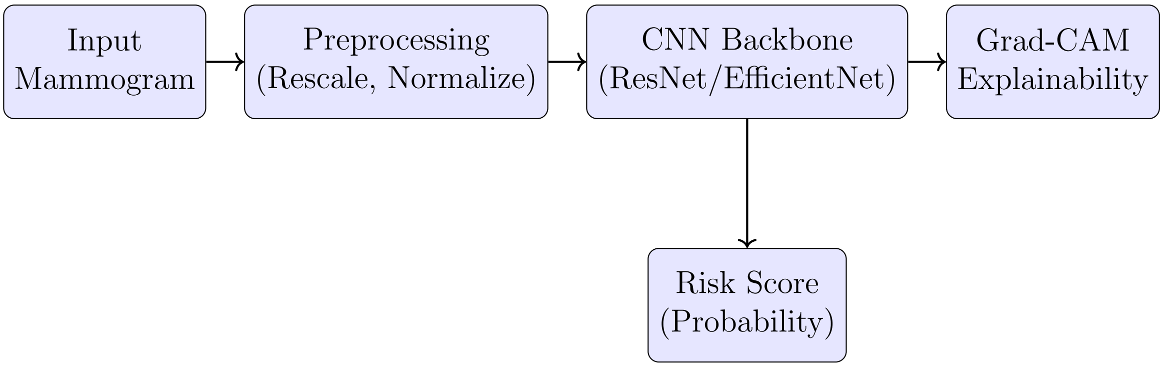

Input Mammogram
Preprocessing (Rescale, Normalize)
CNN Backbone (ResNet/EfficientNet)
Grad-CAM Explainability
Risk Score (Probability)

Figure 1: System architecture for explainable breast cancer risk prediction using convolutional neural networks (CNNs) [1] and Grad-CAM [5].

Pipeline Overview. Figure 1 illustrates the overall pipeline. After preprocessing, images flow through a CNN for feature extraction [1]. One branch produces a risk score; the other passes through Grad-CAM [5] to generate attention heatmaps.

Design Considerations. Each module is deliberately kept modular and swappable. This enables future expansion, such as fine-tuning on Singaporean-specific data, incorporating metadata like patient history, or using multi-view mammography inputs [8]. This architectural flexibility ensures that the system can evolve as screening datasets, clinical requirements, and deployment contexts grow more complex.

In summary, this architecture reflects a balance between predictive strength and clinical usability—ensuring that both radiologists and patients benefit from a risk prediction system that is not only effective, but also transparent and adaptable.

3.3 Dataset Used

Primary Dataset – DDSM. The Digital Database for Screening Mammography (DDSM) [17] serves as the primary dataset. It comprises over 2,600 cases with four standard views per patient (cranio-caudal (CC) and mediolateral oblique (MLO) for each breast), along with verified pathology and ground-truth masks for abnormalities.

This dataset provides a solid foundation for training and evaluating convolutional neural network (CNN)-based models due to its detailed annotation and diversity of tissue presentations. It contains both normal and cancerous cases, enabling the model to learn from contrastive examples.

Auxiliary Dataset – BreaKHis. The Breast Cancer Histopathological Image Classification (BreaKHis) dataset [18] is incorporated for auxiliary testing and domain adaptation experiments. Although histopathology data differ from mammograms, they offer complementary insights into cellular-level features and enable the project to explore cross-domain robustness.

Singapore-Specific Contextualization. Despite their value, both DDSM and BreaKHis originate from Western populations. To simulate Singaporean demographic conditions, stratified testing subsets will be constructed. Breast density and age distributions will be calibrated using data reported by the Singapore Breast Cancer Cohort [6] and the BreastScreen Singapore program [19]. Synthetic augmentation will mimic tissue characteristics more common among local women—such as denser fibroglandular tissue and smaller breast volume. This enables context-aware evaluation and reveals limitations in model generalizability.

Ethical Considerations. All datasets used are publicly available and de-identified, aligning with ethical guidelines for secondary use. In a future deployment scenario, local datasets such as those from National University Hospital (NUH) or Singapore General Hospital (SGH) could be integrated under appropriate ethics approval.

3.4 Feature Engineering

Feature engineering is a critical bridge between raw mammographic data and effective CNN-based learning. As outlined in Figure 2, this stage transforms high-resolution DICOM inputs into standardized, diverse, and clinically enriched representations suitable for training. The goal is to optimize generalization and sensitivity, particularly in dense breast tissue scenarios that are common among Singaporean women.

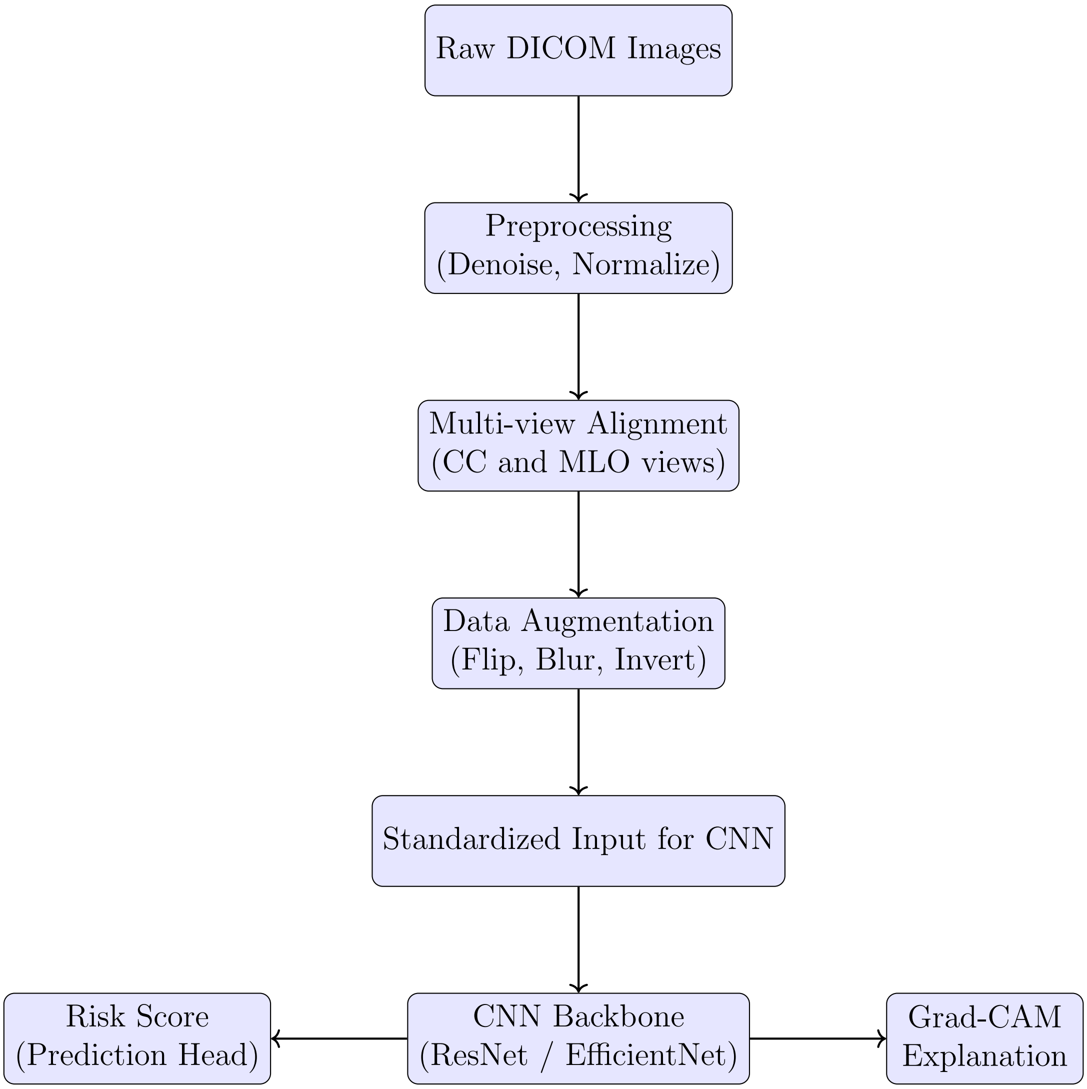

Raw DICOM Images
Preprocessing (Denoise, Normalize)
Multi-view Alignment (CC and MLO views)
Data Augmentation (Flip, Blur, Invert)
Standardized Input for CNN
CNN Backbone (ResNet / EfficientNet)
Risk Score (Prediction Head)
Grad-CAM Explanation

Figure 2: End-to-end pipeline for breast cancer risk prediction, integrating preprocessing, multi-view feature alignment, CNN classification, and visual explainability.

Preprocessing and Augmentation. The initial phase of feature engineering focuses on cleaning and enhancing the diagnostic quality of raw mammograms. This step is essential to ensure consistent model input and eliminate irrelevant imaging artifacts.

- **Denoising:** Median filtering removes impulse noise (e.g., salt-and-pepper artifacts) while preserving fine structural detail—important for detecting subtle anomalies like microcalcifications.
- **Contrast enhancement:** CLAHE (contrast-limited adaptive histogram equalization) [15] enhances local contrast, improving visibility of faint lesions, especially in dense tissue environments.
- **Normalization:** Pixel values are rescaled to $[0, 1]$ or z-normalized to ensure uniform dynamic range, enabling more stable model convergence.

After standardization, augmentation techniques are applied to increase data diversity and improve model robustness:

- **Geometric:** Random flipping, cropping, and small-angle rotations (±15°) simulate positioning variability during imaging.
- **Photometric:** Gaussian blurring and contrast jitter mimic scanner inconsistencies and imaging noise.
- **Semantic:** Intensity inversion introduces exposure variation and histogram shifts, improving cross-device generalization.

These transformations help prevent overfitting and prepare the model for deployment across heterogeneous clinical environments.

Multi-View Spatial Context. Breast screening typically involves multiple views: cranio-caudal (CC) and mediolateral oblique (MLO). Rather than treating each view independently, this project uses a multi-view strategy to leverage spatial correlations across angles. This is based on the approach by Geras et al. [8], who showed that multi-view CNNs improve lesion localization and diagnostic accuracy.

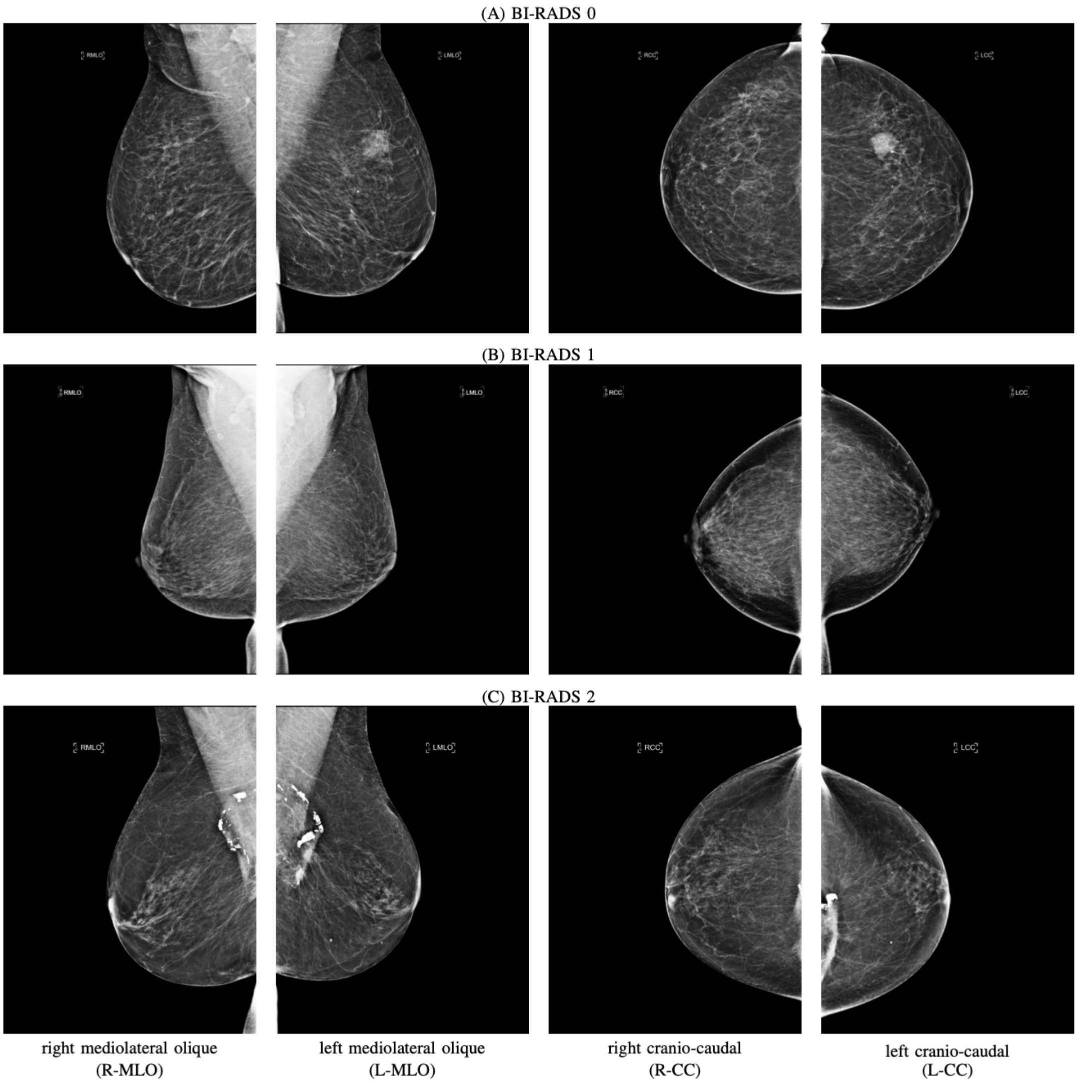


Figure 3: High-resolution mammogram processing via multi-view deep CNNs. Adapted from Geras et al. [8].

Both CC and MLO views are preprocessed and aligned to maintain anatomical consistency. This strategy enables the model to capture bilateral asymmetries, view-dependent lesion projections, and complementary diagnostic signals—critical for real-world screening reliability.

Domain-Specific Feature Emphasis. In addition to automated learning, domain knowledge is embedded to focus the model on clinically meaningful features:

- **Asymmetry detection:** Breast difference maps highlight abnormal structural discrepancies between left and right sides.
- **Texture emphasis:** Gabor filters [20] are used to visualize texture patterns—such as spiculations—that are indicative of malignancy.
- **Density-informed stratification:** BI-RADS (Breast Imaging Reporting and Data System) [16] density levels are approximated via grayscale histograms and used to weight loss functions, placing more emphasis on dense cases that are diagnostically challenging.

These domain-guided strategies strengthen the model's interpretability and diagnostic sensitivity. They also support the project's goal of building a culturally adapted, clinically grounded AI system for breast cancer risk prediction.

3.5 Algorithm Selection

Chosen for its proven performance and balance of depth, ResNet-50 will serve as the initial benchmark. Its residual connections allow deeper gradient flow and faster convergence.

Alternative Models. Two models are benchmarked alongside ResNet:

- **EfficientNet-B0:** Offers comparable accuracy with fewer parameters and reduced computation, ideal for edge deployment.
- **Custom Shallow CNN:** Inspired by Raghu et al. (2019), a smaller architecture trained from scratch will test the hypothesis that transfer learning from ImageNet is not always optimal in medical imaging.

Loss Function. Given the inherent class imbalance in breast cancer screening datasets—where malignant cases are relatively rare—choosing an appropriate loss function is crucial. The model is initially trained using Binary Cross-Entropy (BCE), a standard choice for binary classification tasks:

$$\mathcal{L}_{\text{BCE}} = -\left[y \cdot \log(p) + (1 - y) \cdot \log(1 - p)\right]$$

where $y \in \{0, 1\}$ is the ground truth label, and p is the predicted probability of the positive (malignant) class.

However, BCE treats all samples equally, which may bias the model toward the dominant benign class. To mitigate this and better prioritize harder examples—particularly early-stage or subtle malignancies—a modified objective function, Focal Loss, is introduced:

$$\mathcal{L}_{\text{focal}} = -\alpha_t (1 - p_t)^{\gamma} \log(p_t)$$

Here, p_t denotes the predicted probability for the ground truth class, α_t is a weighting factor to balance class contributions (e.g., higher for malignant), and $\gamma > 1$ is a focusing parameter. Setting $\gamma = 2$ reduces the loss contribution from easy examples and focuses learning on harder, misclassified cases. This is particularly valuable for enhancing sensitivity in dense breast tissue scenarios, where cancer detection is more challenging. The loss function thus directly supports the project's clinical objective of reducing false negatives in high-risk subgroups.

Optimization and Regularization. The Adam optimizer with a cyclical learning rate is used. Dropout layers are placed after the final dense layer to reduce overfitting. Early stopping with patience=10 is implemented.

3.6 Evaluation Metrics

Quantitative Metrics.

- **AUC (Area Under the ROC Curve):** Primary metric to compare model discriminability, aligned with clinical risk score assessments.
- **Sensitivity & Specificity:** Key indicators in screening settings. Sensitivity is emphasized to reduce false negatives.
- **F1-Score:** Provides a harmonic mean of precision and recall, especially relevant in imbalanced settings.

Explainability Evaluation. Grad-CAM heatmaps are overlaid on input mammograms and qualitatively scored by domain experts (when feasible) to assess alignment with tumor regions or diagnostic cues.

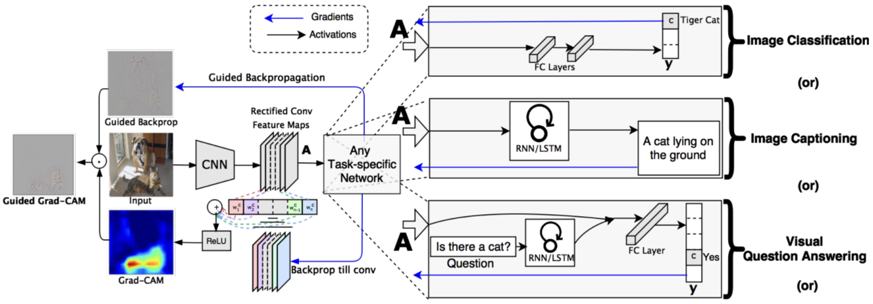


Figure 4: Grad-CAM visualizations showing model attention over mammograms. Red indicates regions of high diagnostic interest.

Qualitative Explainability Assessment. Beyond standard performance metrics, interpretability is essential in medical imaging. Kim et al. (2018) proposed the Data-driven Imaging Biomarker model (DIB-MG), which generates attention maps that highlight risk-relevant regions in mammograms. Their work supports my use of Grad-CAM as a means to validate model behavior against radiological intuition. As shown in Figure 5, attention maps provide visual grounding for predictions, which is key to clinical trust and regulatory validation.

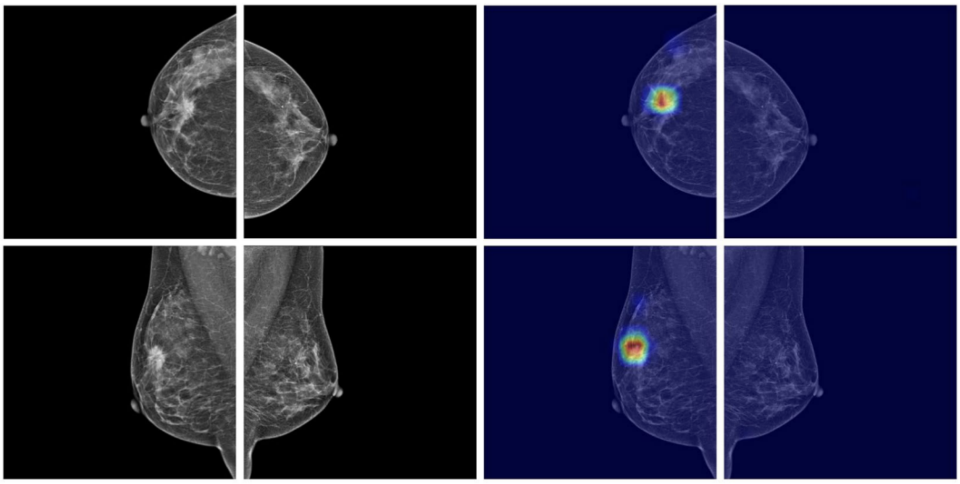


Figure 5: DIB-MG: Attention-based imaging biomarker model localizing high-risk regions. Adapted from Kim et al. (2018).

Risk Stratification. Patients are ranked by predicted risk scores and divided into deciles. Metrics like:

- **Cumulative incidence in top decile**
- **Hazard ratio (HR)**
- **Calibration curves**

are computed to evaluate clinical utility.

Cross-validation. A 5-fold stratified cross-validation is used during training. The final model is evaluated on a held-out test set to prevent data leakage and overfitting.

Conclusion. This project adopts a structured, explainable deep learning architecture validated through standard and custom metrics. By leveraging Grad-CAM, dense tissue-aware training, and Singapore-specific context simulation, the design aligns both technically and clinically with the needs of under-screened populations in Southeast Asia.

3.7 Work Plan and Project Timeline

This project follows a structured timeline that spans from July to early September 2024. It is organized into four major phases:

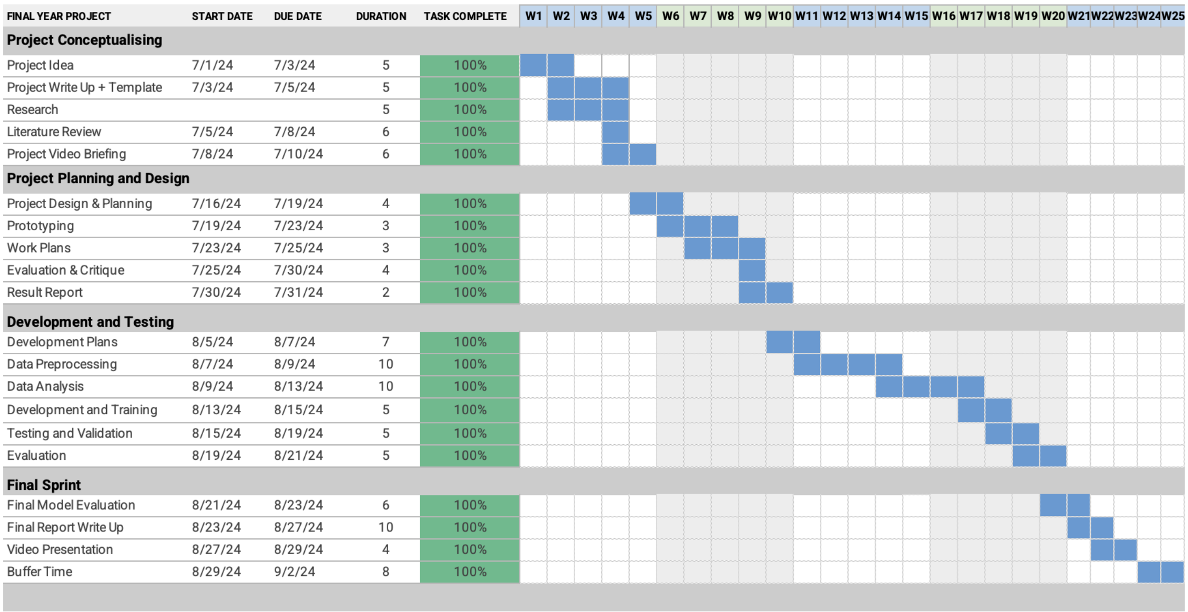
1. **Project Conceptualising (Weeks 1–5):** Includes idea formulation, early research, literature review, and preparation of briefing materials.

2. **Project Planning and Design (Weeks 5–9):** Covers system design, prototyping, drafting of detailed work plans, and early-stage reporting.

3. **Development and Testing (Weeks 10–18):** Encompasses data preprocessing, model development, training, validation, and initial evaluation.

4. **Final Sprint (Weeks 18–25):** Focuses on final model evaluation, report writing, video presentation, and includes buffer time for contingency.

Figure 6 presents the visual Gantt chart outlining key tasks and deadlines, ensuring accountability across all milestones.



FINAL YEAR PROJECT	START DATE	DUE DATE	DURATION	TASK COMPLETE	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14	W15	W16	W17	W18	W19	W20	W21	W22	W23	W24	W25
Project Conceptualising																													
Project Idea	7/1/24	7/3/24	5	100%	■	■																							
Project Write Up + Template	7/3/24	7/5/24	5	100%		■	■	■																					
Research			5	100%		■	■	■																					
Literature Review	7/5/24	7/8/24	6	100%				■																					
Project Video Briefing	7/8/24	7/10/24	6	100%				■	■																				
Project Planning and Design																													
Project Design & Planning	7/16/24	7/19/24	4	100%					■	■																			
Prototyping	7/19/24	7/23/24	3	100%						■	■	■																	
Work Plans	7/23/24	7/25/24	3	100%							■	■	■																
Evaluation & Critique	7/25/24	7/30/24	4	100%									■																
Result Report	7/30/24	7/31/24	2	100%									■	■															
Development and Testing																													
Development Plans	8/5/24	8/7/24	7	100%										■	■														
Data Preprocessing	8/7/24	8/9/24	10	100%											■	■	■	■											
Data Analysis	8/9/24	8/13/24	10	100%														■	■	■	■								
Development and Training	8/13/24	8/15/24	5	100%																	■	■							
Testing and Validation	8/15/24	8/19/24	5	100%																		■	■						
Evaluation	8/19/24	8/21/24	5	100%																			■	■					
Final Sprint																													
Final Model Evaluation	8/21/24	8/23/24	6	100%																				■	■				
Final Report Write Up	8/23/24	8/27/24	10	100%																					■	■			
Video Presentation	8/27/24	8/29/24	4	100%																						■	■		
Buffer Time	8/29/24	9/2/24	8	100%																								■	■

Figure 6: Project Gantt chart showing planned activities from July to September 2024, divided into phases of research, development, evaluation, and reporting.

4 Feature Prototype

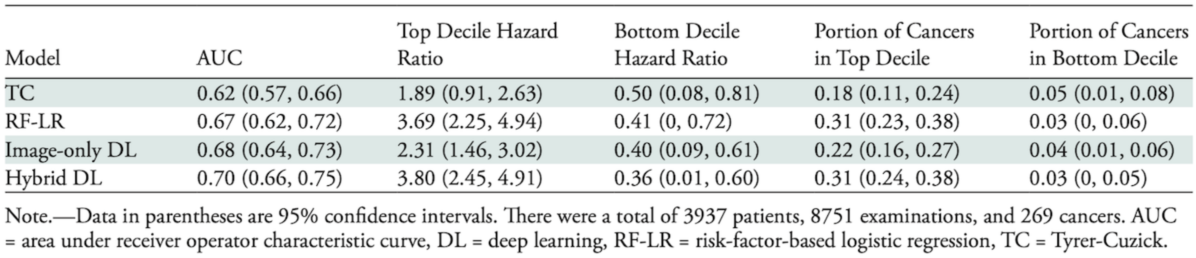
4.1 Development Strategy

4.2 Evaluation

4.3 Improvements and Next Steps

5 Appendices

5.1 Figures Referenced in Section 2.1: Limitations of Traditional Risk Models



Model	AUC	Top Decile Hazard Ratio	Bottom Decile Hazard Ratio	Portion of Cancers in Top Decile	Portion of Cancers in Bottom Decile
TC	0.62 (0.57, 0.66)	1.89 (0.91, 2.63)	0.50 (0.08, 0.81)	0.18 (0.11, 0.24)	0.05 (0.01, 0.08)
RF-LR	0.67 (0.62, 0.72)	3.69 (2.25, 4.94)	0.41 (0, 0.72)	0.31 (0.23, 0.38)	0.03 (0, 0.06)
Image-only DL	0.68 (0.64, 0.73)	2.31 (1.46, 3.02)	0.40 (0.09, 0.61)	0.22 (0.16, 0.27)	0.04 (0.01, 0.06)
Hybrid DL	0.70 (0.66, 0.75)	3.80 (2.45, 4.91)	0.36 (0.01, 0.60)	0.31 (0.24, 0.38)	0.03 (0, 0.05)

Note.—Data in parentheses are 95% confidence intervals. There were a total of 3937 patients, 8751 examinations, and 269 cancers. AUC = area under receiver operator characteristic curve, DL = deep learning, RF-LR = risk-factor-based logistic regression, TC = Tyrer-Cuzick.

Figure 7: Performance comparison between the Tyrer-Cuzick model and deep learning models across subgroups. Adapted from [1].

5.2 Figures Referenced in Section 2.2: Role of Deep Learning in Breast Cancer Imaging

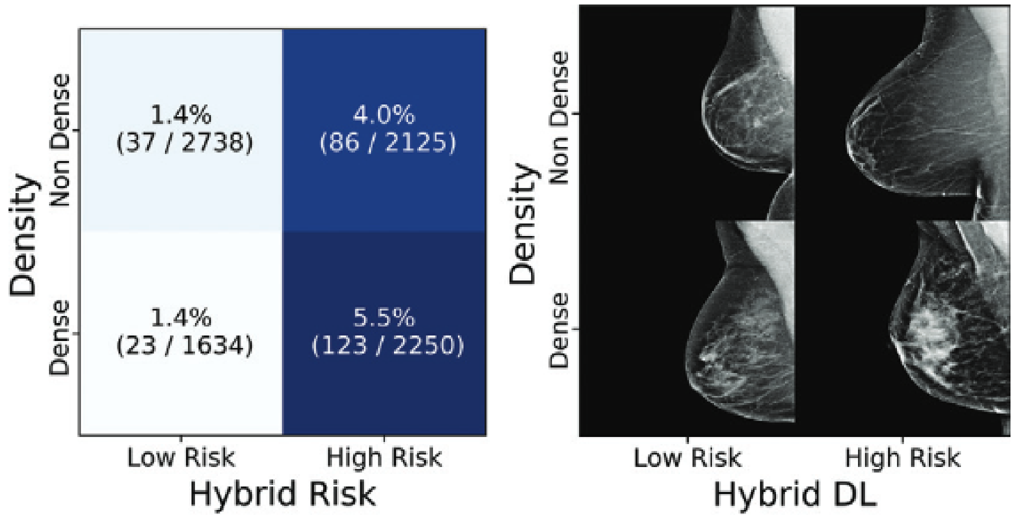


Figure 8: Cancer incidence across density groups and hybrid risk scores. Adapted from [1].

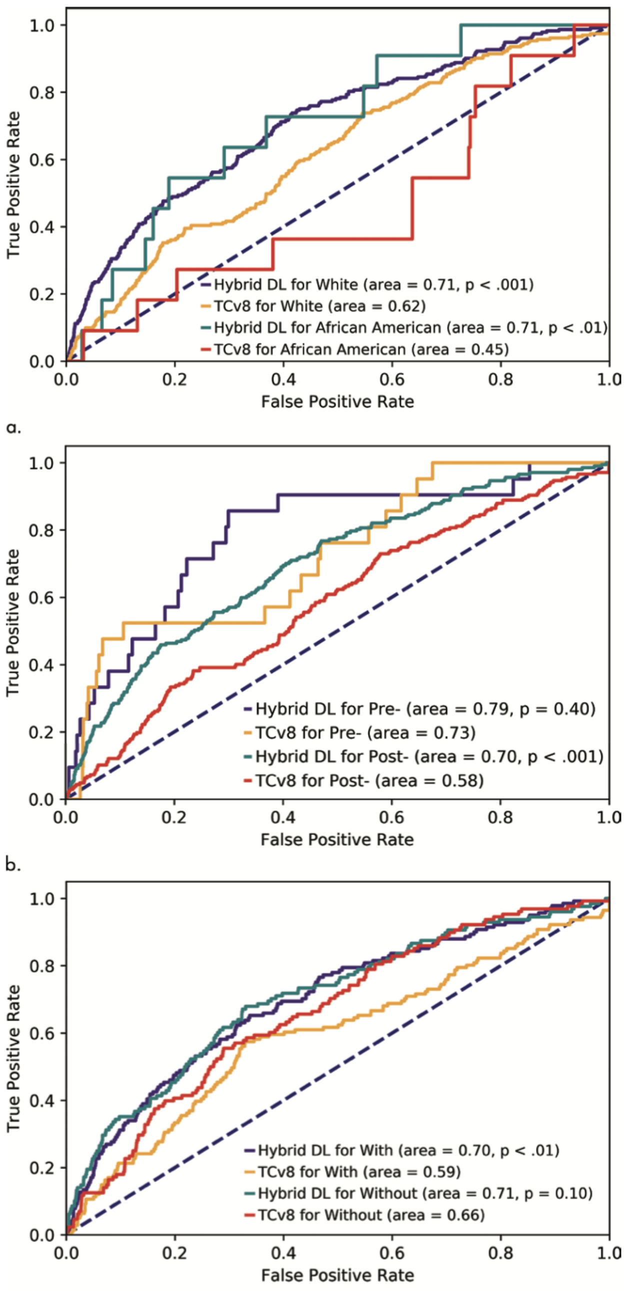


Figure 9: ROC curves comparing Tyrer-Cuzick, image-only DL, and hybrid DL models. Adapted from [1].

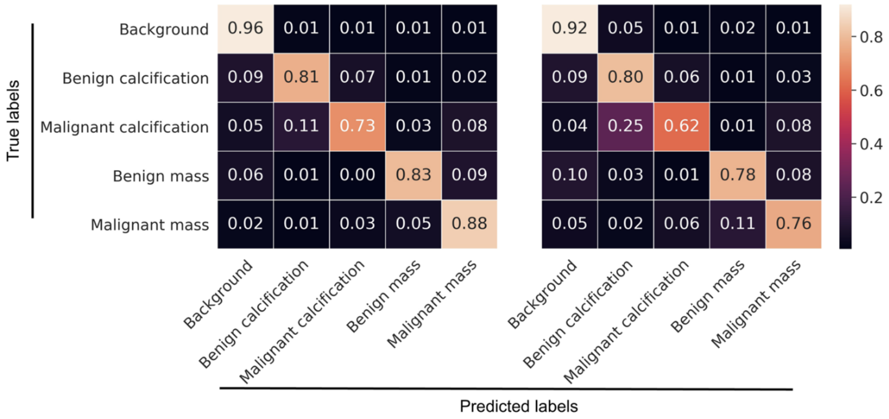


Figure 10: CNN-generated detection probabilities on mammograms. Adapted from [7].

5.3 Figure Referenced in Section 2.3: Explainability in Medical AI

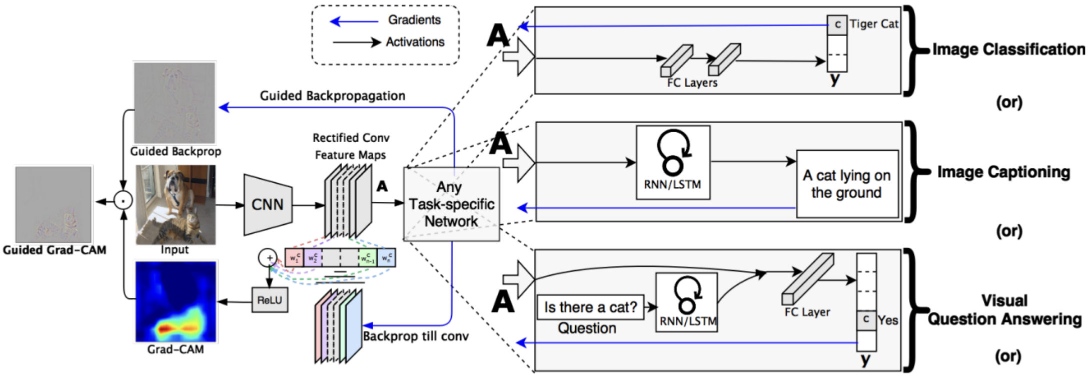


Figure 11: Example of Grad-CAM heatmaps overlaid on mammograms, highlighting regions associated with high-risk predictions. Adapted from [5].

5.4 Figures Referenced in Section 2.4: Dataset Diversity and Generalization

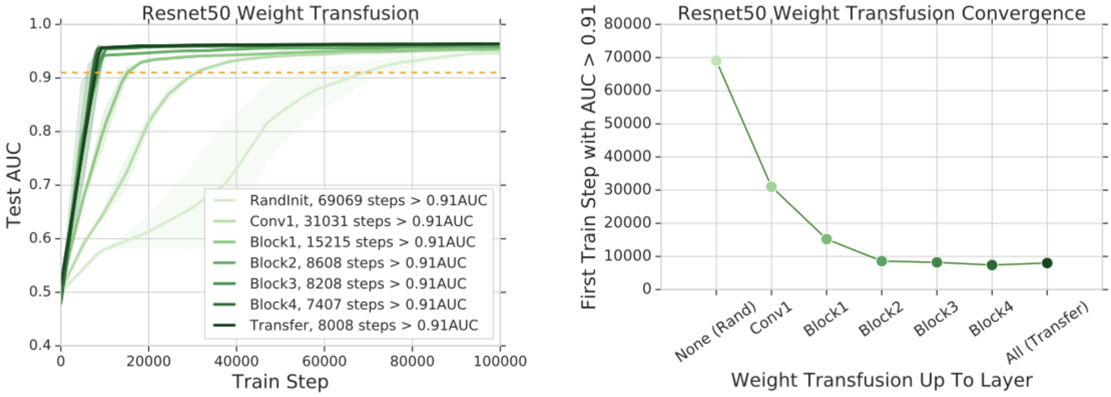


Figure 12: Comparison of pretrained vs. randomly initialized CNNs in medical imaging tasks. Adapted from [2].

5.5 Questionnaire / Lifestyle Simulation Input

5.6 Sample Transcriptions or Audio Prompts

6 References

References

[1] A. Yala, C. Lehman, T. Schuster, T. Portnoi, and R. Barzilay. 2019. A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction. *Radiology* 292, 1 (2019), 60–66. https://pubs.rsna.org/doi/pdf/10.1148/radiol.2019182716

[2] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio. 2019. Transfusion: Understanding Transfer Learning for Medical Imaging. In *Proc. NeurIPS 2019*, 3342–3352. https://proceedings.neurips.cc/paper/2019/file/eb1e78328c46506b46a4ac4a1e378b91-Paper.pdf

[3] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, et al. 2018. Opportunities and obstacles for deep learning in biology and medicine. *Nature Reviews Genetics* 19 (2018), 141–158. https://royalsocietypublishing.org/doi/full/10.1098/rsif.2017.0387

[4] V. Cheplygina, M. de Bruijne, and J. P. W. Pluim. 2019. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis* 54 (2019), 280–296. https://arxiv.org/pdf/1804.06353

[5] R. R. Selvaraju, et al. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In *Proc. ICCV*, 618–626. https://openaccess.thecvf.com/content_ICCV_2017/papers/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.pdf

[6] W. Y. Chau, G. H. Lim, Y. L. Ng, et al. 2021. Breast density and breast cancer risk in Asian women: Evidence from the Singapore Breast Screening Programme. *Cancer Epidemiology* 74 (2021), 101987. https://pmc.ncbi.nlm.nih.gov/articles/PMC5160133/

[7] L. Shen, L. R. Margolies, J. H. Rothstein, et al. 2019. Deep learning to improve breast cancer detection on screening mammography. *Scientific Reports* 9, 1 (2019), 12495. https://www.nature.com/articles/s41598-019-48995-4

[8] K. J. Geras, S. Wolfson, Y. Shen, et al. 2017. High-resolution breast cancer screening with multi-view deep convolutional neural networks. *arXiv preprint arXiv:1703.07047*. https://arxiv.org/abs/1703.07047

[9] H. J. Kim, E. Y. Ko, C. Kim, W. Han, and W. K. Moon. 2018. Applying Data-driven Imaging Biomarker in Mammography for Breast Cancer Screening: Preliminary Study. *Scientific Reports* 8, 1 (2018), 12210. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5807343/

[10] National Registry of Diseases Office. 2021. *Singapore Cancer Registry Annual Report 2021*. https://www.nrdo.gov.sg/docs/librariesprovider3/publications-cancer/cancer2021.pdf

[11] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, et al. 2020. International evaluation of an AI system for breast cancer screening. *Nature*, 577 (2020), 89–94. https://www.nature.com/articles/s41586-019-1799-6

[12] Zebra Medical Vision. 2021. Zebra's AI1 Mammography Solutions. https://www.zebra-med.com/

[13] N. Salim, J. L. Andersson, A. Nordenskjöld, H. Svensson, K. Törnberg, and K. Zackrisson. 2024. Artificial intelligence for breast cancer screening in a real-world, population-based setting: cohort study of 58,000 mammograms. *Nature Medicine*, (2024). https://www.nature.com/articles/s41591-024-03061-0

[14] J. Shen, L. Margolies, J. Rothstein, E. Fluder, R. McBride, and W. Sieh. 2019. Deep Learning to Improve Breast Cancer Detection on Screening Mammography. *Scientific Reports* 9 (2019), 12495. https://www.nature.com/articles/s41598-019-48995-4

[15] K. Zuiderveld. 1994. Contrast Limited Adaptive Histogram Equalization. In P. S. Heckbert (Ed.), *Graphics Gems IV*, 474–485. Academic Press. https://doi.org/10.1016/B978-0-08-050755-2.50065-6

[16] American College of Radiology. 2013. Breast Imaging Reporting and Data System (BI-RADS). *ACR BI-RADS Atlas, Breast Imaging Reporting and Data System.* 5th Edition. American College of Radiology.

[17] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition. In *Proc. CVPR*, 770–778. https://doi.org/10.1109/CVPR.2016.90

[18] M. Heath, K. Bowyer, D. Kopans, R. Moore, and W. P. Kegelmeyer. 2000. The Digital Database for Screening Mammography. In *Proc. 5th International Workshop on Digital Mammography*, 212–218. http://www.eng.usf.edu/cvprg/Mammography/Database.html

[19] I. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte. 2016. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63, 7 (2016), 1455–1462. https://doi.org/10.1109/TBME.2015.2496264

[20] J. Daugman. 1985. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, 2, 7 (1985), 1160–1169. https://doi.org/10.1364/JOSAA.2.001160