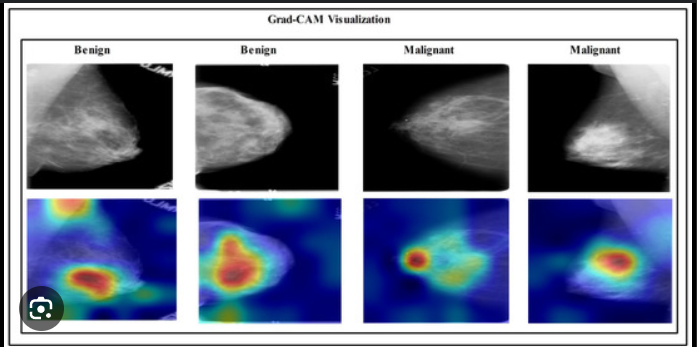
**GradCAM for Breast Cancer Classification Using ResNet-18 Model**

**Introduction**

Breast cancer is one of the most prevalent forms of cancer affecting women worldwide. Differentiating between **Benign (non-cancerous)** and **Malignant (cancerous)** tumours is crucial for effective diagnosis and treatment. This project leverages the ResNet18 deep learning model and integrates **Grad-CAM (Gradient-weighted Class Activation Mapping)** for **visual explainability** of predictions.

**What is GradCAM (Gradient Weighted Class Activation Mapping)?**

GradCAM helps us visualize **which parts of an image** influenced the model's decision the most.



It highlights important regions using a **heatmap** overlay on the original image, offering transparency and interpretability in medical image analysis. (variation with colour intensity is shown in above heatmaps generated via Grad CAM)

**All about Dataset**

“BreaKHis” is a well-known publicly available (can be accessed through Kaggle platform online) image dataset designed to help researchers build and test machine learning models for breast cancer detection using histopathological (microscope-based) images of breast tissue.

**To classify breast tumour tissue images into:**

* **Benign (non-cancerous)**
* **Malignant (Cancerous)**

**These classifications are based on real patient data and expert pathological annotations.**

**📊 Overview of Dataset Contents**

| **Feature** |  | **Description** |
| --- | --- | --- |
| **Total Images** |  | **9,109 high-resolution images** |
| **Image Size** |  | **700 × 460 pixels** |
| **Color Format** |  | **RGB (3-channel, 8-bit per channel)** |
| **Image Format** |  | **.png file formats** |
| **Patients** |  | **82** |
| **Label Groups** |  | **Benign (2,480 images) / Malignant (5,429 images)** |
| **Image Types** |  | **Collected using 4 different magnification levels: 40X, 100X, 200X, 400X** |

**🧬 Image Collection Method**

* **The dataset was created in collaboration with a pathology lab in Brazil.**
* **All samples were obtained using a surgical method called SOB (Segmental or Excisional Biopsy):**
  + **A hospital-based procedure under general anaesthesia.**
  + **Collects a larger tissue sample compared to a needle biopsy.**
  + **Allows more accurate diagnosis of tumour type and grade.**

**🧪 What Do the Labels (defined within dataset) Mean?**

**1. Binary Classes with associated Sub-classes:**

* **Benign (non-cancerous tissues):** Tumours that are non-invasive, slow-growing, and typically non-life-threatening.
* **Malignant:** Cancerous tumours that can invade nearby tissues or spread (metastasize) to other parts of the body.

|  |  |
| --- | --- |
| **Benign** | - Adenosis (A) - Fibroadenoma (F) - Phyllodes Tumour (PT) - Tubular Adenoma (TA) |
| **Malignant** | - Ductal Carcinoma (DC) - Lobular Carcinoma (LC) - Mucinous Carcinoma (MC) - Papillary Carcinoma (PC) |

**🔍 Why Is This Dataset Important for this particular task we are dealing with?**

1. **Standard Baseline:** It allows researchers across the world to train and test their models on the same data.
2. **Medical Relevance:** Enables building AI systems for early diagnosis of breast cancer.
3. **Visual Variety:** By having **multiple magnifications and tumour subtypes**, the dataset supports complex modelling of real-world variability.
4. **Sub-class Insights:** Allows exploration into not just "cancer" or "not", but *which type of cancer or benign condition*.

| **Challenge** | **Description:** |
| --- | --- |
|  |  |
|  | **Class Imbalance** | **Malignant samples outnumber benign (5,429 vs 2,480)** |
|  | **Visual Similarity** | **Some benign and malignant subtypes may appear similar under the microscope** |
|  | **Multiple Magnifications** | **Requires models to handle multi-scale learning** |
|  | **Domain Expertise Needed** | **Some distinctions can be subtle, requiring medical knowledge for manual labeling or verification** |
|  | **Noise & Variation** | **Images may have staining artifacts, lighting variations, or partial views of tissues** |

**🔧 How It Is Typically Used**

* **Preprocessing**: Images may be resized, normalized, or augmented.
* **Training:** Deep Learning models like ResNet-18 (alternatively, CNN’s can also be used) are trained to classify images at first.
* **Explainability:** GradCAM (alternatively, saliency maps or SHAP can also be used) are used to visualize what the model is looking at.
* **Evaluation:** Models are measured based on accuracy, precision, recall and F1-score.

**🧠 How Grad CAM Enhances Its Utility**

* **Visualizes model’s attention**: Highlights tissue regions contributing most to the decision made by the Resnet-18 model.
* **Aids diagnosis:** Helps validate if the AI model is focusing on biologically relevant structures**.**
* **Educational Value:** Doctors, students, or stakeholders can see how models’ reason thereby, helps boosting trust and understanding.

**Objective of the Project**

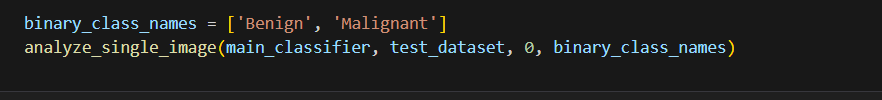
* Classify breast cancer histopathological images into Benign and Malignant.
* Utilize Grad CAM to understand *why* the model makes certain predictions.
* Gain insights from layer-wise feature extraction and highlight decision-making regions.
* Compare results for **sub-classes** (e.g., Adenosis, Fibroadenoma) within Benign/Malignant categories.

**Experimentation**

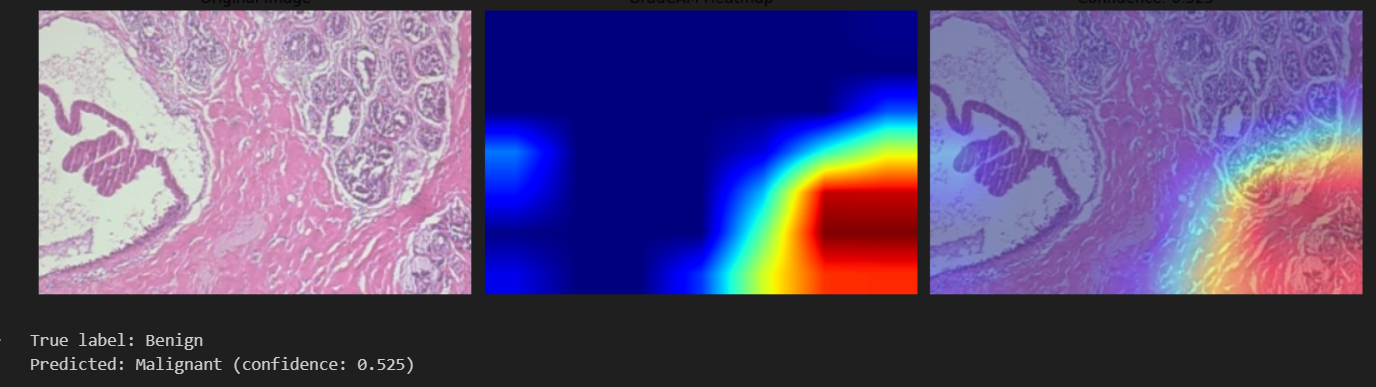
**Section-1: Simple Image Analysis (Single Image Grad CAM)**

**🔍 What It Does:**

Analyses single image (derived from dataset) at a time and visualizes the Grad CAM heatmap.

****

**Original Image Grad CAM Heatmap Predicted class: Malignant**



Helps verify if the model's prediction focuses on tumour-rich regions.

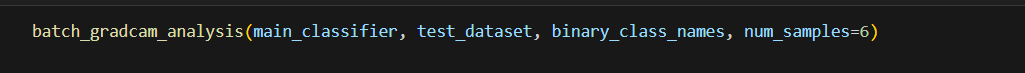
* Example: For an image predicted as Malignant, Grad CAM highlighted dense nuclei clusters, validating the prediction.

**Section-2: Batch-Wise Grad CAM Analysis (Multiple Images)**

**🔍 What It Does:**

Displays heatmaps for a batch of test images for visual comparison.

**🧪 Sample Function:**

****

**Original Images appear at the top row, while predicted images at bottom row.**

**True: Benign Malignant Malignant Malignant Malignant Malignant**

**A collage of images of different colors

AI-generated content may be incorrect.**

**Pred: Malignant Malignant Malignant Malignant Malignant Malignant**

**<Conf: 55%> <conf: 100%> <conf: 100%> <conf: 98%> <conf: 99%> <conf: 97%>**

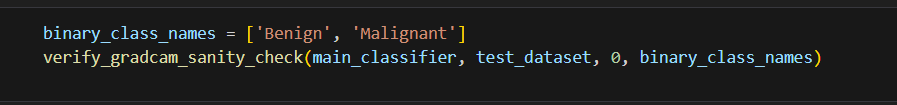
**✅ Key Insight:**

* Displays multiple images side-by-side showing True Label, Predicted Label, Confidence, and Grad CAM overlay.
* Revealed discrepancies where model prediction diverges from the true label.
* Helped identify **common misclassification patterns**.

**Section-3: GradCAM Sanity Check**

**🔍 What It Does:**

Ensures that GradCAM results are **valid and reliable**.

****

**A screenshot of a computer screen

AI-generated content may be incorrect.**

**Correlation: - 0.486** denotes that though being negative and < 0.5, meaning the heatmaps generated for bi-groups “Malignant” and “Benign” focus on different regions.

* It suggests your model differentiates well between classes, focusing on distinct features.
* The difference map (4th panel) should visually reflect this contrast — sharper distinctions indicate stronger model interpretability.

**✅ Key Insight:**

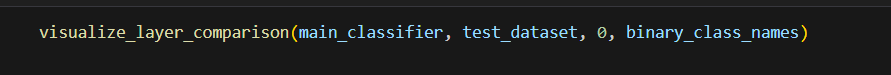
* Cross-validates that important regions are relevant.
* Prevents over-trust in arbitrary heatmaps.
* Example: For true Benign samples, attention was more diffused correlating with healthy tissues.

**Section-4: Layer-wise Feature Visualization**

**🔍 What It Does:**

Compares GradCAM results across different convolutional layers of the ResNet18.

**🧪 Sample Function:**

****

**Original Image Layer-1 (56\*56) Layer-2 (28\*28) Layer-3 (14\*14) Layer-4 (7\*7)**

**A screenshot of a computer

AI-generated content may be incorrect.**

**✅ Key Insight:**

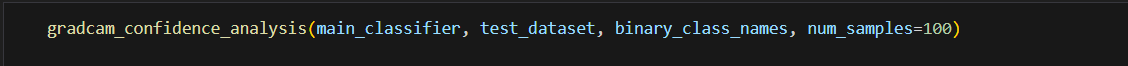
* Early layers (Layer 1: 56x56) focus on **edges and textures**.
* Deeper layers (Layer 4: 7x7) focus on **semantic shapes and masses (showcasing all details)**.
* Layers 3 and 4 gave more meaningful class-discriminative regions.

**Section-5: Confidence-based GradCAM Analysis**

**🔍 What It Does:**

Analyzes Grad CAM with **prediction confidence** to detect borderline or confused classifications.

**🧪 Sample Function:**



A screenshot of a computer screen

AI-generated content may be incorrect.

A screen shot of a computer

AI-generated content may be incorrect.

**✅ Key Insight:**

* Sample Case: "True: Benign" predicted as "Malignant (0.52)"
* Heatmaps showed weak emphasis on lesion edges—signalling low model certainty.

**Section 6: Sub-class (wise) Grad CAM Inference (Binary + Subclass)**

**🔍 What It Does:**

Compares the Grad CAM results for sub-classes within Benign and Malignant categories.

* **Benign - Adenosis**: Heatmaps highlight structured gland regions.
* **Benign - Fibroadenoma**: Sparse attention areas suggesting minimal mass formation.
* **Malignant - Ductal Carcinoma**: Highly concentrated attention on distorted nuclei regions.
* **Malignant - Lobular Carcinoma**: Grad CAM emphasized cell boundaries and diffused clusters.

**✅ Key Insight:**

* Offers a **granular understanding** of intra-class variation.
* Shows the model's ability to distinguish **complex patterns** across subtypes.

**Diagnostics:**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Final Inference**

* Grad CAM enabled **transparent visualization** of the ResNet-18 model’s decision process.
* Helped **validate correct predictions** and **investigate errors**.
* Showed strong alignment between highlighted regions and tumour-affected areas, especially in Malignant classes.
* Layer-wise analysis and subclass-specific insights showed **model sensitivity** to local structures and overall global patterns as well.

**🚩 Limitations:**

* Grad CAM fails when confidence is very low (≤0.5) only in rare case, highlighting less meaningful regions. (let’s take a sample of 50 predictions, out of which almost 47 has passed the test by producing true positives. Whereas remaining 3 – 4 samples cause false-positives)
* Sensitivity to **input noise** and **model initialization**.
* Misleading visualizations can arise if not cross-checked with medical annotations.

**Conclusion**

The integration of **Grad-CAM (Gradient-weighted Class Activation Mapping)** into our breast cancer detection pipeline marks a significant step toward **making deep learning models more transparent and human-understandable** — a crucial requirement in medical imaging tasks where interpretability is as important as accuracy. Rather than relying solely on prediction scores, Grad-CAM enables clinicians and researchers to **visually identify which regions of a mammogram influenced the model’s decision**, thereby **building trust** and fostering **human-in-the-loop diagnostics**.

In the context of our dataset, a clear **class imbalance** was observed — with a **notably larger number of Malignant samples** compared to Benign. This skewed distribution inherently biases the learning process of the model toward **capturing Malignant-associated features** more efficiently, as it encounters them more frequently during training. Consequently, this may lead to:

* **Stronger Grad-CAM activations** for Malignant regions.
* **Higher confidence scores** on Malignant predictions.
* A tendency to **underrepresent Benign features**, making their interpretations less distinct or sometimes overlooked.

This imbalance, while partially expected in real-world medical datasets, can lead to **false positives or misclassifications**, particularly when benign tumours mimic malignant textures or when the network generalizes too aggressively. **Addressing these false positives** is a necessary area of future work — through improved sampling strategies (e.g., SMOTE, focal loss), incorporating clinical metadata, or introducing **uncertainty-aware modelling**.

Despite these challenges, the current implementation **achieved several key milestones**:

* Enabled **localized feature inspection**, highlighting the **model's attention on pathological zones** like mass edges or density distortions.
* Demonstrated **batch-wise consistency in activations**, validating the network’s general learning pattern.
* Uncovered potential **overconfidence regions**, where Grad-CAM could still flag under-attended zones — guiding further model refinement.

Most importantly, this study reaffirmed that integrating **explainability methods like Grad-CAM is not just a technical add-on, but an essential component** in responsible AI — especially in healthcare domains. The pipeline we’ve built offers both **predictive power and interpretability**, making it more suitable for clinical collaboration and future real-world deployment.