S³F-Net: A Multi-Modal Approach to Medical Image Classification via Spatial-Spectral Summarizer Fusion Network

Md. Saiful Bari Siddiqui, Mohammed Imamul Hassan Bhuiyan.

***Abstract*— Convolutional Neural Networks (CNNs) have become a cornerstone of medical image analysis due to their proficiency in learning hierarchical spatial features. However, this focus on a single domain is inefficient at capturing global, holistic patterns and fails to explicitly model an image's frequency-domain characteristics. To address these challenges, we propose the Spatial-Spectral Summarizer Fusion Network (S³F-Net), a dual-branch framework that learns from both spatial and spectral representations simultaneously. The S³F-Net performs a fusion of a deep spatial CNN with our proposed shallow spectral encoder, SpectraNet. SpectraNet features the proposed SpectralFilter layer, which leverages the Convolution Theorem by applying a bank of learnable filters directly to an image's full Fourier spectrum via a computation-efficient element-wise multiplication. This allows the SpectralFilter layer to attain a global receptive field instantaneously, with its output being distilled by a lightweight summarizer network. We evaluate S³F-Net across four diverse medical imaging datasets spanning different scales and modalities: HAM10000 (dermoscopy), BUSI (ultrasound), BRISC2025 (MRI), and Chest X-Ray Pneumonia (radiography), to validate its efficacy and generalizability, and reveal the task-dependent nature of the optimal fusion strategy. Our framework consistently and significantly outperforms its strong spatial-only baseline in all cases, with accuracy improvements of up to 5.13%. With a powerful Bilinear Fusion, S³F-Net achieves a state-of-the-art competitive accuracy of 98.76% on the BRISC2025 dataset. A simpler Concatenation Fusion performs better on the texture-dominant Chest X-Ray Pneumonia dataset, achieving 93.11% accuracy, surpassing many top-performing, much deeper models. Our explainability analysis also reveals that the S³F-Net learns to dynamically adjust its reliance on each branch based on the input pathology. These results verify that our dual-domain approach is a powerful and generalizable paradigm for medical image analysis.**

***Index Terms*— Biomedical imaging, Deep learning, Multimodal fusion, Representation learning, Spectral analysis, Convolutional neural networks.**

I. Introduction

The application of deep learning, particularly Convolutional Neural Networks (CNNs), has led to transformative advances in the field of automated medical image analysis. Architectures such as ResNet and VGG have demonstrated remarkable success in a variety of classification tasks, from detecting pneumonia in chest radiographs to identifying malignancies in dermoscopic images [1], [2], [3]. The primary strength of these models lies in their ability to learn a rich hierarchy of spatial features, progressively building an understanding of complex structures from simple patterns like edges and textures.

However, the operating principle of standard CNNs, which relies on stacked layers of local convolutional kernels, is fundamentally biased towards spatial feature extraction. Although powerful, this approach presents two key limitations that are particularly crucial in the medical domain. Firstly, CNNs are inherently local operators. To capture global, holistic patterns, they must progressively build a large receptive field through significant network depth. This is particularly challenging for high-resolution medical images, where even deep networks may struggle to efficiently integrate long-range dependencies, such as the diffuse textural changes indicative of certain pathologies. A neuron in a deep layer must aggregate information from a long chain of preceding local operations. This hierarchical process is an indirect and computationally inefficient method for perceiving global context. Moreover, the fundamental architecture of a CNN does not explicitly model the frequency-domain characteristics of an image. These spectral properties, describing the periodic and textural nature of an image, can hold crucial diagnostic information but are often lost or represented inefficiently by a purely spatial network.

Historically, attempts to integrate frequency-domain analysis into neural networks have faced their own challenges. Early approaches often focused on using the Fast Fourier Transform (FFT) as a tool for accelerating convolutions [4], [5], rather than as a feature extraction mechanism itself. The prospect of using learnable filters directly in the frequency domain was largely hindered by significant practical hurdles: representing a filter at full spectral resolution required an excessive number of parameters, creating a high risk of severe overfitting, especially on typically smaller medical datasets. However, the core operation of CNNs, spatial convolution, is computationally intensive for both the forward and backward passes. Although the Convolution Theorem states that this expensive convolution in the spatial domain is equivalent to a simple element-wise multiplication in the frequency domain, an operation with a much lower computational complexity of $O(N \log N)$ via the Fast Fourier Transform (FFT) versus $O(N^2)$ for spatial con-

Corresponding Author: Md. Saiful Bari Siddiqui.

Md. Saiful Bari Siddiqui is with the Department of Computer Science and Engineering, BRAC University, Dhaka, Bangladesh (e-mail: saiful.bari@bracu.ac.bd).

Mohammed Imamul Hassan Bhuiyan is with the Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh (e-mail: imamul@eee.buet.ac.bd).

volution [6], harnessing this efficiency in a learnable feature-extraction context remained a challenge. The inherent success and intuitive nature of spatial convolutions, which mimic the local feature detection of the human visual cortex, further solidified their dominance [7].

To bridge these gaps, we introduce the **Spatial-Spectral Summarizer Fusion Network (S³F-Net)**, a novel dual-branch framework that explicitly and efficiently leverages both spatial and spectral information. Our central hypothesis is that by fusing features from these two complementary domains, our model can develop a more robust and comprehensive understanding of medical images. The S³F-Net is comprised of two specialized, asymmetric towers. The first is a deep, standard CNN that acts as a powerful morphological feature extractor. The second, a key contribution of this work, is our novel **SpectraNet** branch. Our work overcomes the historical challenges of learnable spectral models through two key architectural innovations. First is the **SpectralFilter Layer**, which applies a bank of learnable filters directly to an image's entire Fourier spectrum via a parameter-efficient element-wise multiplication, attaining a global receptive field from the very first layer. Second, we introduce the **SpectraNet** branch, a deliberately shallow architecture built upon this layer. **We posit that since a global view is immediately available in the frequency domain, a deep hierarchy is unnecessary.** Instead, the **shallow** SpectraNet branch focuses on summarizing this global information, distilling it through a lightweight "funnel" head comprised of Depthwise Separable Convolutions and dense layers to produce a small but powerful feature vector.

Medical images are a particularly suitable domain for this dual-domain approach. Pathologies are often defined by a confluence of shape and texture; for example, the diagnosis of a skin lesion depends on both its border shape (spatial) and its internal color variegation (textural/spectral). Furthermore, modalities like ultrasound and radiography are often texture-dominant and captured at high resolutions, providing a rich, high-fidelity signal for frequency analysis that is often under-utilized by standard methods.

A central finding of this research is that the optimal strategy for combining the information from the two domains is highly task-dependent. We systematically investigate two primary fusion mechanisms at the vector level. The first is a simple **Concatenation Fusion**, which appends the two feature vectors and allows a subsequent dense layer to learn their relationships. This is a robust and general-purpose strategy. The second is a more powerful **Bilinear Fusion**, which computes the outer product of the two vectors. This mechanism explicitly models the rich, pairwise interactions between every spatial feature and every spectral feature, making it a more suitable choice for complex tasks where the interplay between morphology and texture is critical.

The contributions of this paper are as follows:

- We introduce the **SpectralFilter Layer**, a computationally efficient deep learning component that applies learnable filters directly to an image's Fourier spectrum via element-wise multiplication, replacing the need for computationally expensive spatial convolutions for global feature extraction.
- We present the **S³F-Net**, a flexible dual-domain framework that consistently and significantly outperforms a strong spatial-only baseline across four diverse medical imaging modalities (dermoscopy, ultrasound, MRI, and radiography). The key contribution within this architecture is our proposed lightweight spectral encoder, **SpectraNet**.
- We demonstrate that the optimal fusion strategy for the S³F-Net is task-dependent, with a powerful **Bilinear Fusion** excelling on feature-complex datasets, whereas a simpler **Concatenation Fusion** is superior for texture-dominant tasks.
- We present state-of-the-art competitive results on multiple benchmarks, achieving **98.76%** accuracy on BRISC2025 and **93.11%** accuracy on the Chest X-Ray Pneumonia dataset, validating the S³F-Net as a powerful and data-efficient architecture, even when trained from scratch.
- We develop a custom explainability metric to quantify branch contributions and reveal S³F-Net's ability to dynamically weigh each branch based on the pathology.

II. Related Work

Our research is positioned at the intersection of three key areas in deep learning: the application of CNNs to medical imaging, the use of frequency-domain analysis in neural networks, and the development of multi-modal fusion strategies.

A. Deep Learning in Medical Image Classification

Convolutional Neural Networks (CNNs) have become the de facto standard for a wide array of medical image analysis tasks. Landmark architectures such as VGGNet [1], Inception-v3 [2], DenseNet [8], and particularly ResNet [3], with its residual connections, as well as more recent efficient designs like MobileNet [9] and EfficientNet [10], have established the power of deep networks for feature extraction. In the medical domain, these models, typically pre-trained on large-scale datasets like ImageNet [11], are fine-tuned to achieve state-of-the-art results. This transfer learning paradigm has been successfully applied to diverse tasks, including classifying skin lesions in the ISIC challenges [12], detecting pathologies in chest radiographs [13], [14], and segmenting tumors in brain MRIs [15]. Our work complements this established spatial approach by introducing a parallel, global feature extraction mechanism that is shown to be highly data-efficient, even when trained from scratch.

B. Frequency Domain in Deep Learning

The use of the Fourier Transform in neural networks has a rich history, evolving from a tool for computational acceleration to a mechanism for feature extraction. Early works by Mathieu et al. [5] and Vasilache et al. [4] provided a theoretical and practical foundation, demonstrating that the Convolution Theorem could be leveraged via the Fast Fourier Transform (FFT) to accelerate the training of CNNs on GPUs. However, these approaches treated the frequency domain as a computational shortcut rather than a source of learnable features.

A conceptual shift began with works that explored the representative power of the spectral domain itself. Rippel et al. were pioneers in this area, first proposing the use of spectral representations for weight parameterization and later introducing Spectral Pooling, which performs downsampling by cropping the low-frequency components of the Fourier spectrum, providing a theoretically sound, anti-aliasing alternative to max-pooling [16]. Pratt et al. followed with the Fourier Convolutional Neural Network (FCNN), one of the first end-to-end architectures to learn filters directly in the frequency domain, demonstrating its viability [17].

More recently, the focus has shifted to creating powerful, parameter-efficient spectral layers. The work on Fast Non-Local Neural Networks by Chi et al. used spectral methods for residual learning [18], while other research has explored related domains like wavelets [19]. The recent theoretical analysis of Spectral Neural Networks by Li et al. has begun to shed light on the optimization landscape, making the design of such networks more principled [20]. Practical applications are also emerging, such as lightweight spectral CNNs for specialized tasks like character recognition [21].

The Fast Fourier Convolution (FFC) proposed by Chi et al. [6] also leverages learnable filters in the frequency domain. However, the FFC's design relies on a layer-level channel split, where only a fraction of the cropped feature channels are processed by its global spectral path. This approach, while computationally efficient, means that neither the local nor the global path ever gains access to the full feature representation from the preceding layer. In contrast, our S³F-Net is a macro-level, two-tower architecture where the roles are asymmetric: a spatial feature encoder is fused with a deliberately shallow spectral summarizer, providing the entire input signal to both the spatial and spectral branches independently. This allows each branch to perform a complete and uncompromised analysis of the data from its own domain-specific perspective. We posit that this full-spectrum analysis, combined with our shallow "summarizer" design leveraging only one or two SpectralFilter layers, is a more efficient approach for injecting global context than stacking deep FFC blocks, leading to the strong performance we observe in our experiments.

C. Feature Fusion Strategies

As dual-branch and multi-modal architectures become more common, the strategy for fusing feature representations has become a critical area of research. The simplest method is **Concatenation Fusion**, where feature vectors from different branches are appended and passed to a subsequent fully-connected layer to learn their combined representation [22]. Although robust, this method can be suboptimal when the feature vectors are of disparate scales or levels of abstraction.

More sophisticated techniques aim to model the interactions between features more explicitly. Attention-based mechanisms have become popular [23], where one branch's features are used to modulate or "attend to" the features of another. Methods like FiLM (Feature-wise Linear Modulation) use a global context vector to generate affine transformation parameters (gamma and beta) that are applied to a primary feature map, effectively allowing one branch to guide the other [24]. For fusing two powerful feature vectors, **Bilinear Pooling** has emerged as a state-of-the-art technique, particularly in fine-grained visual categorization [25]. By computing the outer product of two feature vectors, it models every pairwise feature interaction, creating a highly expressive representation. A key finding of our work is the systematic comparison of these strategies, revealing that the optimal choice between these fusion strategies is highly task-dependent.

III. Theoretical Framework

Our proposed S³F-Net is built upon a synthesis of established principles in signal processing and deep learning. This section outlines the theoretical foundations of our approach.

A. The Fourier Domain and the Fast Fourier Transform

Any signal in the spatial domain, such as an image, can be losslessly represented in the frequency domain. For a 2D image $f(x, y)$ of size $M \times N$, its Discrete Fourier Transform (DFT), $F(u, v)$, is defined as:

$$F(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \cdot e^{-j2\pi(\frac{ux}{M}+\frac{vy}{N})} \tag{1}$$

where j is the imaginary unit, (x, y) are the spatial coordinates, and (u, v) are the frequency coordinates. The output $F(u, v)$ is a complex-valued spectrum where each point represents the magnitude and phase of a sinusoidal plane wave of a specific frequency that constitutes the original image. The Fast Fourier Transform (FFT) is an algorithm [26] that computes the DFT with a significantly lower computational complexity of $O(MN \log(MN))$.

B. The Convolution Theorem and the SpectralFilter Layer

The cornerstone of our spectral branch is the Convolution Theorem, which provides a highly efficient alternative to spatial convolution. For two functions $f(x, y)$ (the image) and $h(x, y)$ (the convolutional filter), their spatial convolution $(f * h)(x, y)$ is defined as:

$$(f*h)(x,y) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} f(m,n)h(x-m,y-n) \tag{2}$$

The Convolution Theorem [27] states that this expensive operation is equivalent to a simple element-wise (Hadamard) product in the frequency domain:

$$\mathcal{F}\{f * h\} = \mathcal{F}\{f\} \odot \mathcal{F}\{h\} \tag{3}$$

where $\mathcal{F}$ denotes the Fourier Transform and $\odot$ is the element-wise product. This allows us to compute the convolution by transforming both the image and the filter, multiplying them, and then transforming the result back to the spatial domain via the Inverse Fast Fourier Transform (IFFT), $\mathcal{F}^{-1}$:

$$f * h = \mathcal{F}^{-1}\{\mathcal{F}\{f\} \odot \mathcal{F}\{h\}\} \tag{4}$$

Our proposed **SpectralFilter Layer** leverages this principle directly. Instead of learning a spatial filter h, we learn its

frequency-domain representation $\mathcal{F}\{h\}$ as a set of complex-valued weights. This allows us to replace the computationally intensive spatial convolution with a highly efficient element-wise multiplication in the frequency domain.

C. Instantaneous Global Receptive Field

A key advantage of our spectral approach is its ability to achieve an instantaneous global receptive field, in stark contrast to the incremental field growth of a standard CNN. The receptive field of a neuron defines the specific region of the input image that influences its activation. Its size is therefore a critical measure of the model's ability to perceive and integrate contextual information. Whereas CNNs must gradually expand this field through depth, our spectral layer perceives the entire image context from the outset.

Proof of Global Receptive Field of the SpectralFilter Layer: Let an input image be a 2D signal $f(x, y)$ and our learnable spectral filter be represented in the frequency domain as $H(u, v)$. The output of our **SpectralFilter Layer**, which we denote as the feature map $g(x, y)$, is computed via the Convolution Theorem as:

$$g(x, y) = \mathcal{F}^{-1}\{F(u, v) \odot H(u, v)\} \tag{5}$$

where $F(u, v) = \mathcal{F}\{f(x, y)\}$ and $\odot$ is the element-wise product.

The process involves a forward DFT and an inverse DFT, both of which are global operations. By substituting the definition of the DFT (1) into the inverse DFT and rearranging the summation terms, it can be formally shown that any output pixel $g(x, y)$ is a weighted sum over all input pixels $f(x', y')$:

$$g(x,y) = \sum_{x'=0}^{M-1}\sum_{y'=0}^{N-1} f(x',y') \cdot \left[\frac{1}{MN}\sum_{u=0}^{M-1}\sum_{v=0}^{N-1} H(u,v) \cdot e^{j2\pi(\frac{u(x-x')}{M}+\frac{v(y-y')}{N})}\right] \tag{6}$$

Let the term in the brackets be $W(x, y, x', y')$. This term represents the weight of the contribution of the input pixel at (x', y') to the output pixel at (x, y). Crucially, this weight is non-zero for all pairs of (x, y) and (x', y').

This equation demonstrates that the value of any given output pixel $g(x, y)$ is a weighted sum over *every single pixel* $f(x', y')$ in the entire input image. Therefore, the receptive field of a single **SpectralFilter Layer** is, by definition, the full size of the input image, $M \times N$.

On the other hand, the receptive field (RF) of a neuron in a CNN can be calculated iteratively. For a stack of convolutional layers, the RF of layer l is given by:

$$RF_l = RF_{l-1} + (k_l - 1) \times S_{l-1} \tag{7}$$

where k_l is the kernel size of layer l, and S_{l-1} is the product of all preceding strides. For a standard VGG16 architecture operating on a 224×224 image, the network consists of sequential blocks of 3×3 convolutions and 2×2 max-pooling.

Let's calculate the receptive field at the end of the final convolutional block for a VGG16-style network:

- **Block 1 (2 convs):** RF = $1 + (3-1) + (3-1) = 5$. After Pool: Stride = 2.
- **Block 2 (2 convs):** RF = $5 + (3-1) \times 2 + (3-1) \times 2 = 13$. After Pool: Stride = 4.
- **Block 3 (3 convs):** RF = $13 + (3-1) \times 4 \times 3 = 37$. After Pool: Stride = 8.
- **Block 4 (3 convs):** RF = $37 + (3-1) \times 8 \times 3 = 85$. After Pool: Stride = 16.
- **Block 5 (3 convs):** RF = $85 + (3-1) \times 16 \times 3 = \mathbf{181}$.

Even with a deep architecture like VGG16, the receptive field after all the convolutional layers is only 181×181. This neuron cannot see the entire 224×224 input image; it is blind to the corners and edges. To achieve a global view, such a network must rely on the final, destructive global pooling layer.

In contrast, a single **SpectralFilter Layer**, by the proof above, has an instantaneous receptive field of **224 × 224** (or more depending on the input dimensions), demonstrating its efficiency at capturing global context.

D. Depthwise Separable Convolutions

To efficiently summarize the feature maps produced by the SpectralFilter Layer without incurring a large parameter cost, our SpectraNet branch utilizes Depthwise Separable Convolutions [9], [28]. A standard 2D convolution with kernel size $K \times K$, C_{in} input channels, and C_{out} output channels has a computational cost proportional to $K \cdot K \cdot C_{in} \cdot C_{out}$.

A Depthwise Separable Convolution factorizes this operation into two steps:

1) **Depthwise Convolution:** A single $K \times K$ filter is applied independently to each of the C_{in} input channels. This step learns spatial patterns within each channel but does not combine information across channels. Its cost is proportional to $K \cdot K \cdot C_{in}$.
2) **Pointwise Convolution:** A simple 1×1 convolution with C_{out} filters is then used to project the output of the depthwise layer into the new channel space. This step learns linear combinations of the channels. Its cost is proportional to $1 \cdot 1 \cdot C_{in} \cdot C_{out}$.

The total cost is the sum of these two steps, and the reduction in computation is given by:

$$\text{Reduction} = \frac{K \cdot K \cdot C_{in} + C_{in} \cdot C_{out}}{K \cdot K \cdot C_{in} \cdot C_{out}} = \frac{1}{C_{out}} + \frac{1}{K^2} \tag{8}$$

For a typical 3×3 kernel, this results in an 8-9x reduction in both parameters and computation, making it an ideal choice for a lightweight summarizer head.

E. Bilinear Pooling for Feature Fusion

To effectively fuse the features from our two expert branches, we employ Bilinear Pooling [25]. Given two feature vectors, a spatial vector $\mathbf{x} \in \mathbb{R}^m$ and a spectral vector $\mathbf{y} \in \mathbb{R}^n$, simple concatenation would produce a vector of size $m + n$.

Bilinear Pooling, however, computes the outer product of these two vectors, resulting in a matrix $\mathbf{P} \in \mathbb{R}^{m \times n}$:

$$\mathbf{P} = \mathbf{x} \otimes \mathbf{y} = \mathbf{x}\mathbf{y}^T \tag{9}$$

This matrix is then flattened into a vector $\mathbf{z} \in \mathbb{R}^{mn}$, where each element z_k corresponds to a pairwise product of features

$x_i y_j$. This explicitly models every second-order interaction between the two feature modalities, creating a highly expressive and powerful representation. To stabilize training, the resulting vector is passed through a signed square-root normalization followed by L2 normalization:

$$\hat{\mathbf{z}} = \frac{\text{sign}(\mathbf{z})\sqrt{|\mathbf{z}|}}{||\text{sign}(\mathbf{z})\sqrt{|\mathbf{z}|}||_2} \tag{10}$$

This allows our model to learn the rich, non-linear correlations between the spatial and spectral domains.

IV. Methodology

To address the limitations of purely spatial deep learning models, we propose the **Spatial-Spectral Summarizer Fusion Network (S³F-Net)**, a flexible, two-tower framework designed to learn from both local, spatial-domain features and global, frequency-domain features simultaneously. This section details the datasets used for our evaluation, the core components of our architecture, and the fusion strategies investigated.

A. Datasets

To validate the robustness and generalizability of our proposed framework, we conduct experiments across four diverse and challenging public medical imaging datasets.

- **HAM10000:** The Human Against Machine with 10,000 Training Images dataset [29] is a multi-class dermoscopy dataset for skin lesion classification. It contains over 10,000 images across 7 diagnostic categories. For evaluation, we use the official test set from the ISIC 2018 Challenge [30]. The composition of this dataset, with seven diagnostic categories, presents a severe class imbalance that is highly representative of real-world clinical data.
- **BRISC2025:** The Brain Tumor MRI Dataset for Segmentation and Classification [31] contains over 6,000 T1-weighted brain MRI scans. We utilize the classification task, which is a balanced, 4-class problem (glioma, meningioma, pituitary tumor, and no tumor) with predefined training and test splits.
- **Chest X-Ray (Pneumonia):** This is a large dataset of over 5,000 chest radiographs for the binary classification of Normal vs. Pneumonia [32]. The high resolution of these images provides a high-fidelity signal ideal for frequency analysis. This is critical, as the pathology is often characterized by diffuse textural opacities, making this dataset a direct test of our spectral branch's capacity to capture global patterns that are less apparent to purely spatial models.
- **BUSI:** The Breast Ultrasound Images dataset [33] contains approximately 800 grayscale ultrasound images for a 3-class classification task (normal, benign, and malignant). The smaller scale of this dataset makes it extremely difficult to train deep networks from scratch, making it well-suited to demonstrate the data efficiency of S³F-Net.

B. Overall S³F-Net Architecture

The S³F-Net employs an asymmetric two-tower design, as illustrated in Figure 1. The first tower is a deep spatial encoder responsible for learning a rich hierarchy of morphological features. The second tower is our novel, deliberately shallow **SpectraNet** branch, which efficiently extracts a low-dimensional summary of global, frequency-domain features. The feature vectors from these two branches are then combined using a fusion head, which produces the final output.

C. The SpectralFilter Layer

The foundational component of our spectral branch is the **SpectralFilter Layer**. This layer implements the Convolution Theorem (3) directly, performing a global convolution via an efficient element-wise multiplication in the frequency domain.

The layer takes a 4D tensor (batch, height, width, channels) as input. Its internal weights, W_{real} and W_{imag}, are learnable tensors defined in the frequency domain with a shape corresponding to the real-part FFT of the input. A key implementation detail of the **SpectralFilter Layer** is the handling of the complex-valued learnable filters. As standard deep learning optimizers operate on real-valued tensors, we represent each spectral filter $H(u, v)$ by parameterizing its real and imaginary components as two independent, trainable weight tensors, W_{real} and W_{imag}. During the forward pass, these two tensors are combined to reconstruct the complex-valued filter $H(u, v) = W_{real} + j \cdot W_{imag}$ using a native complex number operation. This reconstructed filter is then multiplied with the complex-valued Fourier spectrum of the input. This approach allows the optimizer to update the real and imaginary parts of the spectral filter independently via standard backpropagation, while still correctly performing complex arithmetic in the forward pass.

The forward pass, detailed in Algorithm 1, is fully differentiable and can be integrated seamlessly into any deep learning framework.

Algorithm 1 SpectralFilter Layer Forward Pass

Input: Image tensor $X \in \mathbb{R}^{B \times H \times W \times C_{in}}$
Parameters: $W_{real}, W_{imag} \in \mathbb{R}^{C_{in} \times C_{out} \times H \times (W/2+1)}$

$X_{perm} \leftarrow \text{Transpose}(X, \text{dims} = [0, 3, 1, 2])$
$X_{fft} \leftarrow \text{FFT}(X_{perm})$
$W_{fft} \leftarrow \text{Complex}(W_{real}, W_{imag})$
$Y_{fft} \leftarrow \text{Einsum}('bchw, cfhw \rightarrow bfhw', X_{fft}, W_{fft})$
$Y_{spatial} \leftarrow \text{IFFT}(Y_{fft})$
$Y_{perm} \leftarrow \text{Transpose}(Y_{spatial}, \text{dims} = [0, 2, 3, 1])$
return $Y_{perm} + \text{bias}$

D. The SpectraNet Summarizer Branch

We posit that since the SpectralFilter Layer has an instantaneous global receptive field, a deep spectral hierarchy is both unnecessary and computationally inefficient. Based on this, we designed the **SpectraNet** branch to act as a lightweight "summarizer" of global information.

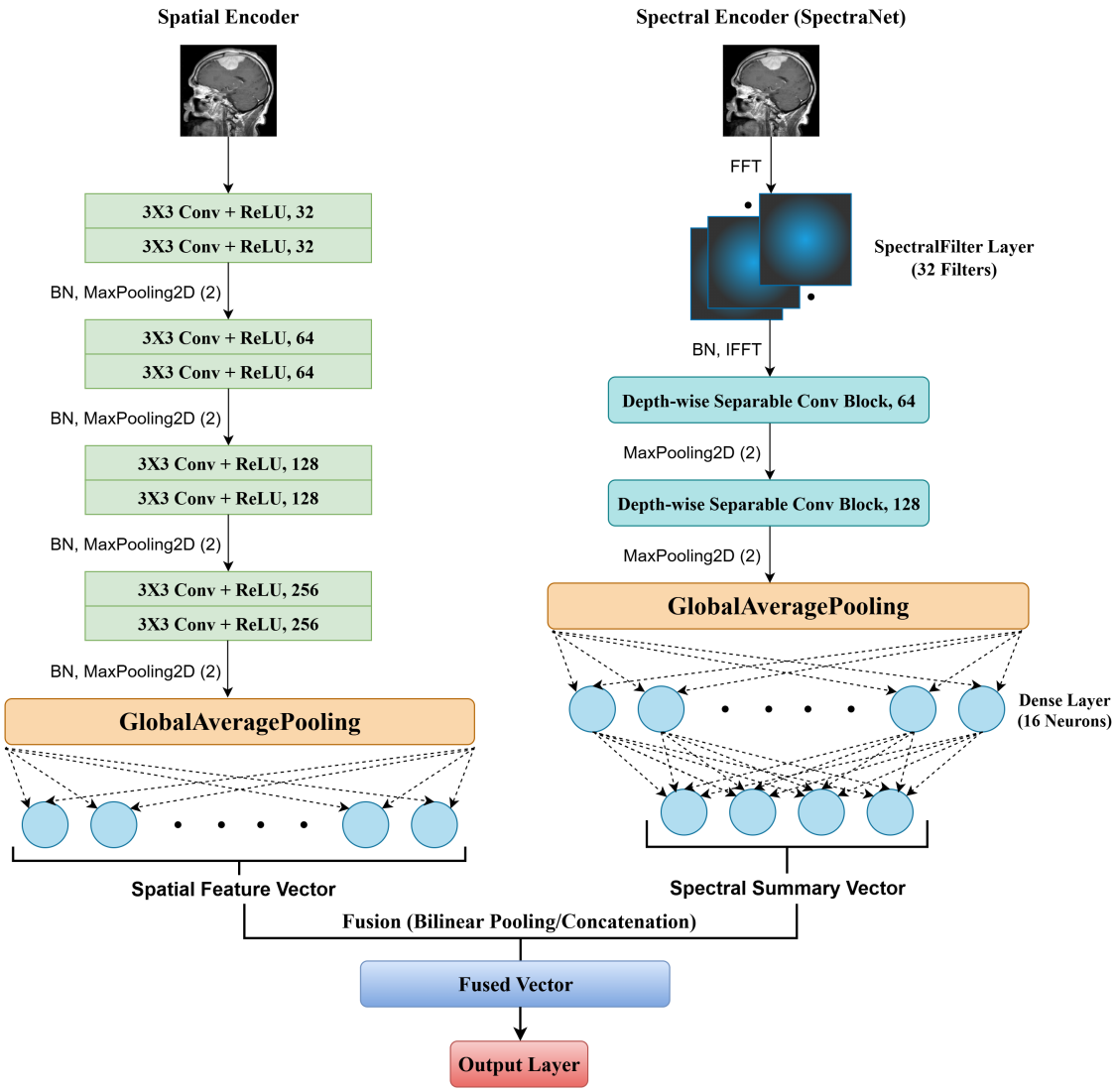

Fig. 1: The architecture of the S³F-Net, showcasing the asymmetric two-tower design with a deeper spatial encoder and a shallow SpectraNet branch, which uses a SpectralFilter layer and an efficient summarizer head to extract a global feature vector. The outputs of both branches are combined in a final fusion head for classification.

We experimented with two shallow variants. The **SpectraNet-1** architecture (the spectral encoder shown in Figure 1), our primary design, utilizes a single SpectralFilter Layer at the full input resolution. Its output feature map is then passed through a lightweight head of Depthwise Separable Convolution blocks to summarize the spatial arrangement of the global frequency patterns. In contrast, the **SpectraNet-2** architecture (Figure 2) stacks two SpectralFilter Layers (32 and 64 filters, respectively), which we found to be beneficial on the texture-dominant BUSI dataset.

A critical component of the summarizer head is the **Depth-wise Separable Convolution block** [9], [28]. The output of the SpectralFilter Layer is passed through these blocks to gradually downsample the feature map while efficiently learning to summarize the spatial arrangement of the global frequency patterns. A Depthwise Separable Convolution is a factorized version of a standard convolution that drastically reduces parameters and computation. The operation, detailed in Algorithm 2, proceeds in two distinct steps.

The first step learns spatial patterns within each input channel in isolation, whereas the second step learns linear combinations across the channels to create a richer output feature representation. In our SpectraNet architecture, each of these blocks is followed by a **MaxPooling** layer, which summarizes the feature maps by reducing the heights and widths. The final stage of the SpectraNet is a **funnel** of two small dense layers. This was a key discovery of our research; this architectural bottleneck forces the model to distill the summarized features into a very small (4-dimensional) but powerful vector, which proved essential for effective fusion and preventing overfitting. A Dropout rate of 0.1875 between the two dense layers proved effective in our experiments.

E. The Spatial Branch Baseline

The spatial branch of the S³F-Net serves two roles: it serves as the spatial encoder in our fusion model, and it acts as our

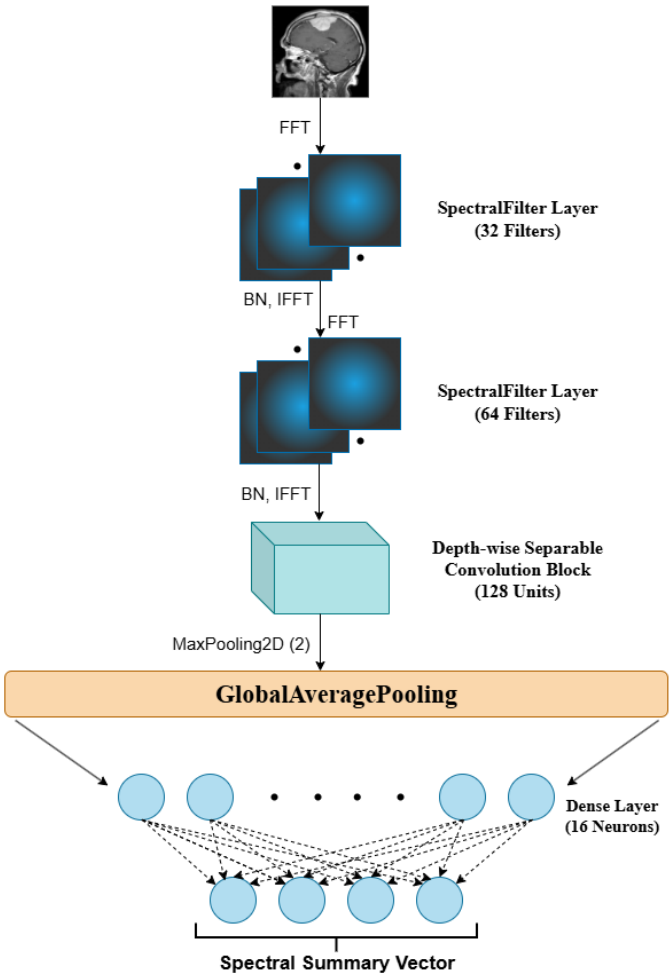

Fig. 2: The architecture of the SpectraNet-2 variation.

primary experimental baseline. We employ a moderately deep, VGG-style architecture with four main blocks, each containing two 3×3 convolutional layers followed by Batch Normalization, a ReLU activation, MaxPooling, and Dropout. The diagram of this spatial encoder has been shown in Figure 1. We deliberately chose this architecture over much deeper models like ResNet50 for two reasons. First, since our SpectraNet branch is already capturing the global receptive field with high efficiency, a massively deep spatial network is less necessary. Second, using a moderately-sized custom CNN allows us to conduct a fair "from-scratch" comparison, isolating the architectural benefits of our fusion strategy without the confounding variable of ImageNet pre-training.

F. Fusion Strategies

A central finding of our work is that the optimal strategy for fusing the refined spatial vector and the summarized spectral vector is highly task-dependent. Let the output of the spatial tower be the vector $\mathbf{v}_s \in \mathbb{R}^{d_s}$ and the output of the SpectraNet tower be the vector $\mathbf{v}_f \in \mathbb{R}^{d_f}$ (where $d_f = 4$). We investigated two primary fusion mechanisms:

1) **Concatenation Fusion:** This is a simple and robust baseline fusion where the two vectors are concatenated to form a single, larger vector $\mathbf{z}_{cat}$.

$$\mathbf{z}_{cat} = [\mathbf{v}_s; \mathbf{v}_f] \in \mathbb{R}^{d_s+d_f} \tag{11}$$

This combined vector is then passed to a subsequent dense layer to learn the relationships between the two feature sets. This proved to be the most effective strategy for the BUSI and Chest X-Ray datasets.

2) **Bilinear Fusion:** To explicitly model the rich, pairwise interactions between the two feature domains, we employ Bilinear Pooling [25]. This process creates a highly expressive feature vector by computing the outer product of the two tower outputs. The operation proceeds in three steps:

(i) First, the refined spatial vector $\mathbf{v}_s \in \mathbb{R}^{d_s}$ and the summarized spectral vector $\mathbf{v}_f \in \mathbb{R}^{d_f}$ are prepared.

(ii) Next, their outer product is computed via matrix multiplication to form an interaction matrix $\mathbf{P}$:

$$\mathbf{P} = \mathbf{v}_s \mathbf{v}_f^T \in \mathbb{R}^{d_s \times d_f} \tag{12}$$

Each element P_{ij} of this matrix represents the multiplicative interaction between the i-th spatial feature and the j-th spectral feature.

(iii) Finally, the matrix $\mathbf{P}$ is flattened (vectorized) to form the final fused feature vector, $\mathbf{z}_{bilinear}$:

$$\mathbf{z}_{bilinear} = \text{vec}(\mathbf{P}) \in \mathbb{R}^{d_s d_f} \tag{13}$$

This vector, which captures the second-order statistics between the two domains, proved to be the top performer on the more complex HAM10000 and BRISC2025 datasets.

Algorithm 2 Depthwise Separable Convolution Block

Input: Feature Map $F \in \mathbb{R}^{B \times H \times W \times C_{in}}$
Parameters:
Depthwise Filters $W_d \in \mathbb{R}^{K \times K \times C_{in}}$
Pointwise Filters $W_p \in \mathbb{R}^{1 \times 1 \times C_{in} \times C_{out}}$

1. Depthwise Convolution (Spatial Filtering)
$F_{intermediate} \leftarrow \text{DepthwiseConv2D}(F, W_d)$
$F_{intermediate} \leftarrow \text{BatchNorm}(F_{intermediate})$
$F_{intermediate} \leftarrow \text{ReLU}(F_{intermediate})$

2. Pointwise Convolution (Channel Mixing)
$G_{out} \leftarrow \text{Conv2D}(F_{intermediate}, W_p)$
$G_{out} \leftarrow \text{BatchNorm}(G_{out})$
$G_{out} \leftarrow \text{ReLU}(G_{out})$

return $G_{out} \in \mathbb{R}^{B \times H \times W \times C_{out}}$

V. Experimental Setup

We evaluate our models on four diverse, publicly available medical imaging datasets: HAM10000 [29], BRISC2025 [31], Chest X-Ray (Pneumonia) [32], and BUSI [33]. We used the predefined publicly available Training and Test sets for all datasets. All models were implemented in TensorFlow and trained from scratch on an NVIDIA A100 GPU. We used the Adam optimizer with an initial learning rate of 3×10^{-4} and

a ReduceLROnPlateau callback. Due to the significant class imbalance present in some datasets, we employed balanced class weights during training and selected the best model based on the validation set's Weighted F1-Score, a metric robust to class imbalance, while balancing overall performance. For our final analysis, we report four key metrics on the unseen test set: Accuracy, Weighted F1-Score, Area Under the Receiver Operating Characteristic Curve (AUC-ROC), and the Matthews Correlation Coefficient (MCC).

VI. Results and Discussions

We conduct a comprehensive set of experiments to validate the efficacy of our proposed S³F-Net framework. Our evaluation is designed to answer five primary research questions: (1) Does the addition of our lightweight spectral branch provide a consistent performance benefit over a strong spatial-only baseline? (2) What is the standalone predictive power of the spectral branch? (3) Which fusion strategy is optimal, and is this choice dependent on the nature of the dataset? (4) How does S³F-Net compare to established SOTA models? (5) How parameter-efficient is S³F-Net?

A. Ablation Study: Standalone Branch Performance

To understand the individual contributions of each domain, we evaluated the spatial-only and spectral-only models in isolation. The Spatial Baseline consists of our CNN-based deep spatial encoder followed by a classification head. The SpectraNet models consist of the spectral summarizer branch fed into the output layer.

A key finding emerged from this ablation study, as shown in Table I. Although the Spatial Baseline performed better across most datasets as expected, the SpectraNet models demonstrated remarkable and surprising performance on the Chest X-Ray dataset. The shallow **SpectraNet-1** model significantly outperformed the baseline CNN in terms of all metrics, as did **SpectraNet-2**. SpectraNet-1, with only 1 SpectralFilter Layer, outperformed the much deeper baseline CNN by a margin of over **3%** in terms of accuracy. We hypothesize this is due to a combination of two factors: the pathology of pneumonia, which often manifests as diffuse textural changes, is inherently well-suited to a global frequency-domain analysis; and the high-resolution nature of the X-ray images provides a high-fidelity signal for the FFT-based *SpectralFilter Layer*. This result supports our hypothesis and provides the strongest possible motivation for a fusion model, as it proves that neither domain is universally superior and that both provide unique, valuable information.

B. Comparison with Baseline CNN and between Fusion Strategies

We now present the core results of our study in Table I. This table compares our two primary S³F-Net fusion architectures against the strong Spatial Baseline across all four datasets. S³F-Net (Concat.) refers to the Concatenation Fusion, whereas S³F-Net (Bilinear) refers to the Bilinear Pooling Fusion. For each fusion method, we tested both the SpectraNet-1 (SN1) and SpectraNet-2 (SN2) variants. We trained and tested each model three or more times and took the average and standard deviation to ensure reproducibility.

The results presented in Table I lead to several key conclusions. Firstly, across all four datasets and all primary metrics, a version of our **S³F-Net consistently and significantly outperforms the strong Spatial Baseline**. This validates our core hypothesis that fusing separate branches trained on spatial and spectral features provides a more powerful and generalizable representation for medical image analysis, providing robust evidence for our model's superior performance, as it achieves simultaneous improvement across four distinct and complementary metrics: Accuracy (overall correctness), F1-Score (balanced precision-recall), AUC-ROC (diagnostic separability), and MCC (correlation on imbalanced classes).

Secondly, the results reveal that the **optimal fusion strategy is highly task-dependent**. For the complex, multi-featured classification tasks of HAM10000 (dermoscopy) and BRISC2025 (MRI), the **Bilinear Fusion** mechanism was the best performer. Its ability to model the pairwise interactions between the spatial and spectral domains proved essential for distinguishing between visually similar classes. For instance, on BRISC2025, the S³F-Net (Bilinear, SN1) achieves a state-of-the-art competitive accuracy of **98.76%**, surpassing the already strong Spatial Baseline by a notable margin.

This advantage is even more apparent on the HAM10000 dataset. The S³F-Net (Bilinear, SN1) achieves the highest W. F1-Score (**0.7038**) and MCC (**0.5136**), surpassing all other variants, including the baseline. This indicates that the Bilinear Fusion is not just improving performance on the majority class in this dataset with severe class imbalance, but is learning a more nuanced decision boundary that leads to better-balanced, class-wise performance. It shows that by explicitly modeling the interaction between a specific morphological feature and a specific textural feature, the Bilinear model is able to make more confident and accurate predictions on the difficult, rare classes, which is critical for a diagnostically useful model.

Conversely, for the datasets that are more dominated by global textural features, namely Chest X-Ray and BUSI, the simpler **Concatenation Fusion** proved to be more effective. On the Chest X-Ray dataset, the S³F-Net (Concat., SN1) achieved a remarkable **93.11%** accuracy, a greater than **5%** absolute improvement over the Spatial Baseline. Similarly, on the data-scarce BUSI dataset, the S³F-Net (Concat., SN2) achieved the highest accuracy of **87.82%**. We hypothesize that for these tasks, the clean separation of features provided by concatenation is a more robust strategy, preventing the model from overfitting on spurious feature interactions that the more powerful Bilinear Fusion might find. The success of the SpectraNet-2 on BUSI also suggests that texture-dominant modalities may benefit from a slightly deeper spectral analysis.

C. Comparison with State-of-the-Art Models

To contextualize the performance of our S³F-Net, we compare our best-performing variants against a suite of well-established deep learning architectures. On the BRISC2025 dataset, we compare against standard ImageNet pre-trained

TABLE I: Main Performance Comparison on Test Sets Across All Datasets. Results are reported as Mean ± Standard Deviation over multiple independent runs.

Dataset	Model	Accuracy	F1-Score	AUC-ROC	MCC
HAM10000	Spatial Baseline	0.6671 ± 0.0180	0.6889 ± 0.0155	0.8884 ± 0.0066	0.4930 ± 0.0025
	SpectraNet-1	0.6310 ± 0.0212	0.6408 ± 0.0243	0.8324 ± 0.0155	0.4074 ± 0.0281
	SpectraNet-2	0.6448 ± 0.0222	0.6402 ± 0.0252	0.8529 ± 0.0133	0.4054 ± 0.0252
	S³F-Net (Concat., SN1)	**0.7011 ± 0.0096**	0.6974 ± 0.0093	0.8820 ± 0.0013	0.4924 ± 0.0077
	S³F-Net (Concat., SN2)	0.6931 ± 0.0112	0.6841 ± 0.0121	**0.8905 ± 0.0027**	0.4765 ± 0.0098
	S³F-Net (Bilinear, SN1)	0.6931 ± 0.0000	**0.7038 ± 0.0013**	0.8892 ± 0.0066	**0.5136 ± 0.0069**
	S³F-Net (Bilinear, SN2)	0.6071 ± 0.0220	0.6311 ± 0.0241	0.8409 ± 0.0163	0.4236 ± 0.0310
BRISC2025	Spatial Baseline	0.9827 ± 0.0007	0.9827 ± 0.0006	0.9982 ± 0.0007	0.9764 ± 0.0010
	SpectraNet-1	0.9446 ± 0.0064	0.9442 ± 0.0063	0.9912 ± 0.0018	0.9244 ± 0.0072
	SpectraNet-2	0.9525 ± 0.0055	0.9524 ± 0.0054	0.9935 ± 0.0015	0.9352 ± 0.0062
	S³F-Net (Concat., SN1)	0.9772 ± 0.0021	0.9772 ± 0.0021	0.9976 ± 0.0008	0.9689 ± 0.0031
	S³F-Net (Concat., SN2)	0.9703 ± 0.0032	0.9703 ± 0.0031	0.9961 ± 0.0011	0.9595 ± 0.0040
	S³F-Net (Bilinear, SN1)	**0.9876 ± 0.0007**	**0.9876 ± 0.0007**	0.9982 ± 0.0001	**0.9831 ± 0.0010**
	S³F-Net (Bilinear, SN2)	0.9792 ± 0.0018	0.9792 ± 0.0019	**0.9985 ± 0.0004**	0.9716 ± 0.0025
Chest X-Ray	Spatial Baseline	0.8798 ± 0.0204	0.9054 ± 0.0117	0.9433 ± 0.0127	0.7440 ± 0.0489
	SpectraNet-1	0.9103 ± 0.0110	0.9314 ± 0.0088	0.9659 ± 0.0052	0.8090 ± 0.0210
	SpectraNet-2	0.9022 ± 0.0122	0.9259 ± 0.0096	0.9573 ± 0.0065	0.7928 ± 0.0243
	S³F-Net (Concat., SN1)	**0.9311 ± 0.0125**	**0.9453 ± 0.0097**	0.9749 ± 0.0067	**0.8524 ± 0.0270**
	S³F-Net (Concat., SN2)	0.9271 ± 0.0079	0.9422 ± 0.0052	**0.9753 ± 0.0040**	0.8443 ± 0.0180
	S³F-Net (Bilinear, SN1)	0.8654 ± 0.0181	0.9007 ± 0.0192	0.8959 ± 0.0152	0.7164 ± 0.0355
	S³F-Net (Bilinear, SN2)	0.8750 ± 0.0170	0.9071 ± 0.0177	0.9286 ± 0.0131	0.7363 ± 0.0311
BUSI	Spatial Baseline	0.8430 ± 0.0046	0.8437 ± 0.0059	0.9505 ± 0.0059	0.7345 ± 0.0131
	SpectraNet-1	0.8269 ± 0.0081	0.8264 ± 0.0091	0.9099 ± 0.0112	0.7029 ± 0.0183
	SpectraNet-2	0.8205 ± 0.0095	0.8175 ± 0.0103	0.9294 ± 0.0093	0.6905 ± 0.0210
	S³F-Net (Concat., SN1)	0.8526 ± 0.0031	0.8496 ± 0.0040	0.9565 ± 0.0059	0.7440 ± 0.0110
	S³F-Net (Concat., SN2)	**0.8782 ± 0.0000**	**0.8771 ± 0.0004**	**0.9601 ± 0.0064**	**0.7911 ± 0.0020**
	S³F-Net (Bilinear, SN1)	0.7949 ± 0.0120	0.7910 ± 0.0131	0.9000 ± 0.0142	0.6601 ± 0.0251
	S³F-Net (Bilinear, SN2)	0.8077 ± 0.0110	0.8031 ± 0.0111	0.8778 ± 0.0189	0.6795 ± 0.0227

The best performing model and the best metrics for each dataset are highlighted in **bold**.

models to demonstrate SOTA-competitiveness, whereas on the Chest X-Ray dataset, we compare against models trained from scratch to highlight data efficiency.

1) Performance on BRISC2025: The BRISC2025 dataset, with its high-quality images, serves as an excellent benchmark for top-end performance. As shown in Table II, our S³F-Net with Bilinear Fusion, despite being trained from scratch, achieves a remarkable accuracy of **98.76%**. This result surpasses deeper and more parameter-heavy models like VGG, ResNet, DenseNet, and Inception variants. Our model also significantly outperforms these complex architectures in all other metrics. Although EfficientNetB0 still holds the top position, our S³F-Net demonstrates that a novel, data-efficient architecture can achieve performance within the elite tier without relying on large-scale pre-training. Also, our model achieves this result with the lowest number of parameters (**3.44M**), making it suitable to employ even in edge devices.

2) Performance on Chest X-Ray Dataset: The Chest X-Ray dataset provides an ideal environment to test our central hypothesis regarding data efficiency. We compare our best-performing S³F-Net with Concatenation Fusion against a range of standard CNNs trained entirely from scratch, using benchmark results from a recent study by Alanazi et al. [34].

The results in Table III are striking. Our S³F-Net not only achieves the highest accuracy (**93.11%**) and the F1 score (**0.9453**) by a significant margin, but it does so with only

TABLE II: S³F-Net Performance on BRISC2025 Compared to Standard SOTA Models as Provided in the Original Paper [31].

Model	# Params	Acc.	F1	Precision	Recall
VGG16	138.4M	0.9692	0.9693	0.9706	0.9693
VGG19	143.7M	0.9629	0.9630	0.9639	0.9630
ResNet50	25.6M	0.9820	0.9820	0.9823	0.9820
ResNet101	44.5M	0.9809	0.9810	0.9815	0.9810
DenseNet169	14.3M	0.5093	0.5710	0.6404	0.5710
InceptionV3	23.8M	0.6198	0.6487	0.7225	0.6487
EfficientNetB0	5.3M	0.9920	0.9920	0.9920	0.9920
MobileNetV3	5.4M	0.9442	0.9447	0.9475	0.9447
S³F-Net	**3.44M**	**0.9876**	**0.9876**	**0.9871**	**0.9891**

7.64M parameters (heavier than the BRISC2025 version due to the high input resolution). It surpasses much deeper models like ResNet50 and DenseNet169, which have 2-3 times the parameter count. S³F-Net also demonstrates a much more balanced Precision-Recall trade-off with a precision of **93.81%** and recall of **95.3%**.

Most remarkably, the performance of our from-scratch S³F-Net is not only superior to other from-scratch models but is also highly competitive with pre-trained architectures. The highest reported accuracy by Alanazi et al. [34] on this dataset using ImageNet pre-training is **93.6%**, achieved by DenseNet169. Our S³F-Net, with a peak accuracy of **93.91%** in one of our runs, surpasses this state-of-the-art benchmark.

TABLE III: S³F-Net Performance on Chest X-Ray Pneumonia Compared to SOTA Models Trained From Scratch. Data for other models sourced from [34], [35].

Model	# Params	Acc.	F1	Precision	Recall
ResNet18	11.7M	0.8960	0.9180	0.9010	0.9360
ResNet50	25.6M	0.8300	0.8610	0.8840	0.8380
VGG16	138.4M	0.9100	0.9310	0.8990	0.9640
MobileNetV2	3.4M	0.8000	0.8300	0.8820	0.7850
InceptionV3	23.8M	0.9120	0.9330	0.8970	0.9720
DenseNet169	14.3M	0.9120	0.9330	0.8920	0.9769
S³F-Net	**7.64M**	**0.9311**	**0.9453**	**0.9381**	**0.9530**

These significant findings suggest that our dual-domain architecture provides such a strong and relevant inductive bias for radiographic images that it can overcome the immense advantage typically afforded by large-scale pre-training.

Furthermore, it is critical to note that a direct comparison of parameter counts can be misleading. The computational complexity of a standard CNN is dominated by its spatial convolution operations, which have a cost of approximately $O(K^2 \cdot H \cdot W \cdot C_{in} \cdot C_{out})$. In contrast, the core of our SpectraNet branch, the **SpectralFilter Layer**, replaces this with an element-wise multiplication in the frequency domain. Via the Fast Fourier Transform (FFT), this operation has a significantly lower asymptotic complexity of $O(C_{in} \cdot C_{out} \cdot HW \log(HW))$. This means that even if a traditional CNN were constructed to have an equal number of parameters to our S³F-Net, our model would still possess a lower computational overhead for both its forward and backward passes. This inherent efficiency is a direct result of leveraging the Convolution Theorem. The results discussed therefore validate that our dual-domain approach, which extracts global, frequency-domain features via the SpectraNet branch, allows the model to learn a more robust representation from limited training data than purely spatial architectures. This inherent architectural advantage makes the S³F-Net a powerful and efficient framework for tasks where large-scale pre-training is not feasible.

The primary scientific comparison in this study is between our proposed S³F-Net and its direct ablation, the Spatial-Only baseline. As both models share an identical spatial encoder, training methodology, and hyperparameters, any observed difference in performance can be directly and causally attributed to the addition of the spectral branch and the fusion mechanism. The subsequent comparisons to state-of-the-art (SOTA) models serve a different but equally important purpose. They are relevant because, as our results showed, the S³F-Net demonstrates a remarkable level of performance. Despite the shallow, lightweight, and parameter-efficient design of our proposed SpectraNet branch, the final fused model produces surprisingly strong results that are highly competitive with, and in many cases, superior to, much deeper, SOTA architectures. This validates not only the efficacy but also the data efficiency of our dual-domain approach.

D. Explainability Analysis: Branch Contributions

To understand *why* our S³F-Net framework is so effective, we perform an explainability analysis on the best-performing Chest X-ray model to quantify the influence of each branch on the final decision. Standard metrics like the L2 Norm are biased by vector dimensionality. On the other hand, per class L2 Norm completely ignores the vector dimensionality. To overcome these challenges and strike a balance, we develop a novel, **balanced contribution score** (C_v) for a feature vector $\mathbf{v} \in \mathbb{R}^d$, defined as the geometric mean of its total magnitude and its per-feature magnitude:

$$C_v = \sqrt{||\mathbf{v}||_2 \cdot \frac{||\mathbf{v}||_2}{d}} = \frac{||\mathbf{v}||_2}{\sqrt{d}} \tag{14}$$

This score fairly compares the influence of vectors with disparate dimensionalities by rewarding both total signal strength and intensity per feature.

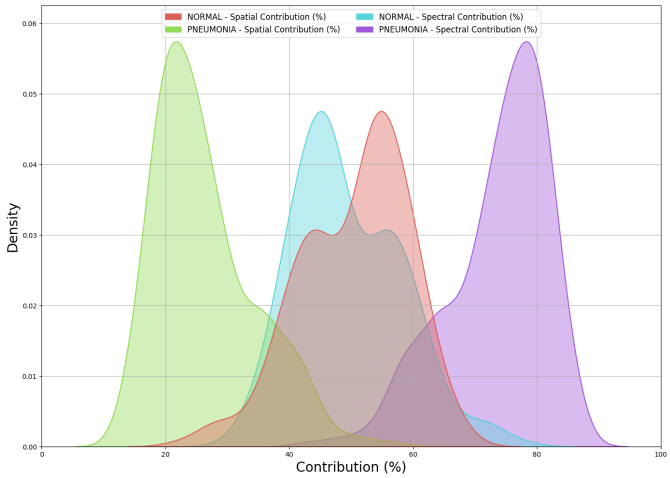

Fig. 3: Combined distribution of the contribution scores for the Spatial and Spectral branches on the Chest X-Ray dataset, separated by true class. It shows different distributions for different classes.

By extracting the refined feature vectors from each tower before fusion, we analyzed their relative contribution scores on the test set. The results, shown in Figure 3, reveal a profound and adaptive division of labor. Overall, the Spectral branch exhibits a dominant average contribution of 64.58%. However, this average masks a more sophisticated strategy: the S³F-Net dynamically alters its reliance on each branch based on the input's class.

For images of **Normal** lungs, the contributions are almost perfectly balanced (50.52% Spatial vs. 49.48% Spectral), suggesting the model requires a consensus from both experts for a "healthy" diagnosis. In contrast, for X-Rays with **Pneumonia**, the model overwhelmingly relies on the Spectral branch, with an average spectral contribution of **73.64%**. This provides powerful quantitative evidence for our hypothesis: the diffuse, textural nature of pneumonia is more effectively captured by the global, frequency-domain analysis of our SpectraNet.

This analysis validates that our spatial-spectral summarizer fusion mechanism is not merely combining signals, but is facilitating a dynamic, data-driven strategy that enables the multi-modal model to learn to weigh different input modalities based on the pathology it observes.

VII. Conclusion

In this work, we addressed two fundamental limitations of standard CNNs in medical imaging: their inherent locality and their dependence on a single modality of signal representation. We introduced the **S³F-Net**, a novel dual-branch framework that learns from both spatial and spectral domains. This successful integration of two complementary domains through the shallow **SpectraNet** branch leads our model to outperform its strong spatial-only baseline across four diverse medical datasets consistently and significantly, validating our core hypothesis. Another key finding of our research is that the optimal fusion mechanism is demonstrably **task-dependent**. This discovery enriches the field by shifting the focus from finding a single "best" fusion method to developing architectures that can adapt their strategy to the nature of the data.

Furthermore, the remarkable performance of the S³F-Net when trained from scratch presents a compelling alternative to the heavy reliance on ImageNet pre-training. By explicitly and efficiently engineering a branch to capture global context, our framework achieves a level of data efficiency that is highly sought after in the often data-scarce medical domain. This opens several promising avenues for future research. The powerful S³F-Net encoder can be directly adapted as a backbone for segmentation and other dense prediction tasks where the interplay of local and global context is critical. Moreover, the task-adaptive nature of our findings suggests a future where models could learn to dynamically select their own optimal fusion strategy, leading to truly versatile and intelligent diagnostic systems.

Data and Code Availability

This research was conducted using publicly available datasets, ensuring full reproducibility. The complete source code for our S³F-Net models and training pipelines are publicly available on GitHub at: `https://github.com/Saiful185/S3F-Net`.

Acknowledgement

The authors gratefully acknowledge the institutional support provided by BRAC University and the Department of Electrical and Electronic Engineering (EEE) at BUET, which was vital for the completion of this work. Additionally, the authors acknowledge the assistance of Google's Gemini, which served as an AI tool for manuscript refinement.

References

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 2014.

[2] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2818–2826.

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

[4] N. Vasilache, J. Johnson, M. Mathieu, S. Chintala, S. Piantino, and Y. LeCun, "Fast convolutional nets with fbfft: A GPU performance evaluation," *arXiv:1412.7580*, 2014.

[5] M. Mathieu, M. Henaff, and Y. LeCun, "Fast training of convolutional networks through FFTs," *arXiv:1312.5851*, 2013.

[6] L. Chi, B. Jiang, and Y. Mu, "Fast Fourier convolution," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 10565–10575.

[7] A. Agrawal, D. Stansbury, J. Malik, and J. L. Gallant, "Pixels to voxels: Modeling visual representation in the human brain," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 4251–4258.

[8] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 4700–4708.

[9] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv:1704.04861*, 2017.

[10] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 6105–6114.

[11] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 248–255.

[12] M. A. Khan, K. Muhammad, M. Sharif, T. Akram, and V. H. C. de Albuquerque, "Multi-class skin lesion detection and classification via teledermatology," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 12, pp. 4267–4275, Dec. 2021, doi: 10.1109/JBHI.2021.3067789.

[13] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," *arXiv:1711.05225*, 2017.

[14] A. Dhere and J. Sivaswamy, "COVID detection from chest X-ray images using multi-scale attention," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 4, pp. 1496–1505, Apr. 2022, doi: 10.1109/JBHI.2022.3151171.

[15] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. J. Pal, P.-M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *Med. Image Anal.*, vol. 35, pp. 18–31, 2017.

[16] O. Rippel, J. Snoek, and R. P. Adams, "Spectral representations for convolutional neural networks," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 28, 2015.

[17] H. Pratt, B. Williams, F. Coenen, and Y. Zheng, "FCNN: Fourier convolutional neural networks," in *Mach. Learn. Knowl. Discovery Databases (ECML PKDD)*, 2017, pp. 52–68.

[18] L. Chi, G. Tian, Y. Mu, L. Xie, and Q. Tian, "Fast non-local neural networks with spectral residual learning," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 1513–1521.

[19] Z. Zhong, T. Shen, Y. Yang, Z. Lin, and C. Zhang, "Joint sub-bands learning with clique structures for wavelet domain super-resolution," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 31, 2018.

[20] C. Li, R. Sonthalia, and N. G. Trillos, "Spectral neural networks: Approximation theory and optimization landscape," *arXiv:2310.00729*, 2023.

[21] I. Y. Alshareef, A. A. H. Ab Rahman, N. Khan, *et al.*, "Lightweight and high-precision spectral convolutional neural network model for custom 94-class ASCII character recognition," *Neural Comput. Appl.*, vol. 37, pp. 16883–16904, 2025, doi: 10.1007/s00521-025-11376-2.

[22] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn. (ICML)*, 2011, pp. 689–696.

[23] Z. Li, J. Zhang, S. Wei, Y. Gao, C. Cao, and Z. Wu, "TPAFNet: Transformer-driven pyramid attention fusion network for 3D medical image segmentation," *IEEE J. Biomed. Health Inform.*, vol. 28, no. 11, pp. 6803–6814, Nov. 2024, doi: 10.1109/JBHI.2024.3460745.

[24] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018.

[25] T. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 1449–1457.

[26] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex Fourier series," *Mathematics of Computation*, vol. 19, no. 90, pp. 297–301, Apr. 1965.

[27] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 3rd ed. Upper Saddle River, NJ, USA: Prentice Hall, 2010.

[28] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1251–1258.

[29] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, Art. no. 180161, 2018.

[30] N. C. F. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection 2018:

A challenge hosted by the International Skin Imaging Collaboration (ISIC)," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2019, pp. 1252–1260.

[31] A. Fateh, Y. Rezvani, S. Moayedi, S. Rezvani, F. Fateh, and M. Fateh, "BRISC: Annotated dataset for brain tumor segmentation and classification with Swin-HAFNet," *arXiv:2406.14318v3*, 2025.

[32] D. S. Kermany *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, pp. 1122–1131, 2018.

[33] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data Brief*, vol. 28, Art. no. 104863, 2020, doi: 10.1016/j.dib.2019.104863.

[34] C. Gu and M. Lee, "Deep transfer learning using real-world image features for medical image classification, with a case study on pneumonia X-ray images," *Bioengineering*, vol. 11, no. 4, Art. no. 406, 2024, doi: 10.3390/bioengineering11040406.

[35] Alshanketi, F., Alharbi, A., Kuruvilla, M., Mahzoon, V., Siddiqui, S. T., Rana, N., & Tahir, A., "Pneumonia Detection from Chest X-Ray Images Using Deep Learning and Transfer Learning for Imbalanced Datasets," *Journal of imaging informatics in medicine*, 38(4), 2021–2040, 2025, doi: 10.1007/s10278-024-01334-0.

[36] Gemini Team, Google, "Gemini: A Family of Highly Capable Multimodal Models," *arXiv:2312.11805*, 2023.