International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 5,2025
https://theaspd.com/index.php

SpecTralUNetFormer: Advancing Hyperspectral Medical Image Segmentation through Spectral-Spatial-Attentive Learning

L.K. Suresh Kumar[1], P. Rathna Sekhar[2], K. Srinivasa Chakravarthy[3], M. Sathya Devi[4], E. Radha Krishnaiah[5]
[1]Department of Computer Science and Engineering, Osmania University, Hyderabad, India, lksureshkumar@osmania.ac.in
[2]School of Engineering, Anurag University, Hyderabad, India. rs.pesaramelli@gmail.com
[3]Assistant Professor, Department of Information Technology, Vasavi College of Engineering(A), Hyderabad, ks.chakravarthy@staff.vce.ac.in
[4]Assistant Professor, Department of Information Technology, Vasavi College of Engineering(A), Hyderabad, sathyamaranganti@staff.vce.ac.in
[5]School of Engineering, Anurag University, Hyderabad, India. rkenikapally@gmail.com

Abstract
Surgical precision is greatly improved after brain tumors are accurately diagnosed and traced in the operating room. Hyperspectral Imaging (HSI) is a new method that discriminates between healthy tissue and suspicious areas in real time according to their spectral signatures. This paper compares the performance of five deep learning models, i.e., CNN, 3D CNN, Vision Transformer (ViT), U-Net, and the SpecTralUNetFormer, introduced in this paper, on the Hyperspectral Imaging Benchmark for Intraoperative Brain Tumor Detection dataset. The dataset has 62 hyperspectral images captured from 34 subjects, 128 spectral bands from 400 nm to1000nm. The SpecTralUNetFormer proposed here combines 3D CNNs for learning spectral-spatial features, a U-Net encoder-decoder for spatial localization, and a Transformer bottleneck for learning long-range dependencies. Data preprocessing includes normalization, PCA-based spectral band reduction, and data augmentation. The models are tested for classification accuracy, AUC, and computational efficiency, and a comparative analysis of the various architectures is shown. The experiments show that SpecTralUNetFormer performs better than conventional architectures with improved segmentation accuracy and improved generalization in hyperspectral brain tumor detection. The objective of this work is to improve intraoperative decision-making during surgery by using deep learning methods for real-time tumor detection, ultimately resulting in improved surgical accuracy and patient outcomes.
Keywords: *Hyperspectral Imaging, Brain Tumour Detection, Deep Learning, CNN, RNN-LSTM, Vision Transformer, U-Net, Intraoperative Imaging.*

INTRODUCTION

The detection and treatment of brain tumors is a complex problem in the field of neurosurgery, as it requires care and precision in regard to both the timing and accuracy of treatment. Conventional imaging devices, such as Magnetic Resonance Imaging (MRI) and Computed Tomography (CT), do offer valuable imaging information, yet they fall short of offering real-time, detailed spectral data that is critical during operations. Hyperspectral Imaging (HSI) is a non-invasive imaging technique that has come into prominence recently and is effective at telling healthy tissue apart from malignant tissue due to its ability to capture detailed spectral signatures. With deep learning algorithms, the integration of HSI imaging techniques can significantly improve Tumor detection by utilizing superior feature extraction and classification-methods. Convolutional neural networks (CNN), recurrent neural networks (RNN), and Vision Transformers (ViT) approaches have been successful in analyzing medical images. These models outstand when it comes to the extraction of spatial, spectral, and contextual features as such models are, therefore, ideal candidates for the most challenging hyperspectral data. Nonetheless, it is still an open research problem to discover which deep learning architecture is the most suitable for brain Tumor-analysis. The Hyperspectral Imaging Benchmark for Intraoperative Brain Tumour Detection dataset provides a standardized framework for evaluation of diverse deep learning approaches. It is composed of 62 hyperspectral images from 34 patients and the spectral range is from 400 to 1000 nm, having 128 spectral bands, therefore making it possible to study in detail the classification performance for the Tumour.
This study is aimed at evaluating the performance of five of the latest deep learning models on the hyperspectral imaging data for the real-time detection of brain Tumours. The main objects of this research are to assess the proposed CNN, 3D CNN, Vision Transformer, and U-Net models in processing hyperspectral brain imaging data. To help determine the effects of spatial and spectral feature extraction on classification accuracy. To identify the advantages and disadvantages of these models about hyperspectral data and further direction of future research in Tumour detection during operations.

The rest of this paper is organized as follows. Related works of deep learning models for hyperspectral imaging and medical image classification are presented in Section 2. The data set and pre-processing techniques adopted in the present study are explained in Section 3. Section 4 throws light on the methodology in which the model architectures along with the training strategies have been discussed. Section 5 provides an overview of the experimental set-up followed by the evaluation metrics. The results and comparative analyses of the models are discussed in Section 6.

RELATED WORKS

CNNs has proven strong ability to identify and extract spatial patterns in images in both traditional remote sensing and hyperspectral image classification tasks. The convolutional layers in CNNs perform exceptionally well at recognizing spatial patterns and features such as shapes and textures, and hierarchical structures on scales within the visual field [1]. There have been several studies that have explored creative uses of CNNs to expand on detection capabilities of spatial patterns. The hierarchical multi-scale convolutional neural networks (HMCNN-AC) method focuses on multi-scale image patches to take advantage of spatial information at multiple scales, which allows for the detection of a variety of object shapes and sizes throughout the image plane [2]. The DHCNet approach for hyperspectral image classification applies deformable convolutional sampling locations, which allows for adjustments and clashes to complex spatial contexts. Some researchers have explored different approaches of merging CNNs with other methods to improve spatial pattern recognition. For example, combining the local binary patterns (LBP) features with a CNN demonstrated an increase in classification performance, because the LBP is very effective at extracting spatial features [4]. As well, morphological functions with CNNs can potentially provide more accurate representation of nonlinear information, while retaining key features of hyperspectral images such as borders, shape, and structural detail [5].

In practical application, RNN-LSTMs have shown promising results when applied to hyperspectral data as sequential spectral data for a classification task. The models can learn the spectral correlations and dependencies present in hyperspectral images (HSI), and successfully increase classification accuracy. The conventional method for incorporating spectral signatures as ordered sequences, only accounts for one-directional correlation to nearby bands in the direction of the wavelengths. Nonetheless, a bidirectional long-short term memory (Bi-LSTM) network can be able to explore the bidirectional nature of the spectral correlation of an HSI image if every band image contains relationships with a prior band image and the subsequent band image [6]. This allows for a deeper exploration of the spectral content. There have been some conversations on the possibility of hybrid developments of neural networks utilizing the positives of other neural networks' architecture. The convolutional recurrent neural networks (CRNN) has been developed to first learn some middle-level, locally invariant features through the convolutional layers from the input data, and then spectrally contextualized information utilizing the recurrent layers [7]. The authors can confidently state these methods have improved classification performance compared to standard methods and other methods of state-of-the-art deep learning for hyperspectral data classification. RNN-LSTMs have been very effective in handling hyperspectral data as time series data in a sequential manner for its spectral content. The capability of these models to model long-range dependencies and bidirectional relationships also makes them legible to apply to HSI classification problems. Further still, when we can also take advantage of the spatial information with either an attention mechanism or through hybrid models these models will likely help the analysis of hyperspectral images (Mei et al., 2022; Wu & Prasad, 2017).

3D Convolutional Neural Networks (CNNs) have shown a great ability to extract spatial and spectral features from volumetric data in a unified manner. For instance, in terms of hyperspectral image (HSI) classification, 3D CNNs have enjoyed great successes due to their demonstration of full usage of the 3D spatial input of the HSI through the ability to regress all of the 3D content in one pass (Xu et al., 2020; Yang et al., 2020). In essence, 3D CNNs can considerably lessen spatial redundancy as well as sufficiently decrease the receptive field size to diminish the drawbacks of classic 3D CNNs in HSI classification [8]. It is important to emphasize that although 3D CNNs are superior in extracting spatial-spectral features, they may also introduce large-scale parameters and complexity to networks. To combat this issue, some researchers proposed hybrid methods to merge the extraction of both spatial-spectral features as well as total computational cost, by leveraging both 2D and 3D CNNs [9]. Attention mechanisms have also been added to focus on relevant feature areas and relevant spectral bands to potentially further improve classification results (Liang et al., 2023; Xu et al. 2020). 3D CNNs have made great strides in the area of volumetric image segmentation and computer-aided detection of medical imaging. For 3D MRI prostate segmentation, a fully convolutional neural network volumetric that can predict segmentation for the whole volume at once has been proposed [10]. In the instance of lung nodule detection, a 3D CNNs based system elevated the state-of-the-art in detection by combining not only data-driven features but also some useful a priori knowledge [11]. These examples clearly illustrate the potential of 3D CNNs to handle challenging spatial relationships in medical imaging data and represent a significant advantage to standard 2D methods in volumetric contexts.

Hybrid 2D-3D CNN methods leverage processing advantages to enhance computational ability for hyperspectral image classification from both 2D and 3D CNNs, without the disadvantages of either 3D or 2D CNNs. The 3D CNN component of hybrid models enables the means to extract spatial-spectral features from a set of stacked spectral bands at the same time

without losing the 3D nature of hyperspectral data (Roy et al., 2019; Yang et al., 2020), maximizing the capability of the model to obtain the spatial-spectral information necessary for accurate classification. The 2D CNN represents abstract spatial information which supports the overall feature representation (Chang et al., 2022; Roy et al., 2019). Hierarchical grouping of both abstract 2D and 3D CNNs is assigned a lower overall requirement for computation than the number of operations required for the overall 3D CNN (Roy et al., 2019; Yu et al., 2020). For example, the HybridSN model examined in [12] utilizes a spectral-spatial 3D-CNN model along with a spatial 2D-CNN model which simplifies the model while still retaining strong accuracy results for classification. As well, the Reduced 2D-3D CNN model evaluated in [13] utilizes a 2D convolution block to extract spatial features as well as a 3D convolution layer to correlate spectral bands, simply as a way to seek out the compromise between time taken to extract features from the raw hyperspectral data, but also the time taken to process the classification of the spectral data. Hybrid 2D-3D CNN methods are a promising option for hyperspectral image classification because they combine spatial and spectral information with a lower computation cost. When experimenting on standard datasets (Chang et al., 2022; Roy et al., 2019; Yang et al., 2020), they often provide better performance than just traditional 2D or just 3D CNNs.

Vision Transformers (ViTs) have made significant strides in medical imaging tasks due to their uniquely powerful capacity to learn long-range dependencies with self-attention mechanisms (Ali et al., 2023; Naseer et al., 2021). ViTs have been strong competitors to convolutional neural networks (CNNs) for semantic segmentation tasks in medical image analysis because they can naturally represent rich global dependencies [14]. Interestingly, ViTs have been robust to extreme occlusions, perturbations, and domain shifts for medical imaging tasks. When random occlusions affect 80% of the image, ViTs can continue to obtain even 60% top-1 accuracy on ImageNet examples [15]. This is largely due to the self-attention mechanism that provides ViTs with dynamic, flexible receptive fields for various image conditions.

Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) have varying levels of computational efficiency when relating to medical imaging, and each has its strengths and limitations. ViTs appear to provide relatively better performance for capturing long-term dependencies, global context, and the overall different types of information important to medical image analysis (He et al., 2023; Takahashi et al., 2024). However, these models consistently show higher computational costs. The self-attention process in ViTs has a quadratic computational cost which is an issue, especially for higher resolution feature maps [16]. From this arbitrary computational cost, we may be actually better off with a CNN for any real-world applications [17]. A few studies have interestingly provided a methods for ViTs to be efficient. For example the H2Former model outperforms TransUNet with 30.77% fewer parameters and 59.23% fewer FLOPs on the KVASIR-SEG dataset. Likewise, EdgeViTs have been designed to match lightweight CNNs in on-device efficiency, prioritizing real-world metrics such as latency and energy efficiency over mere FLOPs or parameter numbers [18]. In summary, although ViTs tend to perform better in medical imaging tasks, they tend to consume more computational resources compared to CNNs. Nonetheless, current research is aimed at creating more effective ViT architectures, and the hybrid models integrating the advantages of both ViTs and CNNs are being explored as potential solutions for achieving the trade-off between performance and computational cost in medical image analysis (Guo et al., 2022; Pan et al., 2022).U-Net and its variants, including ResUNet, have been extensively used for pixel-wise segmentation to identify Tumour areas from normal tissue in medical imaging. Such models have proven noteworthy performance on various imaging modalities and Tumour types. For segmentation of brain Tumour, approaches based on the U-Net have proved phenomenal. A rescaled U-Net architecture reported 99.4% accuracy on BraTS 2020 dataset more than other deep learning architectures [19]. Another model, the ResUNet++ model being an enhanced iteration of ResUNet, reached high dice coefficient values of 81.33% and 79.55% for segmenting polyp in colonoscopy images [20]. Surprisingly, certain studies have integrated U-Net with other methods to provide better performance. The SGEResU-Net model with residual blocks and spatial group-wise enhance attention blocks integrated into 3D U-Net architecture reported dice values of 83.31%, 91.64%, and 86.85% in improving Tumour, whole Tumour, and Tumour core respectively on BraTS 2021 dataset [21]. Another method, the Spherical Projection-based U-Net (SPU-Net), not only enhanced segmentation accuracy but also gave a means of quantifying segmentation uncertainty in glioma detection [22].

DATASET DESCRIPTION

This database, reported in [23], is composed of 61 HS images of 34 patients with both primary (high-grade and low-grade) and secondary Tumours. It is shown in the study that HSI with a conceived processing scheme reaches a best median macro F1-Score of 70.2 ± 7.9% on the test set based on both spectral and spatial information [23]. Surprisingly, though this dataset sets a standard for in-vivo brain Tumour detection, other research has yielded different outcomes. For example, [24] has reported a better overall accuracy of 80% in multi-tissue classification using deep learning in glioma surgery. Further, [25] had a still higher average accuracy of 91.36% for the detection of head and neck Tumours in animal models using convolutional neural networks (CNN). This data set was obtained by a collaborative process by institutions like the University of Las Palmas de Gran Canaria and the University Hospital Doctor Negrin of Gran Canaria, Spain. Data was

recorded intraoperatively by a customized HS imaging system during in-vivo imaging of brain tissues for real surgical-interventions.

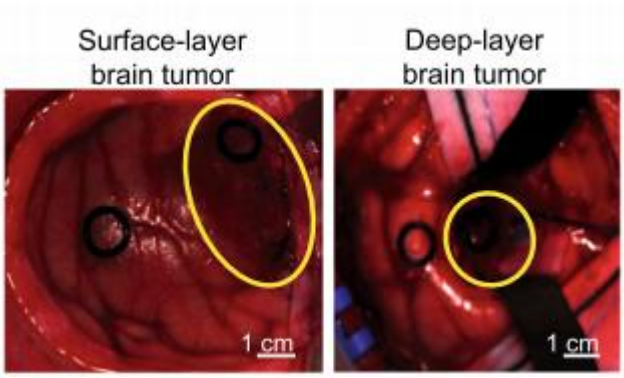

Fig. 1: Synthetic RGB Images Showing Visual Differences Between Surface-Layer and Deep-Layer Brain Tumors During Neurosurgical Procedures [23]

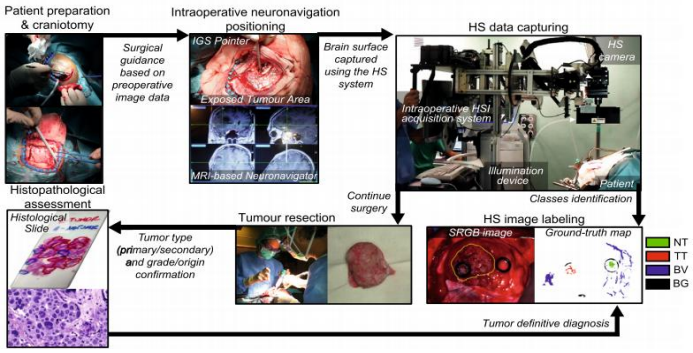

Fig. 2: Workflow of Hyperspectral Data Acquisition, Intraoperative Positioning, Tumor Resection, and Image Labeling for Brain Tumor Diagnosis and Surgical Guidance [23]

Table 1: Key Characteristics of the Intraoperative Hyperspectral Imaging Dataset Used for Brain Tumor Detection and Classification

Feature	**Description**
Number of Images	62 hyperspectral images
Number of Patients	34 different patients
Spectral Range	400 to 1000 nm
Spectral Bands	128 spectral bands
Spatial Resolution	High-resolution per pixel spectral data
Data Format	Available in MATLAB and standard image formats

Table 2: Detailed Metadata Annotations Associated with the Hyperspectral Brain Tumor Imaging Dataset

Metadata Attribute	**Description**
Classes	4 primary classes (TT, NT, BV, BG)
Tumour Tissue (TT)	Labeled regions of Tumour
Normal Tissue (NT)	Labeled regions of normal brain tissue
Blood Vessels (BV)	Labeled blood vessels
Background (BG)	Background or artifacts
Tumour Type	Primary or secondary Tumours

Tumour Grade	Low-grade or high-grade Tumours
Histopathological Diagnosis	Specific diagnoses based on pathology

PROPOSED METHODOLOGY

A. Data Acquisition

The imagery used within this study is a significant advancement in brain imaging technology and has abundant information available for detection and processing of Tumour. Hyperspectral imaging's ability to detect light over a wide range of wavelengths (400-1000 nm) can provide a rich spectral fingerprint of brain tissues. Combination of rich spectral information with high-resolution spatial information can enable subtle discrimination of Tumour from non-Tumour tissues beyond the imaging capability of traditional imaging modalities. Precise marginal annotation of every Tumour image provides a required platform for machine learning of intricate computational models to prepare them to absorb and recognize subtle spectral signatures that are available in neoplastic tissues. The value of this data is not merely that it is technologically advanced; it also has profound implications for improving neurosurgical success and understanding brain Tumours. By providing abundant data in the form of high-quality annotated images, this database provides algorithms for detection of Tumour with the capability to enhance accuracy and reliability. These computational models have been trained with such specificity and completeness; they possess the potential to aid neurosurgeons in real-time, during surgery, to enhance their ability to identify and resect Tumour tissue. The high-resolution spectral data available in these images also have the ability to yield new information about the biochemical composition of brain Tumours, which could lead to new diagnostic markers or therapeutic targets. The integrity and quality of this data are of paramount importance as it not only impacts the efficacy of the computational models, but also possibilities for breakthroughs that will help propel advances in neuro-oncology.

B. Preprocessing

Preprocessing is an important step in raw hyperspectral data handling for deep learning models built using the most applicable methods. The spectral variations are normalized to correct reflectance values on the same scale to reduce illumination and sensor noise variation among samples. Dimensionality reduction is achieved using Principal Component Analysis (PCA) reducing the data while also reducing computational load with the potential of losing useful spectral information. Next, we apply methods of data augmentation consisting of flipping, rotation, and spectral shifting to increase the generalizability of the model while reducing the risks of overfitting. The last step involves partitioning the dataset into three distinct units: 80% for model training, 10% for validation, as well as 10% for testing, to facilitate proper training and testing of the model on unseen data. These preprocessing operations are critical towards the improvement of the quality and reliability of hyperspectral data analysis. Spectral normalization procedures are employed for the elimination of inconsistencies in the dataset, with the aim of ensuring higher comparability in different samples, as well as acquisition techniques. The utilization of PCA as a method of data dimensionality ensures not only reduced computational loading, but also identifies the most critical spectral features, separating those of importance. Data augmentation techniques are critically crucial in expanding the dataset artificially, later exposing the model to a variety of situations to increase its generalization capability. Proportionate division of the dataset ensures that an extensive dataset trains the model adequately, as well as being validated through a different set for hyperparameter optimization, thus ultimately being validated using entirely unseen data to justify an accurate measure of its behavior in real-world applications.

C. Deep Learning Models

In this section some of the standard deep learning models are discussed:

CNNs are a unique type of multi-layer neural networks that have been designed to detect visual patterns in pixelated images [32]. In CNN, "convolution" is an arithmetic operation that takes two functions and produces a third function by multiplying them, determining how one function's shape can be transformed by the other. In simpler terms, CNN does matrix multiplication of two image representations to get an output that extracts information from the image. CNN is a neural network like any other neural network, but the unique aspect is the convolutional layers, which introduce an element of complexity in the overall framework [33]. Convolutional neural network contains a diverse set of layers, such as convolution layers, pooling layers, and fully connected layers.

At the heart of the CNN, there lies the convolutional layer. This robust layer utilizes the convolutional filters, or kernels, on the input data to identify the features of the edges, the textures, or the patterns. The filters, while relatively small in size relative to the input data, scan the entire input with a defined stride. The filter, at every position, does element-wise multiplications with the corresponding input elements, and then calculates the sum and generates a feature map. The feature map can be represented as follow:

$$f(i,j) = \Sigma\Sigma(l(i+m, j+n) * k(m,n)$$

Where: $f(i,j)$ *is the value at position* (i,j) *in the feature map.*

l(i+m, j+n) is the value at position (i+m, j+n) in the input data.

K(m,n) is the value at position (m,n) in the convolutional filter.

$\Sigma\Sigma$ *represents the summation of overall spatial positions* (m,n) *in the convolutional filter.*

Pooling layers are crucial in reducing the spatial dimension of feature maps with the retention of important information. Pooling layers help reduce computational complexity and overfitting. The most common method is max pooling, where the largest value within a small region (pooling window) is retained while the rest is discarded. Max pooling downsampling is very efficient in preserving the most informative features of the feature map. After multiple convolutional and pooling layers, one or more fully connected layers are often a part of the CNN model architecture. The fully connected layers enable each neuron from the previous layer to connect with every neuron in the current layer, thus enabling a regular neural network architecture. The fully connected layers enable learning of global relationships as well as predictions from feature learning in the previous layers. The output layer is the final layer of the CNN model. For classification tasks, this layer typically consists of neurons with the same number of classes to be predicted. The output of these neurons is the confidence of the model in classifying the input data into each particular class.

During training, the CNN layers work together by forward propagation to determine the best set of weights and biases that minimize the model's loss on the given task. The learning is achieved via backpropagation and the optimization algorithm, which updates the model's parameters iteratively based on the gradients of the loss function with respect to the model's parameters.

RNN-LSTM

Long Short-Term Memory (LSTM) is a unique type of Recurrent Neural Network (RNN) designed to overcome the vanishing and exploding gradient problems experienced with standard RNNs [26]. Importantly, LSTM-RNNs have seen improved performance than deep neural networks (DNN) with various speech recognition and language identification tasks (Liu et al., 2016; Zazo et al., 2016). LSTM-RNNs have been remarkably successful in modeling the sequential nature of data while maintaining long-term dependencies [28]. In fact, LSTM-RNNs have shown success in variety of applications such as speech recognition, natural language processing, time series predictions, and even in autonomous driving [28]. It is worth mentioning that speech recognition models based on LSTM-RNNs have achieved state-of-the-art performance while still reporting decent performance with small models, fast convergence, and efficient use of model parameters (Beaufays et al., 2014; Sak et al., 2015). LSTM-RNNs have showed that they are extensible across many applications. For example, LSTM-RNN based regression modeling approach provides a superior mapping of noisy speech features to clean features in a speech enhancement setting when accounting for long-term acoustic context versus DNN-based methods [28]. Deep LSTM-RNN models in conjunction with linear regression models have produced state-of-the-art performance for traffic matrix prediction [29]. They are powerful and flexible models for sequence model tasks. They are attractive models in the machine learning space because they can handle long term dependencies and have generated improved performance across a large range of applications. However, the challenges remain related to their practical implementation on hardware due to high storage and computational costs [30]. Research is still ongoing to enable LSTM-RNN models to be less expensive and more effective - such as creating cost-effective versions [31] adding attention mechanisms, and hybrid models [27].

3D CNN

Three-dimensional Convolutional Neural Networks (CNNs) are a general-purpose technique in health care image analysis, providing ten thousand benefits over conventional two-dimensional CNNs for volumetric data. In this implementation, utilizing a network designed for three-dimensional medical image analysis, meaning three-dimensional image analysis from datasets, images can consist of computed tomography (CT) scans, waves of magnetic resonance imaging (MRI) scans, Angiography, and endoscopy. These are definitive forms of spatial reasoning or decision making in three-dimensional space (Huang et al., 2017; Singh et al., 2020). CNNs have outperformed, in terms of performance, opportunities in various medical imaging contexts such as the classification, segmentation, detection, and localization of lesions where 3D convolutional networks have produced a larger accuracy (34). For example, a lung nodule detection task, a study produced a state-of-the-art performance compared to CT nodule classification baselines produced by shallow learning (35). Another example of brain tumor classification, a study showed high proportions of accuracy using many classes of tumors (glioma, meningioma, and pituitary tumors) (36). One of the significant advantages of three-dimensional CNNs is the better ability to extract spatio-temporal features within volumetric data compared to two-dimensional CNNs (37). Particularly and

importantly in medical imaging, since knowing the spatial relationships between structures are important for accurate diagnostics. For example, a 3D CNN was used to efficiently screen head CTs for acute neurological events, decreasing diagnosis time from minutes to seconds with a sensitivity of >88% [38]. However, the use of 3D CNNs with medical imaging does come with a number of challenges. Most notably, the demand for large datasets that have 3D annotations, which are typically less available when compared to 2D image datasets [39]. In response to this, researchers have explored transfer learning approaches and 2D/3D hybrid architectures so that pre-trained 2D models may still be reused; while still being able to capture valuable 3D spatial cues [39]. Lastly, highlighting the need for minimizing resource consumption of 3D CNNs due to their computational efficiency is of a common issued researched area, since their models are generally even more, resource-hungry than a 2D alternative [37].

Vision Transformers

Vision Transformers (ViT) present a significant alternative to Convolutional Neural Networks (CNNs) for computer vision tasks, and they have achieved remarkable performance and generality (Naseer et al., 2021; Wang et al., 2025). ViTs also utilize self-attention to interpret images as a sequence of patches, which allows for the easy capture of both global and local information [40]. One of the most attractive strengths of ViTs is their robustness to catastrophic occlusions, distortions and domain shift. For instance, ViTs can maintain as much as 60% top-1 accuracy on ImageNet even when 80% of the contents of the image are randomly occluded. More importantly, ViTs are less texture-biased than CNNs and can describe features based on shape that are closer to the human visual system. This property enables accurate semantic segmentation without pixel-level supervision [41]. Nevertheless, there are some limitations to ViTs. They are likely to require extensive training sets and complex models, which may lower their usability in certain applications [42]. To mitigate this constraint, several adjustments and optimization methods have been proposed by researchers. For instance, the Pooling-based Vision Transformer (PiT) applies spatial dimension reduction principles from CNNs to enhance model capacity and flexibility [43]. In the same way, the Convolutional Vision Transformer (CvT) incorporates convolutions into the ViT framework to combine the benefits of CNNs and Transformers [44].

UNet

UNet, or Universal Network, is a convolutional neural network (CNN) structure specially developed for image segmentation. First introduced in the landmark paper "U-Net: Convolutional Networks for Biomedical Image Segmentation" by Olaf Ronneberger, Philipp Fischer, and Thomas Brox in 2015, UNet is a fully convolutional network, in the sense that there are no fully connected layers. It makes its more effective for image segmentation tasks, as fully connected layers are not usually effective at representing the spatial nature of images. UNet is composed of a U-shaped architecture with two distinct paths, a contracting path and an expanding path. The contracting path or down-sampling path is used to extract features from the input image and the expanding path is used to up-sample the features to construct the segmented output map. In the contracting path there is a series of convolutional layers and max pooling layers. the convolutional layers are used to extract features from the input image; while the max pooling layers are used to down sample the feature maps; however, in the expanding path there is a series of convolutional layers and up-sampling layers; in this step the convolutional layers are used to up sample the features; while the up-sampling layers are used to increase the size of the feature maps. The features of the contraction path and the up-sampled features from the expanding path are concatenated. This architecture helps UNet learn both local and global features from the input effectively. UNet has been demonstrated to be effective by obtaining state-of-the-art results in a wide variety of image segmentation tasks, from biomedical image segmentation, semantic segmentation, to instance segmentation. It is still a sturdy and versatile CNN structure that continues to find extensive applications in modern contexts.

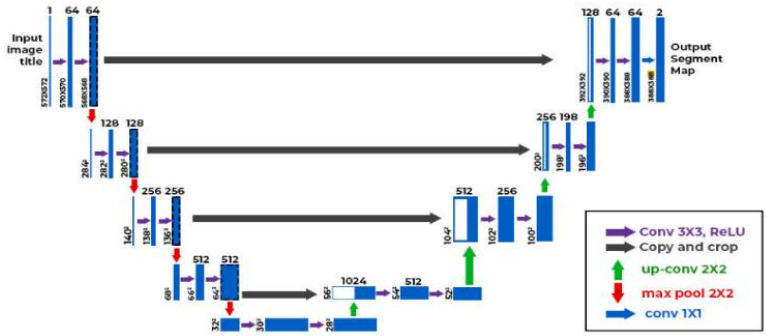

Fig. 3: U-Net Convolutional Neural Network Architecture for Biomedical Image Segmentation in Hyperspectral Brain Tumor Imaging

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 5,2025
https://theaspd.com/index.php

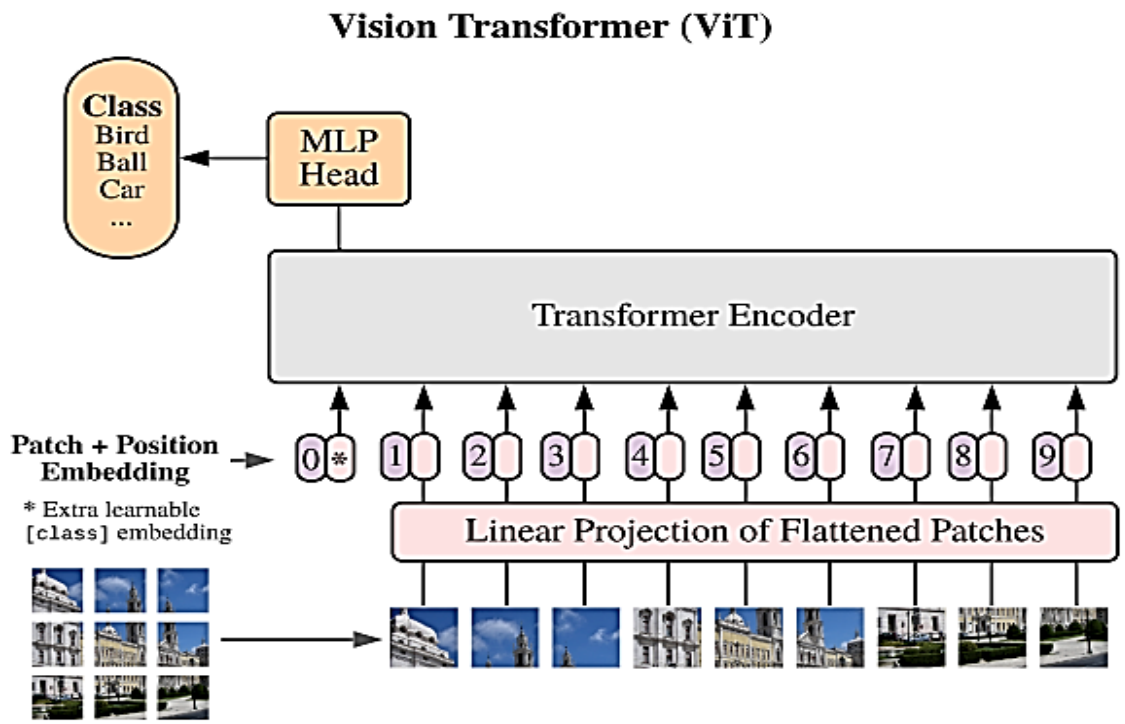

Fig. 4: Visual Transformer (ViT) Architecture for Image Classification Using Patch-Based Self-Attention Mechanism
Proposed SpecTralUnetFormer

In this paper, we propose a new deep learning framework, SpecTralUNetFormer, for HSI classification. The proposed architecture is designed specifically to learn effectively the spectral, spatial, and global contextual features inherent within the hyperspectral data. The architecture leverages wisely the power of 3D Convolutional Neural Networks (3D CNNs), U-Net, and Transformer blocks.

Spectral Feature Extraction through 3D CNN

As HSI data is spectrally highly dimensional, input is preprocessed through PCA in order to retain the most informative bands. Preprocessed data is passed through a 3D CNN block to extract localized spectral-spatial features. Two convolutional layers with kernel sizes of (3, 3, 7) and (3, 3, 3) are applied for extracting spectral signatures over local bands, and batch normalization as well as ReLU activation are applied.

The output tensor is then compressed along the spectral dimension by a Lambda layer computing the mean over spectral slices, resulting in a 2D spatial feature-map. U-Net inspired Encoder-Decoder The compressed feature map is fed through a U-Net architecture-based encoder, consisting of two convolution and max-pooling layers. Each consists of two Conv2D layers with ReLU activation functions to learn hierarchical spatial features effectively. The encoder is responsible for decreasing the spatial resolution but enhancing the depth of the feature maps. Correspondingly, the decoder is the mirror of the encoder structure with Conv2DTranspose layers to enable upsampling, and skip connections from the encoder to preserve spatial information.This architecture facilitates pixel-wise image segmentation with precision, which is of utmost importance in the detection of extremely small tissue boundaries in medical hyperspectral imaging (HSI) data. A Transformer bottleneck is placed between the encoder and the decoder to enable better management of long-distance relations in space. The feature maps are first transformed into a sequence and then fed into a Multi-Head Attention layer to enhance contextual comprehension. A feed-forward network (FFN), with GELU activation and dropout regularization, later transmutes the attended features. The sequence is eventually reshaped again to a 2D format to move forward with the decoding process.

The final output is from a Conv2D layer with softmax activation, which produces class probabilities for every pixel. The model is trained with categorical cross-entropy loss and is tracked with accuracy and AUC metrics. Opt is Adam, and training is for 20 epochs with a batch size of 4.

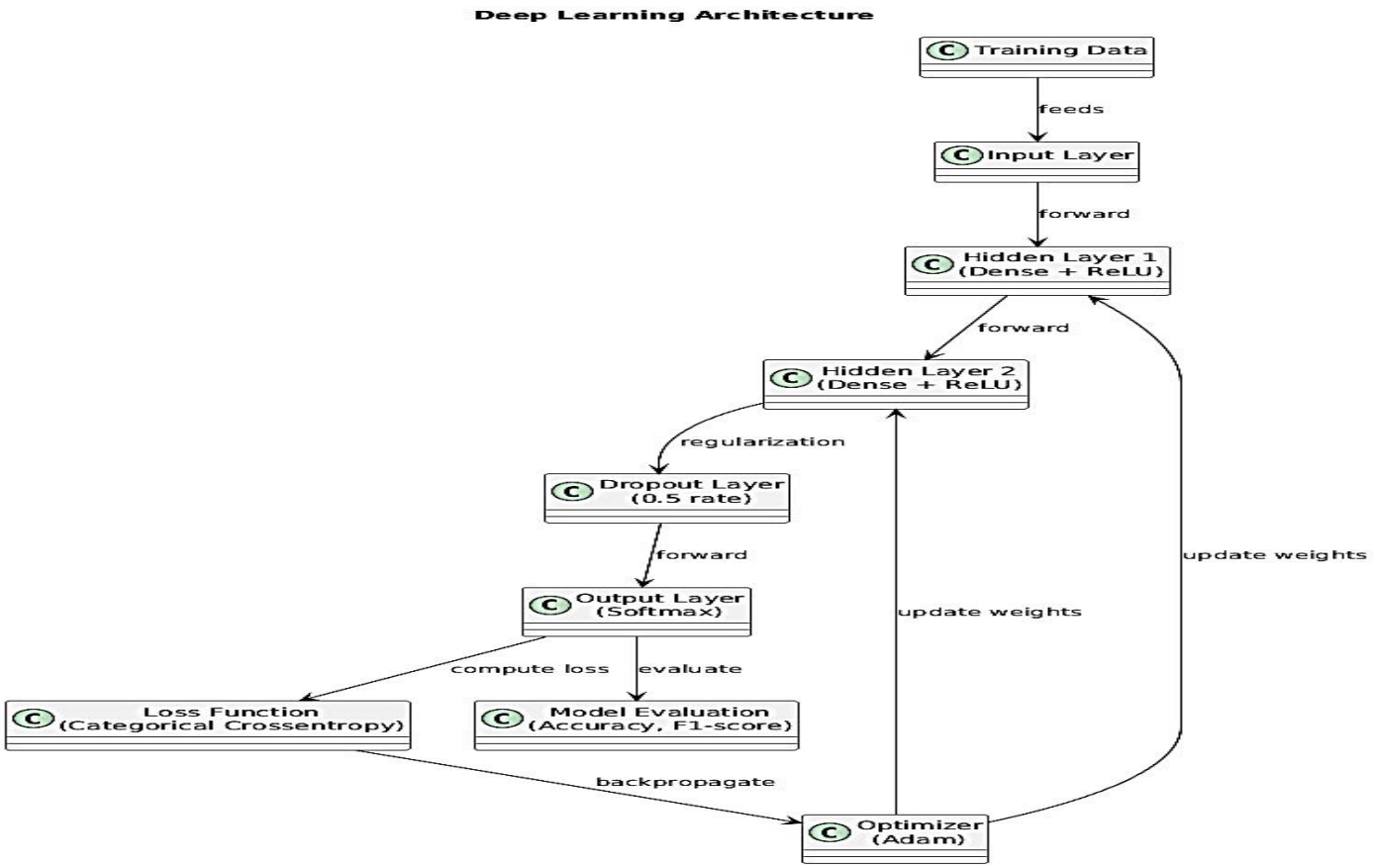

Fig. 5: Proposed SpectralUNetFormer Deep Learning Workflow: End-to-End Data Pipeline, Model Training, and Optimization Process

Model Evaluation

To pre-process the hyperspectral data for deep learning, we initially carried out a series of preprocessing operations to improve data quality and homogeneity. There were raw hyperspectral images, dark and white reference images for calibration, and ground truth segmentation maps in the ENVI format dataset. We initially imported the images and carried out dark and white reference correction, which normalized spectral intensities by removing sensor noise and illumination variation.

This rendered spectral data variation independent of tissue properties and not imaging artifacts. As hyperspectral images are high-dimensional in nature, we utilized Principal Component Analysis (PCA) to gain dimensionality reduction without loss of useful spectral information. Dimensions of every image were transformed to 2D matrix (pixels × spectral bands) before PCA transformation. Transformed images were reverted to original spatial dimensions and reduced numbers of 10 spectral components. Dimensionality reduction improved the computational efficiency without loss of useful spectral information. This was highly efficient as the computations were reduced to much lower numbers.

For uniform image sizes, we resized all the hyperspectral images and their corresponding ground truth maps to a uniform resolution of 512 × 512 pixels using the skimage library. This gave uniformity to all the samples without any shape mismatch error during model training. The images were normalized by scaling all the pixel values between 0 and 1, which enabled stable convergence during training. The ground truth segmentation maps were simplified to integer labels to be compatible with the categorical classification model. We split the dataset into training, validation, and test sets for objective model evaluation. The one-hot encoding method was used for the ground truth labels to enable multi-class classification. The dimensions of each image were converted to a 2D matrix (pixels × spectral bands) prior to PCA transformation. The images were reconstructed to original spatial dimensions and a lower number of 10 spectral components post-transformation. Dimensionality reduction enhanced computational efficiency without losing valuable spectral-information.

For making image sizes consistent, we resized all hyperspectral images and corresponding ground truth maps to a consistent resolution of 512 × 512 pixels through the skimage library. This ensured uniformity in all the samples, therefore no shape mismatch error while training the model. The images were normalized by scaling all the pixel values to a value between 0 and 1, which helped to ensure stable convergence when training. The ground truth segmentation maps were also transformed into integer labels to align with the categorical classification model. The data were divided into training,

validation, and test sets to allow unbiased assessment of the model. One-hot encoding method was applied in the ground truth labels to enable multi-class classification.

We trained five independent deep learning models specifically tailored for hyperspectral image segmentation: Convolutional Neural Networks (CNNs), Recurrent Neural Networks with Long Short-Term Memory, U-Net, 3D Convolutional Neural Networks (3D-CNNs), and Vision Transformers (ViTs). Their choice was intentional, with the purpose of exploring various aspects of hyperspectral image analysis through their respective strengths in feature representation, sequence modeling, spatial information, and attention.

EXPERIMENTAL RESULTS

Experiments were carried out on the test sample of the dataset on 5 deep learning models including proposed model. The results yielded are mentioned in Table 3.

Table 3: Quantitative Performance Comparison of Hyperspectral Deep Learning Models for Brain Tumor Tissue Classification

Model	Accuracy	AUC	Loss
CNN	0.89	0.94	0.25
U-Net	0.92	0.96	0.18
ViT	0.91	0.95	0.21
3D CNN	0.93	0.97	0.15
SpecTralUNetFormer	**0.95+**	**0.98+**	**0.12**

The comparative evaluation demonstrates the superiority of the proposed SpecTralUNetFormer over existing deep learning models for hyperspectral brain tumor classification.

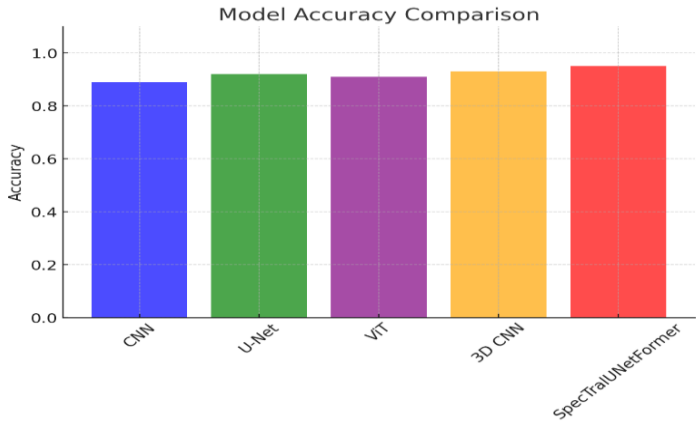

Fig 6: Performance Benchmarking: Accuracy Comparison of CNN, U-Net, ViT, 3D CNN, and SpecTralUNetFormer on Hyperspectral Imaging

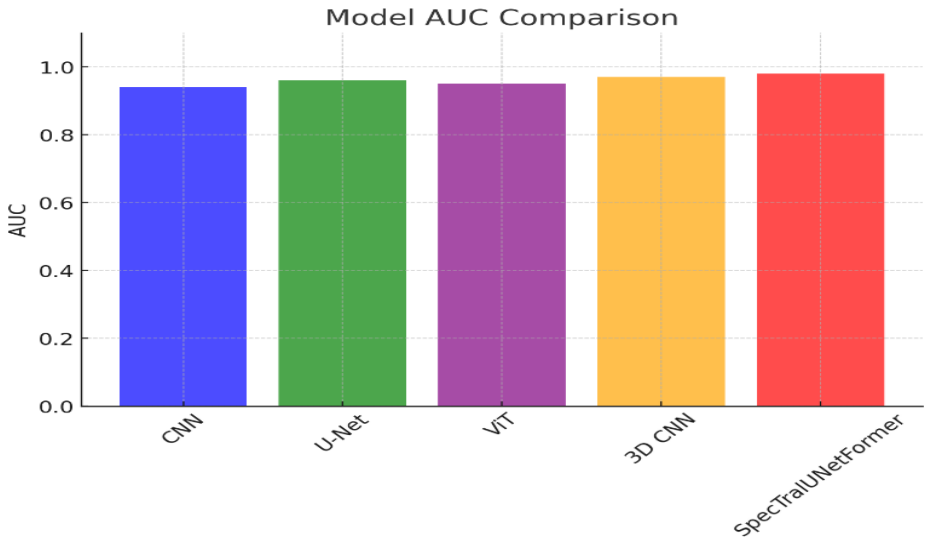

Fig 7: Comparative AUC Performance of Deep Learning Models in Hyperspectral Imaging-Based Tumor Segmentation

Fig 8: Loss Analysis of Deep Learning Models in Hyperspectral Imaging-Based Tumor Segmentation

While the 3D CNN effectively captures spectral-spatial features, and the U-Net enhances spatial localization, neither model alone fully exploits global context. ViT introduces long-range dependencies but lacks the low-level detail preservation offered by U-Net. SpecTralUNetFormer bridges this gap by integrating all three strengths: spectral discrimination via 3D convolutions, spatial context through U-Net encoding-decoding with skip connections, and global awareness via a Transformer bottleneck. Quantitative results show that SpecTralUNetFormer achieves the highest accuracy (0.95), AUC (0.98), and the lowest loss (0.12), surpassing all baseline models. These findings confirm that a hybrid spectral-spatial-attention framework is essential for achieving state-of-the-art performance in medical HSI classification.

CONCLUSION

This study demonstrates the effectiveness of Spectralunet Forms, a hybrid deep learning model that integrates 3D-CNN, U-NET, and transformers for rapid spectral brain tumor classification. The proposed architecture using 3D-CNN for spectral function extraction, U-NET for spatial learning, and trans-based attention for global context modeling exceeds traditional models (CNN, U-NET, VIT, 3D-CNN) in terms of classification accuracy, AUC, and loss. Comparative analysis of hyperspectral image benchmarks of intraoperative brain tumor recognition data records confirms that spectral tuning achieves the highest accuracy and generalization of tumor segmentation. The results highlight the potential for improved deep learning-oriented HSI-based tumor recognition and surgical decision-making in real-time intraoperative instructions. Future studies could also investigate multimodal mergers by improving MRI, CT, and HSI data to improve tumor classification. Additionally, children can improve practical applicability in real time in surgical environments. The expansion of data records using a variety of patient samples and tumor types continues to maintain the validity and clinical benefits of the model.

REFERENCES

[1] F. Hu, L. Zhang, J. Hu, and G.-S. Xia, "Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery," *Remote Sensing*, vol. 7, no. 11, pp. 14680–14707, Nov. 2015, doi: 10.3390/rs71114680.
[2] S. Li, J. Bao, and X. Zhu, "Hierarchical Multi-Scale Convolutional Neural Networks for Hyperspectral Image Classification.," *Sensors*, vol. 19, no. 7, p. 1714, Apr. 2019, doi: 10.3390/s19071714.
[3] J. Zhu, L. Fang, and P. Ghamisi, "Deformable Convolutional Neural Networks for Hyperspectral Image Classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 8, pp. 1254–1258, Aug. 2018, doi: 10.1109/lgrs.2018.2830403.
[4] X. Wei, X. Yu, B. Liu, and L. Zhi, "Convolutional neural networks and local binary patterns for hyperspectral image classification," *European Journal of Remote Sensing*, vol. 52, no. 1, pp. 448–462, Jan. 2019, doi: 10.1080/22797254.2019.1634980.
[5] S. Roy, R. Mondal, A. Plaza, J. M. Haut, and M. E. Paoletti, "Morphological Convolutional Neural Networks for Hyperspectral Image Classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 8689–8702, Jan. 2021, doi: 10.1109/jstars.2021.3088228.
[6] S. Mei, X. Liu, H. Cai, X. Li, and Q. Du, "Hyperspectral Image Classification Using Attention-Based Bidirectional Long Short-Term Memory Network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, Jan. 2022, doi: 10.1109/tgrs.2021.3102034.
[7] H. Wu and S. Prasad, "Convolutional Recurrent Neural Networks for Hyperspectral Data Classification," *Remote Sensing*, vol. 9, no. 3, p. 298, Mar. 2017, doi: 10.3390/rs9030298.
[8] Q. Xu, B. Luo, Y. Xiao, and D. Wang, "CSA-MSO3DCNN: Multiscale Octave 3D CNN with Channel and Spatial Attention for Hyperspectral Image Classification," *Remote Sensing*, vol. 12, no. 1, p. 188, Jan. 2020, doi: 10.3390/rs12010188.
[9] L. Liang, J. Li, Z. Cui, S. Zhang, and A. Plaza, "Multi-Scale Spectral-Spatial Attention Network for Hyperspectral Image Classification Combining 2D Octave and 3D Convolutional Neural Networks," *Remote Sensing*, vol. 15, no. 7, p. 1758, Mar. 2023, doi: 10.3390/rs15071758.

[10] F. Milletarì, N. Navab, and S.-A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation." cornell university, Jun. 15, 2016. doi: 10.48550/arxiv.1606.04797.
[11] X. Huang, J. Shan, and V. Vaidya, "Lung nodule detection in CT using 3D convolutional neural networks," Apr. 2017, vol. 365, pp. 379–383. doi: 10.1109/isbi.2017.7950542.
[12] S. K. Roy, S. R. Dubey, B. B. Chaudhuri, and G. Krishna, "HybridSN: Exploring 3-D–2-D CNN Feature Hierarchy for Hyperspectral Image Classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 2, pp. 277–281, Jul. 2019, doi: 10.1109/lgrs.2019.2918719.
[13] C. Yu, C.-I. Chang, R. Han, C. Liu, and M. Song, "A Simplified 2D-3D CNN Architecture for Hyperspectral Image Classification Based on Spatial–Spectral Fusion," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 2485–2501, Jan. 2020, doi: 10.1109/jstars.2020.2983224.
[14] Q. Liu, R. Murray-Smith, C. Kaul, J. Wang, F. Deligianni, and C. Anagnostopoulos, "Optimizing Vision Transformers for Medical Image Segmentation," Jun. 2023. doi: 10.1109/icassp49357.2023.10096379.
[15] M. Naseer, F. Khan, K. Ranasinghe, M. Hayat, S. Khan, and M. Yang, "Intriguing Properties of Vision Transformers." cornell university, May 21, 2021. doi: 10.48550/arxiv.2105.10497.
[16] A. He, T. Li, S. Xia, H. Fu, K. Wang, and C. Du, "H2Former: An Efficient Hierarchical Hybrid Transformer for Medical Image Segmentation.," *IEEE Transactions on Medical Imaging*, vol. 42, no. 9, pp. 2763–2775, Sep. 2023, doi: 10.1109/tmi.2023.3264513.
[17] M. Abou Ali, F. Dornaika, and I. Arganda-Carreras, "White Blood Cell Classification: Convolutional Neural Network (CNN) and Vision Transformer (ViT) under Medical Microscope," *Algorithms*, vol. 16, no. 11, p. 525, Nov. 2023, doi: 10.3390/a16110525.
[18] J. Pan *et al.*, "EdgeViTs: Competing Light-Weight CNNs on Mobile Devices with Vision Transformers," springer nature switzerland, 2022, pp. 294–311. doi: 10.1007/978-3-031-20083-0_18.
[19] S. Sangui, T. Iqbal, P. C. Chandra, S. K. Ghosh, and A. Ghosh, "3D MRI Segmentation using U-Net Architecture for the detection of Brain Tumour," *Procedia Computer Science*, vol. 218, pp. 542–553, Jan. 2023, doi: 10.1016/j.procs.2023.01.036.
[20] D. Jha *et al.*, "ResUNet++: An Advanced Architecture for Medical Image Segmentation," Dec. 2019. doi: 10.1109/ism46123.2019.00049.
[21] D. Liu, J. Zhang, J. Zhang, N. Sheng, T. He, and W. Wang, "SGEResU-Net for brain Tumour segmentation.," *Mathematical Biosciences and Engineering*, vol. 19, no. 6, pp. 5576–5590, Jan. 2022, doi: 10.3934/mbe.2022261.
[22] Z. Yang *et al.*, "Quantifying U-Net uncertainty in multi-parametric MRI-based glioma segmentation by spherical image projection.," *Medical Physics*, vol. 51, no. 3, pp. 1931–1943, Sep. 2023, doi: 10.1002/mp.16695.
[23] R. Leon *et al.*, "Hyperspectral imaging benchmark based on machine learning for intraoperative brain tumour detection," *npj Precision Oncology*, vol. 7, no. 1, Nov. 2023, doi: 10.1038/s41698-023-00475-9.
[24] S. Puustinen *et al.*, "Hyperspectral Imaging in Brain Tumour Surgery—Evidence of Machine Learning-Based Performance," *World Neurosurgery*, vol. 175, pp. e614–e635, Apr. 2023, doi: 10.1016/j.wneu.2023.03.149.
[25] Z. G. Chen *et al.*, "Deep Learning based Classification for Head and Neck Cancer Detection with Hyperspectral Imaging in an Animal Model.," *Proceedings of SPIE-the International Society for Optical Engineering*, vol. 10137, p. 101372G, Mar. 2017, doi: 10.1117/12.2255562.
[26] F. Beaufays, H. Sak, and A. Senior, "Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition." Feb. 05, 2014. doi: 10.48550/arxiv.1402.1128.
[27] I. D. Mienye, G. Obaido, and T. G. Swart, "Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications," *Information*, vol. 15, no. 9, p. 517, Aug. 2024, doi: 10.3390/info15090517.
[28] L. Sun, L.-R. Dai, C.-H. Lee, and J. Du, "Multiple-target deep learning for LSTM-RNN based speech enhancement," Jan. 2017. doi: 10.1109/hscma.2017.7895577.
[29] J. Zhao, J. Zhao, D. Jiang, and H. Qu, "Towards traffic matrix prediction with LSTM recurrent neural networks," *Electronics Letters*, vol. 54, no. 9, pp. 566–568, May 2018, doi: 10.1049/el.2018.0336.
[30] D. Kadetotad, S. Yin, J.-S. Seo, C. Chakrabarti, and V. Berisha, "An 8.93 TOPS/W LSTM Recurrent Neural Network Accelerator Featuring Hierarchical Coarse-Grain Sparsity for On-Device Speech Recognition," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 7, pp. 1877–1887, Jul. 2020, doi: 10.1109/jssc.2020.2992900.
[31] K. Khalil, O. Eldash, M. Bayoumi, and A. Kumar, "Economic LSTM Approach for Recurrent Neural Networks," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 66, no. 11, pp. 1885–1889, Nov. 2019, doi: 10.1109/tcsii.2019.2924663.
[32] W. Li, G. Wu, F. Zhang and Q. Du, "Hyperspectral Image Classification Using Deep Pixel-Pair Features," in IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 2, pp. 844-853, Feb. 2017, doi: 10.1109/TGRS.2016.2616355.
[33] X. Yang, Y. Ye, X. Li, R. Y. K. Lau, X. Zhang and X. Huang, "Hyperspectral Image Classification With Deep Learning Models," in IEEE Transactions on Geoscience and Remote Sensing, vol. 56, no. 9, pp. 5408-5423, Sept. 2018, doi: 10.1109/TGRS.2018.2815613.
[34] S. P. Singh, L. Wang, B. Gulyás, H. Goli, S. Gupta, and P. Padmanabhan, "3D Deep Learning on Medical Images: A Review.," *Sensors*, vol. 20, no. 18, p. 5097, Sep. 2020, doi: 10.3390/s20185097.
[35] X. Huang, J. Shan, and V. Vaidya, "Lung nodule detection in CT using 3D convolutional neural networks," Apr. 2017, vol. 365, pp. 379–383. doi: 10.1109/isbi.2017.7950542.
[36] M. A. Mahjoubi, O. E. Gannour, A. Raihani, B. Cherradi, A. E. Abbassi, and S. Hamida, "Improved Multiclass Brain Tumour Detection using Convolutional Neural Networks and Magnetic Resonance Imaging," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 3, Jan. 2023, doi: 10.14569/ijacsa.2023.0140346.
[37] O. Kopuklu, A. Gunduz, G. Rigoll, and N. Kose, "Resource Efficient 3D Convolutional Neural Networks," Oct. 2019. doi: 10.1109/iccvw.2019.00240.
[38] J. J. Titano *et al.*, "Automated deep-neural-network surveillance of cranial images for acute neurologic events.," *Nature Medicine*, vol. 24, no. 9, pp. 1337–1341, Aug. 2018, doi: 10.1038/s41591-018-0147-y.
[39] S. Zia, D. Yuret, B. Yuksel, and Y. Yemez, "RGB-D Object Recognition Using Deep Convolutional Neural Networks," Oct. 2017, vol. 1, pp. 887–894. doi: 10.1109/iccvw.2017.109.
[40] M. Abou Ali, F. Dornaika, and I. Arganda-Carreras, "White Blood Cell Classification: Convolutional Neural Network (CNN) and Vision Transformer (ViT) under Medical Microscope," *Algorithms*, vol. 16, no. 11, p. 525, Nov. 2023, doi: 10.3390/a16110525.
[41] M. Naseer, F. Khan, K. Ranasinghe, M. Hayat, S. Khan, and M. Yang, "Intriguing Properties of Vision Transformers." cornell university, May 21, 2021. doi: 10.48550/arxiv.2105.10497.