AI-Based Mammographic

Breast Cancer Detection Using Deep Learning:

Model Development and Validation

Using Kaggle’s RSNA Dataset

1. Project Title

AI-Based Mammographic Breast Cancer Detection Using Deep Learning: Model Development and Validation Using Kaggle’s RSNA Dataset

2. Project Team Members:

Name	Student ID	Contact
Gamal Ehab Gamal Gerges	21071843	ghobrialjamal@gmail.com
Ahmed Jaber Abdelaziz	21049596	ahmedaboenaba33@gmail.com
Tasneem Hussien Aboelsoud	21083319	th1171@fayoum.edu.eg
Omnia Mohamed Kamal	21030547	mostafamohamed332023@gmail.com
Mohamed ElSayed AbdelAzz AbdelAty	21093847	dr.mohamed.abdelaty85@gmail.com

3. Project Advisor:

Eng. Sherif Salem	sherifsalem313@gmail.com

4. Executive Summary:

*This project aims to develop an **AI-based system for breast cancer detection** using the **RSNA Breast Cancer Detection Dataset** hosted on **Kaggle**. Leveraging deep learning and transfer learning techniques such as **Convolutional Neural Networks (CNNs)**, the study will automate mammogram interpretation to improve diagnostic accuracy, reduce radiologist workload, and enhance early cancer detection. Conducted within Kaggle's cloud environment and supported by local GPU computing, the research follows six milestones—from dataset preparation and model development to deployment and explainability analysis. The system's performance will be evaluated using AUC, sensitivity, and specificity metrics, with interpretability ensured through Grad-CAM visualization. The final outcome will be a validated, transparent, and scalable AI model capable of assisting radiologists in real-world breast cancer screening and improving patient outcomes globally.*

5. Background and Rationale:

Breast cancer is the most common malignancy among women globally and a leading cause of cancer-related deaths. Early detection through screening mammography remains the cornerstone of reducing mortality. However, radiologic interpretation of mammograms is time-consuming and prone to inter-observer variability, especially in dense breast tissue.

Advances in **deep learning** have revolutionized image analysis, offering opportunities for automated and accurate cancer detection. The **Radiological Society of North America (RSNA)** has released a large, curated **mammography dataset** through **Kaggle's RSNA Breast Cancer Detection Competition**. This dataset provides labeled mammograms suitable for training AI models to identify cancerous lesions.

Using **Kaggle's free cloud computing environment**, this study aims to design, train, and validate deep learning models that automatically detect breast cancer in mammographic images, comparing their performance with existing radiologic benchmarks.

6. Research Objectives:

6.1 Primary Objective

To develop and validate a convolutional neural network (CNN)-based model for automated detection of breast cancer in screening mammograms using the RSNA Breast Cancer Detection dataset on Kaggle.

6.2 Secondary Objectives

1. *To evaluate and compare the performance of different CNN architectures (ResNet, EfficientNet, DenseNet).*

2. *To apply **transfer learning** to improve model generalization and training efficiency.*

3. *To develop an explainable AI module using **Grad-CAM** visualization to highlight suspicious regions.*

4. *To deploy the best-performing model via an interactive web interface for real-time testing and validation.*

7. Research Questions / Hypotheses:

7.1 Research Questions

- *Can a deep learning model trained on RSNA mammography data accurately detect breast cancer compared to traditional radiologic interpretation?*
- *Does the use of transfer learning improve model performance on mammography datasets?*
- *Can explainable AI enhance clinician trust and usability in breast cancer diagnostics?*

7.2 Hypotheses

- ***H_0:*** *AI models do not significantly outperform baseline radiologist-level accuracy.*

- ***H_1:*** *AI models trained on RSNA mammography data significantly improve detection accuracy and reduce false negatives.*

8. Data Source and Management:

8.1 Dataset

- **Source:** RSNA Breast Cancer Detection Competition on Kaggle https://www.kaggle.com/competitions/rsna-breast-cancer-detection
- **Modality:** Digital Mammography (2D X-ray)
- **Format:** DICOM
- **Size:** ~200,000 images from ~30,000 patients
- **Labels:** Binary (cancer / no cancer), with patient-level metadata and breast laterality
- **Annotations:** Provided by expert radiologists, linked to biopsy-confirmed diagnoses

8.2 Data Preprocessing

1. Convert DICOM to PNG or Tensor format for model compatibility.
2. Normalize image intensity values (0–1 scale).
3. Resize to 512×512 pixels.
4. Apply augmentation (flipping, rotation, brightness/contrast changes).
5. Split into training (70%), validation (15%), and testing (15%).

All preprocessing scripts and model checkpoints will be hosted on **Kaggle Notebooks** and version-controlled via **GitHub**.

9. Methodology:

9.1 Study Design

Experimental study for diagnostic AI model development and validation.

9.2 Workflow Overview

Step	Task	Platform / Tool
1	Dataset acquisition	Kaggle RSNA competition data
2	Data preprocessing & augmentation	Python (OpenCV, Albumentations)
3	Model development	TensorFlow/Keras or PyTorch
4	Model evaluation	ROC-AUC, Precision, Recall, F1-score
5	Explainability visualization	Grad-CAM
6	Deployment	Streamlit / Gradio (Hugging Face Spaces)

9.3 Model Architectures

- Baseline: Custom 5-layer CNN
- Pretrained Models: ResNet50, EfficientNet-B0, DenseNet121
- Fine-tuning: Last 2–3 convolutional blocks + fully connected layers
- Optimizer: Adam (learning rate = 1e-4)
- Loss Function: Binary Cross-Entropy
- Batch Size: 32
- Epochs: 50 (with early stopping)

9.4 Model Evaluation Metrics

- Accuracy
- Sensitivity (Recall)
- Specificity
- F1-Score
- Area Under ROC Curve (AUC)
- Confusion Matrix

10. Expected Outcomes

1. A validated deep learning model capable of identifying cancerous mammographic images with high sensitivity.
2. Comparative performance report between architectures (ResNet vs EfficientNet vs DenseNet).
3. Explainable AI visualization maps (Grad-CAM) for clinical interpretability.
4. A deployable **web-based diagnostic prototype** (Hugging Face / Streamlit).
5. Open-access code and dataset documentation on **Kaggle and GitHub**.

11. Ethical and Legal Considerations

- Data is publicly available and de-identified, compliant with HIPAA and RSNA data-use policies.
- No direct patient contact or clinical intervention is involved.
- Results will be used solely for academic and research purposes.

12. Timeline & deliverables

Phase	***Duration***	***Key Deliverables***
Phase 1: *Literature Review & Dataset Preparation*	*week 1*	*Literature summary, data download, and preprocessing scripts*
Phase 2: *Model Development*	*week 2–3*	*CNN training and architecture optimization*
Phase 3: *Validation & Explainability*	*week 4*	*Performance comparison and Grad-CAM visualizations*
Phase 4: *Deployment*	*week5*	*Web app / API integration for testing*
Phase 5: *Documentation & Dissemination*	*week 6*	*Final report, GitHub repo, and publication submission*

Milestone	Phase Title	Duration	Key Activities	Deliverables
Milestone 1	Project Initiation & Literature Review	Week 1	- Review state-of-the-art research on AI-based breast cancer detection .- Define research scope, objectives, and hypotheses. - Establish project repository (GitHub) and Kaggle workspace.	Project charter, literature summary, initial workflow plan
Milestone 2	Dataset Preparation & Preprocessing	Week 1	- Acquire and organize RSNA Kaggle dataset .- Perform DICOM conversion, normalization, resizing, and augmentation. - Implement data-splitting (train/validation/test).	Cleaned, preprocessed dataset; preprocessing scripts on GitHub
Milestone 3	Model Development & Training	Weeks 2–3	- Design baseline CNN model. - Implement transfer learning using ResNet50, DenseNet121, and EfficientNet-B0. - Conduct hyperparameter tuning and optimization.	Trained CNN models; training logs and model checkpoints
Milestone 4	Model Validation & Explainability Analysis	Week 4	- Evaluate models on test dataset using AUC, sensitivity, specificity, and F1-score .- Generate Grad-CAM visualizations for interpretability. - Compare architectures and select the best-performing model.	Performance report, confusion matrices, Grad-CAM heatmaps
Milestone 5	Prototype Deployment	Week 5	- Integrate the best-performing model into an interactive web interface using Streamlit or Gradio. - Host prototype on Hugging Face Spaces or local server. - Conduct real-time model	Web-based diagnostic prototype; testing feedback summary

			testing.	
Milestone 6	Documentation & Dissemination	Week 6	- Compile final technical report and publish code repository .- Prepare academic manuscript or conference submission .- Summarize findings and ethical compliance.	Final research report, GitHub repository, publication draft

13. Tools and Platforms

Purpose	***Platform / Tool***
Dataset	*Kaggle (RSNA Competition)*
Computing Environment	*Kaggle Notebooks + Local GPU*
Deep Learning Framework	*TensorFlow / PyTorch*
Visualization	*Matplotlib, Seaborn, Grad-CAM*
Deployment	*Streamlit / Gradio / Hugging Face*
Version Control	*GitHub*

14. References

1. Chris Carr, FelipeKitamura, MD, PhD, George Partridge, inversion, Jayashree Kalpathy-Cramer, John Mongan, Katherine Andriole, Lavender, Maryam Vazirabad, Michelle Riopel, Robyn Ball, Sohier Dane, and Yan Chen. RSNA Screening Mammography Breast Cancer Detection.

https://kaggle.com/competitions/rsna-breast-cancer-detection, 2022. Kaggle.

2. RSNA Breast Cancer Detection. Kaggle, 2022. https://www.kaggle.com/competitions/rsna-breast-cancer-detection

3. Clark, K., et al. (2013). *The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository*. *J. Digit. Imaging*, 26(6), 1045–1057.

4. Simonyan, K., & Zisserman, A. (2014). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. arXiv:1409.1556.

5. He, K., et al. (2016). *Deep Residual Learning for Image Recognition (ResNet)*. *CVPR 2016*.