MSHFF-Net: An Explainable Multi-Scale Hierarchical Feature Fusion CNN for Breast Cancer Diagnosis from Histopathological Images

Muhammad Hassaan Ashraf *
Faculty of Computing
Riphah International University
I-14 Islamabad, Pakistan
hassaan.ashraf@riphah.edu.pk

Musharif Ahmed
Faculty of Computing
Riphah International University
I-14 Islamabad, Pakistan
musharraf.ahmed@riphah.edu.pk

Muhammad Nabeel Mehmood
Faculty of Computing
Riphah International University
I-14 Islamabad, Pakistan
nabeel.mehmood@riphah.edu.pk

Muhammad Esham Qureshi
Faculty of Computing
Riphah International University
I-14 Islamabad, Pakistan
mequreshi.cs@gmail.com

Hamed Alghamdi
Department of Computer Science, King Abdulaziz University, Jeddah, 21589, Saudi Arabia
halghamdi0776@stu.kau.edu.sa

***Abstract*: Breast cancer is one of the leading causes of cancer-related deaths among women. Therefore, accurate multi-class grading of histopathological images is crucial for early screening and effective treatment of breast cancer. Given the large amount of intra-class variation and inter-class similarity, traditional Convolutional Neural Networks (CNNs) frequently struggle in multi-class grading. To mitigate these challenges, we present MSHFF-Net, an Explainable Multi-Scale Hierarchical Feature Fusion architecture that combines adaptive multi-scale blocks with a feature-fusion stem to simultaneously capture tissue-level patterns and fine-grained cellular structures. The model processes input images of size 320×320, preserving spatial resolution to enable detailed feature extraction. MSHFF-Net is trained on the publicly available BRACS benchmark dataset using Leaky ReLU activation and sparse categorical cross-entropy loss function. It achieves an overall accuracy of 73.94%, outperforming several state-of-the-art CNN backbones. These results demonstrate the effectiveness of hierarchical multi-scale feature fusion in enhancing automated breast cancer diagnosis from histopathological images. Furthermore, to ensure interpretability, the Explainable AI (XAI) method Gradient-Weighted Class Activation Mapping (Grad-CAM) is employed to visualize the decision-making process of MSHFF-Net, providing greater transparency and clinical trustworthiness.**

Keywords: Breast Cancer Diagnosis, Histopathological Image Analysis, Deep Learning, Convolutional Neural Networks (CNNs), Multi-Scale Feature Fusion, Explainable Artificial Intelligence (XAI), Grad-CAM Visualization, Medical Image Classification, BRACS.

I. INTRODUCTION

Breast cancer is one of the major causes of cancer-related deaths in women worldwide. In 2022, approximately 2.3 million new cases and 670,000 deaths reported [1]. Detecting breast cancer at an early stage is crucial, as timely diagnosis helps doctors start treatment sooner and improve the chance of recovery. Imaging techniques such as mammography, CT, and MRI are widely used for early screening, but they mainly capture structural details. These scans cannot show the cellular or histological subtypes of breast cancer, which are important for choosing the right treatment.

For a confirmed diagnosis, pathologists examine Hematoxylin and Eosin (H&E) stained biopsy samples under a microscope. They examine the shape of the gland structures, and lesion boundaries in Whole-Slide Images (WSIs) to determine whether the tissue is benign or malignant. WSIs are often extremely large, sometimes gigapixel sized, that vary in staining, brightness, and focus, and are obtained via process depicted in Fig. 1. This makes manual review time-consuming and difficult , and can lead to differences in judgment between pathologists. These challenges highlight the need for automated, reliable, and interpretable diagnostic tools that can support pathologists.

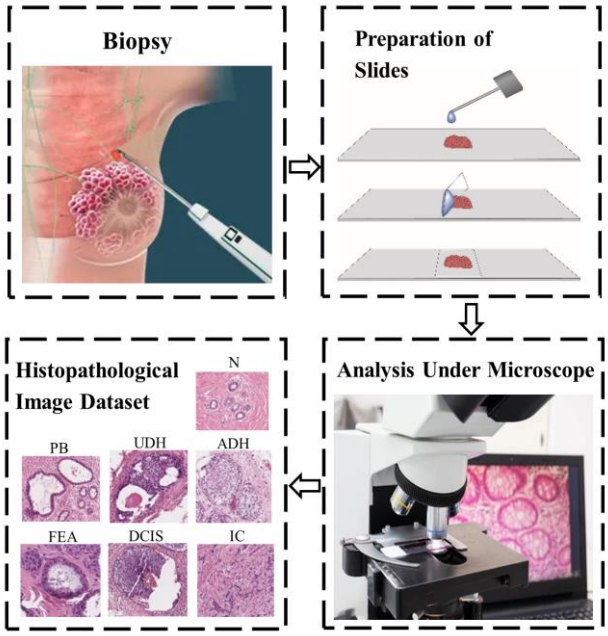

Fig. 1. Histopathological image acquisition for breast cancer diagnosis

Over the past decade, digital pathology integrated with deep learning has changed biomedical image analysis by enabling the automated extraction of useful features directly from raw image [2], [3], [4]. CNNs such as ResNet, DenseNet, and Vision Transformers (ViTs) have shown strong results in tasks such as tumor detection, tissue segmentation, and cancer subtype classification using both fully supervised and weakly supervised learning approaches [5], [6], [7]. Despite these advancements, many existing models still struggle to capture both fine cellular details and wider tissue-level context at the

* Corresponding author

same time. Both levels of information are important for accurate multi-class breast cancer grading. Another limitation is the lack of interpretability, since many deep models operate as "black boxes" which makes it difficult for clinicians to trust their predictions.

In this research, we present MSHFF-Net, a novel Multi-Scale Hierarchical Feature Fusion Network designed to captures both detailed and high-level features while remaining computationally efficient. The model uses a feature fusion stem together with adaptive multiple blocks to learn from multiple resolutions. [3], [8]. The fusion stem gathers information from multiple scales, which helps the network identify different tissue patterns and detect subtle morphological variations without increasing model complexity. Experimental results show that the proposed model outperforms several state-of-the-art architectures. To add transparency, we employed Grad-CAM technique to highlight the regions that influence the model's decisions. Overall, MSHFF-Net demonstrates strong potential as an explainable and scalable deep learning framework for automated breast cancer diagnosis.

II. Literature Review

In this section, we summarize recent key developments in breast cancer diagnosis and highlight the key challenges that still need attention.

Gaia et al. [9] modified the ResNet-50 backbone with Fusion Mamba and Agent Attention modules, achieving superior slide-level consistency on BRACS compared to ABMIL, CLAM SB, and TransMIL. Zhao et al. [10] introduce HDSA-MIL, which jointly trains bag and instance classifiers under a smoothing attention strategy, yielding up to a 6.9% gain over standard MIL in DCIS-vs-invasive splits. Hernández et al. [11] used a data-centric approach by replacing CHOWDER's attention head with stacked 1D convolutions over ViT-based embeddings and applying ANOVA for hyperparameter tuning across all seven BRACS lesion types. Their method improved multi-class accuracy by 3%.

Lightweight and hybrid CNN models address BRACS's large image sizes and slide-level noise. Kausar et al. [12] first apply a discrete wavelet transform to isolate low-frequency bands, then use a depthwise-separable CNN with invertible residual blocks, reaching ~72% multi-class accuracy on all 547 WSIs. Li et al. [13] propose AFFC-Net, a dual-branch CNN + GNN architecture that extracts H&E-normalized patch textures and constructs superpixel graphs processed via GraphSAGE; adaptive fusion yields a weighted F1 of 67.2% on 4,391 WSIs, demonstrating the value of combining local texture and global topology.

Two-stage classification pipelines leverage expert ROI annotations for subtyping. Tafavvoghia et al. [2] train a ResNet-18 tumor detector (F1 = 0.95) on combined TCGA-BRCA and BRACS ROIs to filter non-tumor tiles, then deploy four One-vs-Rest ResNet-18 classifiers balanced per subtype, with XGBoost meta-learning achieving a macro F1 of 0.73 on 221 held-out WSIs. For molecular grading, Kim et al. [14] segment tissue via CLAM SB, tile slides into 224×224 patches, and extract features using a UNI foundation model; their ACMIL framework attains an external F1 of 0.731 and multi-class AUC of 0.835, with predicted grades correlating to five-year survival.

Güler et al. [15] compare traditional classifiers (SVM, Random Forest, k-NN) and modern CNNs (DenseNet, MobileNet), observing >90% accuracy on homogeneous datasets but only ~70% on BRACS due to staining variability and label noise. Brancati et al. [16] established BRACS itself includes 547 de-identified WSIs from 189 patients and 4,539 ROIs labeled by three pathologists preserving real-world artifacts and including rare lesions (ADH, FEA) under patient-wise splits (395/67/85 WSIs; 3 657/312/570 ROIs).

A. Research Gap Analysis

Although deep learning has made strong progress in biomedical image analysis, several challenges are still not fully addressed. Many traditional CNNs face issues such as gradients vanishing and limited feature learning because they rely on fixed-size kernels and simple feed-forward structures [3]. Consequently, they experience difficulty in capturing the intricate morphological variations visible in biomedical images. Accurate classification is further challenging due to high intra-class variation, lesion size diversity, and inter-class similarity [17]. Model stability suffers through the addition of noise from staining variations and real-world artifacts. The practical application of current models in clinical settings is limited due to their inability to efficiently integrate multi-scale contextual information and the computational demands [18]. These drawbacks highlight the need for a scalable, effective, and lightweight architecture capable of robust multi-scale feature extraction and fusion. This motivates the development of MSHFF-Net, which aims to provide a more reliable approach for breast cancer diagnosis.

III. Proposed Methodology

The sections below provide an overview of the proposed framework which covers, preprocessing steps, data acquisition procedures and descriptive explanation of the proposed model's architecture.

A. Data Acquisition

The BReAst Carcinoma Subtyping (BRACS) benchmark is an extensive dataset, encouraging automated diagnosis in computational histopathological breast cancer grading [16]. The data sets include 4,391 WSIs of 189 patients who are aged 16 to 86 years. The digitization of all WSIs was performed by an Aperio AT2 scanner with 40 x magnification and 0.25 -m pixel resolution. The use of this high-resolution scanning is necessary to see the subtle morphological changes that can often mark early or precancerous lesions.

BRACS categorizes breast lesions into seven subtypes, six malignant and one normal. A visual representation of breast lesion subtypes can be seen in Fig. 2.

Three board-certified pathologists used consensus-based methods to annotate each WSI. BRACS sets itself apart as a public dataset because it incorporates rare cases including FEA, ADH, and DCIS along with diagnostic challenges which separates it from other public datasets that are available.

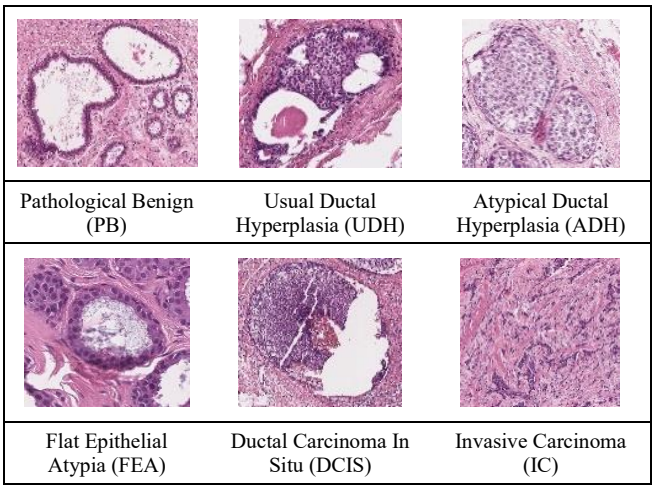

Fig. 2. Visual representations of key breast lesion sub-types.

B. Data Preprocessing

The dataset was divided into three subsets for training, testing and validation procedures by following a predetermined distribution. The grouping was performed on patient level where each patient's slides belong to a single assigned subset which ensures unbiased assessment of the model's ability to generalize across different patient populations. The original distribution of data did not correspond to model's requirements therefore we split it to standard training (70%), testing (20%) and validation (10%) to facilitate effective model training. The count of samples in each class after splitting can be seen in Table I.

TABLE I. DATA SAMPLES IN EACH CLASS AFTER DATA-SPLIT

Classes	Divisions		
	Train (70%)	Test (20%)	Validate (10%)
N	358	103	51
PB	530	153	75
UDH	329	95	47
ADH	397	115	56
FEA	548	157	78
DCIS	524	151	74
IC	385	110	55

The dataset images were resized using bicubic interpolation to a standardized resolution and the model was set to take input shape of 320 x 320. The chosen higher image resolution serves a purpose to preserve spatial features of small abnormalities that could be crucial in early diagnosis and precise subtyping classification from being discarded during preprocessing. The BRACS dataset features paramount importance for multi-class breast cancer subtyping and helps establish fundamental groundwork which will fuel further advancements in digital pathology.

C. Proposed Framework

The proposed Multi-Scale Hierarchical Feature Fusion Network (MSHFF-Net) comprises a hierarchical feature fusion stem that captures feature at various scales using different kernel sizes to minimize the feature loss and multiple layers for capturing the features at different level. Fig. 3 shows the complete architecture diagram of the proposed MSHFF-Net model.

The proposed architecture begins by accepting a 320×320 color input, which is simultaneously processed through four parallel convolutional layers to capture diverse spatial scales. The first path applies a 3×3 convolution with stride 1, producing 320×320×64 feature maps, while the second uses a 3×3 convolution with stride 2 to generate 160×160×64 feature maps. In parallel, a 5×5 convolution with stride 4 extracts 80×80×32 feature maps, and a 7×7 convolution layer with the same stride extracts additional 80×80×32 feature maps. These four outputs, when combined, incorporates hierarchical and multi-scale representations, that serves as the foundation for the feature fusion process of the stem. To integrate these features, the 320×320×64 maps from the first path are down sampled by a 2×2 max pooling layer with stride 2 and concatenated with the 160×160×64 maps from the second path, resulting in 160×160×128. A 1×1 convolutional called "Channel Pooling Layer" then reduces this to 160×160×64. The reduced tensor is split into two branches: one branch passes through a 3×3 convolution (stride 1) followed by 2×2 max pooling (stride 2), while the other undergoes a 3×3 convolution (stride 2). Both outputs are concatenated with the 80×80×32 maps from the 5×5 path, yielding 80×80×160, which a second Channel Pooling Layer compresses to 80×80×64.

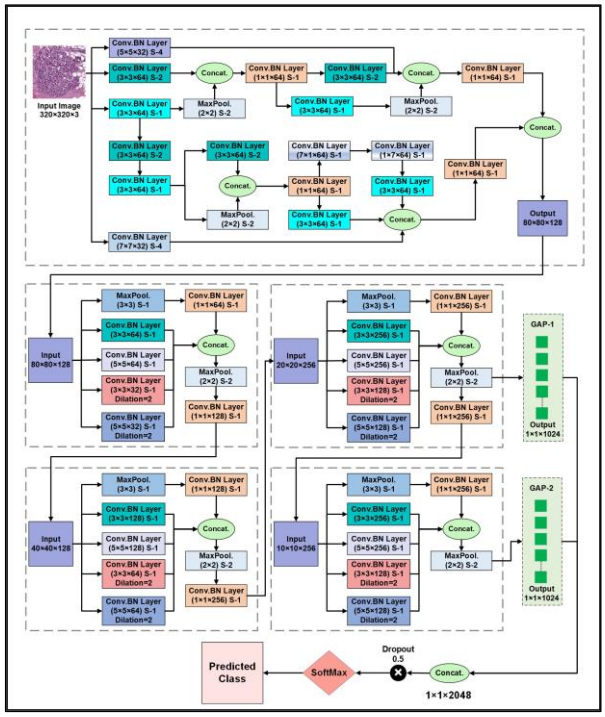

Fig. 3. Proposed MSHFF-Net's architecture.

Meanwhile, the original 320×320×64 output from the first convolution is also processed by a 3×3 convolution (stride 2) to form 160×160×64, then through a standard 3×3 convolution (stride 1). Its result is split between a 3×3 convolution (stride 2) and a 2×2 max pooling operation; their concatenation produces 80×80×128, again reduced to 80×80×64 via Channel Pooling. This compressed output undergoes three parallel multiscale convolutions 1×7, 7×1, and 3×3 (all stride 1) and is concatenated with the 80×80×32 maps from the 7×7 path to yield 80×80×160. After a final Channel Pooling step to

80×80×64 and concatenation with the previous 64 channel branch, the stem outputs an 80×80×128 tensor that preserves fine details with minimal information loss.

This 80×80×128 tensor enters the first hierarchical block, where it is again processed through four parallel convolutions 1×1×64, 3×3×64, 5×5×64 (Dilation 2), and 3×3×32 (Dilation 2) with producing a combined 80×80×256 feature map. A 2×2 max pool and ensuing 1×1×128 convolution reduce the spatial dimensions to 40×40×128. The second block mirrors this structure with doubled channel counts and after parallel convolutions, it yields 40×40×512 on concatenation, which is down sampled to 20×20×256 via max pooling and 1×1×256 convolution.

In the third block, the 20×20×256 input is traversed through by the same parallel scheme of 4 convolutional and a sequence of max pooling and conv. layer, with the doubled feature maps count at each convolutional layer, i.e., 256 and 128 instead of 128 and 64. The third block transforms it into 20×20×1024 feature maps, then reduces it to 10×10×1024 through max pooling. These maps feed into the first Global Average Pooling layer (GAP-1), producing a 1×1×1024 vector, and a subsequent 1×1×256 convolution outputs a 10×10×256 feature map passed onto next block. The fourth block repeats this the third block's pattern on 10×10×256, generating 10×10×1024 that is max pooled and passed through GAP-2 to yield another 1×1×1024 vector.

Finally, the two 1×1×1024 vectors concatenate to form a 1×1×2048 feature vector. After applying a 0.5 dropout, this vector is sent to a Softmax layer, which outputs the predictions for the seven lesion classes. Strong feature extraction and legitimate classification outcomes are assured by this sequential multiscale fusion method.

D. The Activation and Loss Function

In our proposed MSHFF-Net, we utilize the Leaky ReLU activation function, which is defined as:

$$f(x) = \begin{cases} x, & x \geq 0, \\ \alpha x, & x < 0 \end{cases} \tag{1}$$

Here, $\alpha = 0.1$ is the negative slope that keeps a small non-zero gradient for negative inputs. This helps maintain stable gradient flow through the network, improves training smoothness, and supports better convergence. Leaky ReLU is also computationally efficient, which makes it a practical activation choice for MSHFF-Net.

We use sparse categorical cross-entropy as the loss function. It is suitable for multi-class classification tasks on the BRACS dataset. For a dataset with N samples, the loss is:

$$L = -\Sigma_{i=1}^{N}\ log\left(P_{i,y_i}\right) \tag{2}$$

Here, y_i is the true label for sample i, and P_{i,y_i} is the predicted probability for that class. This loss function helps the model learn accurate class probabilities and improves overall classification performance.

E. MSHHF-Net's explainability via Grad-CAM

The proposed framework incorporates Grad-CAM [19] as an XAI technique, that ensures that the model is concentrating on essential patterns by visually validating the model's attentional focal points. Grad-CAM heatmaps are generated to visualize class-specific activation regions, emphasizing the most discriminative features (such as pleomorphic nuclei, irregular chromatin patterns, tumor cells) influencing the MSHFF-Net's prediction. Grad-CAM is applied to the final convolutional block preceding the GAP layer to derive meaningful attention maps.

Grad-CAM utilizes the partial derivatives of the network's pre-softmax score for each spatial location in the feature maps to calculate the channel-wise importance weights. By spatially averaging the gradients, the weight for each channel is determined, indicating how sensitive the class score is to variations in the channel's activations. The class-discriminative localization map (Grad-CAM heatmap) is then created by a weighted sum of the FMs, followed by a ReLU activation preserves features that favorably affect the target class. These heatmaps provides an interpretable illustration of the network's decision-making process.

IV. Experiments and Evaluation

A. Evaluation Metrics

To evaluate the proposed model, we compute the overall performance using the confusion matrix. Let $CM[i][j]$ denote the count of samples whose true label is i but were predicted as j. From this, we derive

$$Precision_i = \frac{CM[i][i]}{CM[i][i]+\sum_{k\neq i} CM[k][i]} \tag{3}$$

$$Recall_i = \frac{CM[i][i]}{CM[i][i]+\sum_{k\neq i} CM[i][k]} \tag{4}$$

$$F1-Score_i = \frac{2\times \text{Precision}_i \times Recall_i}{\text{Precision}_i + Recall_i} \tag{5}$$

$$Accuracy = \frac{\sum_{i=1}^{n} CM[i][i]}{\sum_{i=1}^{n}\sum_{j=1}^{n} CM[i][j]} \tag{6}$$

Here, n is the number of classes, $CM[i][i]$ gives true positive (TP_i) for class i, $\sum_{k\neq i} CM[k][i]$ counts false positives (FP_i) and $\sum_{k\neq i} CM[i][k]$ counts false negatives (FN_i).

The proposed framework was evaluated through precision, recall, F1-score and overall accuracy metrics which enable assessment of specific class discrimination and overall characterization performance.

B. Performance of the Proposed Model

Proposed MSHFF-Net was evaluated on the BRACS test split under a multi-class classification task. Trained for 60 epochs from scratch with Adam optimization and sparse categorical cross-entropy loss, it comprises just 8.7 million parameters substantially lighter than most baselines.

Fig. 4 presents the confusion matrix for the proposed model. High true positive (TP) rates along the diagonal indicate robust detection across both low-risk and high-risk lesion categories. For example, invasive carcinoma and in situ carcinoma exhibit minimal false negatives (FN), underscoring excellent sensitivity for clinically critical classes. Low FN rates reduce the chance of missed malignancies, which is paramount in automated histopathological screening. False positives (FP) are also well controlled, preventing unnecessary alerts for benign or normal tissue. The model's

73.94% accuracy and balanced performance are strengthened by the overall true negative (TN) values.

False negatives (FN) occur when malignant or atypical slides are classified as less severe categories. In clinical screening, these cases are if significant concern. FN values for high-risk classes remain low in our findings, for instance, $FN_{IC} = 2$, $FN_{DCIS} = 8$. This lowers the possibility of overlooking vital diagnoses. Additional follow-up procedures may be necessary due to false positives (FPs). Nevertheless, when considering the overall precision of 73.20%, the FP counts are still acceptable.

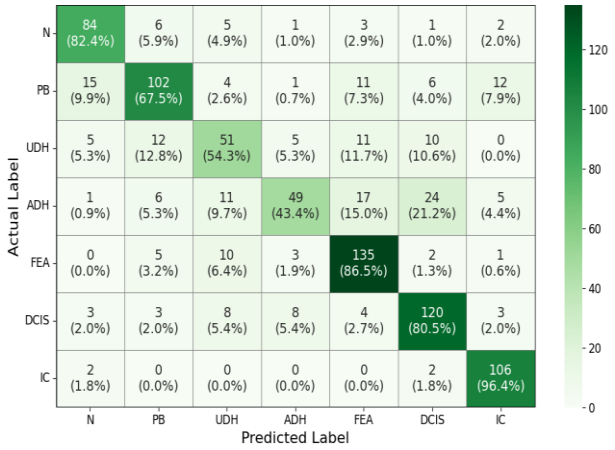

Fig. 4. Confusion matrix for proposed MSHFF-Net

Grad-CAM visualizations of the proposed MSHFF-Net on the BRACS dataset are depicted in Fig. 5. The original histopathological images (left) and the equivalent Grad-CAM heatmaps (right) are presented in each pair, emphasizing discriminative regions that affect model's predictions. From left to right and top to bottom, the samples correspond to each of the seven diagnostic grades: Normal, PB, UDH, ADH, FEA, DCIS, and IC. The MSHFF-Net's emphasis on crucial elements like cell nuclei, architectural distortions, and tissue boundaries is demonstrated by these saliency maps, revealing spots like tumor regions, cellular structures, and tissue morphological irregularities. These components are essential for precise classification and well-informed decision-making.

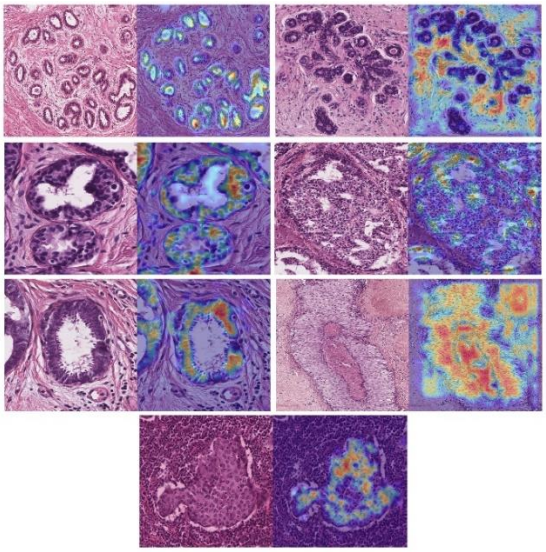

Fig. 5. Grad-CAM results of the proposed MSHFF-Net

C. Comparative Analysis for Multi-Class Classification

Table II summarizes the average precision, recall, F1-score, accuracy, and number of parameters for the CNN models trained on the BRACS dataset. Despite having a lot of parameters, the conventional architectures VGG16 and AlexNet possess poor discriminatory power and yielded accuracy of only about 43%. Similarly, MobileNet records performance metrics in the low 40s, with precision and accuracy at 43.78% and 44.57%, respectively, despite being significantly lightweight at 4.2M parameters. With average F1-Scores of 46.10% and 43.49%, respectively, EfficientNet and XceptionNet likewise fail to achieve significant gains.

TABLE II. COMPARATIVE ANALYSIS OF MULTIPLE CNN MODELS

CNN backbone	Avg. Pr. (%)	Avg. Rec. (%)	Avg. F1-Sc. (%)	Acc. (%)	Parameters (M)
VGG-16	41.62	42.51	41.81	43.42	138
AlexNet	49.22	43.67	40.52	43.54	60
MobileNet	43.78	45.27	43.22	44.57	4.2
EfficientNet	47.26	47.08	46.10	47.77	5.3
XceptionNet	47.83	44.63	43.49	48.11	22.8
ResNet-50	53.93	47.44	44.67	49.02	23
ResNet-18	48.15	49.88	48.44	52.00	11.44
Inception-V3	64.42	61.05	60.84	61.71	23.9
DenseNet-101	69.24	69.32	68.14	70.40	20
Proposed Model	**73.20**	**72.99**	**72.32**	**73.94**	**8.7**

Proceeding on to more contemporary architectures, ResNet-50 and ResNet-18 demonstrate modest gains with accuracy of 49.02% and 52.00%, respectively, with similar Recall and F1 metrics. Nevertheless, the inability of both architectures to capture the finer details needed for this intricate classification task continues to be a challenge.

With an accuracy of 61.71% and an average F1-score of 60.84%, Inception-V3 performs more proficiently, and with an accuracy of 70.40% and an average F1-score of 68.14%, DenseNet-101 further enhances the outcomes. These findings demonstrate how fine-grained features are preserved across layers with through dense connections.

By using only 8.7 million parameters, our proposed MSHFF-Net achieves an average precision of 73.20 %, recall of 72.99%, F1-score of 72.32%, and overall accuracy of 73.94%. These findings demonstrate that MSHFF-Net maintains computational efficiency while providing better diagnostic performance. This shows that careful architectural design can significantly increase prediction accuracy without making the model complex. Compared to larger models, MSHFF-Net is a promising choice for applications that demand quick processing and resource-efficient deployment.

V. CONCLUSIONS AND FUTURE WORK

This word introduces the explainable multi-scale CNN, MSHFF-Net, that can be used for multi-class grading task for breast cancer diagnosis. To characterize fine cellular architecture and global tissue-scale patterns, the architecture incorporates adaptive multi-scale blocks and a hierarchal feature fusion stem. Flow of the gradient is kept constant through Leaky ReLU activation function, and the overfitting is minimized through the usage of GAP layer. MSHFF-Net achieves high diagnostic performance with a total accuracy of

73.94%, while containing only 8.7 million trainable parameters. It performs better than several cutting-edge CNN backbones evaluated under same circumstances. Grad-CAM is utilized to visualize the discriminative regions that influence the model's predictions in order to enhance interpretability. These visual explanations increase the transparency and strengthen the model's suitability for clinical decision support.

In future work, we plan to expand the MSHFF-Net in several directions. To improve localization and classification performance, we will add region-level supervision, which will enable the model to learn from fine-grained lesion annotations. In order to better capture the complete diagnostic context, our second goal is to investigate multi-modal imaging pipelines that combine histopathology with supplementary data, like radiological images or clinical metadata. Third, in order to improve long-range contextual reasoning and scalability to larger WSI, we will move towards the hybrid architectures that combine transformer-based components with the proposed multi-scale CNN backbone. Together, these approaches aim to advance MSHFF-Net into a strong, reliable, and effective framework for digital pathology's computer-assisted breast cancer diagnosis.

REFERENCES

[1] World Health Organization, "Breast cancer," Article. Accessed: May 05, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[2] M. Tafavvoghi *et al.*, "Deep learning-based classification of breast cancer molecular subtypes from H&E whole-slide images," *J Pathol Inform*, vol. 16, Jan. 2025, doi: 10.1016/j.jpi.2024.100410.

[3] M. H. Ashraf and H. Alghamdi, "HFF-Net: A hybrid convolutional neural network for diabetic retinopathy screening and grading," *Biomedical Technology*, vol. 8, pp. 50–64, Dec. 2024, doi: 10.1016/J.BMT.2024.09.004.

[4] M. H. Ashraf *et al.*, "HIRD-Net: An Explainable CNN-Based Framework with Attention Mechanism for Diabetic Retinopathy Diagnosis Using CLAHE-D-DoG Enhanced Fundus Images," *Life*, vol. 15, no. 9, p. 1411, Sep. 2025, doi: 10.3390/life15091411.

[5] M. Xie *et al.*, "Multi-resolution consistency semi-supervised active learning framework for histopathology image classification," *Expert Syst Appl*, vol. 259, p. 125266, 2025, doi: https://doi.org/10.1016/j.eswa.2024.125266.

[6] M. Zubair, S. Wang, and N. Ali, "Advanced Approaches to Breast Cancer Classification and Diagnosis," *Front Pharmacol*, vol. Volume 11-2020, 2021, doi: 10.3389/fphar.2020.632079.

[7] S. Singh *et al.*, "Hybrid Models for Breast Cancer Detection via Transfer Learning Technique," *Computers, Materials and Continua*, vol. 74, no. 2, pp. 3063–3083, 2023, doi: 10.32604/cmc.2023.032363.

[8] M. E. Qureshi, M. H. Ashraf, M. W. Arshad, A. Khan, H. Ali, and Z. U. Abdeen, "Hierarchical Feature Fusion With Inception V3 for Multiclass Plant Disease Classification," *Informatica*, vol. 49, no. 27, Jul. 2025, doi: 10.31449/inf.v49i27.8208.

[9] Y. Gai, J. Hao, Y. Liu, and M. Li, "Cancer diagnosis in smart healthcare: Optimization of the MamCancerX model's multiple instance learning framework," *Alexandria Engineering Journal*, vol. 125, pp. 566–574, Jun. 2025, doi: 10.1016/j.aej.2025.03.103.

[10] J. Zhao, Z. Zhao, X. Song, and S. Sun, "Multiple instance learning with hierarchical discrimination and smoothing attention for histopathological diagnosis," *Applied Intelligence*, vol. 55, no. 6, Apr. 2025, doi: 10.1007/s10489-025-06300-z.

[11] N. Hernandez, F. Carrillo-Perez, F. M. Ortuño, I. Rojas, and O. Valenzuela, "Understanding the Impact of Deep Learning Model Parameters on Breast Cancer Histopathological Classification Using ANOVA," *Cancers (Basel)*, vol. 17, no. 9, p. 1425, Apr. 2025, doi: 10.3390/cancers17091425.

[12] T. Kausar, Y. Lu, and A. Kausar, "Breast Cancer Diagnosis using Lightweight Deep Convolution Neural Network Model", doi: 10.1109/ACCESS.2017.Doi.

[13] L. Li, M. Xu, S. Chen, and B. Mu, "An adaptive feature fusion framework of CNN and GNN for histopathology images classification," *Computers and Electrical Engineering*, vol. 123, Apr. 2025, doi: 10.1016/j.compeleceng.2025.110186.

[14] J. S. Kim *et al.*, "Predicting Nottingham grade in breast cancer digital pathology using a foundation model," *Breast Cancer Research*, vol. 27, no. 1, Dec. 2025, doi: 10.1186/s13058-025-02019-4.

[15] H. Güler, Y. Santur, and M. Ulaş, "Early Diagnosis of Invasive Ductal Carcinoma Breast Cancer using Deep Learning Framework," *Turkish Journal of Science and Technology*, vol. 20, no. 1, pp. 29–40, Mar. 2025, doi: 10.55525/tjst.1483617.

[16] N. Brancati *et al.*, "BRACS: A Dataset for BReAst Carcinoma Subtyping in H&E Histology Images," *Database*, vol. 2022, 2022, doi: 10.1093/database/baac093.

[17] M. H. Ashraf, M. E. Qureshi, A. Khan, and M. Ahmed, "DRD-Net: Diabetic Retinopathy Diagnosis Using A Hybrid Convolutional Neural Network," *International Journal on Robotics, Automation and Sciences*, vol. 7, no. 2, p. 96, 2025, doi: 10.33093/ijoras.2025.7.2.9.

[18] X. L. Pan *et al.*, "EL-CNN: An enhanced lightweight classification method for colorectal cancer histopathological images," *Biomed Signal Process Control*, vol. 100, p. 106933, Feb. 2025, doi: 10.1016/J.BSPC.2024.106933.

[19] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization," 2017. Accessed: Aug. 14, 2025. [Online]. Available: http://gradcam.cloudcv.org