BREAST CANCER CLASSIFICATION AND DETECTION THROUGH ADVANCED PRE-PROCESSING AND DEEP LEARNING TECHNIQUES

PROJECT ID(PCSE25-55)

PROJECT SYNOPSIS

OF MAJOR PROJECT

BACHELOR OF TECHNOLOGY
Computer Science and Engineering

SUBMITTED BY
Sajal Bhilatia (2100290100143)
Rachit Verma (2100290100124)
Harsh Rastogi (2200290109006)

Project Guide
Prof. Gagan Thakral

KIET Group of Institutions, Delhi-NCR,
Ghaziabad (UP)
Department of Computer Science and Engineering

October 2023

INTRODUCTION

Breast cancer is one of the leading causes of cancer-related mortality among women across the globe. Its rising incidence and the severity of its impact on health systems make early detection and diagnosis a matter of urgent concern. Traditional diagnostic practices, although effective to an extent, often suffer from inconsistencies due to human limitations such as fatigue, subjective judgment, and variable experience levels among radiologists. As breast cancer often begins with subtle, hard-to-detect signs, early-stage tumours may go unnoticed or be misclassified, leading to delayed treatment and lower survival rates.

Medical imaging—particularly mammography—has been the cornerstone of breast cancer detection. It enables radiologists to examine breast tissue in high detail, allowing for identification of tumours, calcifications, and architectural distortions. However, the manual interpretation of mammograms is inherently prone to error, especially in cases involving dense breast tissue or inconspicuous lesions. This has prompted significant interest in automated, computer-aided diagnostic (CAD) systems that can support or augment clinical decision-making.

In this project, we address the need for accurate, consistent, and automated detection of breast cancer by designing a deep learning–based system that classifies mammographic images as benign or malignant. Our method integrates **advanced image preprocessing techniques** with **multiple deep learning models**, offering both performance optimization and comparative evaluation.

Preprocessing plays a crucial role in this pipeline. Mammogram images frequently contain artifacts such as pectoral muscles, noise, labels, and inconsistent brightness. If left untreated, these elements can mislead learning models. We apply techniques such as **orientation correction**, **pectoral muscle removal**, **contrast enhancement using CLAHE**, and **morphological operations** to clean and standardize the images.

Once pre-processed, the mammograms are used to train deep learning models, specifically **VGG16**, **ResNet50**, **MobileNet**, and a **Custom CNN**. These architectures were chosen for their proven performance in image classification tasks and their differing levels of depth, complexity, and generalization ability. Each model is rigorously trained, validated, and evaluated using standardized metrics such as **accuracy**, **precision**, **recall**, **F1-score**, and **confusion matrices**.

The outcome of this project is a comprehensive, comparative analysis of these models in the context of breast cancer detection. The proposed system serves as a valuable second-opinion tool that can aid radiologists, particularly in areas with limited access to expert medical personnel or advanced equipment. The project not only provides insights into the performance of various deep learning architectures but also highlights the critical role of preprocessing in medical image analysis.

Moreover, the project contributes to the broader field of **AI in healthcare** by showcasing a practical, scalable solution that could be integrated into real-world diagnostic workflows. By reducing reliance on manual diagnosis, this AI-assisted approach holds the potential to lower operational burdens on healthcare providers and increase diagnostic coverage and accuracy in underserved regions.

RATIONALE

Breast cancer remains a major global health concern, accounting for a significant proportion of cancer-related deaths among women. Despite advances in imaging technology and diagnostic protocols, early and accurate detection continues to pose a challenge—particularly in resource-constrained environments. One of the key issues lies in the variability and subjectivity of radiological interpretations, which can lead to inconsistent diagnoses and, more critically, false negatives. A false negative, where a malignant case is misclassified as benign, may delay timely treatment and substantially reduce the patient's survival prospects.

In light of this, the rationale for this project centres on the urgent need to develop an **automated, accurate, and scalable diagnostic system** to assist radiologists in the early detection of breast cancer. The integration of **deep learning techniques** with **pre-processed mammographic data** provides a powerful approach to enhance diagnostic precision, improve consistency, and minimize human error. Traditional computer-aided diagnosis (CAD) systems often depend on handcrafted features such as texture descriptors, edge detection, or statistical region analysis. While these systems provide some level of automation, they are limited in their adaptability and performance, particularly across diverse datasets. These limitations arise because hand-engineered features may not capture the complex spatial and textural patterns that indicate malignancy in breast tissue. Moreover, manual feature design often requires extensive domain knowledge and may not generalize well to unseen data. **Deep learning**, and in particular **convolutional neural networks (CNNs)**, has transformed the field of image analysis by enabling systems to learn relevant features directly from raw pixel data. Unlike traditional methods, CNNs automatically extract multi-scale, hierarchical patterns, which are especially useful for detecting subtle abnormalities in high-resolution medical images like mammograms. When trained on sufficient data, CNNs have shown performance comparable to, and in some cases exceeding, that of human experts in medical diagnostics.

However, applying CNNs effectively to mammogram images requires more than just feeding data into a model. Mammographic images often contain irrelevant components such as **pectoral muscles**, **scanner labels**, and **noise artifacts** that can mislead deep learning models. These elements introduce noise and distort the true feature space, potentially reducing classification accuracy. Therefore, **robust preprocessing techniques** are necessary to isolate and enhance diagnostically significant features while suppressing irrelevant information.

The rationale behind this project also includes the comparative evaluation of multiple CNN architectures—**VGG16**, **ResNet50**, **MobileNet**, and a **Custom CNN**—to identify the model best suited for the classification of breast cancer in mammograms. Each of these architectures offers unique advantages and trade-offs in terms of accuracy, computational cost, and generalizability. Conducting such a comparative study enables a data-driven selection of the most balanced model for real-world applications, where both speed and reliability are essential.

Another critical justification for this project is its **potential real-world impact**. In rural and under-resourced healthcare settings, access to experienced radiologists is limited. An AI-based diagnostic tool could serve as a **second-opinion system** or even as a **primary screening tool**, helping to prioritize high-risk patients for further evaluation. Such systems can reduce diagnostic delays, increase throughput, and ultimately save lives through early intervention.

In summary, this project is driven by a combination of clinical necessity and technological opportunity. By leveraging deep learning models in conjunction with sophisticated preprocessing pipelines, the system aims to deliver a reliable, interpretable, and scalable solution for breast cancer

detection. It represents a step forward in the application of artificial intelligence to medical diagnostics and aligns with the broader goals of precision medicine, health equity, and technological innovation in healthcare.

OBJECTIVES

The specific objectives are outlined below:

1. To Enhance Mammogram Quality Through Preprocessing

Mammogram images often suffer from poor contrast, noise, and artifacts like labels or pectoral muscles. These anomalies can interfere with the learning process of AI models. Therefore, the first goal is to apply a preprocessing pipeline that includes orientation correction, pectoral muscle removal using Canny edge detection, contrast enhancement using CLAHE, and morphological operations to remove unwanted elements. These steps ensure the model focuses on the diagnostically relevant regions.

2. To Implement and Compare Multiple Deep Learning Models

This project explores and compares four models: VGG16, ResNet50, MobileNet, and a custom-built CNN. Each model offers unique trade-offs in terms of accuracy, training time, and generalization. By comparing their performance, the aim is to identify the most balanced model suitable for clinical use.

3. To Achieve Accurate Binary Classification

The core function of the system is to distinguish between benign and malignant cases from input mammograms. The objective is to train the models to accurately perform this binary classification, reducing errors, especially false negatives.

4. To Evaluate Model Performance Using Key Metrics

Each model will be evaluated using multiple metrics—accuracy, precision, recall, and F1-score—as well as confusion matrices. This multi-metric approach ensures a comprehensive evaluation that goes beyond simple accuracy.

5. To Lay the Foundation for Real-World Deployment

The final goal is to develop a system that could be easily integrated into clinical workflows as a second-opinion diagnostic tool. This includes ensuring modularity, scalability, and potential for future deployment through a user interface.

FEASIBILITY STUDY

1. Technical Feasibility

The project incorporates several open-source libraries, including:

- **TensorFlow/Keras** for designing and training deep learning models.
- **OpenCV** for executing image preprocessing tasks such as orientation correction, contrast enhancement (CLAHE), and morphological noise removal.
- **NumPy** and **Pandas** for data handling and numerical operations.
- **Matplotlib** and **Seaborn** for visualization and graphical evaluation of results.

In terms of hardware, the training and testing of the models are facilitated through **Google Colab**, which provides access to high-performance GPUs (such as NVIDIA Tesla T4) for free. This eliminates the need for costly, high-end local machines. Training is also possible on personal computers with mid-range specifications (e.g., Intel Core i5/i7 with 8–16 GB RAM), making the project accessible to students, researchers, and institutions with limited infrastructure.

The choice of four CNN architectures—**VGG16**, **ResNet50**, **MobileNet**, and a **Custom CNN**—provides both flexibility and depth to the system. Each model has been successfully implemented and trained within the constraints of the development environment, indicating that the tools and computational resources are sufficient for effective model development.

2. Operational Feasibility

The operational feasibility of the project is affirmed by its modular, scalable, and user-friendly design. The system is intended to be deployed in clinical environments as a decision-support tool for radiologists and medical practitioners. Once integrated into a simple web interface (e.g., using Flask or Streamlit), medical personnel will be able to upload mammogram images and receive classification results in real time.

The use of publicly available mammogram datasets also allows the system to be retrained or fine-tuned with new data to adapt to specific population demographics or regional diagnostic standards. Furthermore, since the system can be hosted on cloud-based platforms, operational deployment does not require significant infrastructure investments.

3. Economic Feasibility

One of the strengths of this project lies in its cost-effectiveness. The use of **open-source software** and **free public datasets** eliminates the need for commercial licenses or data procurement. Development and training are conducted on platforms like Google Colab, which provide free access to powerful computational resources.

The initial investment in system integration and training is minimal compared to the long-term benefits in healthcare cost reduction and diagnostic efficiency.

4. Legal and Ethical Feasibility

The project ensures compliance with all relevant ethical and legal standards. The datasets used (e.g., DDSM, INbreast) are publicly available, anonymized, and approved for academic research. No personally identifiable patient data is used or accessed during the development.

In future real-world deployment, the system would need to comply with regulations such as:

- **HIPAA** (Health Insurance Portability and Accountability Act) for data privacy in the U.S.
- **GDPR** (General Data Protection Regulation) for handling personal health information in Europe.

LITERATURE REVIEW

In their ImageNet classification work, Krizhevsky et al. (2021) presented the idea of deep convolutional neural networks (CNNs) for large-scale image recognition. The CNN architecture has now been adjusted for a number of medical imaging applications. Such applications are, the diagnosis of breast cancer through mammograms, after achieving innovative results in image classification tasks.

The VGGNet design, was given by Simonyan and Zisserman in 2022. It showed that using tiny convolutional filters help increase network depth and performance of image recognition. VGG16 is very effective in medical imaging because it can capture very fine visual details in mammography pictures.

He et al. (2020) presented the ResNet architecture which addressed the problem of vanishing gradients in deep networks. ResNet50 has been widely used to label many complicated medical pictures including mammograms. It is known for its efficiency and effectiveness in learning deep hierarchical features.

Howard et al. (2018) presented MobileNet, is a lightweight CNN architecture which was designed for mobile and resource-constrained situations. MobileNet is suitable for real-time mammography analysis as it uses depthwise separable convolutions. This helps in reducing computing costs while also considering the classification accuracy.

Shen et al. (2018) explored the CNN model which was used to extract important features from images. They also focused on the application of deep learning in classification of mammography images. Their study claimed that deep learning models might detect breast cancer with high accuracy, especially when trained on sizable and varied datasets.

Litjens et al. (2017) conducted a study on different deep learning techniques that can be used in medical picture analysis. This study showed important potential of deep learning to reform a number of diagnostic procedures, such as breast cancer screening. Here CNNs have shown strong performance in identifying cancerous regions in mammography pictures.

Esteva et al. (2022) demonstrated the methods of mammography classification for breast cancer screening. These methods achieved comparable results just like the dermatologist-level results in skin cancer classification using deep neural networks. The importance of CNNs in medical imaging, is highlighted by their work, especially where precise classification is essential for early diagnosis.

Liu et al. (2020) explored and showed the classification of mammograms into benign and malignant groups using deep learning. Their work increased the accuracy and reliability of breast cancer classification. This was achieved by integrating feature extraction methods with pre-trained models.

Wang et al. (2020) examined and developed a deep CNN for autonomous breast cancer detection in mammograms with high accuracy. Their study focused on improving pre-processing techniques to enhance the model performance. Processes like contrast augmentation and pectoral muscle removal included and explored.

Rajpurkar et al. (2019) showed the capability of deep learning to accurately detect pneumonia from chest X-rays. Their CheXNet-based approach has been mutated for other medical imaging tasks, which includes mammograms classification. Similar architectures were also used in mammograms to detect breast cancer.

METHODOLOGY

This project follows a systematic methodology consisting of data collection, preprocessing, model training, evaluation, and performance comparison.

Data Collection

Public mammogram datasets such as DDSM and INbreast were used. These datasets offer high-resolution images labelled by expert radiologists. The data was split into training (80%) and testing (20%) sets using stratified sampling to ensure class balance.

Image Preprocessing

Raw mammograms are often unfit for direct training due to noise and artifacts. The following preprocessing steps were applied:

- **Orientation Correction: Standardizes the image alignment.**
- **Pectoral Muscle Removal: Uses Canny edge detection to isolate and remove high-intensity muscle regions.**
- **CLAHE: Enhances local contrast to improve feature visibility.**
- **Morphological Operations: Clean residual noise and artifacts for a smoother image.**

These steps ensure that models focus only on medically relevant image regions.

Data Augmentation

To improve generalization and prevent overfitting, techniques like rotation, flipping, zooming, and cropping were applied during training.

Model Implementation

Four CNN models were used:

- **VGG16: Deep but prone to overfitting without sufficient regularization.**
- **ResNet50: Uses residual blocks to learn complex features more effectively.**
- **MobileNet: Lightweight and efficient, ideal for edge deployment.**
- **Custom CNN: Designed from scratch to test flexibility and adaptability.**

Each model was trained using binary cross-entropy loss and evaluated using various metrics.

Model Evaluation

Accuracy, precision, recall, and F1-score were used for evaluation, along with confusion matrices to visualize classification performance. These metrics provided detailed insights into model reliability, especially in minimizing false negatives.

Model Comparison

The trained models were compared in terms of performance and computational cost. MobileNet was found to offer the best trade-off between accuracy and generalization, while VGG16 achieved high training accuracy but poor generalization due to overfitting.

FACILITIES REQUIRED FOR THE PROPOSED WORK

Hardware Requirements

- **Local Machine: A laptop or desktop computer with at least an Intel Core i5/i7 processor and a minimum of 8GB RAM.**
- **GPU Support: While training was feasible on CPUs, GPU acceleration significantly reduced training time. Google Colab was used for free access to NVIDIA Tesla T4 GPUs.**
- **Storage: At least 10–20 GB of storage was needed to handle image datasets and training outputs.**

Software Requirements

- **Operating System: Windows 10 or Ubuntu 20.04 (or Google Colab's cloud-based environment).**
- **Programming Language: Python 3.8+, for its strong ecosystem of deep learning libraries.**
- **Libraries Used:**
  - **TensorFlow/Keras: For building and training deep learning models.**
  - **OpenCV: For image processing tasks like CLAHE and morphological operations.**
  - **NumPy and Pandas: For data handling and preprocessing.**
  - **Matplotlib and Seaborn: For visualizations.**
- **Notebook Environment:**
  - **Jupyter Notebook: Used for local development and experimentation.**
  - **Google Colab: Used extensively for training due to GPU availability and scalability.**

Dataset Access and Tools

- **Public datasets like DDSM and INbreast were accessed from academic repositories.**
- **Tools like Google Drive and GitHub were used for data storage, collaboration, and version control.**

Optional Tools for Deployment

For potential deployment:

- **Flask or Streamlit: To create a web interface for image uploads and real-time predictions.**
- **Docker: For containerized deployment of the model across different environments.**

This facility setup ensures that the project is both scalable and reproducible, allowing other researchers or institutions to replicate the study with minimal resource constraints.

EXPECTED OUTCOMES

The implementation of this project, *"Breast Cancer Classification and Detection through Advanced Preprocessing and Deep Learning Techniques,"* is expected to yield a number of significant technical, practical, and academic outcomes. These outcomes not only validate the objectives of the study but also contribute to the broader field of computer-aided medical diagnostics.

1. Development of an Accurate Breast Cancer Classifier

The primary outcome of this project is the creation of an automated system capable of classifying mammogram images into **benign or malignant** categories with a high degree of accuracy. By using a carefully constructed pipeline combining **preprocessing** and **deep learning architectures**, the system is expected to demonstrate performance metrics (accuracy, recall, F1-score) that surpass traditional rule-based or feature-engineered models.

Particularly, the integration of preprocessing techniques like CLAHE, pectoral muscle removal, and morphological cleaning should enhance the quality of inputs, leading to improved feature learning and classification accuracy across all CNN models. It is anticipated that at least one of the models—such as **MobileNet**, known for its generalization capability—will emerge as the most effective for deployment.

2. Comparative Analysis of Deep Learning Models

The project includes training and evaluating multiple CNN architectures—**VGG16**, **ResNet50**, **MobileNet**, and a **Custom CNN**—on the same dataset. A key outcome will be a detailed **comparative analysis** of their:

- Training and testing performance
- Overfitting tendencies
- Precision vs. recall trade-offs
- Model complexity and speed
- Suitability for deployment in different environments (e.g., cloud vs. mobile)

Such a side-by-side evaluation will provide valuable insights for future researchers and healthcare developers on choosing the right model architecture based on real-world constraints.

3. Identification and Reduction of Classification Errors

Another expected result is the **error profiling of each model**, particularly focusing on **false negatives**, which are critical in medical diagnostics. The project aims to demonstrate how preprocessing and data augmentation can reduce such errors. Confusion matrices and heatmaps will provide interpretability into where and why models fail, allowing future improvements.

This level of analysis is particularly important in applications involving **patient risk**, as reducing false negatives improves patient safety and increases the clinical trustworthiness of the system.

4. A Scalable, Modular AI Diagnostic Framework

The system is developed with **scalability and modularity** in mind. This means the following can be easily swapped or upgraded in the future:

- Dataset (with new patient data)
- Preprocessing logic (for other imaging modalities)
- Model (using newer architectures like EfficientNet or Vision Transformers)
- Interface layer (for clinical deployment)

This design ensures that the project can act as a **foundation** for more advanced diagnostic systems in the future.

5. Basis for Real-Time or Clinical Integration

With its relatively low compute requirements (especially using MobileNet), the project's solution can be easily integrated into:

- **Web-based systems** for radiologists to upload and analyze scans.
- **Mobile applications** for on-the-go use in remote diagnostics or fieldwork.
- **Hospital information systems** (HIS) to serve as a second-opinion tool.

Although clinical deployment would require further validation and approvals, the project lays the technical groundwork for such integration.

6. Research and Academic Contribution

From an academic perspective, the project contributes to the fields of:

- Deep learning and computer vision
- Medical image analysis
- Explainable AI (XAI) in healthcare (especially with future integration of tools like Grad-CAM)

It provides a reproducible pipeline and comparative results that other students or researchers can build upon. The findings can also be presented at technical conferences or submitted to journals related to medical AI or health informatics.

7. Empowering Early Detection and Decision Support

Perhaps the most important long-term outcome is the **empowerment of early detection systems** through AI. This tool can support doctors by:

- Highlighting high-risk images for closer inspection.
- Flagging suspicious areas with high confidence.
- Acting as a screening tool in resource-constrained settings.

This ultimately contributes to **reducing breast cancer mortality**, aligning with global public health goals and the mission of technology-driven healthcare solutions.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Communications of the ACM, vol. 60, no. 6, pp. 84-90, Jun. 2017, doi: 10.1145/3065386.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. of the Int. Conf. on Learning Representations (ICLR), 2015.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, 2016, pp. 770–778.

[4] A. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, 2018, pp. 1-8.

[5] L. Shen, S. Wang, and M. Jiang, "Deep learning-based breast cancer classification by extracting mammogram features using convolutional neural networks," J. Comput. Sci. Technol., vol. 33, no. 2, pp. 247-257, Mar. 2018.

[6] G. Litjens et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60-88, Dec. 2017.

[7] D. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, pp. 115-118, Feb. 2017, doi: 10.1038/nature21056.

[8] S. J. Liu, K. Plis, J. Gibbons, and J. Carter, "Mammogram classification using deep learning and feature extraction," in Proc. of the Int. Conf. on Machine Learning (ICML), 2020, pp. 1-5.

[9] J. Wang, Y. Xie, and G. Zhang, "Deep convolutional neural networks for automatic breast cancer detection in mammograms," IEEE Access, vol. 8, pp. 25767-25775, Mar. 2020, doi: 10.1109/ACCESS.2020.2970536.

[10] S. Rajpurkar et al., "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," arXiv preprint arXiv:1711.05225, 2017.