Received 08 November, accepted 03 August 2026, date of publication 30 September 2026
www.computingonline.net / computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.25.3.4843

Explainable Histology Images Classification Using an Ensemble of Fractal Geometry, CNNs, and ViT

MOHEB R. GIRGIS, AL HUSSIEN SEDDIK SAAD, MOHAMMED M. TALAAT
Department of Computer Science, Faculty of Science, Minia University, El-Minia, Egypt

Corresponding author: Moheb R. Girgis (e-mail: moheb.girgis@mu.edu.eg).

ABSTRACT Accurately classifying histopathological images is still a fundamental but difficult task in medical diagnostics, because biological tissues are irregular, self-similar, and multi-scale. While effective in tasks involving natural images, convolutional neural networks (CNNs) show limited generalization when confronted with highly heterogeneous tissue structures, small datasets, and variable resolutions. In this paper, we propose a tri-branch ensemble model, called Fractal-CNN-ViT, for histology images classification, which combines the advantages of Vision Transformers (ViTs), CNNs, and fractal geometry. In this model, fractal descriptors—specifically fractal dimension, lacunarity, and percolation—are extracted from histology images, transformed into artificial RGB images, and then processed by a specialized CNN (Fractal-CNN). In order to capture global contextual relationships, a second CNN (ResNet-50) and a ViT simultaneously analyze the original images. The outputs of the three branches are fused through an adaptive weighted ensemble module, which maps the optimal contribution of each stream to the final decision. Experimental results on the BreakHis dataset demonstrated that Fractal-CNN-ViT outperforms standalone deep learning architectures: Fractal-CNN, CNN (ResNet-50) and ViT. It achieved an accuracy of 92.3%, surpassing Fractal-CNN (85.9%), ResNet-50 (87.2%) and ViT (88.5%), confirming the benefits of integrating fractal-based features with deep learning methods. These findings demonstrate how well the proposed ensemble model combines local and global tissue representations, reduces overfitting on small datasets, and generalizes across magnification levels. Finally, Grad-CAM visualization was employed to provide visual interpretability, increasing clinical trust by highlighting image regions most influential to the studied models' classification decisions.

KEYWORDS Histology Image Classification, Fractal Geometry, Convolutional Neural Networks, Ensemble Learning, Deep Learning, Vision Transformers, Grad-CAM Visualization.

I. INTRODUCTION

Classifying histopathological images is essential for cancer diagnosis, prognosis, and treatment planning. Automated techniques are being developed to help pathologists interpret high-resolution tissue images as a result of the growing digitization of pathology slides and the availability of public datasets. The analysis of histology images is computationally complex due to their high levels of heterogeneity, irregularity, and self-similarity [1, 2]. The differences in staining, scale (magnification), and the small size of annotated datasets in actual clinical practice, further compound these difficulties [3].

In both natural and medical image processing tasks, Convolutional Neural Networks (CNNs) have demonstrated exceptional performance [4]. CNNs have been used for histopathological classification in a number of studies. For example, Spanhol et al. [5] used the BreakHis dataset to train deep models on images of breast cancer, and Araujo et al. [6] used residual CNNs for whole-slide classification. But because of local receptive fields and a lack of contextual awareness, these methods frequently have trouble generalizing, particularly when there is a class imbalance or different levels of magnification [7, 8].

Recent studies have concentrated on improving deep models with manually created or domain-specific features in order to get around these restrictions. The intricate, scale-invariant structures found in biological tissues can be mathematically modeled using fractal geometry [9]. It has been demonstrated that characteristics like fractal dimension, lacunarity, and percolation rate enhance the description of histology textures. Fractal features alone were used in some studies for classification; however, these have limited scalability and lack hierarchical abstraction.

Hybrid models have been proposed that combine the high representation power of CNNs with the interpretability of fractal descriptors. For example, Roberto et al. [10] have

proposed a double-CNN model with one branch for raw images and another for synthetic RGB images derived from fractal features. However, there was no mechanism for modeling global context in the fusion method, which was merely a sum-rule.

However, through self-attention mechanisms, Vision Transformers (ViTs) have recently become powerful models in image classification that can capture global relationships [11]. Although they usually require large datasets and might miss low-level spatial textures that CNNs are excellent at detecting, their use in medical imaging is expanding [12].

Beyond fixed-size patch embedding, recent work on edge-based semantic segmentation has shown that adaptive patch reshaping, guided by multi-scale feature-variance analysis, helps ensure that important object features and boundary information are not overlooked when images are split into patches, which is a relevant consideration for any patch-based model such as a ViT [28].

In this paper, we propose a tri-branch ensemble model, called Fractal-CNN-ViT, for the classification of histology images, which combines the benefits of ViTs, CNNs, and handcrafted fractal features. In the proposed architecture:

- A CNN trained on the original histopathological image,
- a second CNN trained on an RGB image derived from fractal dimension, lacunarity, and percolation metrics, and
- a ViT model pretrained on large-scale data that extracts global attention features from the original image.
- A weighted ensemble strategy that was optimized during training is used to fuse the outputs of all three branches.

Unlike earlier methods, this structure enables our model to provide a more comprehensive view of tissue structure by successfully integrating local textures, handcrafted complexity, and global context.

Furthermore, the proposed approach works especially well in real-world clinical settings with small datasets, high stain variability, and magnification. For instance, examining slides at various scales to find structural abnormalities is frequently necessary for the diagnosis of breast cancer [3]. In order to tackle these problems, the proposed model combines transfer learning-based representations with multi-scale fractal features.

The key contributions of this research are:

- Introducing a fractal-based feature extraction technique that enhances hierarchical texture representation in histology images.
- Integrating CNNs and ViTs to exploit their complementary strengths in local and global feature extraction.
- Empirical validation on the BreakHis dataset, proving the effectiveness of the proposed ensemble compared to standalone deep learning architectures: Fractal-CNN, CNN (ResNet-50) and ViT.
- Utilizing explainable AI (XAI) for interpretability: Deployment of Grad-CAM (Gradient-weighted Class Activation Mapping) [13] for visual interpretability, increasing clinical trust by highlighting regions most influential to the studied models' classification decisions

The rest of this paper is organized as follows: Section 2 gives a review of relevant literature; Section 3 provides a technical background; Section 4 presents the proposed methodology; Section 5 presents the experiments that were conducted to evaluate the performance of the proposed ensemble model; Section 6 gives the conclusion of the work presented in this paper.

II. RELATED WORK

In medical diagnostics, classifying histopathological images is essential, particularly for conditions like breast cancer. Many methods have been used, from deep learning models and hybrid frameworks to texture-based analysis. Below, we classify and examine these techniques, focusing on how they are specifically used in medical image analysis.

A. FRACTAL GEOMETRY IN MEDICAL IMAGE ANALYSIS

Complexity, irregularity, and spatial self-similarity—all important traits in pathological tissues—are captured by fractal geometry, which offers quantitative measurements of tissue morphology that can be interpreted.

Based on spatial complexity, fractal dimension (FD) is used to distinguish between benign and malignant tumors. For example, Roberto et al. [14] used FD on lymphoma slides to measure tumor architecture. Lacunarity (Λ) is used to measure texture heterogeneity. For example, Ivanovici and Richard [15] classified biopsy images into various tissue types using lacunarity. Roberto et al. [14] used the percolation theory to quantify the porosity and cluster connectivity in cancerous tissues, exposing variations in the spatial distribution of aberrant cells. In order to feed these descriptors into classifiers, they are usually taken from grayscale images and converted into feature vectors or artificial RGB images

B. CONVOLUTIONAL NEURAL NETWORKS IN HISTOPATHOLOGY

CNNs are now the industry standard for directly learning features from unprocessed histology images. Araujo et al. [6] used a patch-wise aggregation technique to train CNNs on breast histology patches in order to categorize images into four tumor classes. They demonstrated that deep features perform better than conventional handcrafted ones. To enhance generalization across variable image scales, Bayramoglu et al. [16] proposed a CNN that is independent of magnification and was trained on breast histology slides at various resolutions. Spanhol et al. [5] developed the BreakHis dataset and used full-resolution histopathological scans to show how CNNs can reliably distinguish between benign and malignant tissues. Talo [17] proposed deep CNN models using transfer learning technique for the classification of histopathology images. The author used two well-known pre-trained CNN models, ResNet-50 and DenseNet-161, as deep models, and tested them on color and grayscale histopathology images.

C. VISION TRANSFORMERS

In order to capture spatial relationships in histological structures, Dosovitskiy et al. [11] developed the Vision Transformer (ViT), which divides images into patches and applies transformer encoders. Transformers, which use self-attention to model global context, have recently become strong substitutes for CNNs in vision tasks. ViTs are utilized in histopathology to process Whole Slide Images (WSIs) by dividing them into patch embeddings, which allows for the capture of contextual and cellular tissue information. In their

review of transformer applications in medical imaging, Khan et al. [12] demonstrated that ViTs can perform better than CNNs in tasks requiring global structural awareness, like tumor segmentation and large tissue section classification.

D. HYBRID CNN-FRACTAL MODELS IN HISTOLOGY

The goal of hybrid models is to combine the high representation power of CNNs with the interpretability of fractal descriptors. Roberto et al. [10] proposed an approach to classify histological images based on an ensemble of two CNNs, one of them processes the original image, while the other one processes artificial images generated from fractal features. Although their findings showed promising results across several datasets, the approach was based on static fusion (sum rule) and lacked mechanisms to capture long-range dependencies, which are essential for tissue organization.

The majority of hybrid models are limited in their capacity to learn interdependencies between modalities due to their lack of context modeling and reliance on basic fusion mechanisms (such as summation). By introducing global reasoning via ViTs, multi-view learning via dual CNNs, and fractal-aware representation learning, the proposed model overcomes significant shortcomings in earlier methods. Moreover, our dynamic fusion module assigns weights to each branch adaptively, leading to robust performance across magnifications and tissue types.

III. TECHNICAL BACKGROUND

This section outlines the main ideas and mathematical instruments utilized in the development of the proposed Fractal-CNN-ViT model. We go over the fundamentals of fractal geometry, the conversion of numerical features into artificial images, CNNs, and ViTs.

A. FRACTAL GEOMETRY

Fractal geometry was first presented by Benoît Mandelbrot [9]. It offers a mathematical framework to describe the complexity and self-similarity of biological tissues. Fractal geometry captures the fragmented and recursive nature of objects that are difficult to model using traditional shapes, in contrast to Euclidean geometry, which assumes regular, smooth, and integer-dimensional structures. There is a growing use of fractal analysis as a mathematically sound and biologically significant method for measuring diagnostic characteristics in histological image analysis.

Several studies have utilized fractal descriptors to quantify tissue irregularity and texture granularity. The most relevant fractal descriptors are the fractal dimension (FD), lacunarity (Λ) and percolation (Pc), which are described below.

A.1 FRACTAL DIMENSION

Fractal dimension (FD) estimation has been extensively studied as a tool for analyzing self-similar structures in medical imaging, texture classification, and pattern recognition. Traditional fractal dimension estimation methods, such as Box-Counting and Differential Box-Counting, have been widely used due to their simplicity and efficiency. The Box-Counting method [9] is frequently used to calculate the FD, which measures how the number of boxes $N(\varepsilon)$ needed to cover a structure scale with the box size ε. Based on the relationship between box counts $N(\epsilon)$ and ϵ, the FD can be estimated by using Eq. (1).

$$FD = \lim_{\epsilon \to 0} \frac{\log\big(N(\epsilon)\big)}{\log(1/\epsilon)} \quad . \tag{1}$$

A.2 LACUNARITY

Lacunarity (Λ) is a very useful metric for the multi-scale analysis of the images that exhibit fractal properties. It is a statistical descriptor that enhances the FD as it measures the distribution of voids, gaps, or "texture holes" in a spatial pattern. Lacunarity captures the difference between two structures that may have the same fractal dimension but have very different visual textures [18, 19]. It is a measure of gappiness or heterogeneity of the texture. A low lacunarity suggests a more homogeneous, uniform structure, while a high lacunarity denotes a heterogeneous texture with large and unevenly distributed gaps.

Lacunarity is defined as the ratio of the variance (σ^2) to the square of the mean (μ^2) of the pixel mass distribution within moving box windows across the image, as shown in Eq. (2).

$$\Lambda = \frac{\sigma^2}{\mu^2} \quad . \tag{2}$$

Lacunarity has biological significance in the analysis of histological images. Normal and benign tissues, for instance, typically have low lacunarity and regular glandular and nuclear patterns. Malignant tumors, on the other hand, cause this organization to be disrupted, resulting in irregular gaps between cell clusters and nuclei, which raises the lacunarity values [20].

Scale and spatial organization may affect lacunarity. A multi-resolution texture profile can be created by calculating it across several box sizes (scales). Because of this, lacunarity is a perfect feature for capturing architectural heterogeneity in histopathology, particularly when combined with fractal dimension in classification tasks.

A.3 PERCOLATION

Percolation is a statistical physics-based idea that was first applied to simulate fluid flow through porous materials. The idea is modified for use in image analysis to examine the spatial organization and connectivity of pixel clusters in binary or segmented images. When a continuous path (or connected cluster) crosses the image, usually from top to bottom or left to right, the structure is said to "percolate." In histological images, where aberrant cellular growth frequently results in increased connectivity and structural invasion, this is particularly pertinent [14].

In percolation-based image analysis, the input image is first binarized, i.e. RGB values are converted to binary values, using a threshold Δ, and then separated into L×L boxes. A cluster labelling algorithm is used to identify clusters of connected pixels labelled as 1 within each box β_i, as shown in Fig. 1. The following features can be obtained from these labelled maps: [10]

1. The average number of clusters per box, C(L), is obtained by Eq. (3):

$$C(L) = \frac{\sum_{i=1}^{T(L)} c_i}{T(L)} \quad , \tag{3}$$

where T(L) is the total number of boxes of size L and c_i is the number of clusters in box β_i.

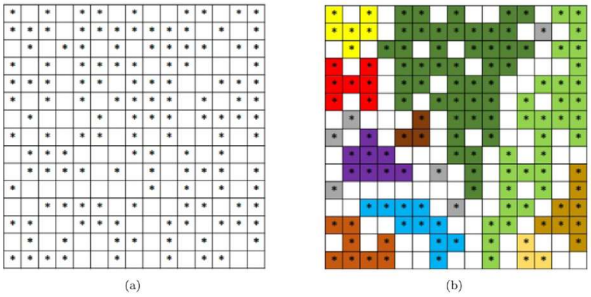

Figure 1. Matrix before (a) and after (b) the application of a cluster labelling algorithm, where the symbol * indicates pixels labelled as 1 [10].

2. The average coverage of the largest cluster per box, M(L), is obtained by Eq. (4):

$$M(L) = \frac{\sum_{i=1}^{T(L)} \frac{m_i}{L^2}}{T(L)}, \quad (4)$$

where m_i is the pixel count of the largest cluster in box β_i.

3. Binary percolation indicator q_i:
A box is considered percolating if the fraction of pixels labeled 1 exceeds the percolation threshold (≈ 0.59275 for 2D square lattices [21]). Let Ω_i be the number of pixels labelled as 1 in box βi with size $L \times L$, such box can be determined whether it is percolating or not according to Eq. (5):

$$q_i = \begin{cases} 1, & if \ \frac{\Omega_i}{L^2} \geq 0.59275 \\ 0, & if \ \frac{\Omega_i}{L^2} < 0.59275 \end{cases} \quad (5)$$

where $q_i = 1$ indicates that the box is percolating.

4. Percolation rate Q(L):
Percolation rate, Q(L), is the average occurrence of percolation on a scale L, and can be obtained by Eq. (6):

$$Q(L) = \frac{\sum_{i=1}^{T(L)} q_i}{T(L)} \quad (6)$$

This feature captures the proportion of boxes that are percolating at a given scale L, providing a robust measurement of local connectivity.

The fractal dimension (FD) features are used as a measure for evaluating the irregularity and the complexity of a fractal. On the other hand, the lacunarity (Λ) features are used as a measure for evaluating how the space is filled. Features related to percolation theory may complement such measures, mainly when providing a greater distinguishing capacity of vascular structures on histological images [22]. Percolation theory can aid in image quantification and classification within diagnostic support systems by providing insights into potential cluster characteristics present in the images. [14] These descriptors are typically converted into features vector or reshaped into artificial RGB images to be compatible with deep learning models.

B. CONVOLUTIONAL NEURAL NETWORKS (CNNS)

CNN is a type of deep learning model that automatically and adaptively learn spatial hierarchies of features from low- to high-level patterns on grid-shaped data, such as images. CNNs are especially good at simulating the intricate, recurring, and locally correlated structures of nuclear and glandular tissue patterns in histopathology. CNN is typically composed of three types of layers: convolution, pooling, and fully connected layers, as shown in Fig. 2. Convolution and pooling layers are responsible for extracting features, while the fully connected layer translates these extracted features into a final output, such as a classification result. [4]

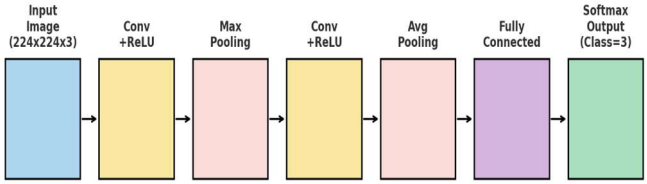

Figure 2. An example of a CNN pipeline.

As illustrated in Fig. 2, a CNN combines three standard layer types: convolutional layers, which extract features using learnable kernels followed by a non-linear activation (typically ReLU) [23]; pooling layers, which downsample the feature maps (via max or average pooling) to reduce dimensionality and add translation invariance, which is particularly useful in histology images where structures vary slightly between samples; and fully connected layers, which combine the extracted features into a final class prediction via a softmax output. As these operations are standard and well documented in the literature, their detailed mechanics are not repeated here.

Despite their effectiveness, CNNs suffer from local receptive field constraints and struggle to generalize over global spatial patterns, especially in histology images with multiscale textures [1, 8].

C. VISION TRANSFORMERS (VITS)

Vision Transformers (ViTs) symbolize a change in thinking about visual representation learning. ViTs, which were first presented by Dosovitskiy et al. [11], modify the transformer architecture, which has been widely used in natural language processing, to process 2D image data. ViTs employ self-attention mechanisms to capture global relationships within an image, whereas CNNs depend on local receptive fields and convolutional kernels. This allows them to capture contextual relationships and long-range interactions, which are crucial in complex visual data like histological tissue. ViTs offer scale-independent representation, global tissue context understanding, and long-range feature modeling.

C.1 ViTs Architectural Overview

Vision Transformers adopt the standard transformer architecture originally designed for 1D text sequences. As shown in Fig. 3, the ViT model for processing images includes the following steps [11]:

- *Split the image into image patches.* The 2D image is split into fixed-size patches, for example, P×P pixels each, which are then flattened into vectors. If the image has dimensions HxW with C channels, the total number

of patches is N = HxW / PxP, which is the effective input sequence length for the Transformer.

- *Get initial patch embeddings.* The flattened patches are passed through a linear projection layer, mapping them into a fixed-dimensional space DD, referred to as patch embeddings.
- *Prepend trainable "class" embedding to patch embeddings*. A special learnable token, analogous to the [CLS] token in BERT, is inserted at the start of the patch embedding sequence.
- This token encodes a global representation of the image, which is later used for classification.
- *Add learned positional embeddings to patch embeddings*. Positional embeddings are incorporated into the patch embeddings to encode positional information, enabling the model to capture the spatial structure of the image.
- *Pass the sequence of embeddings through the Transformer encoder*. The Transformer encoder is composed of alternating layers of Multi-Headed Self-Attention (MSA) and a Multi-Layer Perceptron (MLP) block, which functions as a feedforward neural network. In each layer, Layer Normalization (LN) is applied before these operations, and residual connections are added afterward to enhance training stability. The encoder's output—namely, the final state of the [CLS] token—acts as the representation of the image.
- For classification tasks, attach a simple prediction head to the final [CLS] token. This classification head is implemented by an MLP with one hidden layer at pre-training time and by a single linear layer at fine-tuning time.

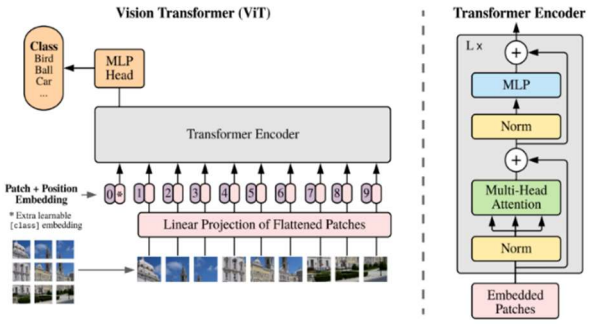

Figure 3. An overview of Vision Transformer (on the left) and the details of Transformer encoder (on the right) [11].

This architecture enables ViTs to capture global relationships among patches and fully leverage self-attention for image understanding.

While ViTs excel at capturing long-range dependencies, they may lack sensitivity to low-level texture features unless combined with CNNs or domain-specific features [11, 12].

IV. METHODOLOGY

A. CALSSIFICATION APPROACH OVERVIEW

In order to improve histopathological image classification, the proposed approach utilizes a tri-branch ensemble framework called Fractal-CNN-ViT. This framework is specifically designed to use both handcrafted fractal features and deep hierarchical representations. In contrast to traditional CNN-based models or dual-branch hybrids, the proposed architecture incorporates biologically-informed fractal descriptors and combines the global contextual modeling power of ViTs with the discriminative capabilities of CNNs. Fig. 4 shows a flow diagram illustrating the proposed tri-branch ensemble framework.

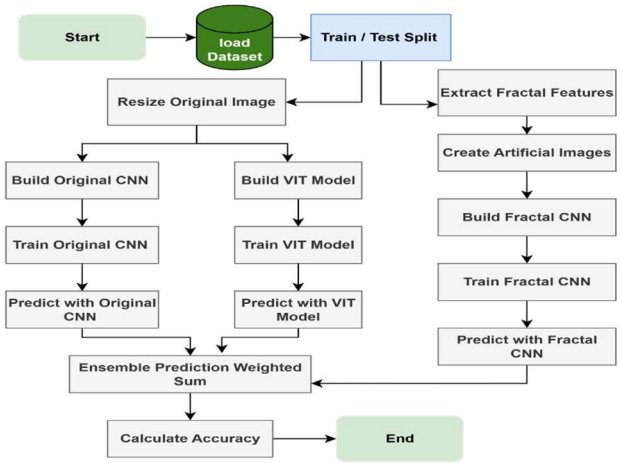

Figure 4. Flow diagram of the proposed tri-branch ensemble framework Fractal-CNN-ViT

The three parallel branches that make up the proposed ensemble framework are:

- A CNN that is trained on the original histopathology image with the goal of extracting spatially local features pertaining to color variations, glandular textures, and nuclei shape.
- A CNN that is trained on an artificial RGB fractal image, which is created from localized maps of fractal dimension (FD), lacunarity (Λ), and percolation rate (Q(L)). These descriptors measure irregularity, connectivity, and texture complexity of histopathological images.
- A Vision Transformer (ViT) stream that has been fine-tuned to model the long-range tissue structure relationships in histopathological slides after being pretrained on natural image datasets, such as ImageNet.

The outputs of the three branches are fused through an adaptive weighted ensemble module, which maps the optimal contribution of each stream to the final decision.

B. FRACTAL FEATURE CONSTRUCTION AND RGB IMAGE GENERATION

Fractal Geometry provides a strong mathematical framework for analyzing the intricate self-similar patterns seen in histopathological images—patterns that are frequently missed by traditional Euclidean descriptors [9, 20]. The proposed approach uses a multiscale fractal-based analysis to extract three key descriptors: Fractal Dimension (FD), Lacunarity (Λ), and Percolation Metrics (Q, C, M). Different facets of the morphological and spatial heterogeneity in tissue regions are captured by these features.

Each RGB image is first converted into grayscale, before the multiscale Gliding-Box algorithm [10] is applied, in order to maintain consistency and minimize computational complexity. To examine local pixel intensity variations, a gliding box, β, of size L × L then moves across the image. Each time the box β is moved, a multidimensional analysis of intensity similarity is performed for every pixel inside it. This

is done by calculating an intensity distance Δ between a central pixel (f_c) and its neighbors (fi) for each box. In the proposed approach, three distance metrics are used: Chessboard Distance Δ_h, Euclidean Distance Δ_e, and Manhattan Distance Δ_m, which are calculated using Eqs. (7) to (9), respectively.

$$\Delta_h = max(|f_i - f_c|), \quad (7)$$

$$\Delta_e = sqrt(\sum(f_i - f_c)^2), \quad (8)$$

$$\Delta_m = \sum|f_i(k) - f_c(k)|. \quad (9)$$

Pixels that satisfy $\Delta \leq L$ are labeled as 1 (active); others are labeled as 0. This binary representation is used to construct a Probability Matrix P(m, L), which captures the probability that m pixels in a box of size L are active.

Using this structure, the following fractal descriptors are computed:

- Fractal Dimension (FD) (Eq. (1)), which captures self-similarity across scales [9, 20].
- Lacunarity (Λ) (Eq. (2)), which measures the distribution and gaps in textures [18; 19].
- Percolation metrics (C, M, and Q) (Eq. (3) to Eq. (6)), which reflect connectivity and cluster structures [22].

In the proposed approach, a sequence of 5 values is assigned to the scale L, and for each value of L, the 5 descriptors ($FD(L)$, $\Lambda(L)$, $C(L)$, $Q(L)$, and $M(L)$) are computed. This results in 25 features for each of the three distances Δ, i.e. a set of 75 features is obtained. Then, the mean, Max, Std, and Min values are calculated for these features, which result in obtaining a set of 300 features. Once extracted, the features are reshaped into a 10×10×3 matrix to create an artificial RGB image, where: R-channel encodes FD values, G-channel encodes Lacunarity, and B-channel encodes Percolation descriptors.

In the proposed ensemble architecture, this artificial RGB image is fed into a specific Fractal CNN branch. This conversion enables deep learning over interpretable representations by enabling traditional CNNs to process handcrafted features as structured visual inputs [10].

C. FUSION STRATEGY AND ENSEMBLE INTEGRATION

Among the most inventive features of the proposed Fractal-CNN-ViT framework is the design of a trainable fusion mechanism that integrates three different but complementary streams of feature representations: (1) the global semantic context from a ViT, (2) the local spatial detail from a CNN trained on original histological images, and (3) the complexity-based structure derived from fractal feature maps fed to a second CNN.

Unlike earlier approaches, such as [10], [24], which employed fixed-rule decision fusion (e.g., averaging or voting), the proposed approach uses a dynamic learnable ensemble fusion. Each model contributes probabilistic output scores (softmax activations), combined as in Eq. (10):

$$\hat{y} = \alpha \cdot y_{CNN_org} + \beta \cdot y_{CNN_{fractal}} + \gamma \cdot y_{ViT}, \quad (10)$$

where α(x), β(x), γ(x) are learnable, sample-dependent ensemble weights produced by a gating network applied to input x, subject to the constraint: α(x) + β(x) + γ(x) = 1 for every input, y_{CNN_org} is the prediction vector resulted from applying CNN on original image, $y_{CNN_fractal}$ is the prediction vector resulted from applying CNN on fractal RGB image, and y_{ViT} is the prediction vector resulted from applying ViT.

Explicitly, a Softmax function is applied to the three raw outputs of the gating layer: this guarantees, by construction, that all three weights are strictly non-negative (Softmax outputs lie in the range (0, 1)) and that they always sum exactly to one, so no additional clipping, normalization, or positivity constraint is required. The criterion that determines the optimal contribution of each stream is the training loss itself: the weights α, β, γ are not fixed, handcrafted, or global constants; they are produced per input image by a small gating network (a linear layer applied to the concatenation of the three branches' intermediate feature representations). This gating network is initialized so that every image receives equal weights (α = β = γ = 1/3) at the start of training, and is then updated jointly with the three branch networks through end-to-end backpropagation, driven solely by minimizing the sparse categorical cross-entropy loss computed on the fused prediction ŷ. Because the gate conditions on each image's own features, the resulting weights are sample-specific rather than a single triple shared across the dataset: the module increases the contribution of whichever branch (Fractal-CNN, CNN, or ViT) most reduces classification error for that particular image, so the "optimal" contribution of each stream is learned in a data-driven and image-specific manner rather than prescribed in advance.

Backpropagation is used to optimize the fusion weights during training, enabling the system to dynamically modify the emphasis on each branch based on tissue subtype or complexity. This approach is similar to multimodal deep learning's attention-weighted fusion [25].

Our adaptive model is more generalizable to changes in image magnification and class imbalance than conventional static fusion techniques. It produces a more resilient ensemble by utilizing the domain-specific fractal descriptors, local detail of CNNs, and global capacity of ViTs.

The architecture of the proposed Fractal-CNN-ViT model is shown in Fig. 5. Through supervised learning, the outputs from every stream are combined in a weighted manner to generate a single class prediction.

D. IMPLEMENTAION DETAILS

The implementation of the proposed framework was done in Python 3.10 using PyTorch 2.0 and Hugging Face Transformers.

D.1 DATASET PREPARTION AND FEATURE EXTRACTION

We utilize the publicly available BreakHis dataset [5], which includes histopathological images of benign and malignant breast tumors at various magnification levels (40×, 100×, 200×, and 400×).

Each input image undergoes the following preprocessing:

- ***Resizing*** to 224×224 pixels for compatibility with CNN/ViT input layers. This resolution was chosen because it is the native input size required by the pre-trained backbones used in the original-image branches, ResNet-50 and ViT-B/16, both of which were pre-trained on ImageNet at 224×224; using their native resolution allows the pre-trained weights to be reused effectively through transfer learning without architectural modification. In addition, since the experiments were carried out on a CPU-based machine

rather than a GPU/TPU (as noted in Section V), keeping the input resolution at this standard, moderate size was necessary to keep training and inference time computationally manageable. The fractal descriptors are unaffected by this choice, as they are computed separately from a fixed sliding-window analysis of the grayscale image rather than from the resized RGB input.

- ***Grayscale conversion*** to facilitate fractal-based feature extraction.
- ***Fractal feature computation*** including:
  - Fractal Dimension (FD**)** using the box-counting method (Eq. 1).
  - Lacunarity (Λ) to capture texture gaps (Eq. 2).
  - Percolation-based metrics: Average clusters C(L) (Eq. 3), maximum coverage M(L) (Eq. 4), and percolation threshold Q(L) (Eq. 6).
- ***Distance transforms*** using Euclidean Δ_e, Chessboard Δ_h, and Manhattan Δ_m metrics.

All features are combined into a 300-dimensional vector, then reshaped into a 10×10×3 synthetic RGB image, which is used as input for the fractal CNN stream.

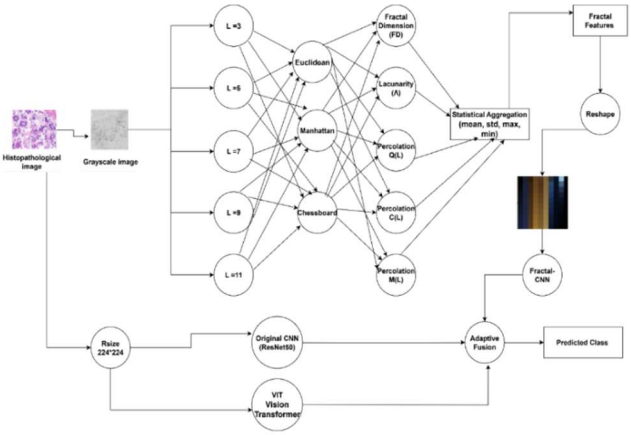

Figure 5. The architecture of the Fractal-CNN-ViT model.

D.2 PROPOSED MULTI-STREAM DEEP LEARNING MODEL (FRACTAL-CNN-ViT) ARCHITECTURE

The proposed framework is composed of three complementary models, as shown in Figure. 5, a CNN trained on real histological images, a CNN trained on artificial RGB images generated from fractal features, and a ViT to capture global contextual cues.

Original CNN Model

This model is based on a pre-trained ResNet-50 architecture. The inputs to this model are the original 224×224 RGB histology images. It consists of the following layers: convolutional blocks → global average pooling → fully connected layers → softmax.

Fractal CNN Model

This model is a custom CNN designed to process the 10×10×3 artificial RGB images created from fractal features. It consists of the following layers:

- 3 Conv2D layers (filters: 32, 64, 128) + ReLU
- 3 MaxPooling layers
- Flatten + Dense (128) → Dropout (0.5) → Dense (Softmax)

Vision Transformer Model

This model is based on pre-trained ViT-B/16 transformer. The inputs to this model are the original 224×224 RGB histology images segmented into 16×16 patches. It consists of the following layers: Processing → Patch embedding → positional encoding → multi-head self-attention → MLP head → softmax.

Fusion and Ensemble Integration

The outputs of the three branches y_{CNN_org}, $y_{CNN_{fractal}}$, and y_{ViT} are combined using a weighted ensemble strategy, where the ensemble output $\hat{y}$, which represents the final class prediction, is computed using Eq. (10).

Model Training and Optimization

- All models are trained using the **AdamW optimizer**.
  - CNNs (Original & Fractal) were trained for 10 epochs
  - ViT: was trained for 5 epochs (due to pretraining)
- Batch size: 32
- Early stopping and learning rate reduction are applied to avoid overfitting.

Table 1 shows a summary of the parameters of the proposed Fractal-CNN-ViT model.

Table 1. Summary of the parameters of the proposed model (Fractal-CNN-ViT)

Parameter	Value
Dataset	BreakHis (Benign and Malignant images)
Fractal Features	FD, Lacunarity, Percolation (Q, C, M) + Distance Transforms ($\Delta_h, \Delta_e, \Delta_m$)
CNN (Original)	ResNet-50 (pre-trained on ImageNet) → Global Average Pooling → Dense(512) + ReLU → Dropout(0.5) → Dense(2) + Softmax
CNN (Fractal)	3 Conv2D layers (32, 64, 128) → 3 MaxPooling → Flatten → Dense (128) → Dropout (0.5) → Dense (2) → Softmax **Input size:** 10×10×3 (artificial image from 300D fractal vector)
ViT	ViT-B/16 (patch size 16×16) → Patch embedding (768D) + Positional encoding → 12 Transformer Blocks (12 heads) → MLP Head → Dense (2) → Softmax **Input size:** 224×224×3 (RGB histology image)
Loss Function	sparse categorical cross-entropy loss
Learning Rate	0.001
Fusion	$\hat{y} = \alpha \cdot y_{CNN_org} + \beta \cdot y_{CNN_{fractal}} + \gamma \cdot y_{ViT}$ Learned per-sample gating network (softmax-weighted sum; α, β, γ vary per image)
Optimizer	AdamW
Batch size	32
Epochs	10 Epochs for the CNN (original + fractal), 5 Epochs for the ViT
Evaluation Metrics	Accuracy, Precision, Recall, F1-score, AUC

The steps of the proposed framework are depicted in Algorithm (Fractal-CNN-ViT Model), shown in Fig. 6.

V. EXPERIMENTS

This section describes the experiments the we have conducted to evaluate the performance of the proposed Fractal-CNN-ViT model. TensorFlow and PyTorch frameworks were used in the experiments, which were carried out on an Intel Core i7 computer running at 2.3GHz.

We have used the BreakHis dataset [5], which consists of 7909 breast cancer histology images with magnifications of 40×, 100×, 200×, and 400×, categorized into benign (2,480) and malignant (5,429) tumor types. The model was trained, validated and tested, using 64% of the dataset for training, 16% for validation, and 20% for testing. In absolute terms, this corresponds to 5,062 images used for training, 1,265 images for validation, and 1,582 images held out for testing, with the class balance (benign/malignant) preserved across the three splits.

Algorithm (Fractal-CNN-ViT Model)

Input: Histopathological image dataset D = $\{(x_i, y_i)\}$
Output: Predicted class label $\hat{y}$

Begin

1. Load and split dataset D into training, validation, and testing sets
2. For each image x_i in the training set:
   - 2.1 Resize RGB image to 224×224 (for CNN/ViT)
   - 2.2 Convert image to grayscale for fractal feature extraction
   - 2.3 Apply sliding window at scales L ∈ {3, 5, 7, 9, 11}
   - 2.4 Compute FD, Lacunarity (Λ), Percolation (Q, C, M)
   - 2.5 Compute distance metrics: Euclidean, Manhattan, Chessboard
   - 2.6 Generate 300D feature vector and reshape it to 10×10×3 RGB image
3. Initialize models:
   - 3.1 Original CNN ← pre-trained ResNet50 (224×224 input)
   - 3.2 Fractal CNN ← adapted CNN for 10×10 input
   - 3.3 ViT ← pre-trained ViT-B/16 (224×224 input)
4. Train each model independently using AdamW optimizer
5. Apply early stopping and learning rate scheduling
6. For each test sample x_i:
   - 6.1 y_{CNN_org} ← prediction from Original CNN
   - 6.2 $y_{CNN_{fractal}}$ ← prediction from Fractal CNN
   - 6.3 y_{ViT} ← prediction from ViT
   - 6.4 Fuse predictions: $\hat{y} = \alpha \cdot y_{CNN_org} + \beta \cdot y_{CNN_{fractal}} + \gamma \cdot y_{ViT}$
7. Evaluate predictions using accuracy, precision, recall, F1-score

End.

Figure 6. The steps of the Fractal-CNN-ViT Model.

A. EVALUATION METRICS

The metrics used to evaluate the model's performance are accuracy, precision, recall, F1-score, and AUC-ROC In the calculation of these metrics a Confusion Matrix is used, which is a table with four alternative combinations of predicted and actual values, as shown in Fig. 7, where: TN (True Negatives) refers to the correct number of classifications of negative instances, TP (True Positives) refers to the correct number of classifications of positive instances, FP (False Positives) refers to the incorrect number of classifications of negative instances, and FN (False Negatives) refers to the incorrect number of classifications of positive instances. It shows how well the proposed model classifies benign and malignant cases.

The evaluation metrics used in this study — accuracy, precision, recall, F1-score, and ROC-AUC — are computed from the confusion-matrix components (TP, TN, FP, FN) shown in Fig. 7 using their standard, widely known definitions. Briefly, accuracy is the proportion of correctly classified instances among all instances; precision is the proportion of predicted positives that are truly positive; recall is the proportion of actual positives that are correctly identified; F1-score is the harmonic mean of precision and recall; and ROC-AUC is the area under the curve obtained by plotting the true positive rate against the false positive rate across classification thresholds [26]. A perfect classifier has AUC = 1, while a random classifier has AUC = 0.5 [27].

	Predicted positive	Predicted negative
Actual positive	TP	FN
Actual negative	FP	TN

Figure. 7 The Confusion Matrix

B. EXPERIMENTAL RESULTS

We have evaluated the performance of the proposed model (Fractal-CNN-ViT) compared to the standalone models: Fractal-CNN, CNN (ResNet-50), and Vision Transformer (ViT), in terms of the accuracy, precision, recall, F1-score, and AUC-ROC metrics. The comparison results are summarized in Table 2. These metrics are computed on the held-out test subset only (1,582 images, 20% of the dataset, as described in Section V.A), following standard practice for reporting generalization performance.

Table 2. Performance Comparison of Different Models

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC
CNN (ResNet-50)	87.2	85.6	84.9	85.2	0.91
Fractal CNN	85.9	84.3	83.8	84.0	0.90
Vision Transformer (ViT)	88.5	86.7	86.2	86.4	0.92
Proposed (Fractal-CNN-ViT)	**92.3**	**91.1**	**90.8**	**90.9**	**0.95**

It can be seen from Table 2 that:

- The Fractal-CNN-ViT model outperforms Fractal-CNN, CNN (ResNet-50), and ViT, achieving the highest scores for all metrics on the BreakHis dataset.
- Achieving F1-score of 90.9% indicates the balanced performance of Fractal-CNN-ViT model between precision and recall, reducing the risk of false positives and false negatives.
- The high AUC-ROC value of 0.95 achieved by the proposed model confirms its ability to distinguish between different histological classes with high confidence.
- Roberto et al. [10] reported a maximum classification accuracy of 89.7% using fractal-based dual CNNs on a different and very small breast cancer dataset, which is UCSB. Although direct comparison is limited due to dataset differences, our model demonstrates superior generalization and robustness across all magnification levels in breast cancer histology images.

Fig. 8 shows the ROC Curves for the proposed model (Fractal-CNN-ViT) and the standalone models: Fractal-CNN, CNN (ResNet-50), and ViT, when applied on the BreakHis Dataset. This figure demonstrates that:

- Fractal-CNN-ViT shows the best performance with AUC = 0.95.

- ViT with AUC = 0.92 performs slightly better than CNN (ResNet-50) with AUC = 0.91, and Fractal-CNN with AUC = 0.90.
- The diagonal line (gray) represents a random classifier (AUC = 0.50), meaning all models perform significantly better.

Fig. 9 and Table 3 shows the confusion matrices, resulted from applying the four models on the BreakHis Dataset. Unlike Table 2, which reports performance on the held-out test subset only, the confusion matrices in Fig. 9 and Table 3 were obtained by applying each trained model to the entire BreakHis dataset (all 7,909 images, spanning the training, validation, and test subsets), in order to give a complete picture of each model's classification behavior across the full range of magnifications and tissue samples. As a result, the accuracy, precision, and recall values implied by Table 3 are not directly comparable to the test-set-only metrics in Table 2, and should not be expected to match them numerically; the recall values are close between the two tables because the malignant-case counts are similar in proportion, but readers should treat Table 2 as the primary measure of generalization performance and Table 3 as a complementary, whole-dataset view of the models' error patterns.

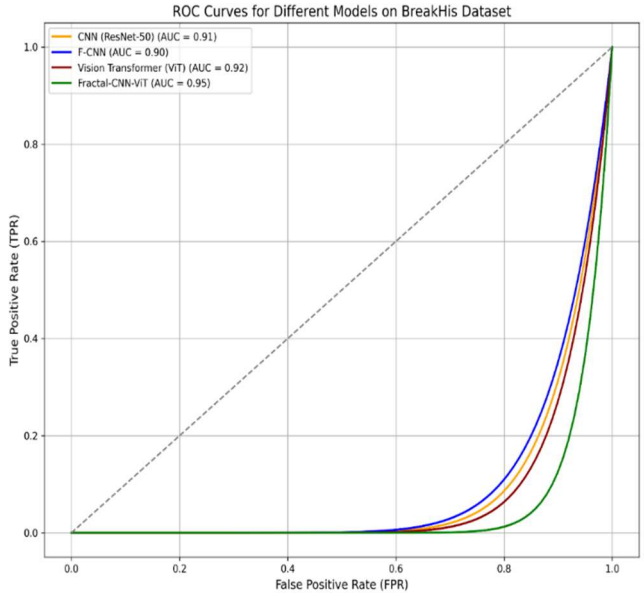

Figure 8. ROC Curves for the four models.

It can be seen from the confusion matrices that:

- Fractal-CNN and CNN (ResNet-50) produce more false negatives and false positives than the other models, meaning they more frequently misclassify malignant cases as benign and vice versa.
- ViT performs slightly better, reducing false negatives (753) and false positives (331).
- Fractal-CNN-ViT is the best model, with the lowest FN (501) and FP (218), indicating fewer misclassifications across the full dataset

These results collectively demonstrate the effectiveness of the Fractal-CNN-ViT ensemble for histology image classification.

Table 3. Confusion matrices for the four models on BreakHis Dataset (full dataset, N = 7,909; see Table 2 for held-out test-set metrics)

Model	TN	FP	FN	TP
Fractal-CNN	2012	468	861	4568
CNN (ResNet-50)	2123	357	814	4615
Vision Transformer (ViT)	2149	331	753	4676
Proposed (Fractal-CNN-ViT)	2262	218	501	4928

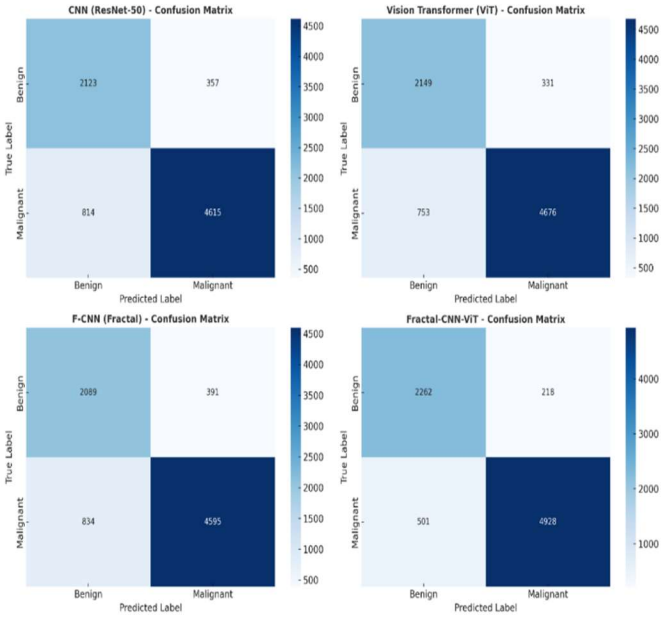

Figure 9. Confusion matrices for the four models

Figs. 10(a & b) to 13(a & b) show the training and validation (loss & accuracy) curves, during the training of Fractal-CNN, CNN (ResNet-50), ViT, and Fractal-CNN-ViT models on the BreakHis Dataset, respectively. Initially, the loss curves show high values and gradually decrease as the training progresses, indicating that the models are improving their performance and making better predictions. Similarly, the accuracy curves begin with low values and increase as the training progresses, indicating that the models are making more correct predictions, thereby enhancing their overall performance.

Despite training each model for only 10 epochs, the loss and accuracy values were tracked at the mini-batch level to capture finer-grained learning behavior. This approach allowed us to better visualize the convergence pattern and stability of training, and observe performance fluctuations.

C. EXPLAINABLE AI (XAI) ANALYSIS FOR ALL MODELS

Gradient-weighted Class Activation Mapping (Grad-CAM) [13] was employed across all classification models to enhance interpretability and highlight the image regions most influential in decision-making. Grad-CAM heatmaps were generated for each branch (CNN, Fractal-CNN, and ViT), then fused to form a unified Grad-CAM visualization for the ensemble model (Fractal-CNN-VIT) to highlight the discriminative regions influencing the model's decision. These visualizations provide qualitative evidence that the model's focus aligns with histopathological structures relevant to benign and malignant tumor classification.

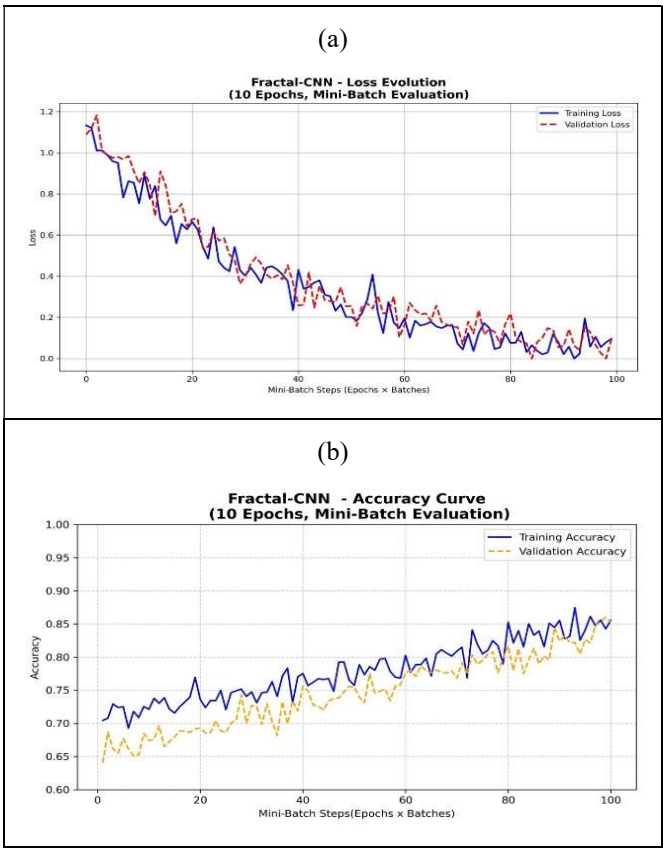

Figure. 10 (a) The training and validation loss curves, (b) the training and validation accuracy curves of the Fractal CNN model.

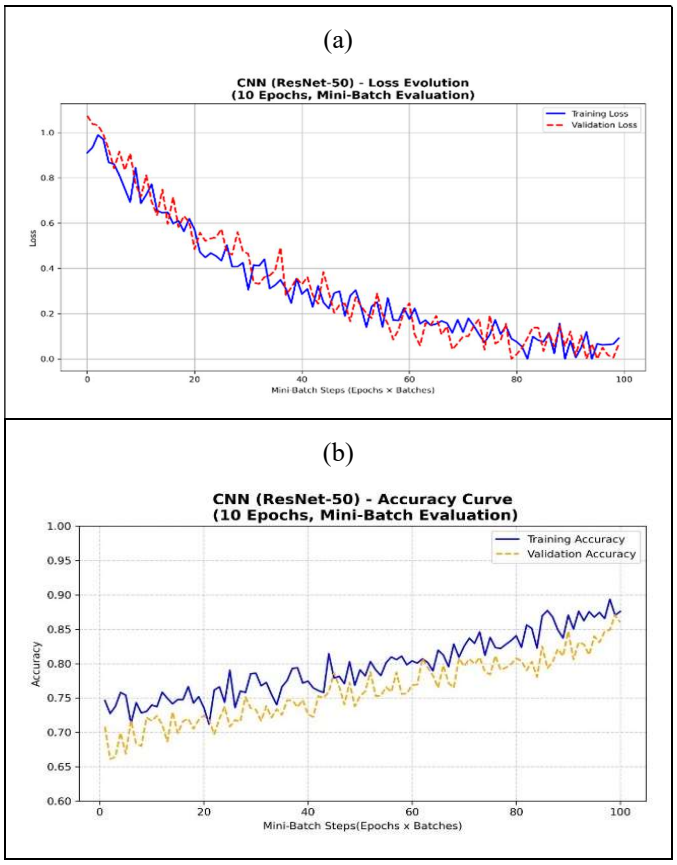

Figure. 11 (a) The training and validation loss curves, (b) the training and validation accuracy curves of the CNN (ResNet-50) model.

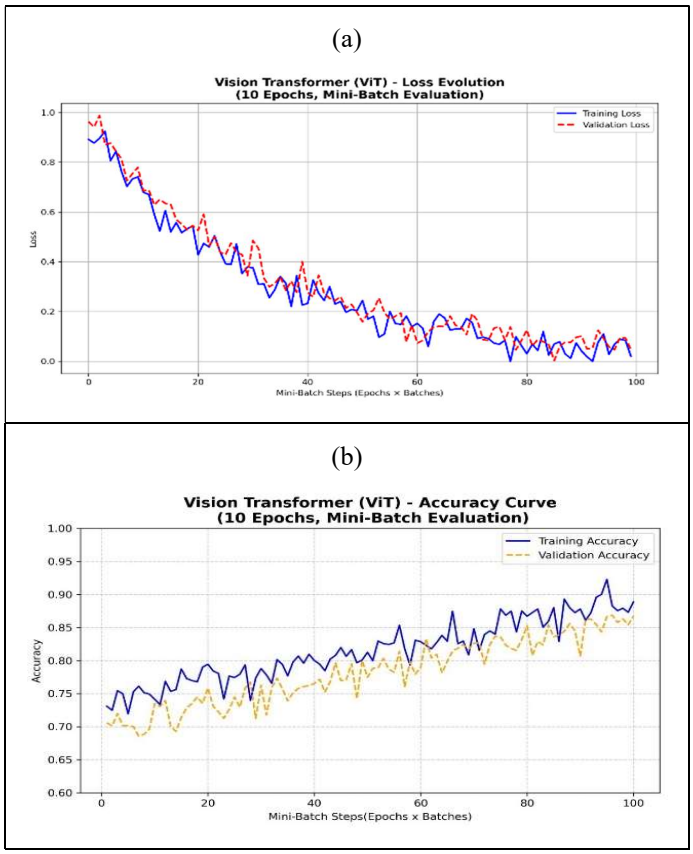

Figure. 12 (a) The training and validation loss curves, (b) the training and validation accuracy curves of the ViT model.

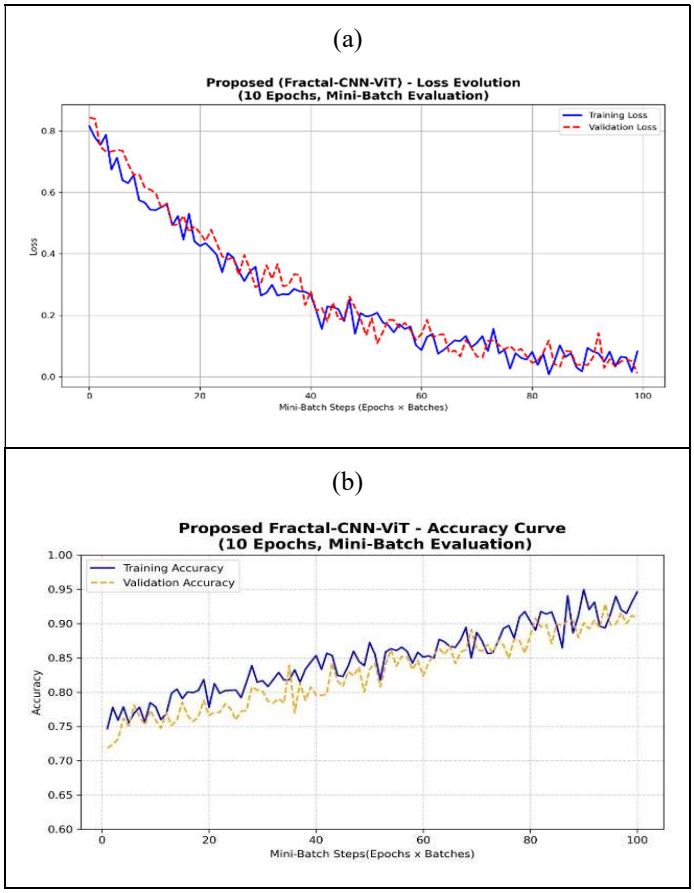

Figure. 13 (a) The training and validation loss curves, (b) the training and validation accuracy curves of the Fractal-CNN-VIT model.

Fig. 14 presents a comparative analysis of Grad-CAM visualizations generated by four deep learning models: CNN, FractalCNN, ViT, and the proposed Fractal-CNN-ViT ensemble. Each model exhibits distinct attention patterns when applied to benign and malignant histology samples, revealing its interpretive focus and decision rationale. It can be seen from this figure that:

- **CNN Grad-CAM** highlights nuclei edges and structural boundaries.
  - **Benign**: Regular and well-defined boundaries suggest organized tissue architecture.
  - **Malignant**: Disrupted boundaries and irregular nuclei indicate architectural distortion.
- **FractalCNN Grad-CAM** emphasizes texture irregularities and lacunarity.
  - **Benign**: Diffuse activation reflects low fractal complexity.
  - **Malignant**: Strong focus on complex textures and branching patterns aligns with increased fractal dimension.
- **ViT Grad-CAM** provides a global attention map across spatial regions.
  - **Benign**: Distributed attention with no dominant region suggests absence of discriminative features.
  - **Malignant**: Concentrated focus on spatially abnormal zones reflects global disorganization.
- **Fractal-CNN-ViT Grad-CAM** fuses local and global cues into a unified attention map.
  - **Benign**: Soft, boundary-focused activation integrates weak signals from all models.
  - **Malignant**: Sharp, centralized activation over structurally and texturally abnormal regions enhances interpretability and diagnostic relevance.

It is worth noting that the Grad-CAM visualizations of the ViT branch appear smoother and less localized, which is consistent with the global self-attention mechanism of ViTs. Unlike CNN-based Grad-CAMs that highlight local activations, ViTs distribute attention across the entire image, resulting in more diffuse but contextually global heatmaps. Overall, the Grad-CAM analysis of Fractal-CNN-ViT demonstrates superior visual clarity and alignment with known histopathological features. It effectively combines the strengths of CNN and ViT, offering both structural precision and contextual awareness.

To make this improvement explicit, the peak-activation region(s) in each Grad-CAM panel of Fig. 14 are now marked with a white ellipse, detected automatically from the heatmap's pixel intensities. On the malignant sample, CNN, Fractal-CNN, and Fractal-CNN-ViT all show one or more clearly localized hot regions, whereas ViT shows no comparably dominant region; on the benign sample, only the ViT panel lacks a marked region, consistent with the "distributed attention with no dominant region" pattern described above for benign tissue. This gives a qualitative, region-level counterpart to the quantitative accuracy gap reported in Table 2, where Fractal-CNN-ViT reaches 92.3% accuracy versus 87.2% for CNN (ResNet-50), corresponding to a 5.1 percentage-point improvement. However, the Grad-CAM examples are qualitative and are not themselves statistical proof of the accuracy improvement; the quantitative performance difference is established separately over the full 1,582-image test set summarized in Table 2.

VI. CONCLUSION

This paper presented a proposed tri-branch ensemble model, called Fractal-CNN-ViT, for histology image classification. The proposed model is composed of three complementary models: a CNN trained on real histological images (CNN (ResNet-50)), a CNN trained on artificial RGB images generated from fractal features (Fractal-CNN), and a Vision Transformer (ViT) to capture global contextual cues. The goal was to enhance feature extraction by combining local texture analysis (CNNs), global contextual understanding (ViT), and multi-scale representation learning (fractal-based feature extraction technique). The outputs of the three branches are fused through an adaptive weighted ensemble module, which maps the optimal contribution of each stream to the final decision.

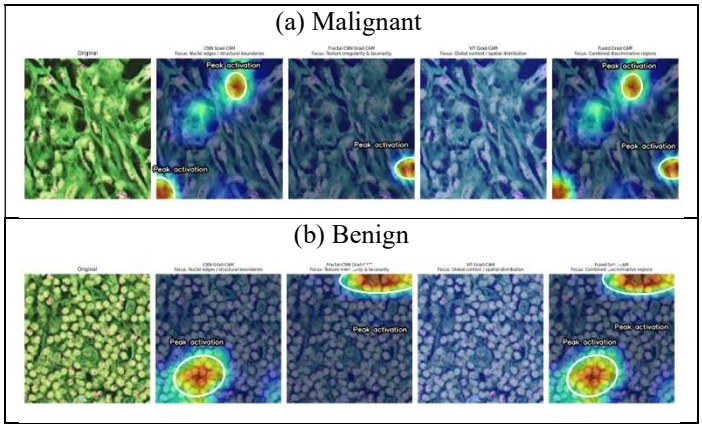

Figure 14. Comparative analysis of Grad-CAM visualizations of CNN, Fractal CNN, ViT, and Fractal-CNN-VIT models applied to (a) malignant and (b) benign tissue samples. White ellipses mark the region(s) of peak activation in each heatmap, identified automatically from the pixel intensities of the Grad-CAM overlay, to make the model's focus explicit; panels with no marked region (e.g., ViT on the benign sample) did not show a clearly dominant activation area.

Experimental results on the BreakHis dataset demonstrated that the proposed Fractal-CNN-ViT ensemble outperforms standalone conventional deep learning architectures: Fractal-CNN, CNN (ResNet-50), and ViT. The proposed model achieved an accuracy of 92.3%, surpassing Fractal-CNN (85.9%), CNN (ResNet-50) (87.2%), and ViT (88.5%), confirming the benefits of integrating fractal-based features with deep learning methods. The model also showed superior precision (91.1%), recall (90.8%), F1-score (90.9%), and AUC-ROC (0.95), ensuring reliable classification performance. These results demonstrate the effectiveness of the Fractal-CNN-ViT ensemble in the classification of histology images.

Moreover, the integration of Grad-CAM visualizations in this study has demonstrated the model's ability to highlight diagnostically relevant regions, enhancing transparency and supporting its potential deployment in clinical decision support systems.

It should be noted that the proposed Fractal-CNN-ViT architecture involves three complementary branches and a trainable fusion module and therefore has higher computational requirements than individual CNN, Fractal-CNN, or ViT models. The experiments in this study were conducted using the training configuration described in the implementation

details, and the selected configuration achieved the reported performance on the held-out test set. The adaptive gating mechanism further enables the contribution of the three branches to be learned automatically from the input features rather than manually specified for each sample. Future work will investigate automated hyperparameter optimization, lightweight backbones, and model-compression techniques to further improve computational efficiency and facilitate deployment in resource-constrained clinical environments.

In the classification of histology images, the proposed Fractal-CNN-ViT ensemble model has shown encouraging results. Nonetheless, a number of topics can be investigated for additional enhancements and expansions:

- Model Optimization: Utilize knowledge distillation, quantization, and pruning to lower computational costs.
- Generalization: Use the model for multi-class classification, MRI, CT scans, and other medical imaging modalities.
- Explainability: Use Grad-CAM, SHAP, and feature attribution techniques, to improve interpretability.
- Multi-Modal Fusion: Integrate histology pictures with molecular, genomic, or clinical textual information.
- Higher-Resolution Inputs: Explore whether input resolutions above 224×224, enabled by GPU/TPU acceleration rather than the CPU-based setup used in this study, further improve diagnostic accuracy, particularly for fine-grained morphological features such as nuclear boundaries.

References

[1] G. Litjens, T. Kooi, B. E. Bejnordi, et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60-88, 2017. https://doi.org/10.1016/j.media.2017.07.005.

[2] A. Janowczyk and A. Madabhushi, "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases," *Journal of Pathology Informatics*, vol. 7, issue 1, 2016. https://doi.org/10.4103/2153-3539.186902.

[3] G. Aresta, T. Araújo, S. Kwok, et al., "BACH: Grand challenge on breast cancer histology images," *Medical Image Analysis*, vol. 56, pp. 122–139, 2019. https://doi.org/10.1016/j.media.2019.05.010.

[4] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into Imaging*, vol. 9, issue 4, pp. 611–629, 2018. https://doi.org/10.1007/s13244-018-0639-9.

[5] F. A. Spanhol, L. S. Oliveira, C. Petitjean and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering*, vol. 63, issue 7, pp. 1455-1462, 2016. https://doi.org/10.1109/TBME.2015.2496264.

[6] T. Araújo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polónia and A. Campilho, "Classification of breast cancer histology images using convolutional neural networks," *PLoS One*, vol. 12, issue 6, 2017. https://doi.org/10.1371/journal.pone.0177544.

[7] A. Madabhushi and G. Lee, "Image analysis and machine learning in digital pathology: Challenges and opportunities," *Medical Image Analysis*, vol. 33, pp. 170-175, 2016. https://doi.org/10.1016/j.media.2016.06.037.

[8] C. Wemmert, J. Weber, F. Feuerhake, and G. Forestier, "Deep learning for histopathological image analysis," in Mourad Elloumi (Ed.), *Deep Learning for Biomedical Data Analysis. Techniques, Approaches, and Applications*, Springer, pp.153-169, 2021. https://doi.org/10.1007/978-3-030-71676-9_7.

[9] B. B. Mandelbrot, *The Fractal Geometry of Nature*, third ed., New York, NY, USA: W. H. Freeman, 1983. https://doi.org/10.1119/1.13295.

[10] G. F. Roberto, F. A. Faria, H. Bragança, A. A. Santos, J. P. Papa, and D. Menotti, "Fractal neural network: A new ensemble of fractal geometry and convolutional neural networks for the classification of histology images," *Expert Systems with Applications*, vol. 166, p. 114103, 2021. https://doi.org/10.1016/j.eswa.2020.114103.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *International Conference on Learning Representations (ICLR)*, Vienna, Austria, May 04 2021.

[12] S. Khan, M. Naseer, M. Hayat, et al., "Transformers in vision: A survey," arXiv:2101.01169v1 [cs.CV] 4 Jan 2021. https://doi.org/10.1145/3505244.

[13] R. R. Selvaraju, M. Cogswell, A. Das, et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision (IJCV)*, vol. 128, pp. 336–359, 2020. https://doi.org/10.1007/s11263-019-01228-7.

[14] G. F. Roberto, L. A. Neves, M. Z. Nascimento, T. A. Tosta, L. C. Longo, A. S. Martins, and P. R. Faria, "Features based on the percolation theory for quantification of non-Hodgkin lymphomas," *Computers in Biology and Medicine*, vol. 91, pp. 135–147, 2017. https://doi.org/10.1016/j.compbiomed.2017.10.012.

[15] M. Ivanovici, and N. Richard, "Fractal dimension of color fractal images," *IEEE Transactions on Image Processing*, vol. 20, issue 1, pp. 227–235, 2011. https://doi.org/10.1109/TIP.2010.2059032.

[16] N. Bayramoglu, J. Kannala, and J. Heikkilä, "Deep learning for magnification independent breast cancer histopathology image classification," *Proceedings of the 23rd International Conference on Pattern Recognition (ICPR)*, 04-08 December 2016, Cancun, Mexico, pp. 2440–2445. https://doi.org/10.1109/ICPR.2016.7900002.

[17] M. Talo, "Automated classification of histopathology images using transfer learning," *Artificial Intelligence in Medicine*, vol. 101, 101743, 2019. https://doi.org/10.1016/j.artmed.2019.101743.

[18] R. E. Plotnick, R. H. Gardner, W. W. Hargrove, K. Prestegaard, and M. Perlmutter, "Lacunarity analysis: A general technique for the analysis of spatial patterns," *Physical Review E*, vol. 53, issue 5, pp. 5461–5468, 1996. https://doi.org/10.1103/PhysRevE.53.5461.

[19] C. Allain and M. Cloitre, "Characterizing the lacunarity of random and deterministic fractal sets," *Physical Review A*, vol. 44, issue 6, pp. 3552–3558, 1991. https://doi.org/10.1103/PhysRevA.44.3552.

[20] A. Di Ieva, F. Grizzi, H. Jelinek, A. J. Pellionisz, and G. A. Losa, "Fractals in the neurosciences, part I: General principles and basic neurosciences," *The Neuroscientist*, vol. 20, issue 4, pp. 403–417, 2014. https://doi.org/10.1177/1073858413513927.

[21] N. Bird and E. Perrier, "Multiscale percolation properties of a fractal pore network," *Geoderma*, vol. 160, issue 1, pp. 105–110, 2010. https://doi.org/10.1016/j.geoderma.2009.10.009.

[22] J. W. Baish and R. K. Jain, "Fractals and cancer," *Cancer Res*, vol. 60, issue 14, pp. 3683–3688, 2000.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

[24] L. Nanni, S. Ghidoni, and S. Brahnam, "Handcrafted vs. Non handcrafted features for computer vision classification," *Pattern Recognition,* vol. 71, pp. 158–172, 2017. https://doi.org/10.1016/j.patcog.2017.05.025.

[25] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. Morency, "Tensor fusion network for multimodal sentiment analysis," *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, (EMNLP'2017)*, pp. 1103–1114, Copenhagen, Denmark, September 7–11, 2017, Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1115.

[26] S. Narkhede, "Understanding AUC-ROC curve," *Towards Data Science*, vol. 26, issue 1, pp. 220-227, 2018.

[27] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, "Densely connected convolutional networks," *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 21-26 July 2017, pp. 2261-2269. https://doi.org/10.1109/CVPR.2017.243.

[28] R. Kosarevych, O. Lutsyk, B. Rusyn, et al., "Adaptive patch reshaping for edge-based semantic segmentation in remote sensing," *IEEE Access*, vol. 14, pp. 38951–38964, 2026, https://doi.org/10.1109/ACCESS.2026.3666289.

***Moheb R. Girgis** received his B.Sc. Degree from Mansoura University, Egypt, in 1974, M.Sc. degree from Assiut University, Egypt, in 1980, and Ph.D. degree from the University of Liverpool, England, in 1986. He is a professor of computer science at Minia University, Egypt. His research interests include software engineering, software testing, information retrieval, evolutionary algorithms, image processing, computer networks, and bioinformatics.*

Al Hussien Seddik Saad *received the B.Sc. degree in Computer Science from Minia University, Egypt, in 2006, and the M.Sc. and Ph.D. degrees in Image Steganography from Minia University, Egypt, in 2012 and 2015, respectively. He is an Assistant Professor in Computer Science Department, Minia University. His research interests include Covered and Coverless Steganography, data hiding, Information security and Internet of Things (IOT).*

Mohammed M. Talaat *received his B.Sc. Degree from Minia University, Egypt, in 2006, M.Sc. degree from Minia University, Egypt, in 2018. He is a Ph.D. Student at Department of Computer Science, Faculty of Science, Minia University, Egypt. His research interests include Image Processing and Software Development.*