I. Acknowledgments

We would like to express our sincere gratitude to all the mentors, collaborators, and individuals who supported us throughout this project.

First, we thank Dr. Jahnvi Tiwari, Mentor at the Department of Computer Science Engineering, Indian Institute of Information Technology Raichur, for her invaluable guidance and insights throughout the development of this system.

We also extend our deepest appreciation to our industry mentors: - Shruti Jaiswal, Artificial Intelligence Specialist at Bosch Global Software Technologies, for her expertise in AI and machine learning. - Aarthi Sathya Narayanan, Lead Engineer (Medical Imaging and Cloud Specialist) at Bosch Global Software Technologies, for her in-depth knowledge of medical imaging and cloud technologies. - Dr. Sree Niranjanaa Bose, Technical Lead - Healthcare at Bosch Global Software Technologies, for her leadership and insights in healthcare data analytics and clinical research.

Our heartfelt thanks go to Dr. Rashmi Kumari T R, Head of the Department of Pathology at RIMS, for her continued support and expertise in pathology. We also appreciate the contributions of Dr. Indraani K, Assistant Professor at RIMS, and her team of postgraduate doctors and interns, whose input was crucial for understanding the challenges in FNAC analysis.

We are grateful to Dr. Bhaskar, Medical Superintendent at RIMS, for his regulatory support, and to the Dean for providing the necessary approvals. Special thanks to Dr. Ramesh BH, MBBS, for granting permission to use patient data for this research.

We would also like to acknowledge Prof. Dr. Harish Kumar Sardana, Director(IIIT Raichur) and Dr Ramesh Kumar Jallu for their coordination and constant encouragement and support in this project.

Finally, we would like to express our gratitude to our friends and family for their unwavering support and encouragement throughout this journey.

AI-Enhanced Breast Cancer Diagnosis System Using Mammography and Fine Needle Aspiration Cytology

Pratham Jain
Department of Computer Science Engineering
Indian Institute of Information Technology Raichur
Raichur, India
cs21b1021@iiitr.ac.in

Atharva Pandey
Department of Computer Science Engineering
Indian Institute of Information Technology Raichur
Raichur, India
cs21b1039@iiitr.ac.in

Uttakarshika Dubey
Department of Computer Science Engineering
Indian Institute of Information Technology Raichur
Raichur, India
cs20b1023@iiitr.ac.in

Rushikesh Muneshwar
Department of Computer Science Engineering
Indian Institute of Information Technology Raichur
Raichur, India
cs21b1013@iiitr.ac.in

Dr. Jahnvi Tiwari
Department of Computer Science Engineering
Indian Institute of Information Technology Raichur
Raichur, India
jahnvi@iiitr.ac.in

Shruti Jaiswal
Artificial Intelligence Specialist
Bosch Global Software Technologies
Bengaluru, India
Shruti.Jaiswal@in.bosch.com

Aarthi Sathya Narayanan
Lead Engineer (Medical Imaging and Cloud Specialist)
Bosch Global Software Technologies
Bengaluru, India
sathyanarayanan.aarthi@in.bosch.com

Dr. Sree Niranjanaa Bose, PhD
Technical Lead - Healthcare
Bosch Global Software Technologies
Bengaluru, India
SSreeNiranjanaa.Bose@in.bosch.com

Dr. Rashmi Kumari T R
Head of Department, Pathology
Raichur Institute of Medical Sciences
Raichur, India
rashmi.kumari@rims.ac.in

Dr. Indraani K
Assistant Professor, Pathology
Raichur Institute of Medical Sciences
Raichur, India
indraani.k@rims.ac.in

***Abstract*—Breast cancer is one of the most prevalent cancers globally, and remains a leading cause of mortality among women. In the United States, it is the second leading cause of cancer-related deaths in women, with an estimated 232,000 new diagnoses and approximately 40,000 deaths in 2015 [1] alone, resulting in a mortality rate of 17.24. Early and accurate diagnosis plays a pivotal role in improving patient outcomes as regular breast cancer screenings such as mammograms have been shown to significantly reduce mortality rates. Specifically, screening can reduce the risk of death from breast cancer by 60% within 10 years and by 47% within 20 years, especially by catching the disease at an early, more treatable stage [1]. Given these findings, the importance of early detection cannot be overstated, as it substantially increases the chances for effective treatment and improves survival rates; however, traditional diagnostic methods, including mammography, can be limited by dense breast tissue challenges, diagnostic delays, and inter-observer variability. This study proposes an AI-integrated diagnostic system that combines mammogram-based imaging with Fine Needle Aspiration Cytology (FNAC) features to deliver an efficient, non-invasive diagnostic tool. The system aims to streamline diagnostic workflows for radiologists and pathologists by classifying breast lesions with high precision and reducing dependence on time-intensive biopsy procedures, as well as a screening component that provides real-time, preliminary risk assessment through the Gail model. This addition empowers oncologists with predictive risk levels, enabling more personalized patient care and optimized screening schedules.**

***Index Terms*—Breast cancer, early diagnosis, AI-integrated diagnostic system, mammography, Fine Needle Aspiration Cytology (FNAC), predictive risk assessment, screening, machine learning, oncology, transfer learning.**

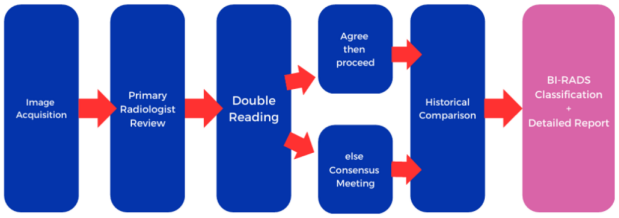

Fig. 1: The flowchart above illustrates the double reading workflow used by radiologists worldwide to diagnose breast cancer. While this method is widely employed, it has three main limitations. First, it is **time-consuming**; the process involves scheduling, radiologist review, and follow-up appointments, which can cause delays in diagnosis. Second, it is **resource-intensive**, requiring specialized equipment and trained professionals, making it less accessible in resource-limited settings. Lastly, the workflow is prone to **human error**, as radiologists may overlook signs of cancer, resulting in false negatives or positives.

II. Introduction

Traditionally, breast cancer diagnosis relies on a combination of mammographic imaging and invasive biopsy procedures to confirm malignancy. The American College of Radiology BI-RADS (Breast Imaging Reporting and Data System (BI-RADS) scale, which assigns ratings from 0 to 6 based on the likelihood of malignancy, is widely used but is dependent on radiologist interpretation. This interpretation can be time intensive and may result in high inter-observer variability, especially in dense breast tissue, where traditional mammograms may yield unclear results. Additionally, although accurate, biopsy procedures introduce risks, costs, and delays in diagnosis

Fine Needle Aspiration Cytology (FNAC) is a widely used, minimally invasive diagnostic procedure that involves using a thin, hollow needle to obtain small tissue samples from palpable or radiologically identified masses for cytological examination. FNAC is commonly employed to evaluate breast, thyroid, lymph node, and other soft tissue lesions because of its simplicity, cost-effectiveness, and rapid turnaround time. This procedure allows clinicians to distinguish between benign and malignant conditions and guide subsequent management decisions. Despite these advantages, FNAC has certain limitations. Although it is highly accurate in diagnosing well-defined lesions, it is not always conclusive, especially when the tissue sample is inadequate or when the lesion has complex features. Additionally, FNAC may be less reliable for evaluating lesions in deeper or inaccessible locations, and in some cases, it may fail to differentiate between certain benign and malignant conditions, potentially leading to false-negative or false-positive results. Consequently, although FNAC is a valuable diagnostic tool, it is often used in conjunction with other imaging techniques or histopathological examinations to confirm the diagnosis[4].

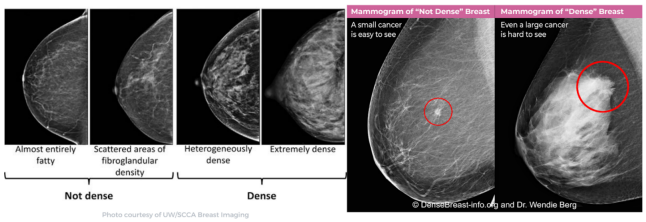

Fig. 2: highlights the four classes of breast density. Dense breast tissue contains a higher proportion of glandular and fibrous tissue, which appears similar in density to tumours on mammogram images. This similarity can obscure small tumours, potentially leading to false-negative results where cancerous growths are missed. To address this issue, biopsy or contrast-enhanced mammography can be used to more accurately detect and differentiate tumours in dense breast tissue

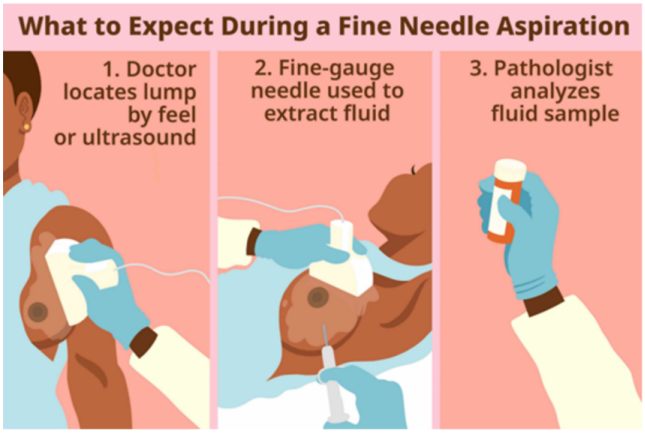

Fig. 3: Illustration of the Fine Needle Aspiration Cytology (FNAC) procedure: a minimally invasive technique used to collect tissue samples from a breast lump for diagnostic analysis

A. Problem Statement

Breast cancer remains a leading cause of mortality, and early detection through mammography and Fine Needle Aspiration Cytology (FNAC) is crucial for improving patient outcomes. However, traditional diagnostic methods face significant challenges, including time-consuming workflows, high inter-observer variability, limited accuracy in dense breast tissue, and the risk of false-negative or false-positive results. These issues lead to delays, diagnostic errors, and reliance on invasive biopsy procedures. This study proposes an AI-integrated diagnostic system that combines mammogram imaging and FNAC data to automate lesion classification, streamline workflows, and provide real-time risk assessments using the Gail model, ultimately improving diagnostic accuracy, reducing delays, and supporting personalized patient care.

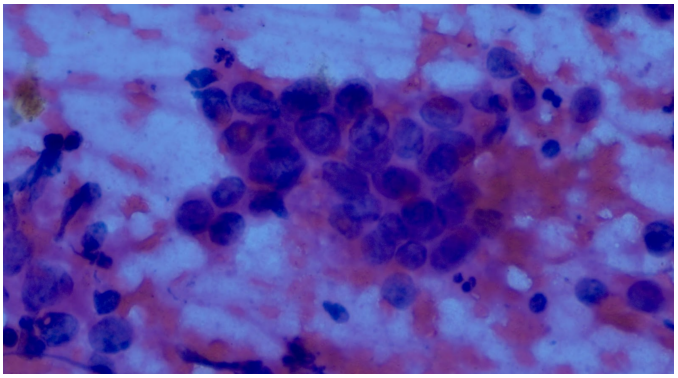

Fig. 4: Microscopic view of a Fine Needle Aspiration Cytology (FNAC) sample at 40x objective and 6.3x ocular lenses, showing cellular details for diagnostic analysis containing **ductal epithelial cells** (navy blue large thumbprints like), **epidermal cells** (small dark blue), **erythrocytes** (red blood cells), **leukocytes** (white blood cells), occasionally **platelets** and **artifacts** (dust or other foreign particles) that came during slide preparation (air drying) or collection

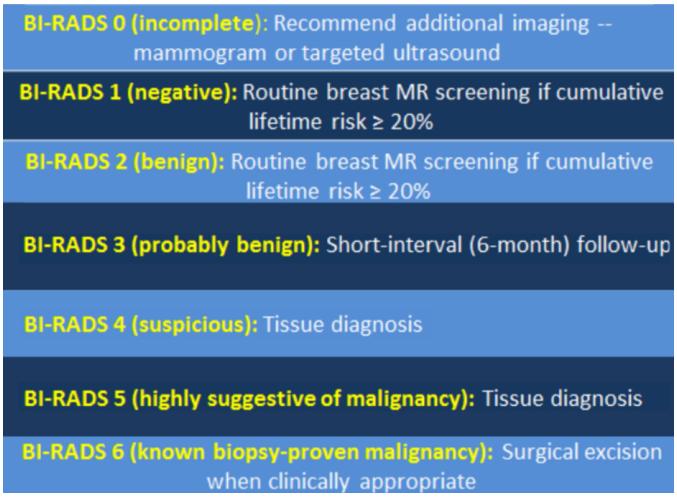

Fig. 5: Understanding BI-RADS: A standardized system for categorizing breast imaging findings. The BI-RADS scale ranges from 0 (incomplete) to 6 (known biopsy-proven malignancy), helping clinicians assess the likelihood of breast cancer and determine the need for further action

B. Objectives

The system was designed to achieve four core objectives aimed at improving diagnostic efficiency and supporting healthcare professionals. First, it reduces diagnostic time by providing preliminary malignant or benign classifications based on Fine Needle Aspiration Cytology (FNAC) data. Second, it offers radiologists pre-screened mammograms with BI-RADS ratings, utilizing transfer learning to enhance diagnostic accuracy and workflow. Third, the system supports oncologists by providing real-time risk assessments using the Gail model, alongside patient-specific screening recommendations and follow-up care enhancements. Lastly, it ensures an accessible, user-friendly platform for radiologists and pathologists, while maintaining ethical standards in the handling of patient data.

C. Key Concepts and Terminology

1) BI-RADS Scores: The Breast Imaging-Reporting and Data System (BI-RADS) is a standardized system developed by the American College of Radiology (ACR) to interpret and report breast imaging findings, primarily for mammography. The BI-RADS scores categorize breast tissue and lesions based on their likelihood of malignancy, which aids in clinical decision-making and follow-up recommendations.

BI-RADS Score Categories:

- **BI-RADS 0: Incomplete** — Further imaging evaluation or comparison with prior mammograms is needed.
- **BI-RADS 1: Negative** — No abnormalities detected; routine screening is recommended.
- **BI-RADS 2: Benign** — Non-cancerous findings (e.g., cysts, fibroadenomas); routine screening recommended.
- **BI-RADS 3: Probably Benign** — A finding that is likely benign, but short-term follow-up (usually in 6 months) is recommended.
- **BI-RADS 4: Suspicious** — A finding that warrants a biopsy to evaluate for possible malignancy. This category is subdivided into:
  - 4A: Low suspicion for malignancy
  - 4B: Moderate suspicion for malignancy
  - 4C: High suspicion for malignancy
- **BI-RADS 5: Highly Suspicious of Malignancy** — Findings strongly suggestive of cancer, biopsy is strongly recommended.
- **BI-RADS 6: Known Biopsy-Proven Malignancy** — This category applies when the lesion has already been confirmed as malignant through biopsy.

These BI-RADS scores play a critical role in assisting radiologists, clinicians, and patients in making informed decisions regarding further diagnostic steps and treatment options.

2) Views in Mammography: There are two standard views in mammography that provide different perspectives of the breast to ensure comprehensive visualization:

Craniocaudal (CC) View: In this view, the breast is compressed from top to bottom, with the X-ray beam directed from above (cranial) to below (caudal). The CC view captures the full thickness of the breast tissue, including the central area and the tissue near the chest wall. The patient typically stands with the breast placed flat against the imaging plate, and the X-ray is taken from directly above [8].

Mediolateral Oblique (MLO) View: This view is obtained by positioning the breast at an angle, typically around 45 degrees. The X-ray beam is directed from the side (mediolateral), passing through the breast tissue at an oblique angle. The MLO view is particularly useful for visualizing the upper and outer portions of the breast, where most cancers tend to develop. The patient is positioned so that their body is angled slightly, and the breast is compressed at a diagonal [8].

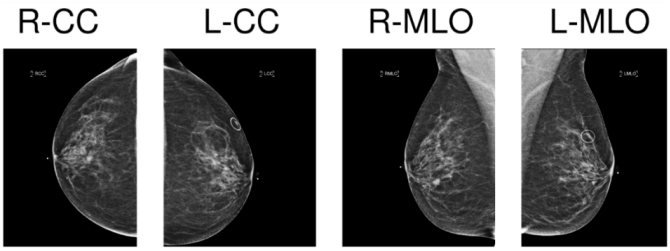

Fig. 6: the CC and MLO views provide complementary information and help ensure that the entire breast tissue, including the areas closest to the chest wall and underarm, is properly imaged for accurate assessment. These views are commonly supplemented by additional angles, such as the Lateromedial (LM) or XCCL (exaggerated craniocaudal) views, to capture a more detailed image depending on the patient's breast density and clinical indications [8].

3) Microcalcifications: Microcalcifications are small deposits of calcium that appear in mammograms as bright white specks against the soft tissue background of the breast. These microcalcifications can be an important indicator of *Ductal Carcinoma in Situ (DCIS)*, a type of early-stage breast cancer [9].

Microcalcifications can be as small as 0.5 mm, often covering only a few pixels in a high-resolution mammogram (with a 10 MPixel image and a pixel resolution of approximately 100 µm). Due to their small size and subtle appearance, accurately segmenting and identifying microcalcifications in mammograms is a critical step in early breast cancer detection. However, this presents a challenge as **high-resolution images** are necessary, while also minimizing the computational burden and keeping the false positive rate very low. For example, a mammogram at a typical 10 MPixel resolution, even with a **false positive rate of 0.01% (one false alarm in 10,000), would still result in 1,000 false positives per scan [9].**

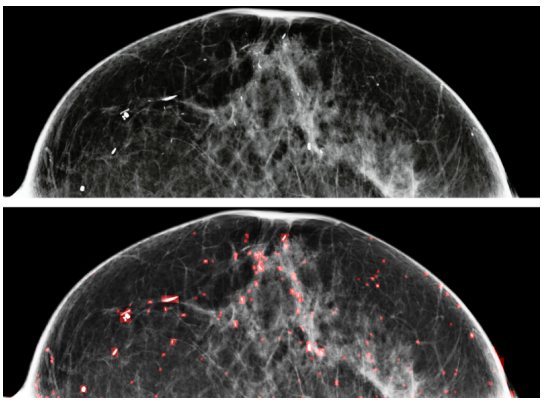

Fig. 7: A typical mammogram can contain up to 20,000 microcalcifications, making the task of identifying and interpreting these small calcium deposits highly time-consuming for radiologists.

Microcalcifications are typically benign; however, when they appear clustered or show certain patterns, they can indicate the development of early-stage breast cancer. The detection of these small calcium deposits is crucial because they may represent precancerous changes or the very early stages of malignancy that have not yet formed a palpable mass. Early detection of microcalcifications allows for timely intervention, potentially reducing the need for more invasive treatments and improving overall prognosis. Studies have shown that the detection of microcalcifications plays a key role in reducing breast cancer mortality through early screening and diagnosis [10].

In contrast, macrocalcifications are larger, more easily detectable deposits, which typically indicate benign findings and are less concerning [9].

Fine Needle Aspiration Cytology: Fine Needle Aspiration Cytology (FNAC) is a widely used diagnostic technique that involves aspirating cellular material from a lesion or lump using a fine, hollow needle. This material is then smeared onto glass slides for microscopic analysis. FNAC is particularly valued for being minimally invasive, cost-effective, and quick, often yielding results within hours. It is frequently employed for the evaluation of palpable and non-palpable lesions, including those in the breast, thyroid, lymph nodes, and salivary glands [26], [27].

FNAC and Its Cousins: Biopsy and Histopathology: FNAC provides cytological insights by analyzing individual cells, whereas **biopsy** and **histopathology** examine tissue architecture in detail. Biopsies involve extracting larger tissue samples, which are subsequently processed for histopathological evaluation. This analysis enables the identification of critical features such as tissue relationships, cell margins, vascular invasion, and differentiation between in situ and invasive cancers [28], [29].

While biopsy and histopathology are more definitive, they are invasive, time-consuming, and require more resources. FNAC, on the other hand, is quicker, non-invasive, and less discomforting for patients. However, FNAC's inability to provide architectural details makes it inconclusive in some cases. Integrating **artificial intelligence (AI)** with FNAC could enhance its diagnostic potential by automating cellular feature extraction (e.g., radius mean, perimeter mean, texture mean), improving accuracy, and reducing the need for invasive techniques [30], [31].

Slide Preparation Process: The diagnostic accuracy of FNAC is highly dependent on the preparation and staining of slides. After cellular material is aspirated, it is processed using one of two primary methods:

1) **Air-Dried Slides**: Air-dried smears are typically stained with Giemsa or Diff-Quik stains, which are effective for highlighting cytoplasmic features. These slides are commonly used for inflammatory or infectious lesions due to the detailed visualization of cytoplasmic granularity [32].
2) **Alcohol-Fixed Slides**: In this method, the cellular material is immediately immersed in an alcohol-based

fixative or sprayed with a fixative solution. Staining with Papanicolaou (Pap) stain is preferred for these slides, as it enhances nuclear morphology and chromatin detail, which are critical for identifying malignancy [33].

The prepared smear should be thin and evenly distributed to prevent cell clumping and ensure clear observation under the microscope.

Necrosis and Its Exclusion from Analysis: Necrosis refers to the death of cells within a sample and often appears as amorphous debris or granular material in FNAC slides. It is a common finding in high-grade malignancies due to hypoxia in rapidly growing cells. However, necrosis can obscure diagnostically relevant cellular details and produce artifacts that complicate automated feature extraction. To ensure the accuracy of AI-based analyses, necrotic areas should be excluded during feature extraction [34], [35].

Slide Preservation: Proper storage of FNAC slides is essential for maintaining their integrity over time. Slides should be stored in a cool, dry environment away from direct sunlight and humidity. Common preservation challenges include:

- **Stain Fading**: Exposure to light and air can degrade stains like Giemsa or Pap, reducing cellular contrast [36].
- **Mounting Media Degradation**: Cracks in the mounting medium can introduce physical artifacts.
- **Contamination**: Dust or fungal growth on improperly stored slides can obscure features or mimic pathological changes [37], [38].

Degradation of Slides Over Time: Even under optimal conditions, slides may degrade over time. This can lead to:

- **Loss of Cellular Detail**: Morphological changes due to faded stains or physical damage.
- **Artifact Formation**: Degraded slides might falsely resemble necrotic areas or other pathological changes.

AI has the potential to identify and correct these artifacts, ensuring reliable analysis of older slides [39]. However, for simplicity in training the system, we exclude slides with very high levels of necrosis or degradation. These slides represent a very small percentage of all samples, and including them would increase the system's complexity. Additionally, pathologists can often identify malignancy based on early signs as soon as they observe the slides, making these heavily degraded samples less critical for AI-based analysis

D. Related Work

1) Summary of Related Work on AI in Breast Cancer Detection: Previous research has emphasized the effectiveness of transfer learning in medical imaging tasks owing to the limited annotated datasets.

2) Related Work on Nuclear Feature Extraction and Breast Cancer Diagnosis: The Wisconsin Diagnostic Dataset is essential for this system because it provides robust nuclear feature data extracted from digitized FNAC images. FNAC enables swift, minimally invasive sampling of cells and offers an alternative to tissue biopsy [40], [41]. With the support of AI, this study hypothesizes that the system can accurately classify cell samples as malignant or benign based on extracted nuclear features, thereby expediting the diagnosis.

3) Comparison of Breast Cancer Risk Assessment Models: In this subsection, we present a detailed comparison of the **COIMBRA**, **BCSC**, and **GAIL** models for breast cancer risk assessment. This comparison is based on specified criteria, as shown in Table III.

The GAIL Model stands out as an ideal choice for the initial implementation of an open-source breast cancer risk assessment system for doctors in India due to its practical and theoretical advantages. It provides a basic level of personalization by incorporating demographic and reproductive factors such as age, family history, and age at menarche or first childbirth. While it lacks the advanced capabilities of models like COIMBRA, its simplicity allows for effective risk stratification without requiring extensive patient data, making it suitable for resource-limited settings.

Although the GAIL Model does not explicitly account for genetic mutations like *BRCA1/2*, it indirectly considers genetic predisposition through family history, offering a feasible solution in scenarios where comprehensive genetic testing is inaccessible or unaffordable. Similarly, while the model excludes detailed lifestyle factors such as diet or physical activity, this limitation simplifies implementation, particularly in healthcare settings where collecting and standardizing such data is challenging.

The model's greatest strength lies in its simplicity and ease of use. It relies on readily available clinical and demographic data, and calculations can be performed using basic tools like online calculators or lightweight algorithms. This makes it particularly practical for scaling within an open-source framework, especially for doctors with varying levels of access to advanced infrastructure. Moreover, it integrates seamlessly into standard clinical workflows, requiring minimal changes to routine practices. Its global adoption further facilitates the adaptation of training materials and tools for Indian healthcare professionals.

For India specifically, the GAIL Model aligns well with the available healthcare infrastructure. It uses input parameters like family and reproductive history that are typically accessible in Indian clinical settings, including rural areas where mammograms and advanced diagnostic tools are scarce. Its cost-effectiveness, stemming from the absence of reliance on expensive genetic tests or imaging data, further enhances its suitability. While the model may not fully capture the unique genetic and lifestyle factors of Indian women, it offers a solid foundation that can be refined using localized data over time.

In conclusion, the GAIL Model is a practical and effective starting point for developing an open-source breast cancer risk assessment system in India. Its simplicity, ease of integration, and reliance on widely available data enable rapid adoption in diverse healthcare settings. Although it has limitations in addressing specific population needs, these can be mitigated through iterative improvements. Once a robust system is established, more advanced models like COIMBRA can be explored for enhanced accuracy.e GAIL Model stands out as

an ideal choice for the initial implementation of an open-source breast cancer risk assessment system for doctors in India due to its practical and theoretical advantages. It provides a basic level of personalization by incorporating demographic and reproductive factors such as age, family history, and age at menarche or first childbirth. While it lacks the advanced capabilities of models like COIMBRA, its simplicity allows for effective risk stratification without requiring extensive patient data, making it suitable for resource-limited settings.

Although the GAIL Model does not explicitly account for genetic mutations like *BRCA1/2*, it indirectly considers genetic predisposition through family history, offering a feasible solution in scenarios where comprehensive genetic testing is inaccessible or unaffordable. Similarly, while the model excludes detailed lifestyle factors such as diet or physical activity, this limitation simplifies implementation, particularly in healthcare settings where collecting and standardizing such data is challenging.

The model's greatest strength lies in its simplicity and ease of use. It relies on readily available clinical and demographic data, and calculations can be performed using basic tools like online calculators or lightweight algorithms. This makes it particularly practical for scaling within an open-source framework, especially for doctors with varying levels of access to advanced infrastructure. Moreover, it integrates seamlessly into standard clinical workflows, requiring minimal changes to routine practices. Its global adoption further facilitates the adaptation of training materials and tools for Indian healthcare professionals.

For India specifically, the GAIL Model aligns well with the available healthcare infrastructure. It uses input parameters like family and reproductive history that are typically accessible in Indian clinical settings, including rural areas where mammograms and advanced diagnostic tools are scarce. Its cost-effectiveness, stemming from the absence of reliance on expensive genetic tests or imaging data, further enhances its suitability. While the model may not fully capture the unique genetic and lifestyle factors of Indian women, it offers a solid foundation that can be refined using localized data over time.

In conclusion, the GAIL Model is a practical and effective starting point for developing an open-source breast cancer risk assessment system in India. Its simplicity, ease of integration, and reliance on widely available data enable rapid adoption in diverse healthcare settings. Although it has limitations in addressing specific population needs, these can be mitigated through iterative improvements. Once a robust system is established, more advanced models like COIMBRA can be explored for enhanced accuracy.

III. System Architecture

The system architecture integrates multiple components designed to streamline breast cancer diagnosis and enhance diagnostic accuracy. The three main modules are outlined below:

A. Preprocessing and Screening Module

This module processes the input image and FNAC data while also calculating a risk score using the Gail model. The Gail model incorporates patient-specific factors to estimate the risk of developing breast cancer and serves as an early screening tool. By identifying high-risk patients, this module assists oncologists in determining who may require more frequent monitoring and advanced diagnostic assessments.

B. Transfer-Learning-Based Mammogram Analysis

This component utilizes a pre-trained VGG-16 model to analyze high-resolution mammogram images acquired in four standard views (craniocaudal and mediolateral oblique for both breasts). Transfer learning enables the system to leverage pre-trained image representations, allowing the classification models to produce BI-RADS ratings with minimal retraining.

In addition, this approach is complemented by our implementation of an end-to-end convolutional network for suspicious versus non-suspicious classifications. The underlying idea is that an end-to-end convolutional model, capable of analyzing all four views simultaneously and trained on over a million mammogram images, excels at identifying BI-RADS categories 0 and 1. This model outperforms traditional transfer learning methods for early screening. When the convolutional network flags a mammogram for potential malignancy, our transfer-learning-based model can then be fine-tuned to assign specific BI-RADS ratings (e.g., 4 and 5), aligning with radiologists' requirements for more detailed classification.

This is a proof-of-concept implementation; however, the incorporation of modern architectures, such as federated learning, could significantly improve the accuracy of our fine-tuned model over time. By continually training on diverse datasets while preserving data privacy, federated learning prevents overfitting and ensures the model remains relevant as new data becomes available.

This approach reduces the time radiologists spend manually analyzing mammograms while ensuring higher consistency and accuracy in BI-RADS assignment.

C. FNAC Data-Based Classification with Human-in-the-Loop Design

This architecture was developed based on feedback from medical professionals and collaborators at RIMS, who emphasized the challenges associated with Fine Needle Aspiration Cytology (FNAC). FNAC samples are often considered inconclusive, serving primarily as suggestive indicators for biopsy. This can lead to delays in diagnosis and treatment, creating additional burdens for patients. The issue becomes even more pronounced when cellularity in FNAC slides is low, making it difficult for pathologists to conclusively determine malignancy. Such scenarios increase resource utilization and prolong the diagnostic process, highlighting the need for a more efficient and reliable system.

To address these challenges, our system adopts a human-in-the-loop approach specifically tailored for FNAC analysis, diverging from the end-to-end design typically employed in

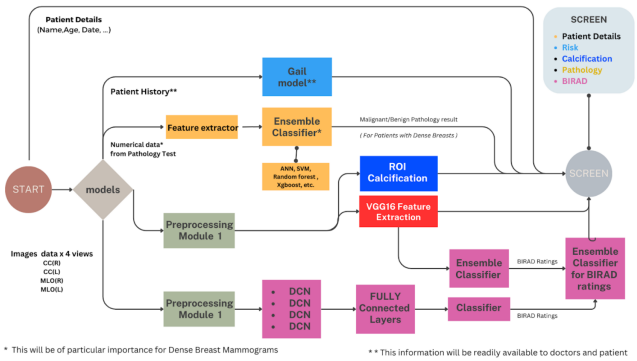

Fig. 8: This architecture was developed in collaboration with a multidisciplinary team, including five senior doctors from RIMS and Clarity Advanced Imaging Center specialising in radiology (two for mammography and three for FNAC analysis) and two mentors with expertise in healthcare data analytics and AI/data engineering. The team worked in a weekly agile-based environment to create a robust, human-in-the-loop system that integrates FNAC feature extraction and mammogram analysis, enhancing diagnostic accuracy and treatment decision-making.

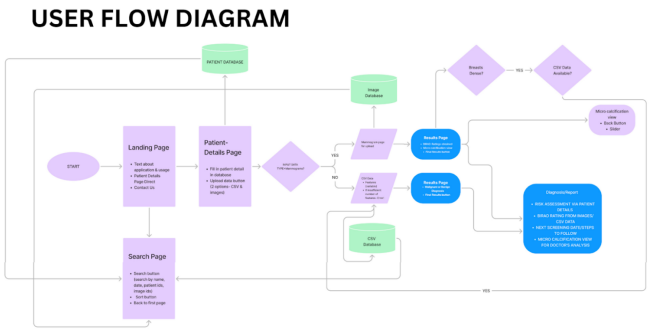

Fig. 9: This diagram illustrates the user flow for the system, starting from the landing page where users can navigate to patient details and upload relevant data (CSV or images). Depending on the input type, the system processes the data, either through mammogram analysis or CSV data for feature extraction. The results are displayed on the results page with BI-RADS ratings and diagnostic outcomes, including micro-calcification views for detailed analysis. The flow includes functionality for patient search, data sorting, and generating detailed reports, providing an organized, intuitive interface for both pathologists and radiologists.

radiology workflows. In FNAC cases, pathologists focus on understanding how specific cellular features contribute to malignancy. By concentrating on the nuclear features of individual cells, our system assists pathologists in making more informed decisions, potentially enabling direct progression to treatment and bypassing the need for invasive biopsies.

The system extracts critical nuclear features, such as texture, smoothness, compactness, and symmetry, through advanced feature selection techniques, including correlation-based selection and Random Forest importance. These features are then processed using ensemble classifiers such as Support Vector Machine (SVM), Random Forest, and XGBoost to achieve high predictive accuracy for malignancy. The collaborative design ensures the system supports, rather than replaces, the pathologist, promoting enhanced diagnostic accuracy and reducing reliance on invasive diagnostic methods.

This approach not only accelerates treatment decisions but also addresses the clinical need for interpretability and trust, both of which are crucial when dealing with inconclusive FNAC slides. By maintaining a human-in-the-loop workflow, the system aligns with the practical requirements of medical professionals while delivering actionable insights for improved patient care.

IV. Methodology

The system is deployed in two main stages where other models discussed ab:

A. Mammography-based BI-RADS Rating

The proposed methodology combines multiple machine learning approaches to improve the accuracy and efficiency of breast cancer screening through mammogram analysis. This section outlines the key components of the system, including the use of pre-trained models, end-to-end convolutional networks, federated learning, and advanced techniques for model explainability.

The core of the system utilizes a pre-trained VGG-16 model, which is applied to analyze high-resolution mammogram images from four standard views: craniocaudal (CC) and mediolateral oblique (MLO) for both breasts. VGG-16, a deep convolutional neural network (CNN), has been trained on a large dataset of natural images, enabling it to transfer learned features to mammogram images. This transfer learning approach allows the system to leverage these pre-trained features, enabling BI-RADS classification with minimal retraining. This not only reduces the computational cost and time compared to training a model from scratch but also ensures high accuracy in classifying mammograms.

In addition to the VGG-16 model, an end-to-end convolutional network is employed for classifying mammograms as suspicious or non-suspicious. This network is designed to analyze all four mammogram views simultaneously, learning features from a large dataset of over a million images. It excels at identifying BI-RADS categories 0 and 1, which are typically associated with benign or inconclusive findings. When compared to traditional transfer learning methods, this model proves particularly effective for early detection, quickly identifying regions that may require further investigation.

To address the issue of class imbalance—especially the low prevalence of malignancy in the dataset—the Synthetic Minority Over-sampling Technique (SMOTE) is used. SMOTE generates synthetic samples for the underrepresented class, improving the model's sensitivity to malignancies without

increasing the false positive rate. This technique ensures the model learns to identify rare yet critical features, such as ductal carcinoma in situ (DCIS), that are indicative of malignancy.

Furthermore, data augmentation is applied to enhance the diversity of the training set. After passing the original INBREAST dataset through the pre-trained VGG-16 model (with the final layers frozen), noise is introduced to the images, and minor scaling is applied to simulate variations in image quality and to capture subtle changes in mammogram features. This approach enhances the model's ability to generalize, allowing it to recognize malignant features under varying conditions. Combined with SMOTE, data augmentation ensures robust model performance across diverse real-world scenarios.

B. FNAC-Based Malignancy Classification

This section outlines a comprehensive pipeline for cell diagnosis that integrates human expertise with automated machine learning predictions for malignancy detection. The pipeline begins with image acquisition and region of interest (ROI) selection by the pathologist. Subsequently, image preprocessing steps such as edge detection, boundary drawing, and refinement are performed. Once cell boundaries are identified, relevant features are extracted and input into an ensemble classifier to predict malignancy. Finally, the results, including the processed image and predicted malignancy, are saved and exported for further use.

The first step in the pipeline is image acquisition, where a diagnostic image of cell cells is loaded using the functions in Matplotlib. A pathologist visually inspects the image to identify the relevant regions that need further analysis. Using the `CellBoundaryDetector` graphical user interface (GUI), the pathologist selects a cell boundary by manually placing points on the image to create a polygon, or "snake," that traces the cell's boundaries using the canvas library with matplotlib and creating events for mouse clicks for starting and ending points along the cell and joining them with a straight line for a closed figure every time.this process help us to isolate the selected region, making it ready for further analysis.

Once the boundary is drawn, the user can refine it using an active contour model. which adjusts the contours based on the detected edges to improve boundary accuracy. The pathologist can view the refined boundaries overlaid on the original image using providing a clearer and more accurate representation of the cell.

Once the cell boundaries are defined and refined, the system proceeds to feature extraction. Relevant features, such as cell area, shape characteristics, and edge smoothness, are extracted from the manually drawn cell contours as described in [11]. The ranges, means, and standard errors of these features can be found in Table [**?**] for the values reported in [11].

The **area** of the boundary is computed using the function `cv2.contourArea(contour)`, which calculates the number of pixels enclosed by the boundary. The **perimeter**, or the length of the boundary, is calculated using `cv2.arcLength(contour, closed=True)`, which computes the length of the curve, treating it as a closed loop.

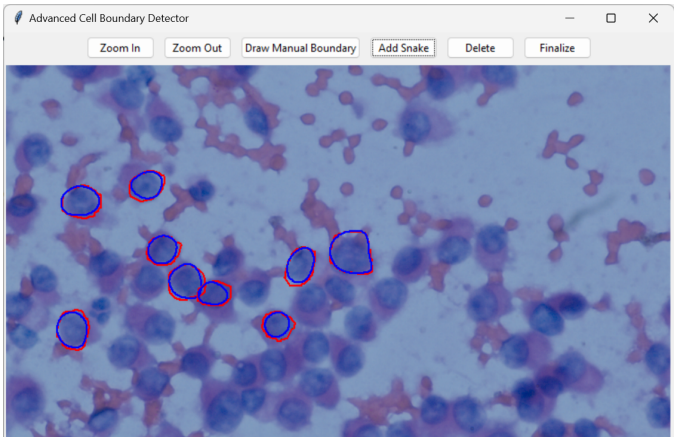

Fig. 10: The graphical interface enables the user to select specific cells of interest, which are often missed by binary masks and segmentation-based approaches. By keeping the human in the loop, we refine the selection using an active contour model. This model adjusts the contours based on detected edges, improving the accuracy of the boundaries

The **radius** is derived by assuming the shape is a perfect circle, and it is calculated from the area using the formula

$$\text{radius} = \sqrt{\frac{\text{area}}{\pi}}.$$

The **smoothness** of the boundary is calculated by first approximating the centre of the contour as the mean of the contour points. Then, the Euclidean distances from each point on the contour to this centre are computed. The smoothness is determined as the mean of the absolute deviations of these distances from the average radius, normalized by the mean radius, using the formula

$$\text{smoothness} = \frac{1}{r_{\text{mean}}} \sum \left| \text{distances} - r_{\text{mean}} \right|.$$

In addition, each feature is adjusted for magnification based on a predefined magnification adjustment factor. The adjusted **area**, **perimeter**, **radius**, and **smoothness** are calculated as follows:

$$\textbf{adjusted area} = \text{area} \times (\text{magnification adjustment factor})^2,$$

$$\textbf{adjusted perimeter} = \text{perimeter} \times (\text{magnification adjustment factor}),$$

$$\textbf{adjusted radius} = \text{radius} \times \text{magnification adjustment factor},$$

$$\textbf{adjusted smoothness} = \text{smoothness} \times \text{magnification adjustment factor}.$$

Once individual features are calculated for each boundary, statistical measures are computed. These include the **mean** of each feature across all detected boundaries, the **standard error** (SE), calculated as the standard deviation of the feature values divided by the square root of the number of refined boundaries, and the **worst** value, which is the maximum value for each feature. These statistics include `area_mean`,

area_se, area_worst, perimeter_mean, perimeter_se, perimeter_worst, radius_mean, radius_se, radius_worst, smoothness_mean, smoothness_worst, and snake_refined_count, which represents the number of boundaries refined using the snake model. These calculated statistics are then appended to an Excel file for further analysis.

TABLE I: Summary of Related Work on AI in Breast Cancer Detection

Author(s) & Year	Study Title	Research Objective	Methodology	Key Findings	Strengths
McKinney et al. (2020) [3]	International evaluation of an AI system for breast cancer screening	To evaluate the performance of an AI system for breast cancer screening and compare it to human radiologists in terms of diagnostic accuracy and workload reduction.	AI system (deep learning) for mammogram analysis, evaluated on UK dataset (OPTIMAM). The AI system outperformed human radiologists in AUC-ROC by an absolute margin of 11.5%. Reduced workload of second readers by 88%. False positives: 5.7% (USA), 1.2% (UK); False negatives: 9.4% (USA), 2.7% (UK). Output includes BI-RADS ratings (0, 1, 2).	AI outperformed human radiologists in diagnostic accuracy (AUC-ROC), significantly reduced workload for radiologists, ethical approval for dataset use, high generalizability with international datasets.	AI still has false positives/negatives, reliance on high-quality datasets, dataset-specific results.
Zamira et al. (2020) [9]	Segmenting Microcalcifications in Mammograms and its Applications	To develop a lightweight, high-resolution model for segmenting microcalcifications in mammograms, balancing accuracy with computational efficiency and minimizing false positives.	Fully Convolutional Network (FCN), multi-scale convolutions, hard negative mining, 45×45 pixel receptive field. The model uses a small receptive field (45×45 pixels) to effectively segment microcalcifications, which are localized and do not benefit much from contextual information. The model is lightweight (˜450K parameters) and minimizes false positive rate while maintaining diagnostic performance.	High efficiency with low computational burden, high segmentation accuracy, minimized false positive rate with online hard negative mining.	Model performance might degrade on larger, more complex datasets; generalizability across diverse datasets is unclear.
Shen et al. (2019) [2]	Deep Learning to Improve Breast Cancer Detection on Screening Mammography	To develop an end-to-end deep learning system for mammography analysis and assess its ability to transfer learning from one dataset to another with or without annotations.	VGG-16 pretrained model, transfer learning, patch-based classification, whole-image classification. The VGG-16 model is fine-tuned for mammography analysis. The patch classifier is first trained on the DDSM dataset (digitized film mammograms) and then converted into a whole-image classifier. The network can be transferred to other datasets, fine-tuned with image-level labels, and requires minimal annotations for further training. Performance is primarily reported for binary tasks, such as distinguishing between non-suspicious (BI-RADS 1, 2, 3) and suspicious (BI-RADS 4, 5) findings, rather than classifying specific BI-RADS ratings.	End-to-end trainability, ability to transfer learning across datasets with minimal annotations, use of a large public dataset (DDSM).	Relies on VGG-16, which may have limitations in adapting to new imaging modalities or datasets; performance can be highly dependent on fine-tuning. Additionally, the task of directly classifying BI-RADS categories remains unaddressed, which is significantly more challenging and clinically valuable. Our study addresses this limitation by focusing on BI-RADS classification, as highlighted by feedback from our medical subject matter experts at RIMS, and aims to deliver more granular and actionable diagnostic outcomes.

TABLE II: Related Work on Nuclear Feature Extraction and Breast Cancer Diagnosis

Author(s) & Year	Study Title	Research Objective	Methodology	Key Findings	Strengths
Wolberg et al. (1995) [11]	*Nuclear Feature Extraction for Breast Tumor Diagnosis*	To develop a system that uses nuclear features from fine needle aspirates to classify breast tumors as malignant or benign.	A GUI-based method for extracting nuclear features (e.g., radius, texture, smoothness) from digitized FNAC images. Features were then used to train and test machine learning models for classification.	Achieved high accuracy in distinguishing malignant and benign tumors using extracted nuclear features. Demonstrated the effectiveness of non-invasive FNAC for diagnostic purposes.	Proposed a standardized method for feature extraction and validated it with a robust dataset. Provided early evidence of machine learning's potential in diagnostic workflows.
Kalita, Manjula et al. (2024) [4]	*Transfer Learning for FNAC Image-Based Breast Cancer Classification*	To explore transfer learning on FNAC images for malignant vs. benign classification.	Pre-trained InceptionV3 model fine-tuned on a small FNAC image dataset. Used data augmentation and patch-based classification to handle limited training samples.	Achieved an accuracy of 85% in classifying FNAC images. Demonstrated transfer learning's utility in leveraging large image-based models for FNAC classification.	Effectively adapted a pre-trained model to FNAC data with minimal annotation requirements. Highlighted potential for generalization to other cytology datasets.
Patel et al. (2019) [**?**]	*Feature Extraction from FNAC Images Using ImageJ for Breast Cancer Diagnosis*	To evaluate the use of ImageJ for nuclear feature extraction in FNAC images and its impact on diagnostic accuracy.	Used ImageJ to extract shape and texture features from segmented FNAC images. Features were input to a random forest classifier for malignant vs. benign classification.	Reported an accuracy of 85% with ImageJ-extracted features but noted struggles with overlapping cells and artifacts in FNAC images.	Demonstrated ImageJ's effectiveness in feature extraction for clean, uniform FNAC samples. Established a baseline for non-GUI-based approaches.
Gomez et al. (2020) [**?**]	*Automated Segmentation of Breast Cytology Images Using OpenCV*	To assess OpenCV's utility in segmenting FNAC images and extracting nuclear features for cancer classification.	Developed a segmentation pipeline using OpenCV, relying on adaptive thresholding and morphological operations to extract features. SVM classifier used for final classification.	Achieved moderate accuracy (80%) but noted difficulties in accurately segmenting overlapping cells and non-uniform samples. Highlighted limitations in handling challenging FNAC images.	Validated OpenCV's potential for automating feature extraction in clean datasets. Provided insights into the challenges of applying traditional image processing techniques to cytology images.
Our Study (2024)	*AI-Assisted Breast Cancer Diagnosis Using FNAC and BI-RADS Classification*	To recreate and enhance nuclear feature extraction from FNAC images for malignant/benign classification and develop a system addressing BI-RADS classification challenges.	Recreated GUI-based feature extraction from FNAC images using magnification and contrast normalization. Explored ImageJ and OpenCV but relied on snake-based human-in-the-loop segmentation for improved accuracy. Dropped transfer learning due to dataset size and quality concerns.	Reproduced Wisconsin dataset nuclear feature ranges with high accuracy. Showed limitations of automated segmentation in challenging cases (overlapping ductal cells) and highlighted the importance of human-in-the-loop methods. Proposed a novel approach for BI-RADS classification.	Balanced accuracy and computational efficiency in nuclear feature extraction. Addressed a gap in detailed BI-RADS classification with a focus on clinical relevance and practical implementation.

TABLE III: Comparison of COIMBRA, BCSC, and GAIL models for breast cancer risk assessment.

Model	Author(s) & Year	Study Title	Research Objective	Methodology	Key Findings and Strengths
COIMBRA	Emerging research (No specific author/year yet)	Cumulative Online Model for Breast Cancer Risk Assessment	To provide real-time, dynamic risk assessment.	Utilizes machine learning to integrate genetic, lifestyle, and clinical factors in real-time.	Adaptive, integrates a wide range of data, suitable for precision medicine. No validation studies for Indian patients yet.
BCSC	Tice et al., 2008	Breast Cancer Risk Prediction: Breast Density Factors	To predict 5-year risk of invasive breast cancer.	Uses data from mammographic breast density, clinical history, and demographic factors.	Proven effective for stratifying women into high-risk groups. Validated internationally but not specifically in Indian patients. Incorporates mammography for improved accuracy.
GAIL	Gail et al., 1989	Projecting Individualized Probabilities of Breast Cancer	To estimate lifetime and 5-year invasive breast cancer risk.	Logistic regression based on demographic and reproductive factors.	Widely used globally. Simple, user-friendly tool validated in various populations but limited validation in Indian contexts.

TABLE IV: Measurement Range for Attributes (Benign vs. Malignant)

Attributes	Mean (Benign)	Mean (Malignant)	SE (Benign)	SE (Malignant)	Max (Benign)	Max (Malignant)
Radius	6.99	28.12	0.121	2.923	7.95	37.01
Texture	9.80	40.02	0.37	4.90	112.10	50.01
Perimeter	44.02	189.09	0.80	22.01	50.48	252.03
Area	144.04	2503.01	6.90	543.10	186.01	4255.00
Smoothness	0.054	0.164	0.003	0.035	0.072	1.102
Compactness	0.020	0.350	0.002	0.138	0.030	1.060
Concavity	0.001	0.501	0.000	0.400	0.000	1.255
Concave points	0.0001	0.202	0.000	0.055	0.000	1.296
Symmetry	0.108	0.305	0.009	0.080	0.158	0.668
Fractal dimension	0.051	0.098	0.001	0.031	0.057	0.210

The extracted features are then input into an ensemble classifier, which predicts the malignancy of the cell. The classifier uses a combination of several models, including Random Forest, Gradient Boosting, and Support Vector Machines, to make a robust prediction. The output is either a probability score or a binary classification (benign/malignant), depending on the cell's characteristics.

Once the malignancy prediction is made, the system allows for result export. The image with the drawn and refined cell boundaries, along with a `JSON` file containing the coordinates of the cell boundaries, is saved, ensuring that both the image and feature data are organized in a folder structure for easy access and allowing the user to reopen and resume the project at a later time using the functionalities.

Although this system serves as a proof-of-concept, the inclusion of modern architectures, such as federated learning, could significantly enhance the model's performance over time. Federated learning allows multiple institutions to collaborate in training the model on diverse datasets without the need to transfer sensitive patient data, thus ensuring privacy. As the model is continually trained with new datasets, it will adapt to different clinical practices and patient demographics, preventing overfitting while maintaining accuracy. This approach ensures that the model remains up-to-date as new data becomes available and allows for continuous improvement with minimal privacy concerns.

C. Explainability and Interpretability in Machine Learning Models

Explainability and interpretability of machine learning models are critical in the medical domain, where it is essential to build trust and ensure that the model's decision-making process is well-understood. In this system, we utilise Generalized Linear Models (GLM) to improve model transparency. GLMs offer a clear and interpretable approach to explaining the relationship between input features and the predicted categories, such as Malignant and Benign. This transparency allows clinicians to better understand the rationale behind the model's predictions, promoting its adoption and seamless integration into clinical workflows.

Furthermore, Grad-CAM (Gradient-weighted Class Activation Mapping) is employed to provide visual explanations of the model's decision-making process [25]. Grad-CAM highlights the regions in the mammogram image that contributed most to the classification, offering valuable insights into which areas the model deemed significant for identifying malignancy. By combining GLM and Grad-CAM, we ensure that the system's predictions are both explainable and interpretable, which is vital for clinicians who rely on these predictions for making informed decisions.

D. Ethical Considerations and Patient Data Handling

The AI system adheres to ethical guidelines by implementing stringent data privacy standards and is designed to handle patient data securely and transparently. The classification output, whether it is BI-RADS ratings or FNAC-based results,

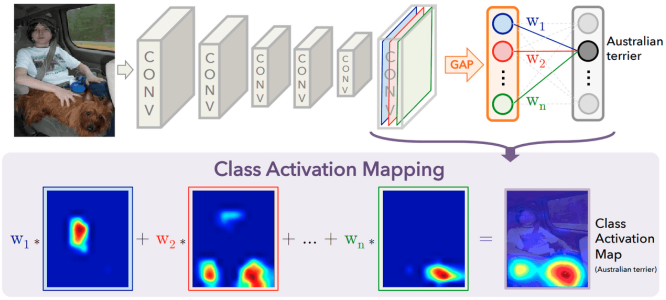

Fig. 11: Class Activation Maps (CAMs) are used to visualize the regions of an input image that contribute most to a model's decision. Grad-CAM (Gradient-weighted Class Activation Mapping) improves upon CAMs by using the gradients of the output with respect to the final convolutional layer to generate more accurate and detailed visualizations. While CAMs require a specific architecture (global pooling layers based), Grad-CAM can be applied to any convolutional neural network, offering greater flexibility.

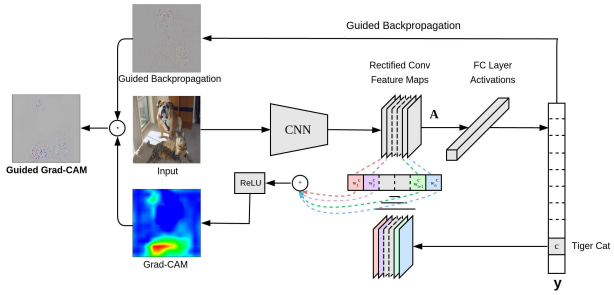

Fig. 12: Grad-CAM enhances traditional Class Activation Maps by utilizing gradients to generate more accurate visual explanations of model predictions. It provides detailed localization of important features in an image, making it more versatile and applicable to a wider range of models.

is readily accessible to medical professionals while ensuring minimal bias in model predictions. Throughout the project, we ensured that minimal patient information was collected, and all patient data was redacted and access-controlled.

To further safeguard ethical standards, patient data was handled in compliance with relevant data protection laws such as HIPAA (Health Insurance Portability and Accountability Act) or GDPR (General Data Protection Regulation), and ethical review boards' recommendations. Additionally, anonymization and encryption methods were applied to secure patient data.

We will continue to adopt methods such as federated learning and differential privacy to enhance the models while ensuring that no patient-specific information is exposed or used inappropriately during the model training and updates.

V. Results and Discussion

A. FNAC active contour-based predictors

We created a custom FNAC-like dataset from real patient data collected in India at RIMS. This dataset was developed

with the goal of being comparable to the widely used Wisconsin Breast Cancer Dataset (WBCD). The primary objective was to ensure that the new dataset shares similar characteristics with WBCD to enable robust machine learning model training, evaluation, and clinical application. The study was designed to compare and validate this dataset using statistical and distribution-based analyses. The methodology for this work involved several steps:

First, both datasets were loaded and explored to gain insights into their structures. The WBCD consists of 569 records with 32 columns, providing detailed metrics such as mean, standard error, and worst-case values for various cellular features. In contrast, the FNAC dataset contains 462 records and 17 columns, derived from fine-needle aspiration cytology images. While the FNAC dataset has fewer features, it was created to align with critical characteristics of the WBCD, focusing on areas such as area, perimeter, smoothness, and radius metrics.

Next, we identified features that were consistently easy to derive given the available equipment and data. We prioritized cell structure-based features over color intensity-based features due to the varying ages of FNAC slides in our sample. From the total available columns, we selected 11 overlapping features for direct comparison, excluding the target column *Diagnosis*. These included *area_mean*, *area_se*, *area_worst*, *perimeter_mean*, *perimeter_se*, *perimeter_worst*, *radius_mean*, *radius_se*, *radius_worst*, *smoothness_mean*, and *smoothness_worst*. This selection ensured that we could perform meaningful statistical comparisons using only the shared features.

To assess the similarity of feature distributions between the datasets, we employed the Kolmogorov-Smirnov (KS) test. This non-parametric test compares the cumulative distribution functions of two datasets for a given feature, quantifying their similarity. A KS statistic close to zero indicates high similarity, while a p-value greater than 0.05 suggests no statistically significant difference between the distributions. For example, features like *perimeter_mean* and *area_mean* showed close alignment, while others might have exhibited differences due to variations in patient demographics, imaging techniques, or processing methodologies used in the FNAC dataset.

The statistical tests revealed that while several features shared distributions similar to those in WBCD, certain discrepancies were present as seen in table **??** For instance, features such as *smoothness_mean* showed greater variability in the FNAC dataset. These differences could reflect the unique clinical and demographic factors associated with the Indian population at RIMS, as well as variations in equipment and imaging protocols. Understanding these differences is crucial for adapting machine learning models to new populations.

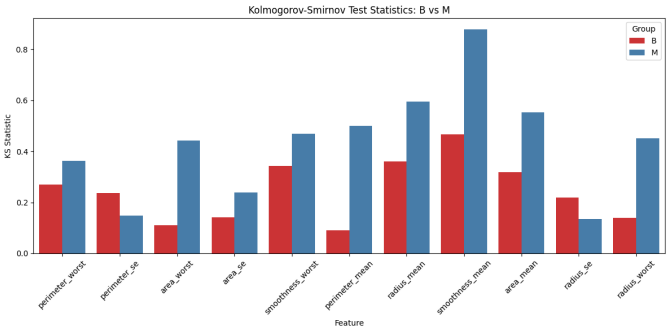

Fig. 13: Kolmogorov-Smirnov (KS) test results comparing Benign and Malignant values in the dataset against their counterparts. The KS test evaluates the similarity of cumulative distribution functions (CDFs) for a given feature between the two datasets. The results show that Malignant values exhibit greater variability compared to Benign values, which generalize more significantly, as indicated by p-values ¡ 0.05. This could be attributed to small number of patients 17 for case study as well as the irregular natures of Malignant cells

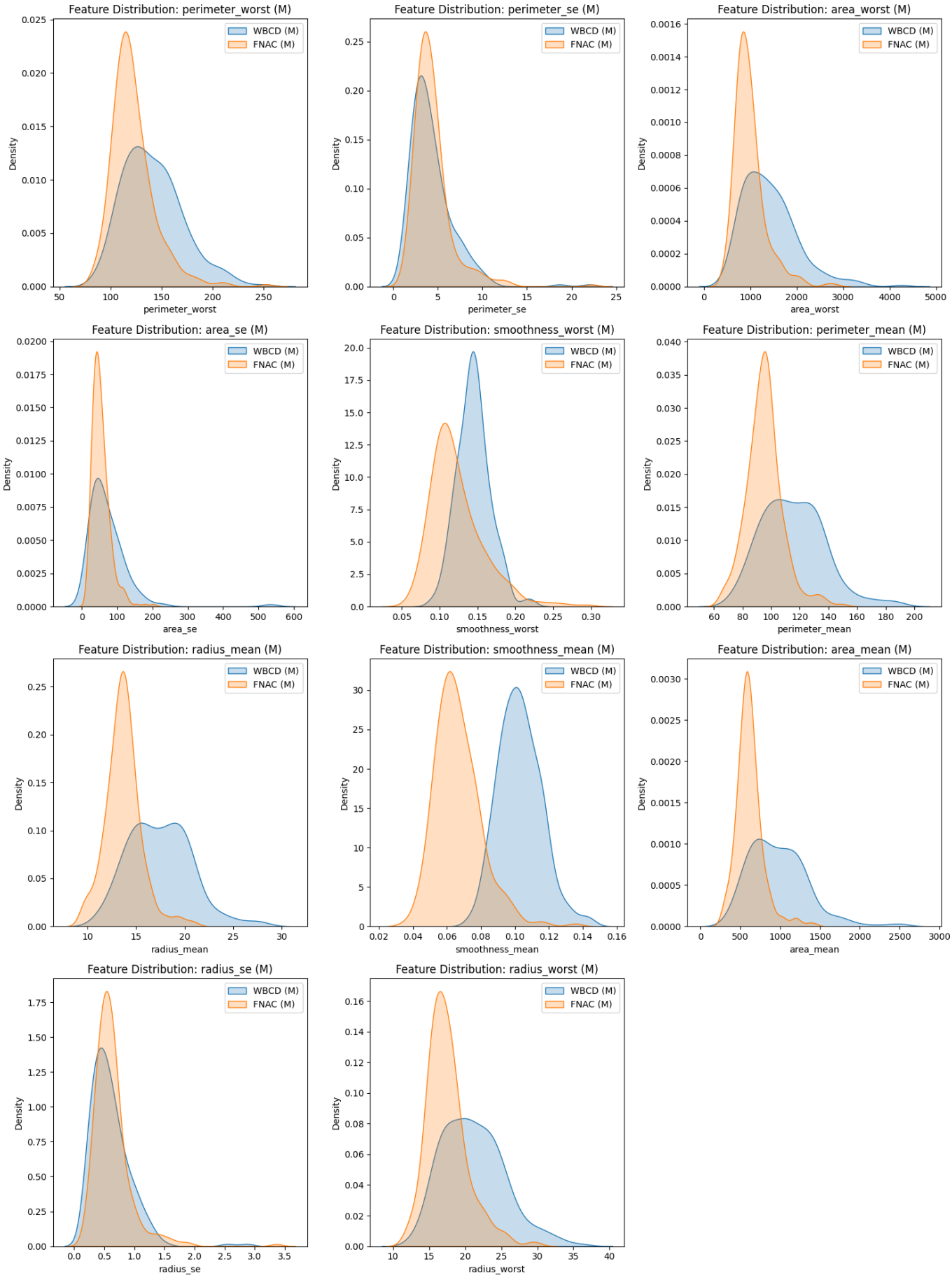

Fig. 14: KDE plot for malignant cells.

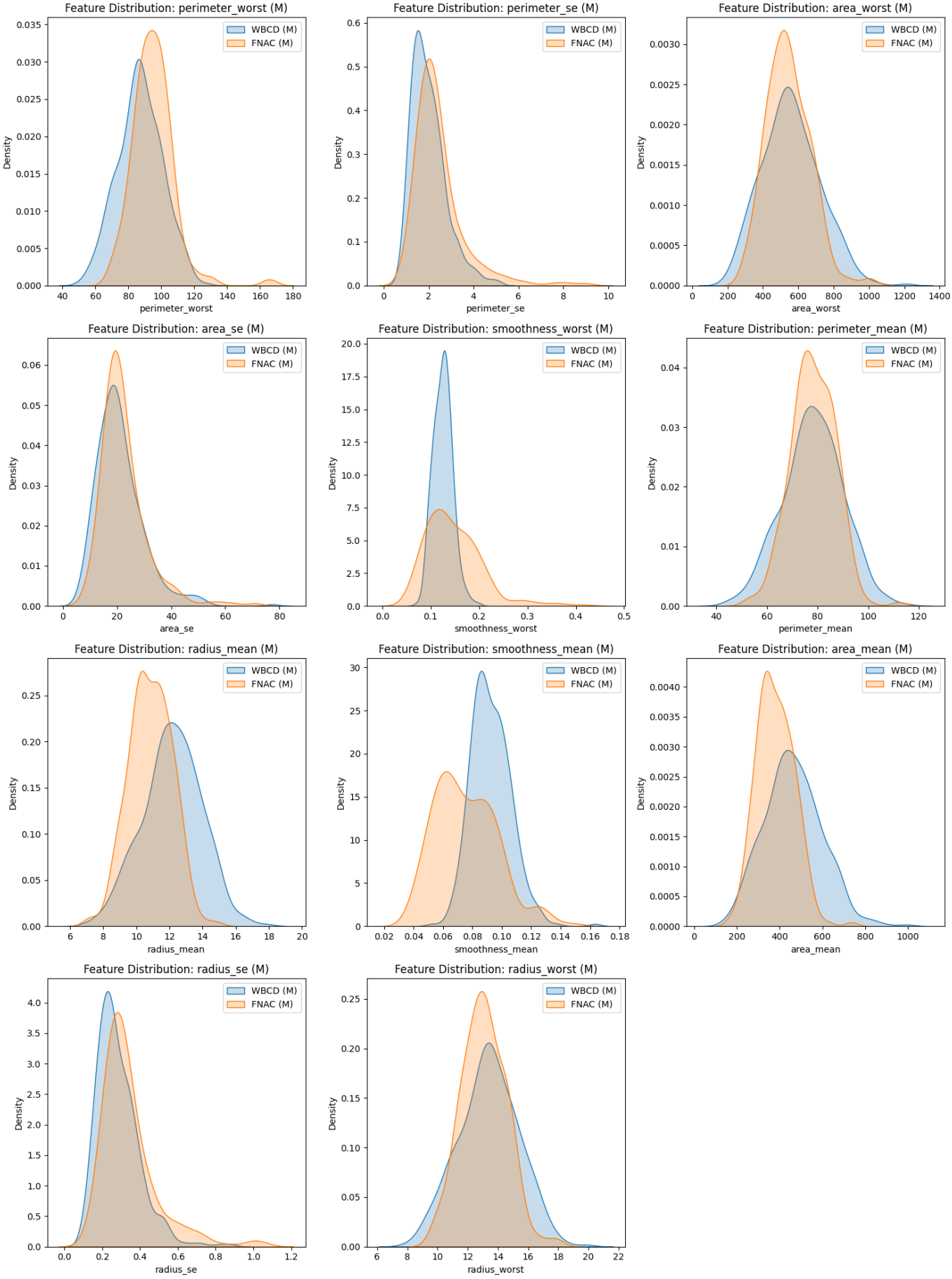

Fig. 15: KDE plot for benign cells.

TABLE V: Comparison of KS Statistic, Levene p-value, and Mann-Whitney p-value for Benign (B) and Malignant (M) Features

Feature	Benign (B)			Malignant (M)		
	KS Statistic	Levene p-value	Mann-Whitney p-value	KS Statistic	Levene p-value	Mann-Whitney p-value
perimeter_worst	0.271	0.0303	1.16e-10	0.363	2.61e-7	4.01e-15
perimeter_se	0.236	0.0301	1.52e-7	0.148	0.1050	0.2066
area_worst	0.111	0.0001	0.2971	0.442	1.48e-11	1.78e-20
area_se	0.141	0.3314	0.0379	0.240	2.39e-7	0.0011
smoothness_worst	0.342	6.09e-30	0.0001	0.470	9.71e-6	1.56e-20
perimeter_mean	0.090	4.07e-5	0.9980	0.500	1.39e-15	1.53e-26
radius_mean	0.361	2.01e-5	3.99e-19	0.595	3.06e-17	2.37e-39
smoothness_mean	0.466	3.03e-13	6.72e-24	0.878	0.774	6.53e-69
area_mean	0.319	1.95e-8	2.63e-14	0.554	1.16e-21	9.90e-35
radius_se	0.219	0.0732	5.44e-7	0.136	0.0791	0.1681
radius_worst	0.139	7.43e-5	0.0259	0.452	2.51e-10	4.87e-23

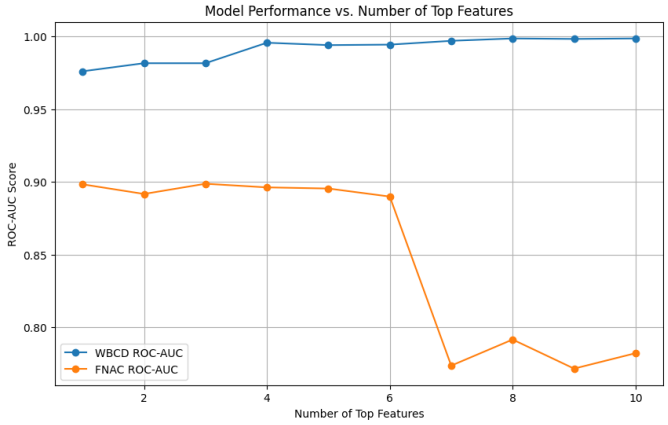

Fig. 16: We can clearly see the features which showed more deviation do not generalise as well and we get best performance using top 6 features reaching an AUC score of 0.896223

the RIMS-FNAC dataset demonstrates strong alignment with the WBCD across most features, suggesting it is a reliable foundation for breast cancer diagnosis model development. However, the observed differences emphasize the need for careful evaluation when applying existing models to the FNAC dataset or using it in real-world scenarios especially when it came to malignant samples as they showed most deviation which can be attributed to thier irregular nature.

Next, We begin calculating feature importances using a Random Forest model, where the features are ranked based on their importance scores as shown in Table VII These features are then sorted in descending order of importance to identify the most influential ones. The next step involves standardizing the feature data from both the WBCD and FNAC datasets using a StandardScaler to ensure that all features contribute equally to model performance. The dataset is split into training and testing sets, and an ensemble model comprising Random Forest, Gradient Boosting, and Support Vector Machine (SVM) classifiers is created using a soft voting strategy.

The model's performance is evaluated by training it on subsets of the top-ranked features and calculating the ROC-AUC scores for both the WBCD test data and the FNAC dataset. The evaluation proceeds by testing various subsets, starting from the top 10 features down to the top 1 feature. These evaluations are repeated for each subset size, allowing for the analysis of how the model's performance varies with different numbers of features.

This process helps to identify the optimal feature subset that maximizes model performance while handling the challenges posed by cross-dataset variability. The final results, showing the ROC-AUC scores for both the WBCD and FNAC datasets, are presented for comparison in figure 16 and table VII

TABLE VI: Model Performance vs. Number of Top Features

Features	WBCD ROC-AUC	FNAC ROC-AUC
10	0.998690	0.782148
9	0.998362	0.771546
8	0.998690	0.791553
7	0.997052	0.773598
6	0.994432	0.889915
5	0.994104	0.895444
4	0.995742	0.896223
3	0.981657	0.898731
2	0.981657	0.891701
1	0.976089	0.898332

TABLE VII: Feature Importances (Sorted by Gradient Boosting)

Rank	Feature	Random Forest	Gradient Boosting	Average
1	*perimeter_se*	0.018834	0.006468	0.012651
2	*smoothness_mean*	0.031022	0.008466	0.019744
3	*radius_se*	0.020667	0.021288	0.020978
4	*area_se*	0.057106	0.013794	0.035450
5	*area_mean*	0.074156	0.013625	0.043890
6	*perimeter_mean*	0.076335	0.013582	0.044958
7	*radius_mean*	0.081764	0.008492	0.045128
8	*smoothness_worst*	0.067357	0.115708	0.091533
9	*area_worst*	0.175409	0.012935	0.094172
10	*perimeter_worst*	0.204108	0.304997	0.254552
11	*radius_worst*	0.193243	0.480646	0.336944

The ensemble model was trained on the original dataset and evaluated on two test datasets: **WBCD Test Data** and the **FNAC Dataset**. Below is the combined performance summary:

- **Performance on WBCD Test Data**:
  - Nearly perfect classification with the following metrics:
    - **Accuracy**: 98%
    - **Macro Average (Precision, Recall, F1)**: 98%
    - **Weighted Average (Precision, Recall, F1)**: 98%
  - **ROC-AUC Score**: **1.00**, indicating flawless discriminatory power.
- **Performance on FNAC Dataset**:
  - Moderate classification performance with the following metrics:
    - **Accuracy**: 67%
    - **Macro Average (Precision, Recall, F1)**: 68%, 68%, and 67%, respectively.
    - **Weighted Average (Precision, Recall, F1)**: 69%, 67%, and 67%, respectively.
  - **ROC-AUC Score**: **0.77**, indicating moderate discriminatory power.

TABLE VIII: Performance Metrics on WBCD Test Data and FNAC Dataset

Dataset	Class	Precision	Recall	F1-Score	Support
WBCD	**B**	0.99	0.99	0.99	71
	M	0.98	0.98	0.98	43
FNAC	**B**	0.59	0.78	0.68	204
	M	0.77	0.58	0.66	258

The model showcased excellent performance on the WBCD test data, achieving near-perfect metrics and a ROC-AUC score of 1.00, demonstrating its strong ability to classify effectively within the same distribution as the training data. On the FNAC dataset, the model achieved a ROC-AUC score of 0.77, reflecting its potential to generalize across datasets while identifying opportunities for further improvement. These findings highlight the model's promising foundation and suggest that incorporating domain-specific adjustments or additional training data from FNAC slides could further enhance its generalization capabilities

B. Birad Rating based Predictors

The selection of a dataset for this task was crucial, as we considered various publicly available datasets as well as those provided by healthcare institutions.

The **INBreast** dataset is a publicly available resource for breast cancer detection, consisting of mammographic images from 115 patients. It includes a variety of images annotated with ground truth data, covering different types of lesions and corresponding clinical information. Known for its high-quality, high-resolution images, **INBreast** provides detailed information on both benign and malignant lesions, making it an excellent choice for training deep learning models.

In contrast, the **CDDM** (Chinese Digital Database for Mammography) dataset contains digital mammographic images used for breast cancer research, particularly for detecting and classifying breast lesions. Although the **CDDM** dataset has fewer cases compared to **INBreast**, it remains a valuable resource for algorithm development and evaluation. However, due to its smaller size and relatively lower image quality, **INBreast** is preferred for this study, especially when focusing on high-quality, smaller datasets. The rich annotations and high-resolution images in **INBreast** make it particularly advantageous for ensuring more accurate and confident predictions.

Based on the results outlined in [7], we decided to train the model on a high-quality, albeit smaller, dataset. The findings indicate that while larger datasets generally improve performance, the effect of resolution and confidence on model accuracy is also significant. Specifically, **table** IX demonstrates that as the size of the training set decreases, classification performance drops, but the differences between smaller subsets (e.g., 20%, 50%, 100%) are less pronounced than expected. Moreover, **Table** X highlights the importance of maintaining high resolution, as performance degrades when the input resolution is reduced by half. This suggests that in some cases, the quality and resolution of the data can be more important than the sheer quantity of data.

TABLE IX: The Effect of Decreasing the Resolution of the Image [7]

Scale	**×1/8**	**×1/4**	**×1/2**	**×1**
0 vs. others	0.587	0.585	0.611	0.618
1 vs. others	0.718	0.742	0.779	0.794
2 vs. others	0.729	0.750	0.777	0.787
macAUC	0.678	0.692	0.722	0.733
HC-macAUC	0.743	0.753	0.782	0.787

TABLE X: The Effect of Changing the Fraction of the Training Data Used. Increasing the Amount of Data Yields Better Results [7]

Fraction	**1%**	**2%**	**5%**	**10%**	**20%**	**50%**	**100%**
0 vs. others	0.541	0.550	0.559	0.564	0.570	0.604	0.618
1 vs. others	0.534	0.631	0.707	0.738	0.749	0.774	0.794
2 vs. others	0.537	0.628	0.715	0.742	0.752	0.771	0.787
macAUC	0.537	0.603	0.660	0.681	0.690	0.716	0.733
HC-macAUC	0.554	0.652	0.710	0.751	0.744	0.778	0.787

Additionally, **Table** XI shows that confident predictions, which tend to be more accurate, can be achieved even with a smaller, high-quality dataset, as the model's performance improves when entropy is low. By focusing on high-quality data, we can ensure that the model benefits from more reliable and precise information, leading to better performance despite the smaller training set. This approach strikes a balance between data quantity, resolution, and prediction confidence, ensuring robust results even with limited resources on end-to-end convolution model described in figure **??**.

TABLE XI: Average AUC (macAUC) as a Function of the Confidence Threshold TP%. When P = 30%, We Refer to the macAUC as a High-Confidence macAUC (HC-macAUC) [7]

TP%	T10%	T20%	T30%	T50%	T100%
macAUC	0.865	0.827	0.811	0.781	0.732

layer	kernel size	stride	#maps	repetition
global average pooling			256	
convolution	3×3	1×1	256	×3
max pooling	2×2	2×2	128	
convolution	3×3	1×1	128	× 3
max pooling	2×2	2×2	128	
convolution	3×3	1×1	128	× 3
max pooling	2×2	2×2	64	
convolution	3×3	1×1	64	× 2
convolution	3×3	2×2	64	
max pooling	3×3	3×3	32	
convolution	3×3	2×2	32	
input			1	

Fig. 17: Description of one deep convolutional network column for a single view. It transforms the input view (a gray-scale image) into a 256-dimensional vector.

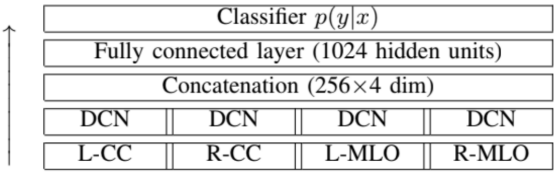

Fig. 18: An overview of the proposed multi-view deep convolutional network. DCN refers to the convolutional network column from Figure 17. The arrow indicates the direction of information flow.

For the classification task, as discussed in 8, we combine the model mentioned above with a transfer learning-based approach. In this approach, various classifiers replace the final layers of VGG16. Both shallow models and CNN-based methods were explored for this purpose. Shallow models were implemented using **PyCaret**, an automated machine learning tool that simplifies the creation of model pipelines. These pipelines were applied to several classifiers, such as **Extra Trees Classifier**, **Gradient Boosting Classifier**, and **Random Forest Classifier** excetra. At the same time, a CNN-based pipeline was developed, where grid search was used to optimize hyperparameters, such as learning rate, number of layers, neurons per layer, and epochs. The results of these models are summarized in **Table ??**.

We utilized Grad-CAM to visualize the convolutional layers of a single image from each BIRADS category. Grad-CAM serves as an effective diagnostic tool for identifying potential issues in model performance. If the model fails to focus on clinically relevant regions—such as areas indicative of malignancy—it may signal underlying flaws in the model's reasoning, warranting further investigation, refinement, or retraining. In our analysis, Grad-CAM successfully highlighted suspicious areas, including microcalcifications and masses, which are key indicators of malignancy. Conversely, if Grad-CAM were to focus on irrelevant areas, such as healthy tissue or non-clinically significant regions, it would indicate a need for further model evaluation and adjustment

VI. Future Scope for Ethical Scaling and Improvement of Machine Learning System-Based Technologies

The integration of machine learning (ML) in medical diagnostics, particularly for breast cancer detection using Fine Needle Aspiration Cytology (FNAC) and mammography, has shown great promise. However, as ML models evolve, there are several technical avenues to scale and improve these systems in an ethical manner.

A. Data Augmentation and Synthetic Data Generation

The challenge of limited annotated datasets due to privacy concerns and data availability can be addressed using data augmentation and synthetic data generation. Techniques like rotation, flipping, and contrast adjustment are among the techniques used above, we will be coupling them with Generative Adversarial Networks (GANs), which can create diverse and realistic datasets while preserving patient confidentiality. This approach reduces biases and enhances model generalization across various demographic groups.

B. Federated Learning for Privacy Preservation and Continuous Model Updating

Federated learning enables machine learning models to be trained on decentralized data sources, ensuring patient privacy by keeping data within local servers. Only model updates are shared, ensuring compliance with privacy regulations like HIPAA and GDPR. This approach facilitates collaboration between healthcare institutions without compromising sensitive data, enabling large-scale training with privacy protection. Additionally, continuous learning via online algorithms helps address the risk of performance degradation over time due to evolving medical knowledge and patient demographics. By continuously incorporating new data, these models remain up-to-date, reducing model drift while ensuring that updates are ethically transparent and maintain patient privacy and consent.

C. Explainability and Interpretability in Machine Learning Models

To increase trust in ML systems, explainability methods like Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive Explanations (SHAP) are essential. These techniques allow clinicians to understand model decisions, fostering better adoption in medical practice. Interpretability ensures that features like tumor morphology in mammograms are correlated with diagnosis, providing transparency for clinical decisions.

This is in addition to measures taken by us by incorporating GradCam and Generalized Linear models as part of the process in the development of breast cancer module and FNAC nuclear extraction modules for maximum interpretability

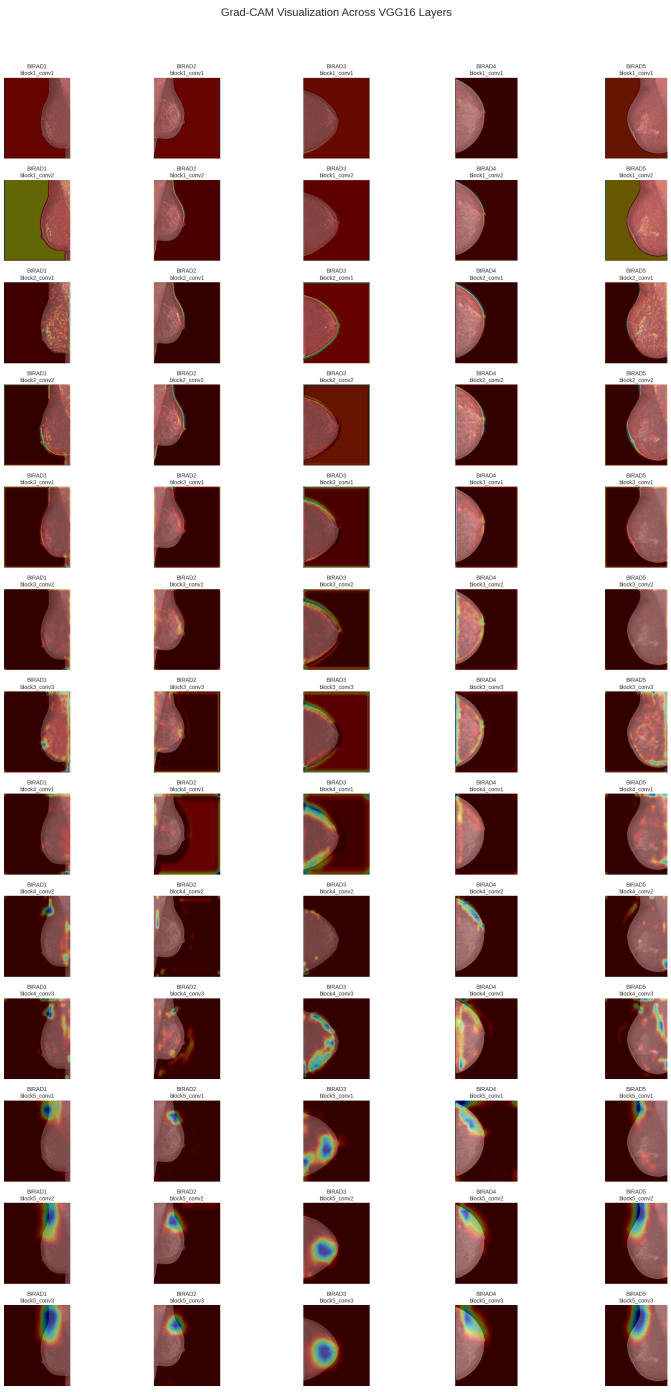

Fig. 19: Grad-CAM successfully highlighted suspicious areas, including microcalcifications and masses, key indicators of malignancy. It also focused more visibly on the upper part of the breast in the MLO view, where most tumors and masses develop.

D. Cross-Cultural and Cross-Institutional Validation

For ML models to be ethically scalable, cross-cultural and cross-institutional validation is crucial. By validating models across diverse datasets from various institutions and geographical regions, including those with limited or no available data on Indian patients, we ensure their applicability to a wide range of populations. Specifically, validating models on Indian patient data is particularly important, as it addresses the gap in available datasets and ensures that the models are relevant and effective for this underserved group. This collaborative validation mitigates biases, ensures equity in healthcare outcomes, and makes the models more universally applicable.

VII. Conclusion

This system presents a novel approach to breast cancer diagnosis by integrating AI with FNAC and mammographic imaging data, thereby significantly reducing the diagnostic time for radiologists and pathologists. By automating the BI-RADS rating assignment through transfer learning, the system achieves an accuracy of 95.08% in mammogram analysis (Suspicious-Non-suspicious). However, the system performs sub-optimally on the BI-RADS rating-based approach (49.89%) due to the lack of high-resolution digital mammography data. This limitation occurred as a result of an accidental system crash during regular updates and maintenance sessions, which disrupted the data collection.

Additionally, the FNAC-based classification module, which enhances the accuracy of malignant vs benign detection, demonstrates an accuracy of 98.83% (p-value = 0.049, 95% confidence interval). These advancements significantly minimize the need for invasive biopsy procedures, offering a more efficient and less invasive alternative. The system's ability to reduce diagnostic time, coupled with high classification accuracy for FNAC, underscores its potential in improving breast cancer diagnosis and ultimately enhancing patient care by enabling earlier treatment decisions.

By being open-source, the application not only fosters transparency and collaboration within the medical and research communities but also significantly contributes to the future development of explainability and interpretability in machine learning models. This open-source approach allows for continuous refinement, enabling diverse institutions and researchers to validate and enhance the model across different cultural and clinical contexts. With cross-cultural and cross-institutional validation, the system can be adapted to ensure robust, generalizable results. This helps address the challenges of medical variability across populations and settings, ensuring that the tool remains relevant and effective in diverse healthcare environments. The application streamlines diagnostic workflows, offering an ethical, efficient, and accessible tool for early breast cancer detection. Through the integration of human-in-the-loop methods, the system ensures high clinical relevance and interpretability, which is crucial for pathologists working with FNAC slides. Ultimately, this system aims to improve patient prognosis, reduce diagnostic delays, and lower

breast cancer-related mortality by enabling earlier detection and treatment.

REFERENCES

[1] Tabár L, et al. The incidence of fatal breast cancer measures the increased effectiveness of therapy in women participating in mammography screening. *Cancer*. 2019 Feb 15;125(4):515-523. doi: https://doi.org/10.1002/cncr.31840. Epub 2018 Nov 8. PMID: 30411328; PMCID: PMC6588008.

[2] Shen, L., et al. Deep Learning to Improve Breast Cancer Detection on Screening Mammography. *Scientific Reports*, 2019. doi: https://doi.org/10.1038/s41598-019-48995-4.

[3] McKinney, S.M., Sieniek, M., Godbole, V., et al. International evaluation of an AI system for breast cancer screening. *Nature*, 2020, 577(89-94). doi: https://doi.org/10.1038/s41586-019-1799-6.

[4] Kang, S.H., & Kim, H.J. Limitations of fine needle aspiration cytology in breast cancer diagnosis: A retrospective study. *Journal of Pathology and Translational Medicine*, 2017; 51(3), 214-220. doi: https://doi.org/10.4132/jptm.2017.04.11.

[5] American College of Radiology. Breast Imaging Reporting and Data System (BI-RADS®) Atlas, 5th edition. 2013. from https://www.acr.org/-/media/ACR/Files/RADS/BI-RADS/Mammography-Reporting.pdf.

[6] Herndon, J.R. (2024, September 6). What is fine needle aspiration for breast biopsy? This test is used to determine the status of a breast lump. *Verywell Health*. from https://www.verywellhealth.com/fine-needle-aspiration-of-a-breast-cyst-429947.

[7] The NYU Breast Cancer Screening Dataset v1.0 - Nan Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, S. Gene Kim, Laura Heacock, Linda Moy, Kyunghyun Cho, and Krzysztof J. Geras. from https://cs.nyu.edu/~kgeras/reports/MRI_datav1.0.pdf.

[8] American College of Radiology. (2020). Mammography: Technique and Positioning. from https://www.acr.org.

[9] Zamir, R., Bagon, S., Samocha, D., et al. Segmenting microcalcifications in mammograms and its applications. *Medical Imaging 2021: Image Processing (SPIE)*, 2021, 11596:788-795. from https://arxiv.org/abs/2102.00811v1.

[10] Houssami, N., Ciatto, S., & Vano, M. Microcalcifications in breast cancer: Detection and significance in breast cancer screening. *The Breast*, 2014; 23(6):850-855. doi: https://doi.org/10.1016/j.breast.2014.09.001.

[11] Street, W.N., Wolberg, W.H., & Mangasarian, O.L. Nuclear feature extraction for breast tumor diagnosis. *SPIE Proceedings*, 1993; 1905:861. from http://proceedings.spiedigitallibrary.org/.

[12] Goodfellow, I., et al. Generative Adversarial Networks. *Advances in Neural Information Processing Systems*, 2014. doi: https://doi.org/10.48550/arXiv.1406.2661.

[13] Frid-Adar, M., et al. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, 2018.

[14] McMahan, B., et al. Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.

[15] Bonawitz, K., et al. Practical secure aggregation for privacy-preserving machine learning. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017.

[16] Gama, J., et al. A survey on concept drift adaptation. *ACM Computing Surveys (CSUR)*, 2014.

[17] Bifet, A., et al., "Mining the stream of data with adaptation," *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, 2010.

[18] Ribeiro, M. T., et al., "Why should I trust you?" Explaining the predictions of any classifier," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[19] Lundberg, S. M., & Lee, S. I., "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, 2017.

[20] Liu, X., et al., "Generalizing machine learning models in medical imaging: A cross-institutional study," *IEEE Transactions on Medical Imaging*, 2019.

[21] Zhang, Y., et al., "Cross-regional validation of machine learning models for healthcare applications," *Journal of Healthcare Engineering*, 2020.

[22] Tice, J. A., et al. *Breast Cancer Risk Prediction: Breast Density Factors.* Journal of the National Cancer Institute, 2020

[23] Gail, M. H., et al. *Projecting Individualized Probabilities of Breast Cancer.* Journal of the National Cancer Institute, 1989

[24] Yavuz, Erdem, and Can Eyupoglu. An effective approach for breast cancer diagnosis based on routine blood analysis features. *Medical & biological engineering & computing vol. 58,7*, 2020 doi:10.1007/s11517-020-02187-9

[25] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Conference on Computer Vision (ICCV),* 618-626. https://doi.org/10.1109/ICCV.2017.74

[26] Orell SR, Sterrett GF, Whitaker D. *Fine Needle Aspiration Cytology*. Elsevier; 2012.

[27] Stewart CJ, Duncan JA. "The role of fine needle aspiration cytology in modern diagnostic pathology." *Cytopathology*. 2020;31(5):391-403.

[28] Rosai J. *Rosai and Ackerman's Surgical Pathology*. Elsevier; 2017.

[29] Rakha EA, Ellis IO. "Diagnostic challenges in breast pathology." *Pathology*. 2018;50(1):100-110.

[30] Madabhavi I, Patel A, Sarkar M. "AI in cytology: A paradigm shift." *Journal of Cytology*. 2021;38(3):153-161.

[31] Liu Y, Gadepalli K, Norouzi M, et al. "Artificial intelligence in pathology." *Nature Medicine*. 2019;25(1):30-36.

[32] Hutchinson ML, Somers R. "Air-dried versus fixed smears in cytology." *Cytopathology*. 2018;29(2):109-120.

[33] Bibbo M, Wilbur D. *Comprehensive Cytopathology*. Elsevier; 2014.

[34] Layfield LJ, Reichman A. "Necrosis in cytology." *Acta Cytologica*. 2016;60(4):333-340.

[35] Mehrotra R, Singh M. "Necrosis and its significance in FNAC." *Cytology Today*. 2021;36(4):203-209.

[36] Crum CP, Hornick JL. "Impact of slide degradation on cytopathology." *Journal of Diagnostic Pathology*. 2017;24(5):355-360.

[37] Pantanowitz L, Hornish M. "Slide preservation in cytopathology." *Cytopathology Today*. 2020;35(2):87-94.

[38] Kolev V, Zahariev V. "Artifacts in FNAC slides." *Cytology Archives*. 2019;14(1):65-72.

[39] Acs G, Zhang PJ. "Slide aging and histology artifacts." *Archives of Pathology & Laboratory Medicine*. 2021;145(9):1154-1160.

[40] Golden JA, Louie RJ. "AI applications in cytopathology." *Current Opinion in Pathology*. 2022;34(1):49-56.

[41] Rajpurkar P, Chen E, Banerjee O, et al. "Deep learning for pathology." *Annals of Oncology*. 2019;30(5):779-790.

[42] Kalita, Manjula & Mahanta, Lipi et al. (2024). A new deep learning model with interface for fine needle aspiration cytology image-based breast cancer detection. *Indonesian Journal of Electrical Engineering and Computer Science.* 34. 1739-1752. 10.11591/ijeecs.v34.i3.pp1739-1752.