Research Article

Spectral-Spatial Optimization of Deep Networks: A Frequency-Domain VGG Framework For Wildfire Detection

D. Crystal Jaba Kani[1] and S. Saudia[2]

[1,2]Centre for Information Technology and Engineering, Manonmaniam Sundaranar University,Tirunelveli-12, Tamil Nadu, India.
Email Id: [1]crystaljebakani@yahoo.com and [2]saudiasubash@msuniv.ac.in

Abstract

Wildfires pose an escalating global threat, inflicting severe ecological devastation and substantial economic liability annually. While conventional spatial-domain deep learning frameworks have advanced automated fire detection, their efficacy remains heavily constrained by environmental noise, variable illumination, and atmospheric occlusion. This paper addresses these vulnerabilities by introducing a novel frequency-domain deep learning paradigm for predictive wildfire classification. Utilizing the Fire and No Fire Images (FFNI) from the DeepFire benchmark dataset, it deploys the 2D Fast Fourier Transform (FFT) to map spatial visual data into spectral matrices, thereby isolating robust structural textures and high-frequency pixel transition dynamics. The design-rigorously evaluate a series of customized Convolutional Neural Network (CNN) topologies, specifically benchmarking standard VGG16 architectures against structurally optimized variants tailored with localized receptive fields, adaptive filter counts, and modified kernel sizes. Quantitative empirical evaluations demonstrate that the proposed FFT-driven modified VGG16 framework achieves a classification accuracy of 94.0% and an optimized Area Under the Receiver Operating Characteristic (AUC-ROC), substantially outperforming the baseline spatial-domain VGG16 configuration which achieved 62.0%. These findings establish that spectral feature extraction via FFT effectively mitigates ambient spatial noise, offering a computationally efficient, highly resilient trajectory for real-time computer vision deployments in disaster mitigation frameworks.

Keywords: Computer Vision, Fast Fourier Transform, Frequency-Domain Analytics, Hyperparameter Optimization, Wildfire Prediction.

1. Introduction

Wildfires represent one of the most volatile and destructive environmental threats of the modern era. Over the past several decades, the frequency, intensity, and geographic spread of these blazes have escalated dramatically due to shifting global climate patterns, prolonged droughts, and increasing temperatures [1]. The resulting crises demand a fundamental shift from reactive firefighting to proactive, technology-driven prediction models. The economic toll is staggering, costing billions annually in suppression, healthcare and infrastructure repair, while the ecological consequences include irreversible biodiversity loss, soil erosion and massive carbon dioxide feedback loops [2].

Conventional wildfire monitoring heavily relies on human observation, ground patrols, and basic satellite imagery. However, human lookouts suffer from fatigue and satellite arrays face spatial or temporal resolution gaps [3]. Computer Vision systems driven by Deep Learning offer automated hazard detection, but standard spatial-domain (RGB) models encounter significant technical hurdles outdoors [4]. Natural environments present unpredictable variables such as shifting sunlight, dense fog, moving shadows and dust storms. Standard networks struggle to differentiate these benign atmospheric anomalies from actual fire signatures, yielding critical vulnerabilities and high false-alarm rates [5]. To contextualize these vulnerabilities, it is essential to evaluate the fundamental trade-offs between spatial-domain and frequency-domain image processing within Deep Learning architectures.

In the spatial domain, Deep Learning algorithms produce classification [6] models by directly learning from pixel intensity values localized in a continuous coordinate space, f (x, y). Standard Convolutional Neural Networks (CNNs), such as the baseline VGG16 architecture, rely on localized receptive fields (typically 3x3 kernels) to build a hierarchy of features moving from low-level edges to high-level semantic shapes [7]. While highly effective for object recognition under controlled conditions, spatial representations are intrinsically coupled with ambient illumination, color space distributions and local pixel perturbations. In chaotic open-air environments, phenomena like shifting sunlight, dense fog and moving shadows alter the local pixel values drastically. Because spatial CNNs look for explicit visual patterns (like smoke color or flame geometry), they often confuse these benign atmospheric anomalies with active fire signatures, resulting in high false-positive rates [8].

Conversely, transforming spatial data into the frequency domain via the Fast Fourier Transform (FFT) completely decouples image semantics from localized pixel coordinate variations. The 2D FFT decomposes an image into its sinusoidal sine and cosine components, isolating structural textures and high-frequency pixel transition dynamics into a global spectral matrix [9]. This shift introduces unique mathematical properties and trade-offs like Noise and Illumination Invariance wherein global illumination changes such as shifting sunlight or lens flare manifest primarily as low-frequency shifts near the center of a shifted spectrum [10]. Meanwhile, chaotic environmental noise like scattered dust or mist spreads across high-frequency ranges. This separation allows a network to bypass superficial color fluctuations and lock onto the core structural textures of a scene. In global versus local context, spatial convolutions operate on a strictly local scale, requiring many deep layers to establish a global field of view. In contrast, every single coefficient in a frequency spectrum contains information derived from the entire spatial image. This global context allows shallower networks to capture macro-level texturing without requiring massive structural depth [11]. The primary disadvantage of a pure frequency representation is the loss of spatial localization. Because the spectrum maps how often an intensity changes rather than where it changes, conventional deep networks optimized for spatial RGB arrays struggle to converge when fed abstract, unaligned frequency maps. By applying a logarithmic scaling transformation and quadrant-shifting [14] operations, we centralize the low-frequency components and control the extensive dynamic range of the spectrum. Passing single-channel frequency maps into a standard, deeply stacked VGG16 architecture frequently results in severe feature dilution, gradient saturation~~,~~ or total failure to converge [12]. To resolve this structural mismatch, this paper introduces a frequency-domain pre-processing approach that mitigates ambient spatial noise while adapting the downstream [13] neural architecture to preserve fragile spectral coefficients. The remainder of this paper is structured into 6 sections: Section 1, is the Introduction; Section 2 delineates the 2. Proposed Methodology; Section 3 details the proposed Methodology Section 4 details on the Experimental setup; Section 5 is a discussion and analysis of Results and Section 6 concludes the study.

2. Proposed Methodology

The proposed methodology rather than relying on standard spatial configurations, engineers a Modified VGG16 topology. This custom architecture scales down block depths to prevent feature dilution, expands the kernel sizes to 5 x 5 to accommodate global frequency mapping and applies adaptive layer-wise dropout to insulate the network from over-parameterization. Utilizing the Fire and No Fire Images (FFNI) from the DeepFire benchmark dataset [15], it proposes to establish a noise-insulated framework that dramatically elevates predictive accuracy, offering a highly resilient asset for real-time Wildfire prediction systems as shown in Fig.1.

2.1 Mathematical Modelling & Dataset Architecture

The empirical foundation for this study is derived from the DeepFire Benchmark [15] Dataset consisting of two structurally balanced classes: Fire and No Fire Images (FFNI). Sample images are shown in Fig.2. The dataset is subjected to a rigorous 80-20 partition strategy, yielding highly isolated training and validation cohorts.

Frequency features describe periodic patterns and are obtained by transforming the image from the spatial domain using discrete transforms [16]. The analytical breakthrough of this research lies in the structural mapping of raw visual data into the complex spectral domain prior to neural representation [17]. The mathematical framework leverages the Two-Dimensional Discrete Fourier Transform (2D DFT) [18] accelerated via the Fast Fourier Transform (FFT) algorithm. Given an input spatial-domain image f (x, y) of dimensions MxN, where x and y represent discrete spatial coordinates, the transformation into the frequency domain F(u, v) is formally defined by the equation (1) [19] as follows:

$$F(u,v) = \sum_{X=0}^{M-1}\sum_{Y=0}^{N-1} f(x,y).\, e^{-j2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)} \qquad (1)$$

where u and v denote spatial frequencies along the horizontal and vertical axes, respectively and j = $\sqrt{-1}$. To optimize this input for representation within convolutional layers, the complex spectrum is decomposed into its magnitude component |F(u,v)|, which captures structural power [20] distribution as defined in equation (2).

$$|F(u,v)| = \sqrt{}(\mathrm{Re}\{F(u.v)\}^2 \ln\{F(u,v)\}^2 \qquad (2)$$

where Re stands for the Real part of a complex number. In signal processing, image processing (such as analysing frequency domains), and mathematics, a Fourier Transform or frequency-domain function F(u,v) results in a complex number for any given coordinates (u, v). To handle the extensive dynamic range of the frequency components and prevent numerical gradient saturation during backpropagation, a logarithmic scaling transformation [21] is applied as in equation (3).

$$|F(u,v)| = \sqrt{}(\mathrm{Re}\{F(u.v)\}^2 \ln\{F(u,v)\}^2 \qquad (3)$$

The resulting spectral matrices D(u, v) undergo a quadrant shift operation to centralize the low-frequency components, shifting the zero-frequency component to the centre of the spectrum and are shown the Fig.3. This centralized representation generates the final spectral images used for network training.

3. Proposed Neural Topologies & System Design

To process the extracted frequency maps, two distinct structural configurations of Convolutional Neural Networks are evaluated: the Baseline VGG16 architecture [22] as shown in Figure 4 and the proposed modified VGG16 configuration. Standard VGG16 relies on uniform 3x3 convolutional filters stacked deeply across 5 blocks, built primarily to extract hierarchical spatial features from RGB scenes. However, passing transformed single-channel frequency maps into a heavy, deep network results in massive feature dilution [21], causing the network to stall or fail to converge effectively on abstract spectral input patterns.

To resolve this limitation, the proposed modified VGG16 optimizes the architecture by scaling down the network depth, adjusting filter distributions and applying localized receptive fields tailored specifically to high-frequency centre shifting. This custom tuning eliminates structural redundancy while protecting fragile spectral coefficients. The structural comparison between the two network topologies is synthesized in Table 1.

4. Experimental Setup & Optimization Metrics

All computational models are implemented utilizing the TensorFlow/ Keras framework and executed on an enterprise-grade hardware stack comprising an NVIDIA H100 Tensor Core GPU with 156 GB RAM and 64-Core Processor. Optimization was driven uniformly across architectures via the Adaptive Moment Estimation (Adam) optimizer. The objective cost function is calculated using binary cross-entropy loss [23], formulated as in equation (4).

$$L_{BCE} = -\frac{1}{N}\sum_{1=1}^{N}[Y_i \log(\text{Yhat}_i) + (1 - \text{yi}) \log 1 - \text{yhat}_i\ ] \quad (4)$$

Where, Y-hat is the model's predicted probability (between (0.0) and (1.0). L_{BCE} is total binary cross-entropy loss value. Y_i is the actual target binary indicator and y_i represents the continuous predicted scalar probability. N is total number of samples in your dataset. Hyperparameter tuning curves utilize a learning rate schedule initialization of $\alpha_0 = 10^{-4}$ coupled with a callback-driven decay coefficient if validation loss plateaus for three consecutive epochs.

5. Results & Comparative Discussion

Accuracy is a metric that measures the percentage of correct predictions made by a neural network out of the total number of predictions [24]. It is primarily used to evaluate classification models and is calculated using the following formula defined in equation (5):

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \quad (5)$$

In the context of a confusion matrix, which tracks True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), the formula is written as in equation (6). The values for the two models of baseline VGG16 are tabulated in Table 2 along with the other top-ranking models in literature. The accuracy results favour the modified VGG 16 model.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

The baseline VGG16 model while processing standard spatial-domain visual signals struggled significantly, stabilizing at a poor accuracy rate of 62.0% and modified version produces an accuracy of 94.0 %. The baseline model demonstrates an extreme vulnerability when confronted with outdoor background variations, mistaking lighting shifts, clouds and complex shadows for active wildfire elements due to spatial feature confusion. Conversely, the proposed FFT-based modified VGG16 configuration achieved a remarkable classification performance threshold of 94.0%. By downsizing the block depths and adjusting the filter capacities to match the single-channel frequency maps, the network effectively avoided feature dilution. The inclusion of adaptive dropout successfully insulated the network from over-parameterization, allowing it to focus exclusively on discriminating high-frequency texture components. To validate the significance of these findings, the proposed framework was benchmarked against existing frequency-domain wildfire detection literature, as summarized in Table 2. Table 2 is a listing of accuracy values towards Wildfire prediction produced by different multivariate DL models likes YOLO v8 backbone, TD Fusion UNet, ResNet18 and Temporal Sequence Descriptors. The accuracy values are lesser when compared to the modified version 94% accuracy produced by VGG16 for Wildfire prediction.

The proposed framework outperforms existing state-of-the-art models by a margin of at least 9.0%, demonstrating that localized, shallow feature blocks equipped with a larger receptive field (5 x 5 kernels) map abstract spectral properties significantly better than alternative hybrids or unoptimized deep architectures. The models were evaluated comprehensively against classification accuracy, computational inference speed, and structural robustness. Performance curves were constructed using Area Under the Receiver Operating Characteristic (AUC-ROC) to isolate model sensitivity thresholds [24] under complex feature visibility constraints in Fig.5.

6. Conclusion

This paper has demonstrated the high efficacy of shifting computer-vision wildfire classification pipelines from traditional spatial RGB arrays into localized frequency-domain matrices. By employing a 2D Fast Fourier Transform framework combined with a customized, depth-optimized VGG16 topology, this system minimized ambient environmental noise interference while maintaining high architectural efficiency. The significant increase in classification precision from 62.0% to 94.0% has proven that uncalibrated deep spatial architectures fail to parse messy natural environments, whereas specialized spectral-domain mapping provides absolute resilience. This advancement validates the operational potential of deploying spectral-spatial deep learning frameworks onto real-time processing techniques and technologies, to drastically improve early wildfire detection and disaster management globally.

References

1. Waring RH, Schlesinger WH. Forest ecosystems, analysis at multiples scales, Vol. 55. ISBN: 978-0-12-735443-9.
2. Doolin DM, Sitar N. Wireless sensors for wildfire monitoring. InSmart structures and materials 2005: sensors and smart structures technologies for civil, mechanical, and aerospace systems 2005 May 17 (Vol. 5765, pp. 477-484). SPIE. DOI.org/10.1117/12.605655.
3. Voulodimos A, Doulamis N, Doulamis A, Protopapadakis E. Deep learning for computer vision: A brief review. Computational intelligence and neuroscience. 2018; 2018 (1):7068349. DOI.org/10.1155/2018/7068349.
4. Goudie AS. Human impact on the natural environment: Past, present and future. John Wiley & Sons; 2018 Aug 27. DOI: 10.1109/CSSE.2008.837.
5. Brownlee J. Deep learning for computer vision: image classification, object detection, and face recognition in python. Machine Learning Mastery; 2019 Apr 4.
6. Wang S, Cao J, Philip SY. Deep learning for spatio-temporal data mining: A survey. IEEE transactions on knowledge and data engineering. 2020 Sep 22; 34(8):3681-700. DOI: 10.1109/TKDE.2020.30255 80.
7. Geusebroek, J-M., Rein Van den Boomgaard, Arnold W. M. Smeulders, and Hugo Geerts. "Color invariance." IEEE Transactions on Pattern analysis and machine intelligence 23, no. 12 (2001): 1338-1350. DOI: 10.1109/34.977559.
8. Cerna M, Harvey AF. The fundamentals of FFT-based signal analysis and measurement. National Instruments, Junho. 2000 Jul; 54.
9. Hess DJ, Foss DJ, Carroll P. Effects of global and local context on lexical processing during language comprehension. Journal of Experimental Psychology: General. 1995 Mar; 124(1):62. DOI.org/10.1037/0096-3445.124.1.62
10. Cao S, Lu W, Xu Q. Grarep: Learning graph representations with global structural information. In Proceedings of the 24th ACM international on conference on information and knowledge management 2015 Oct 17 (pp. 891-900). DOI.org/10.1145/2806416.28065.
11. Boashash B, Khan NA, Ben-Jabeur T. Time–frequency features for pattern recognition using high-resolution TFDs: A tutorial review. Digital Signal Processing. 2015 May 1; 40: 1-30. DOI. 10.1016/j.dsp.2014.12.015
12. Stathatos S, Hobley M, Perona P, Marks M. Saved: Learning to denoise low-snr video for improved downstream performance. In2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2026 Mar 6 (pp. 6851-6861). IEEE. DOI: 10.1109/WACV61042.2026.00662
13. Froehlich GK. Optical information processing experiments for undergraduate engineers (Doctoral dissertation, Texas Tech University).
14. Khan A, Hassan B, Khan S, Ahmed R, Abuassba A. DeepFire: A novel dataset and deep transfer learning benchmark for forest fire detection. Mobile Information Systems. 2022;2022(1):5358359.DOI.org/10.1155/2022/5358359.
15. Boashash B, Khan NA, Ben-Jabeur T. Time–frequency features for pattern recognition using high-resolution TFDs: A tutorial review. Digital Signal Processing. 2015 May 1; 40:1-30. DOI. 10.1016/j.dsp.2014.12.015
16. Fumero M, Pegoraro M, Maiorca V, Locatello F, Rodolà E. Latent functional maps: a spectral framework for representation alignment. Advances in Neural Information Processing Systems. 2024 Dec 16; 37:66178-203. DOI: 10.52202/079017-2115.
17. Sundararajan D. The discrete Fourier transform: theory, algorithms and applications. World Scientific; 2001. ISBN: 981-02-4521-1.
18. Weng A, Li D, Yang Y, Li S, Li J, Li S. Transforming a time-domain electromagnetic signal to a frequency-domain electromagnetic response using regularization inversion. Geophysics. 2017 Sep 1; 82(5):E287-95. DOI.org/10.1190/geo2016-0505.1.
19. Guo H, Marfurt KJ, Liu J. Principal component spectral analysis. Geophysics. 2009 Jan 1; 74(4):P35-43. DOI.org/10.1190/1.3119264

20. Lee HM, Chen CM, Huang TC. Learning efficiency improvement of back-propagation algorithm by error saturation prevention method. Neurocomputing. 2001 Oct 1;41(1-4):125-43. DOI.org/10.1016/S0925-2312(00)00352-0
21. Qassim H, Verma A, Feinzimer D. Compressed residual-VGG16 CNN model for big data places image recognition. In2018 IEEE 8th annual computing and communication workshop and conference (CCWC) 2018 Jan 8 (pp. 169-175). IEEE. DOI: 10.1109/CCWC.2018.8301729.
22. Ruby U, Yendapalli V. Binary cross entropy with deep learning technique for image classification. Int. J. Adv. Trends Comput. Sci. Eng. 2020 Aug 25; 9(10):5393-7. DOI.org/10.2139/ssrn.4944298
23. Hestness J, Ardalani N, Diamos G. Beyond human-level accuracy: Computational challenges in deep learning. InProceedings of the 24th symposium on principles and practice of parallel programming 2019 Feb 16 (pp. 1-14). DOI.org/10.1145/3293883.329571
24. Ma X, He Y, Du P, Lv W, Yang Y. Wildfire Detection from a Drone Perspective Based on Dynamic Frequency Domain Enhancement. Forests. 2025 Oct 21; 16(10):1613. DOI.org/10.3390/f16101613
25. Luo Y, Rong S, Watts A, Cetin AE. U-net with hadamard transform and dct latent spaces for next-day wildfire spread prediction. arXiv preprint arXiv:2602.11672. 2026 Feb 12. DOI.org/10.48550/arXiv.2602.11672.
26. Lin X, Li Z, Chen W, Sun X, Gao D. Forest fire prediction based on long-and short-term time-series network. Forests. 2023 Apr 10; 14(4):778. DOI.org/10.3390/f14040778
27. Zhang Z, Zhao J, Zhang D, Qu C, Ke Y, Cai B. Contour based forest fire detection using FFT and wavelet. In2008 International conference on computer science and software engineering 2008 Dec 12 (Vol. 1, pp. 760-763). IEEE. DOI: 10.1109/CSSE.2008.837
28. Ghibeche Y, Sellam A, Nouri N, Khaldi A, Harrane A, Ghibeche I. Machine learning for forest fire prediction: a case study in north algeria. Ingénierie des Systèmes d'Information. 2024 Feb 1; 29(1):337. DOI:10.18280/isi.290133

Figures

Fig.1: Architecture of the Proposed System.

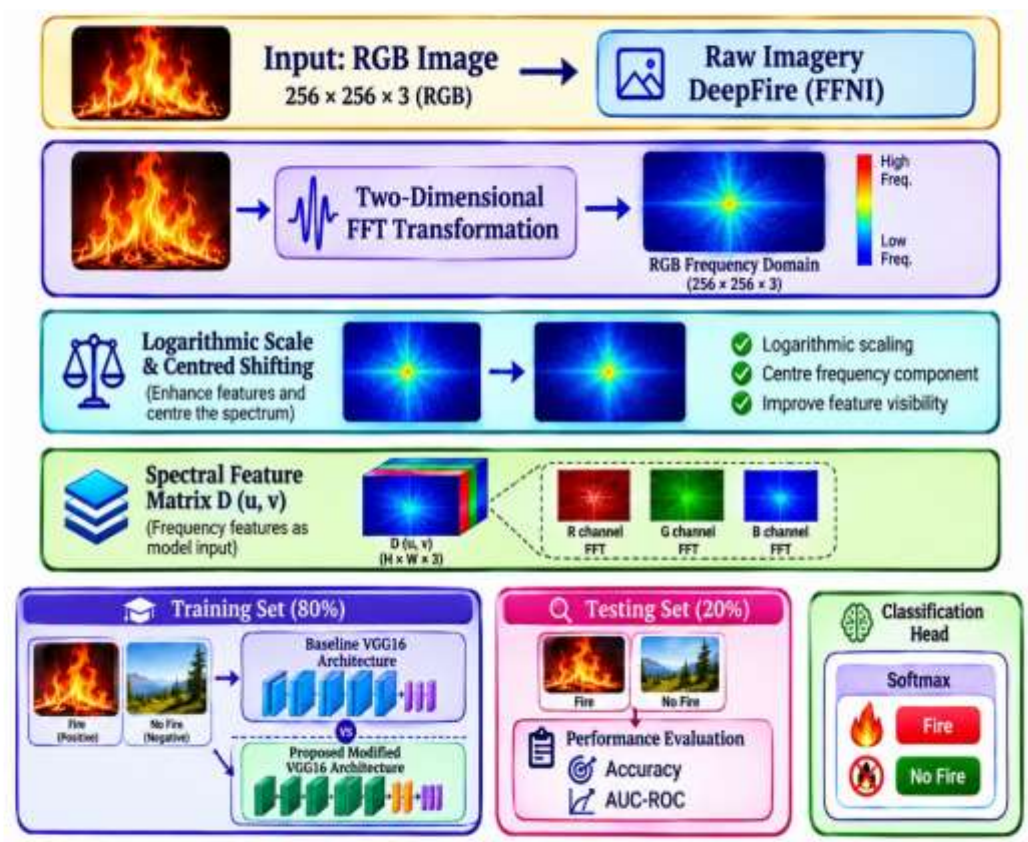
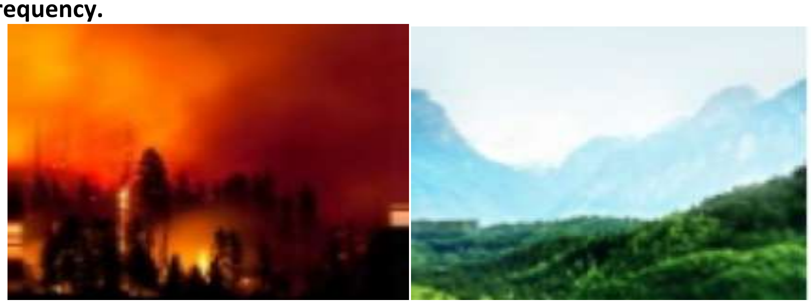


Fig. 2: Images from FFNI dataset (A) Fire Image and (B) NoFire image.

(A) (B)

Fig. 3: (C) Fire and (D) NoFire grayscale frequency images from FFNI after Fast Fourier Transform (A) Fire and (B) NoFire grayscale frequency.



(A) **(B)**

(C) **(D)**

Fig .4: Architecture of the Baseline VGG16 Model.

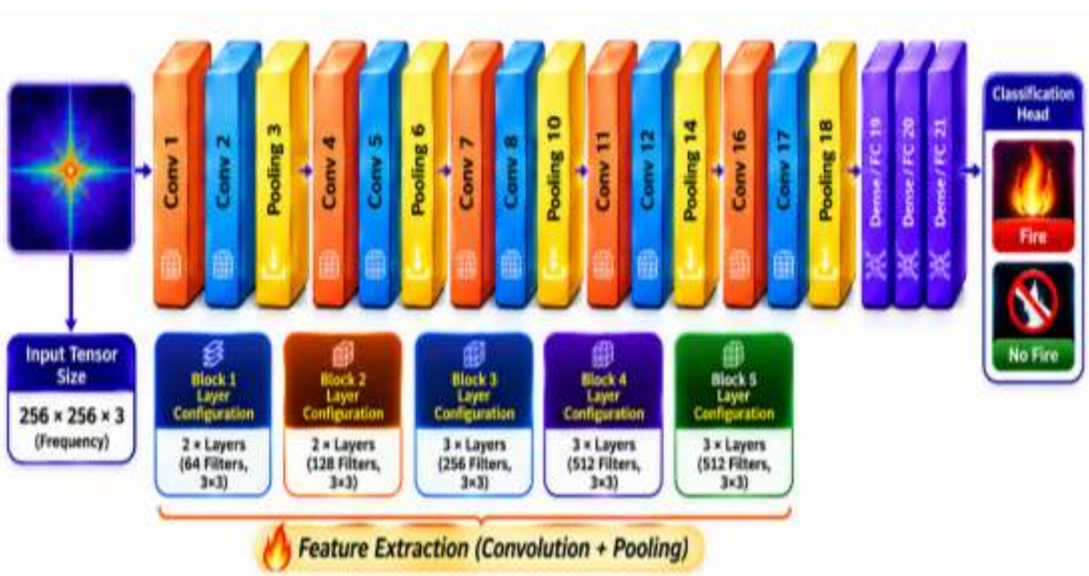
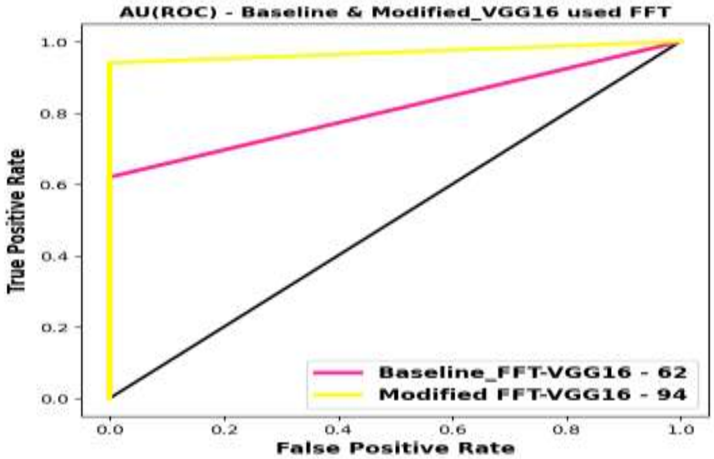


Fig .5: AUROC curve of VGG16 Networks.



Tables

Table 1: Structural Attributes of Baseline vs. Modified VGG16 Architectures

Structural Attribute	Baseline VGG16 Architecture	Proposed Modified VGG16 Architecture
Input Tensor Size	256 x 256 x 3 (Frequency- 3x3 filters)	256 x 256 x **1** (FFT Shifted Map-5x5)
Block 1-Configuration	2 Layers (64 Filters, 3x3)	2 Layers (8 Filters, 5 x 5)
Block 2-Configuration	2 Layers (128 Filters, 3 x3)	2 Layers (12 Filters, 5 x 5)
Block 3-Configuration	3 Layers (256 Filters, 3x3)	2 Layers (16 Filters, 5 x 5)
Block 4-Configuration	3 Layers (512 Filters, 3 x 3)	3 Layers (20 Filters, 5 x 5)
Block 5-Configuration	3 Layers (512 Filters, 3 x 3)	3 Layers (24 Filters, 5 x 5)
Classification Head	2 (Dense Layers, 84 Neurons) Dropout Regularization	2 Dense layers Adaptive Layerwise Dropout (1093 Neurons)
	Softmax layer -84 Neurons	Softmax Output Dense layer- 304 Neurons

Table 2: Accuracy Comparison of modified VGG16 Architecture and DL algorithms in literature

Models	Features	Accuracy
Ma, Xiaohui et al., (2025) [25]	YOLOv8 backbone & 2D FFT	88.4% mAP50
Luo, Yingyi et al., (2026) [26]	2D Discrete Cosine Transform branch used ResNet18-UNet	0.591 – 0.702 F1-Score
Lin, Xufeng et al., (2023) [27]	FFT used TD-Fusion U Net	70.2%
Zhang et al., (2008) [28]	Wavelet Multi-scale Coefficients + Temporal Sequence Descriptors	75% -79%
Ghibeche.Y. and Sellam et al., (2024) [29]	Frequency & RF	74–87%