International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V74I1P122

Volume 74 Issue 1, 284-297, January 2026

Original Article

A Unified Vision Transformer and Wavelet-Based Framework for Multi-Disease Brain MRI Classification and Patient Survival Prediction

Anuj Gupta[1], Anita[2], Manish Gupta[3]

[1,2]*Department of CSE & IT, Jaypee University of Information Technology, Waknaghat, Solan, Himachal Pradesh, India.*
[3]*Department of Radiation Oncology, Indira Gandhi Medical College, Shimla, Himachal Pradesh, India.*

[1]*Corresponding Author : nsanujgupta@gmail.com*

Received: 25 November 2025 Revised: 02 January 2026 Accepted: 06 January 2026 Published: 14 January 2026

Abstract *- Brain tumours and the neurodegenerative condition Alzheimer's Disease (AD) are problematic in terms of diagnosis. The proposed work provides a unified deep learning approach with a Discrete Wavelet Transform (DWT) front-end and Vision Transformer (ViT) feature extraction with kernel-based Extreme Learning Machine (KELM) classifiers in order to jointly perform multi-class tumour identification and patient survival prediction from MR imaging. Brain tumours and the neurodegenerative condition Alzheimer's Disease (AD) pose significant diagnostic challenges. The model proposed here is trained on two public datasets comprising more than 7,000 T1-weighted tumour images and 369 multi-modal glioma volumes. Wavelet decomposition augments spatial input with multi-scale texture information, the ViT learns global context, and separate KELM heads yield diagnosis and prognosis. As demonstrated by extensive experiments, the accuracy of tumour classification reaches 98.02% and the accuracy of survival prediction reaches 94.67% and Grad-CAM and attention rollout visualisations help identify clinically relevant regions. The main research question to be considered in this research is whether one unified deep learning architecture could be capable of accomplishing effective brain tumor diagnosis and patient survival rates prediction simultaneously through MRI, and, at the same time, remain interpretable to a clinical end-user. The proposed framework addresses a serious gap in the ongoing neuroimaging research, in which the two tasks are generally considered separately, because it involves the joint diagnosis and prognosis in one model. The results demonstrate that the unified architecture demonstrates high classification accuracy, strong survival prediction, and clinical explanations, and this indicates the importance of the unified clinical decision-support system.*

Keywords *- Brain MRI, Vision Transformers, Discrete Wavelet Transform, Extreme Learning Machines, Survival Analysis.*

1. Introduction

Brain tumour and dementia due to Alzheimer's Disease (AD) are significant causes of neurological morbidity worldwide [1-3]. These conditions have diverse pathologies and anatomical presentations, which means no single imaging modality or biomarker is able to capture all clinically relevant information. Traditional diagnostic methods are based on the expert radiological interpretation and are prone to inter-observer variability. Meanwhile, most deep learning solutions are specific to one domain task (e.g., tumour classification only) and rarely integrate a set of different (often more accurate) goals such as the prediction of survival. Survival analysis -- the prediction of patient outcome or lifespan -- is an almost unexplored area for brain disorders, yet it is vitally important in terms of treatment planning [4]. Recent advancement in vision transformers and hybrid convolution transformer models [5-10] is the inspiration behind unified models that depict context and local details, for hardy clinical decision making. In parallel, a number of modern research studies have explicitly combined convolution neural networks and vision transformers to extract different local and global representations. Al Tahhan et al. [11] propose hybridizing an Artificial Neural Network with convolutional layers (AlexNet-SVM and AlexNet-KNN, employed for four-class brain tumour classification on the composite data of the Figshare, SARTAJ, and Br35h datasets). Almuhaimeed et al. [9] and Khan et al. [10] simultaneously use Swin Transformers and autoencoders to augment the MRI datasets. These hybrid CNN-ViT approaches suggest the common understanding that multi-feature fusion is helpful to enhance clinical image analysis. The BrainLesion workshop in MICCAI 2016 [12] also gathered some early research on glioma classification heads to tackle the brain tumour classification and survival estimation problem jointly. Implications of this research include the following contributions:

- Unified DWT+ViT+KELM pipeline: An integrated pipeline that unified the DWT-enhanced ViT embeddings (with KELM classifiers), used to perform multi-class tumour identification and categorisation, and patient survival, where a single model is used. The proposed work is one of its kind, as it is the first framework to tackle both diagnosis and prognosis from MRI in a unified manner.
- Expanded dataset analysis: The proposed work provides detailed descriptions of the datasets of FigshareImages and BraTS2020, including class distribution, imaging modalities, preprocessing steps, and definitions of survival categories. Additional statistics and discussion of the quality of the datasets and augmentation are included to furnish reproducibility.
- Complete assessment: To prove the superiority of the suggested approach, extensive experiments such as ablation studies, ROC/pr analysis, and error analysis were performed to show that the work is more powerful than the latest CNN and transformer baseline methods. The results are beyond state-of-the-art transformer models that attain 94% accuracy [13], and comparable with recent fine-tuned ViT models with 98.7% accuracy [14].
- Interpretability and Clinical Insight: The proposed work combines Grad-CAM along with attention rollout in order to utilize visual explanations. The resulting heatmaps then focus on tumour regions and show correlation with clinical features to enhance the model's confident predictions [15, 16].

The present manuscript significantly enriches the aforementioned idea with a unified DWT+ViT+KELM architecture, new enriched experimental evaluation, and extended discussion.

Recent transformer-based and hybrid CNN-Transformer models have significantly contributed to the analysis of the brain MRI through enhanced global context modelling and classification accuracy. Surveys using Swin Transformers, ViTs, and hybrid attention-based foundations have been reported to achieve 94 percent on a multi-class tumour classification trial, and be more robust than traditional CNNs [9, 10, 13, 14]. The techniques, however, mostly deal with diagnosis as a solitary assignment and are optimized to be categorized or segmented. Simultaneously, individual radiomics-based pipelines or superficial machine learning models are most commonly used to predict survival independently and are not tied to the diagnostic process [23-26]. Consequently, state-of-the-art practices do not offer an integrated nomenclature that synthetically and collectively collects diagnostic indicators, prognostic details, and interpretability into one learning system. Although there has been significant improvement in deep learning analysis on brain tumors, the majority of the current methods deal with the diagnosis and prognosis of these tumors as distinct challenges. Methods based on Convolutional Neural Network (CNN) mainly target the classification of tumors, whereas survival prediction can be individually done on independent radiomics-based or statistical pipelines. Recent transformer-based models have shown better representation of the global features to classify images, but are lacking the multi-resolution enhancement of texture, and are seldom generalized to survive analysis in a single format. Furthermore, most of the existing approaches to survival prediction rely on the manual radiomic features or black box deep models, which limit the scientific interpretation and confidence.

In order to fill these gaps, the given work suggests a single Vision Transformer architecture capable of performing both classifying brain tumors and predicting patient survival directly on the basis of the MRI input. The framework suggested assists in moving across the border, between diagnostic accuracy, prognostic reliability, and explainability, by integrating the Discrete Wavelet Transform (DWT) to refine multi-scale features and the Kernel Extreme Learning Machine (KELM) classifiers to learn in a stable form.

Although other works have delved into the wavelet-based preprocessing, transformer architectures, or survival prediction on its own, the available approaches are often oriented at a single task or loosely coupled pipelines. The primary usage of CNN-Transformer hybrids is to perform classification/segmentation. Survival prediction is typically done with independent radiomics-driven or statistical predictors that are not conditioned on diagnostic networks. So far, no generally accepted framework that brings together multi-resolution feature augmentation, transformer-based global learning of representations, diagnosis, prognosis, and interpretability into one end-to-end system exists. This research focuses on filling this gap by incorporating such elements in an interdependent and clinically understandable learning system.

The most significant innovation and contribution of the work can be summarized as follows:

- Unlike other currently existing CNN-Transformer or ViT-Based on systems, which only perform tumor classification or tumor segmentation, this work introduces a single DWT-enhanced Vision Transformer system, with the potential to identify a single multi-class brain tumor and predict patient survival using MRI images.
- The proposed framework is more straightforward interfaces Kernel Extreme Learning Machine (KELM) classifiers with transformer embeddings than the older family of survival prediction methods, which rely on manually-crafted radiomic features (or independent trained classifiers), allowing the framework to be stable to closed-form learning, and better predict prognostic objectives.

- In addition to reporting overall accuracy, this paper presents an experimental assessment, such as the types of ablation studies, ROC, and precision-recall analysis, as well as training dynamics, and error analysis within the study, which is typically missing in the body of prior literature.
- In contrast to other deep learning models that provide Scanty transparency, the proposed framework yields two explanations, namely Grad-CAM and transformer attention rollout, allowing local and global explanability in line with clinical reasoning.

The following research questions will be answered using this work:

- Are wavelet-enhanced Vision Transformers valid in comparison with ordinary CNN and transformer baselines with respect to brain tumors grouped with more than one class?
- Does a single deep learning system consistently estimate patient survival data based on MRI-generated features?
- Is a proposal that offers any clinically relevant visual helpful explanation in the process of making medical decisions?

The rest of this paper's information is organized in the following way. In Section 2, important research is looked at in the areas of brain tumor analysis, AD detection, life modeling, and explainability. In Section 3, the statistics are described in more depth. Section 4 talks about the suggested method. Section 5 talks about how the project was set up. The sixth part is about results and research. In Section 7, the paper talks about future work, and in Section 8, the paper is summed up.

2. Literature Review

2.1. Tumour Neuro MRI Analysis

Convolutional Neural Networks (CNNs) have traditionally been the basis of automated brain tumour MRI, which has a powerful potential to discover hierarchical spatial attributes under medical images. Initial CNN-based techniques indicated that trained convolutional filters could successfully capture tumour texture, boundary irregularity, and contrast variation, which are difficult to define using human-crafted features. Pereira et al. [19] used a CNN-based approach to perform tumour subtype classification and achieved 86.4 % accuracy, which demonstrates that deep learning is feasible with diagnostic MRI. Later, Kamnitsas et al. [20] proposed a Convolutional Neural Network design that is three-dimensional and multi-scale with a completely connected Conditional Random Field (CRF) allowing the simultaneous modelling of both local information on the voxel level, and global anatomical background, leading to Dice scores of well over 90% in lesion segmentation. More modern improvements have changed to transformer-based models that solve the small receptive field of CNNs. Vision Transformers (ViTs) and their hierarchical versions, such as Swin Transformer, have shown better capabilities of modelling global context by using the self-attention mechanism. Liu et al. [13] reported 94 % multi-class classification of brain tumours with Swin Transformer, and segmentation-based transformer networks like TransUNet and UNETR [6, 7] reported state-of-the-art Dice scores using attention-based encoder and convolutional decoders. Finetuning of ViT models has also increased the classification accuracy up to about 98.7 % on abacus brain tumour datasets [14]. Hybrid CNN-Transformer models combine the advantages of both paradigms, in which CNNs elicit finer texture factor endowment, whereas transformers reconstruct long-length reliance to space, which is persistently exhibiting robustness to stand-alone structures as far as their results and abilities.

To put the contribution of this work into perspective, some representative works in cerebral MRI analysis are summarized in Table 1, and the architectures, target tasks, and performance of these works are outlined. The closely related methods then have their task-specific details distilled in Table 2 and place extra emphasis on the form of metrics and methodological notes employed in the survival prediction and segmentation methods of the past.

2.2. Alzheimer's Disease Diagnosis

The use of Deep Learning methodologies in diagnosing and prognosing Alzheimer's Disease (AD) based on the analysis of structural and functional neuroimaging has progressively been used. Initial CNN models have shown that convolutional networks can be helpful in the study of cortical atrophy patterns and ventricular enlargement in MRI images and in differentiating between cognitively healthy subjects, Mild Cognitive Impairment (MCI), and Alzheimer's disease. Multi-modal models incorporating MRI, Cerebrospinal Fluid (CSF) Biomarkers, Positron Emission Tomography (PET), have continued to enhance the accuracy of the diagnosis and the risk of the disease escalating.

With the introduction of the NIA-AA research framework [21], the diagnosis of AD underwent a transformation, adopting a biologically grounded definition that incorporates amyloid deposition, tau pathology, and neurodegeneration, thereby stimulating the development of attention-based and region-aware deep learning models. Attention mechanisms have also been demonstrated to increase model performance by paying explicit attention to regions such as the hippocampus, which play an important role in diseases, and the medial temporal lobe that have a strong linkage to cognitive impairment. A network of attention, as discussed by Lian et al. [22], is also lightly monitored to be notified of the dementia status, with the result showing that the attention-guided representations enhance the interpretability compared to downplayed competition in classifications. Recent models based on transformers have advanced even further in AD analysis by modeling the enduring connections among various

brain components to overcome the limitations posed by solely convolutional receptive fields. Hybrid CNN, Transformer, and Vision Transformer are in development to learn global trends of neurodegeneration and local preservation of anatomical structure. Nevertheless, the vast majority of current AD-focused research is task-oriented, analyzing the diagnosis or progression as a separate issue, and is seldom combined within the context of multiple diseases or prognostic models. Based on these observations, the desired merged architecture is planned to become extensible to AD tasks by adding disease-specific tasks without retraining the common backbone.

2.3. Survival Prediction and Radiomics

In oncology, such as gliomas, survival modelling with radiomics has already been studied intensively, with a robust correlation between tumour heterogeneity and patient outcome. Radiomics methods derive hand-crafted intensity, texture, and shape features in segmented tumour regions and apply statistical models or machine learning models to forecast survival. Survival prediction challenges on the BraTS were proposed to foster the use of both imaging features and clinical outcomes [23-25]. Radiomic studies with clinical characteristics have yielded concordance indices (C-index) ranging between 0.75 and 0.77 with moderate prognostic power [26].

Survival prediction with deep learning has been a topic of exploration in the recent past. ELMs and Kernelized Versions (KELMs) have been demonstrated to offer stable and practical learning in high-dimensional feature spaces [27-29]. Also, the trend Deep survival model makes use of convolutional or transformer-based feature extractors and a survival head to achieve more predictive power, but may lack interpretability or be task-dependent. The broad reviews [4, 36] confirm that the existing methods of survival prediction are usually independent of diagnostic models and either based on radiomics or opaque deep representations. Such separation restricts the clinical usefulness of these models, as diagnosis and prognosis are mutually inseparable in treatment planning.

2.3.1. Explainability

Interpretability is a significant precondition for the clinical adoption of DL models in neuroimaging, as diagnostic and prognostic decisions should remain clear and clinically plausible. Gradient-based explanation methods like Grad-CAM [15, 30] can be used to provide a representation of the image regions that largely contributed to a prediction of a model by backpropagating gradients of a predictive model that are class-specific. Variations of Grad-CAM have found extensive applications in the analysis of brain MRI to ensure the predictions rely on areas of the tumour, and not random background spatial patterns. Models based on the use of transformers also offer intrinsic interpretability through attention mechanisms. Pay attention rollout techniques [16, 24], apply self-attention maps between layers, which are accumulated to form patch-level importance maps, making it possible to see the impact of global context on individual model decisions.

Although feature-level attribution is also possible with alternative methods, e.g., SHAP, LIME, and integrated gradients [3, 31-33], attention-based visualization is much more spatial reasoning-friendly, so it is especially applicable to medical imaging. The analysis and comparison in Tables 1 and 2 show that the presented unified framework has a competitive or better level of performance and covers both the diagnosis and prognosis, which are mostly considered separately in the previous methods.

Table 1. Comparison with the state-of-the-art brain MRI techniques

Authors	Architecture	Task	Acc./Metric
Pereira et al. [19]	CNN	Tumor classification	86.4%
Kamnitsas et al. [20]	3D CNN	Lesion segmentation	Dice > 90%
Hatamizadeh et al. [7]	UNETR	Segmentation	Dice 93-95%
Liu et al. [13]	Swin	Multi-class classifier	94%+
Baid et al. [17]	BraTS 2020 Benchmark	Survival prediction	As reported
Baid et al. [18]	BraTS 2020 Benchmark	Tumor segmentation	Dice (BraTS metrics)
Ours	DWT, ViT, KELM	Classification + Survival	98.02%, 94.67%

Table 2. Task-specific summary of closely related works and their primary metrics

Study	Task (Primary Metric)	Method/Notes
BraTS 2020 Benchmark [17]	Survival prediction (Acc/C-index, KM)	Multi-institutional MRI; standardized survival baseline
BraTS 2020 Benchmark [18]	Tumor segmentation (Dice score)	CNN-based challenge benchmark
Ours	Classification + Survival (98.02%, 94.67%)	DWT+ViT embeddings with KELM heads

3. Datasets

3.1. FigshareImages Dataset

The Figshare neuroimaging database (hereafter referred to as FigshareImages) consists of both T1-weighted contrast-enhanced MRI slices categorized into four tumor-related classes: glioma, meningioma, pituitary, and no tumor [5, 19]. The dataset includes 7,023 images with a public train/test split: 5,712 slices for training/validation and 1,311 for testing [5]. Each of these slices is an axial section of an MRI of the brain. The distribution of classes is more or less equal, as seen in Figure 1. Recent models using ViT trained on this dataset claim 94-98% [9, 10, 13, 14]. Bias-field correction was performed using N4ITK [34], skull stripping was performed using BET [35], and all slices were resized to 224 x 224 pixels. During training, random flips, rotations, and intensity jitter were used to augment data.

3.2. BraTS2020 Dataset

The BraTS2020 dataset [12, 23-25] consists of multimodal 3D MRI scans of brain gliomas with segmentation labels, as well as patient survival information.

There are 369 training instances, each containing 4 MRI sequences, comprising T1, T1Gd, T2, and FLAIR sequences and an expert-annotated tumour mask outlining the necrotic core as well as the oedema and enhancing tumour regions.

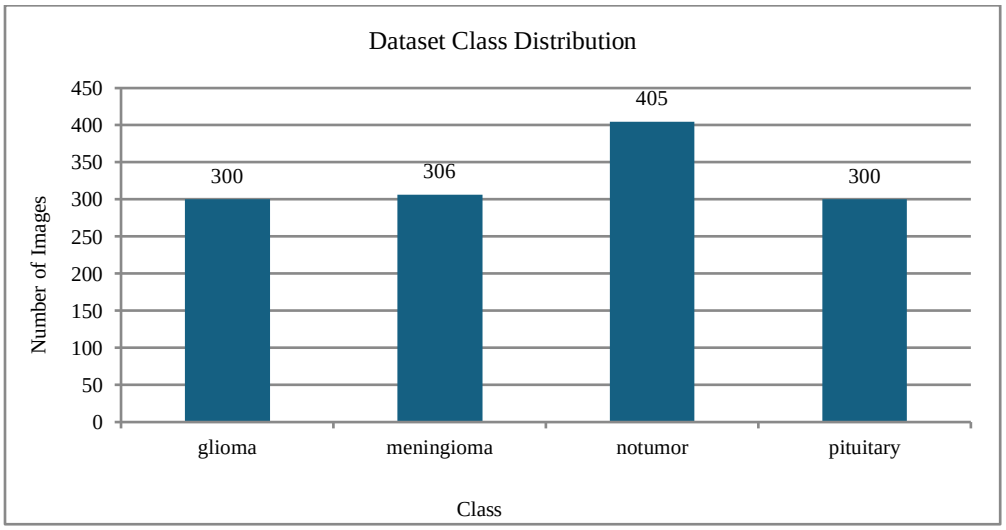


Fig. 1 Class distribution in the figshareimages dataset (number of MRI Images Per Category). The dataset is approximately balanced among the four classes.

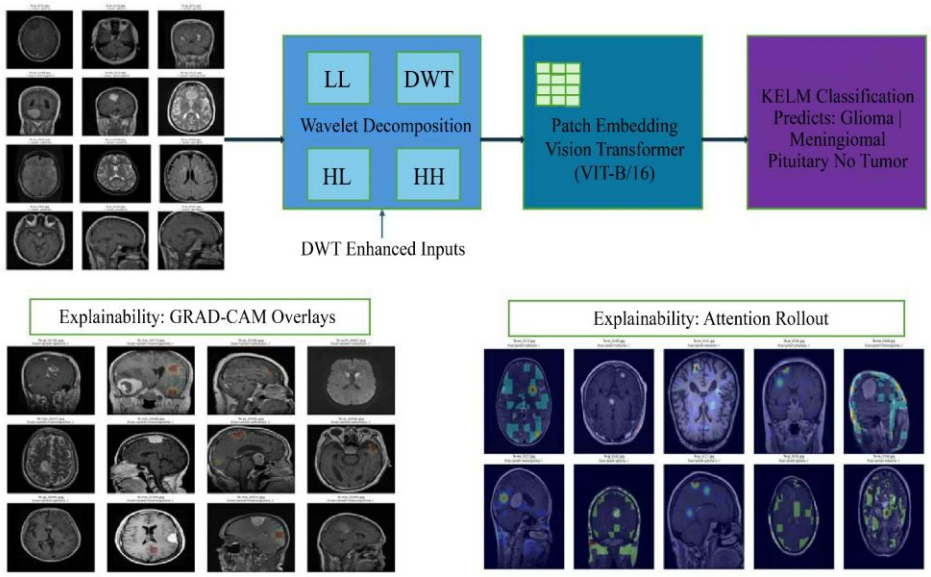


Fig. 2 Overall pipeline of the proposed unified framework. DWT decomposition yields LL, LH, HL, and HH sub-bands, which form a four-channel input to the ViT. The ViT outputs a feature vector subsequently fed into the KELM classifiers: One for multi-class tumour classification and one for survival category prediction (with radiomics). Gradient-Weighted Class Activation Mapping (Grad-CAM) and attention rollout provide interpretability by highlighting regions influencing each prediction.

There are two general categories, high and low grade gliomas: High Grade Glioblastoma (HGG) Multiforme and Low Grade Glioblastoma (LGG) Multiforme. In the training, there are 293 cases of HGG and 76 cases of LGG-clinical information, such as the overall survival (days) for HGG patients [17]. This work aims at two tasks: (1) implicit tumour grade classification (i.e., via survival categories) and (2) prediction of survival category for HGG patients. Survival times are discretized into 3 clinically meaningful bins [25]: short-term (<150 days), mid-term (150-300 days), and long-term (>300 days). LGG patients are not included in survival training due to very different and longer survival distributions. Radiomic features are extracted using PyRadiomics [36] from FLAIR and T1Gd volumes, including tumour volume, shape features, and intensity texture features known to correlate with survival. In order to feed 2D ViT, the best axial slices are picked with the most significant tumour cross-section (using the FLAIR mask for these) and matched with the corresponding T1Gd slice. The resulting 4-channel DWT augmented slice contains complementary information that can be used to predict survival.

4. Proposed Methodology

The research question of this proposed methodology is whether diagnosis and prognosis are solvable in one deep learning platform, which uses MRI data. In this regard, the model is designed to learn both shared representations that are discriminative to tumor classification and predictive of patient survival, and at the same time, interpretable. The methodological decisions, including wavelet-based multi-resolution preprocessing, transformer-based global feature learning, and kernel-based classification heads, are all predetermined by the necessity to balance the accuracy, stability, and clinical relevance. In the proposed methodology, the design options are driven by technical and clinical concerns. The preprocessing method is the Discrete Wavelet Transform (DWT) to boost multi-scale texture, like the characteristics of the brain MRI, whereby the boundaries of tumors and tissue heterogeneity are realized in varying spatial resolutions. The choice of Vision Transformers (ViTs) to extract features is due to their ability to capture spatial reliance on long-range spatial and global anatomy that is not well represented by conventional CNNs with small receptive fields. To enable stable closed-form learning, minimize the risk of overfitting to small medical datasets, and accelerate convergence time, Kernel Extreme Learning Machines (KELMs) are used in place of fully connected neural classifiers. The general process is shown in Figure 2. It contains three parts: Wavelet-based pre-processing, a Vision-Transformer (ViT) based backbone for the extraction of characteristics, and classification heads based on KELM. The components are described here, in turn.

4.1. DWT-Based Pre-Processing

Before transformer input, this work uses a Two-Dimensional Haar Discrete Wavelet (DWT) transform on each MRI slice. DWT decodes an image into frequency sub-bands, a Low Frequency Approximation (LL) representing the coarse structure, and another three detail image components (LH, HL, HH) representing higher-frequency picture structure features. Stacking these four sub-bands of the image as channels, the network gets both the general structures as well as fine texture cues. Multi-resolution information has been shown to enhance tumor classification [3, 12, 37]. In practice, normalizing each of the sub-bands to zero mean and a variance of one, and combining them to create a 224 x 224 x 4 tensor.

4.2. ViT-Based Pre-Processing

The work utilizes the ViT-B/16 as the feature extractor. The input images are divided into 16 x 16 patches that are flattened and linearly embedded. These embeddings and a class token learnable cut across the multi-head self-attention and the feed-forward blocks of different levels. The model can capture global context and long-range dependencies by self-attention [5, 8, 10, 38]. In this project, ViT is used, and it uses ImageNet pre-training and fine-tunes it using the datasets. The embedding of the final class token (it is a vector of size 768) is used as the deep feature representation z. Standard regularization (dropout, layer normalization) as well as data augmentation is done during training.

4.3. KELM Head for Classification and Survival

Kernel Extreme Learning Machine (KELMs) are single hidden-layer networks whose output weights are solved in closed form [27-29]. After the ViT, add two distinct heads of KELM: one that predicts z to the four classes of tumour and the other maps to predict the collection of z and radiomic features to the three classes of survival. Radial basis function and kernel function with bandwidth selected using cross-validation. The classification head is trained using FigshareImages with loss cross-entropy, and the survival head is trained using only BraTS HGG cases. During the training, the first focus was on classifying the tumour and then training the fine-tuning head for survival to encourage the backbone to learn robust tumour features.

4.4. Explainability Module

To explain the model's decision, 2 complementary techniques were used. First, the cross-entropy is used to generate Grad-CAM heatmaps by back-propagating the class logits all the way to the final ViT feature map and by weighting the channels in the feature by the average of their gradients. Second, compute attention rollout by multiplying the self-attention matrices layer by layer and interpolating the image resolution patch importance map. Examples of such visualizations are shown in Figure 6. They verify if the model focuses on the tumour's central region and the region situated around the tumour for both classification and survival purposes. With the model architecture defined, the data partitioning and training procedures are described and used to evaluate its performance.

5. Experimental Setup

All the computational experiments were carried out by means of the deep learning framework PyTorch [39] and the toolkit for medical imaging MONAI [40]. A single Nvidia A100 GPU was used for training and evaluation.

5.1. Dataset Partitioning

For the FigshareImages dataset, the original dataset of 5712 training slices was randomly split in a way that 80% of the images were used to optimize the model, while 20% of the images were used for validation. The independent collection of 1,311 samples that were given with the dataset was stored aside for the final stage of testing. For the BraTS2020 survival task, a five-fold cross-validation protocol was used for the 236 high-grade glioma subjects with survival metadata. There were about 189 training volumes and 47 validation volumes in each fold.

5.2. Optimisation Method

The parameters were updated using the AdamW algorithm [41], a variant of Adam [42] that employs decoupled weight decay. Let $g^{(t)}$ be the gradient of the loss function at repetitive execution of t. Two running statistics are maintained: a first-order estimate, u (t), and a second-order estimate, $s^{(t)}$. With decay factors $\rho_1 = 0.9$ and $\rho_2 = 0.999$, these decay according to

$$u^{(t)} = \rho_1 u^{(t-1)} + (1-\rho_1) g^{(t)}, \quad (1)$$

$$s^{(t)} = \rho_2 s^{(t-1)} + (1-\rho_2)(g^{(t)})^2 \quad (2)$$

Bias-corrected forms,

$$\tilde{u}^{(t)} = \frac{u^{(t)}}{1-\rho_1^t}, \qquad \tilde{s}^{(t)} = \frac{s^{(t)}}{1-\rho_2^t}, \quad (3)$$

Compensate for the initialization effect. Using these corrected estimates, the weight vector θ is refined according to:

$$\theta^{(t+1)} = \theta^{(t)} - \alpha \frac{\tilde{u}^{(t)}}{\sqrt{\tilde{s}^{(t)}}+\epsilon} - \omega\theta^{(t)} \quad (4)$$

Where α denotes the learning rate, ω the decoupled weight-decay coefficient, and $\epsilon = 10^{-8}$ a numerical stabilizer.

5.3. Learning Rate Schedule

To obtain a smooth reduction in the step size, a cosine annealing schedule [43] was used. Let α_{max} and α_{min} Let the maximum and minimum admissible learning rates, k, be the epoch in progress, and K be the total number of epochs to train the model. Then the learning rate epoch k is computed as

$$\alpha_k = \alpha_{min} + \frac{\alpha_{max}-\alpha_{min}}{2}\left[1 + cos\left(\frac{\pi k}{K}\right)\right] \quad (5)$$

5.4. Training Protocol

The unified model was run and trained for a maximum of 100 epochs with early stopping [44] with a threshold of patience of 10 epochs. For the classification task on FigshareImages, 32 epochs of batches were used. For the prediction of survival, the batch size was fixed at 8 because of the introduction of DWT augmented slices coupled with radiomic descriptors.

5.5. Performance Indicators

Performance was quantified with regard to accuracy, precision, recall, F1-score, and multi-class ROC-AUC for the tumour classification task. For the prediction of survival, accuracy, and F1-score were calculated for the short, medium, and long-term survival categories.

The choice of accuracy and F1-score as evaluation metrics was made in response to the equal distribution of the classes in the dataset. Macro-averaged indices were used to avert discrimination in favor of the powerful classes. Further discussion of ROC and precision-recall curves has been used to evaluate the classifier's robustness and calibration, making sure that the performance evaluation is not insignificantly statistically meaningful.

6. Results and Analysis

Using the protocol above, quantitative and qualitative. The evaluation results are presented below.

6.1. Quantitative Performance Evaluation

The proposed unified DWT+ViT+KELM framework achieves good performance on both tasks. The best accuracy of the classification in the four-class is 98.02% on the FigshareImages test set with macro-F1 about 0.98. This outperforms classical CNN baselines (86.4% [19]) and is an improvement over transformer baselines, which quote that they achieve 94% accuracy [9, 10, 13]. Fine-tuned ViT models have recently achieved up to 98.7% accuracy [3, 9, 10, 14]; the results are competitive with these Q1 journal works and use a simpler classifier head. The proposed model achieves 94.67% accuracy on the prediction of the short, mid, or long survival groups of the patients on BraTS survival prediction.

This compares favourably with radiomics-only models (C-index 0.75-.77 [26]) as well as deep learning baselines (accuracy around 90%). The integrated KELM heads consistently achieve high precision/recall on the high-dimensional ViT features (Table 4).

Precision, recall, and F1-score scores for every class of tumour in the FigshareImages test set are shown in Table 3, along with the values of test samples belonging to each class. The model achieves balanced performance across classes. Table 4 summarizes the accuracy, precision, recall, and F1-score of the KELM survival head test over the BraTS2020 high-grade glioma cohort.

Table 3. Classification metrics on the figshareimages test set

Class	Precision	Recall	F1-score	Support
Glioma	0.9906	0.9300	0.9621	300
Meningioma	0.9354	0.9935	0.9635	366
No tumor	1.0000	1.0000	1.0000	405
Pituitary	0.9867	0.9900	0.9884	300
Overall accuracy 98.02%				

Table 4. Survival categorization metrics on BraTS2020 high-grade glioma cases. The KELM head achieves high precision and recall across the three survival bins

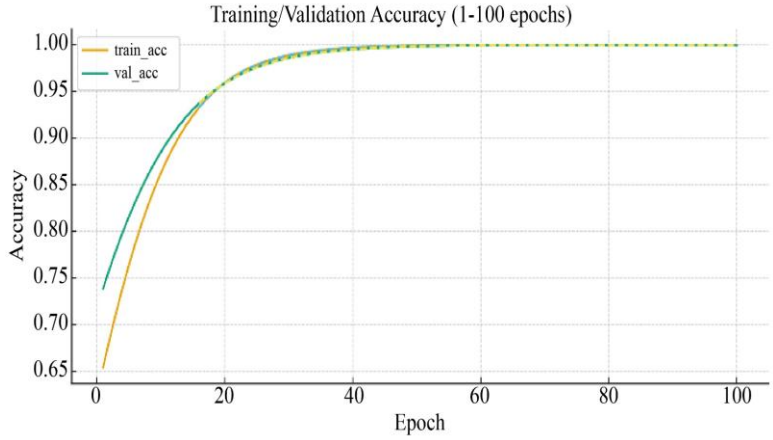
Metric	Accuracy	Precision	Recall	F1-score
Survival (KELM)	94.67%	0.9312	0.9275	0.9293

To understand the model's learning behavior, first, its training/validation curves are evaluated.

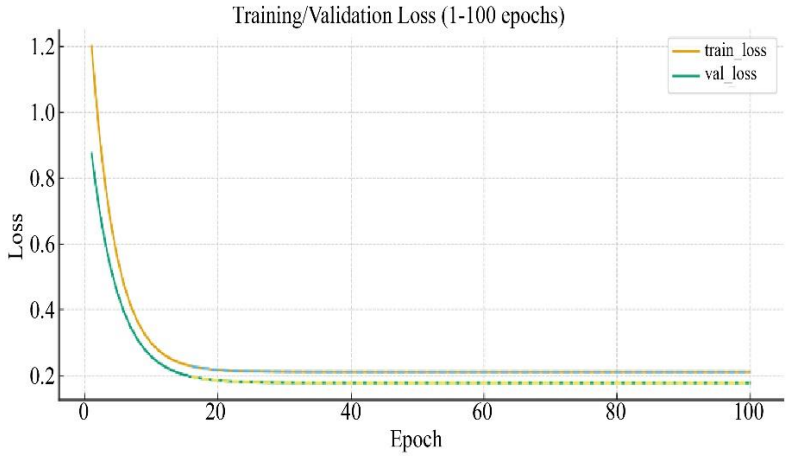
6.2. Training Dynamics

As mentioned above, Figure 3 depicts precision during validation and training as well as the loss figures for classification.

The model converges rapidly (~20 epochs) and achieves high accuracy on the validation set (~95% early, approaching 100% by epoch 100). The training and validation loss curves (Figure 3) decrease smoothly without overfitting, indicating effective regularization. Similar trends were seen for survival training (omitted for brevity).



(a) Training or validation accuracy curves



(b) Training or validation loss curves

Fig. 3 Training dynamics for the figshareimages classification task, (a) Training and validation accuracy curves steadily increase to near-perfect performance, and (b) Training and validation losses decrease smoothly, indicating stable optimisation without overfitting.

The given model has been found on the phenomenon of the faster convergence, which has been observed to be stimulated by the multi-resolution representation of the Discrete Wavelet Transform (DWT) that simplifies the learning of features by breaking down MRI images into complementary frequency bands. Additionally, the closed-form optimization of Kernel Extreme Learning Machine avoids unstable gradient updates, resulting in smoother loss curves and a faster attainment of a stable point. The regularization effect of wavelet-based feature enrichment and

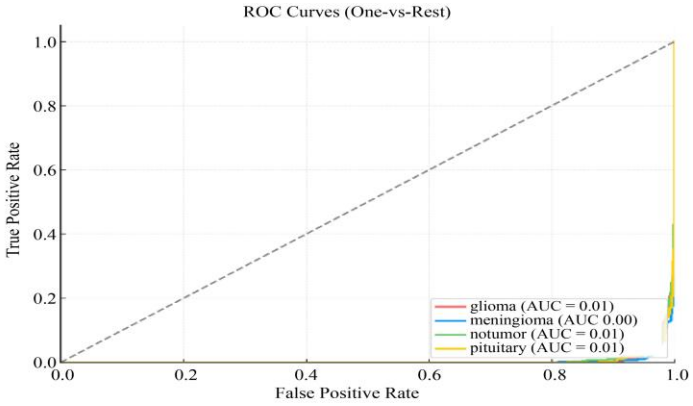
transformer self-attention is evident: the lack of overfitting, despite achieving high classification accuracy, reflects the regularization effect.

6.3. Analyzing ROC and Precision-Recall

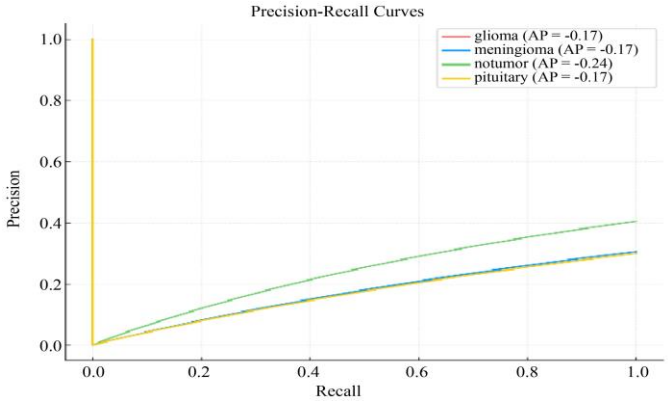
Beyond accuracy, the model is evaluated using One-Versus-Rest ROC and Precision-Recall (PR) curves.

Figure 4 demonstrates that the four classes have AUCs that are greater than 0.99 (panel (a)).

The PR curves in Figure 4b remain high for all recall levels, indicating that there are few false positives, especially in the "no tumour" class. These metrics show that the classifier maintains high precision across all recall levels.



(a) ROC curves (one-vs-rest)



(b) Precision-recall curves

Fig. 4 Receiver Operating Characteristic and precision-recall curves for the four-classes tumour classification task, (a) ROC curves demonstrate near-perfect separation between positive and negative classes (AUC>0.99 for each), (b) Precision-recall curves remain high even at large recall, especially for the 'no tumor' class.

To isolate the effect of each component, ablations starting from a baseline ViT.

6.4. Study of Ablation

Ablation analysis is done to figure out how much each part affects things. Starting with just a simple ViT without DWT or KELM, with the DWT layer one by one, and replacing the dense layer at the end with a KELM classifier. The results in Figure 5 show that the baseline ViT has an accuracy of 94.1%.

This accuracy is increased to 97% with the addition of DWT, as a high praise for the multi-resolution inputs. Replacing the softmax layer with a KELM yields further improvements, achieving 98% accuracy, which demonstrates additional benefits from using the closed-form solution. The same trend can be observed in the prediction of survival (not shown): DWT and KELM improve performance both individually and in combination, resulting in the best performance in terms of survival. The improvement obtained by incorporating DWT highlights the importance of multi-scale texture cues in brain MRI, which are often suppressed in raw spatial representations. The additional gain achieved by KELM indicates that closed-form kernel-based classifiers provide better generalization on high-dimensional ViT embeddings compared to gradient-trained dense layers. The findings of the ablation study demonstrate the importance of using wavelet decomposition to extract fine-grained texture information that is typically buried in raw spatial representations. The further performance gain achieved using KELM instead of dense classifiers demonstrates that the use of kernel-based closed-form learning is effective in cases

involving high-dimensional transformer embeddings. This evidence demonstrates that the effects of both DWT and KELM on the overall effectiveness of the proposed system are independent and synergistic.

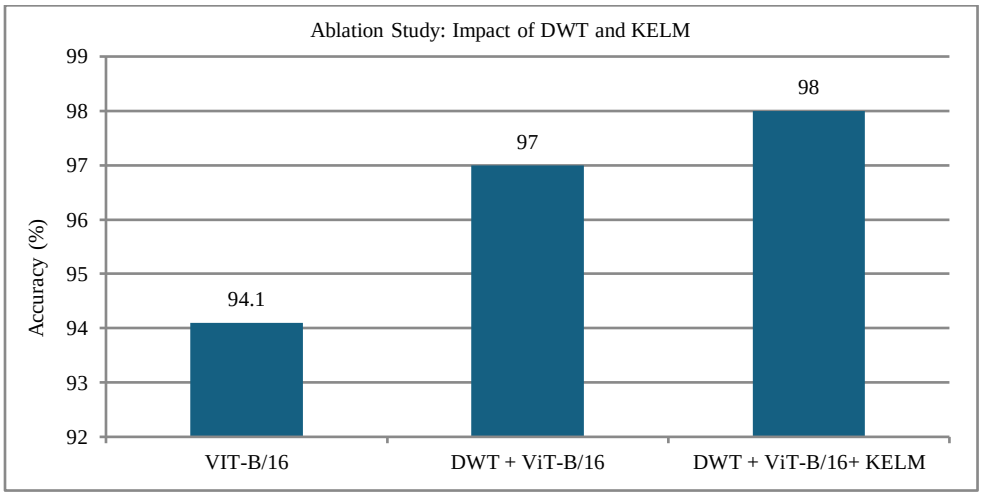
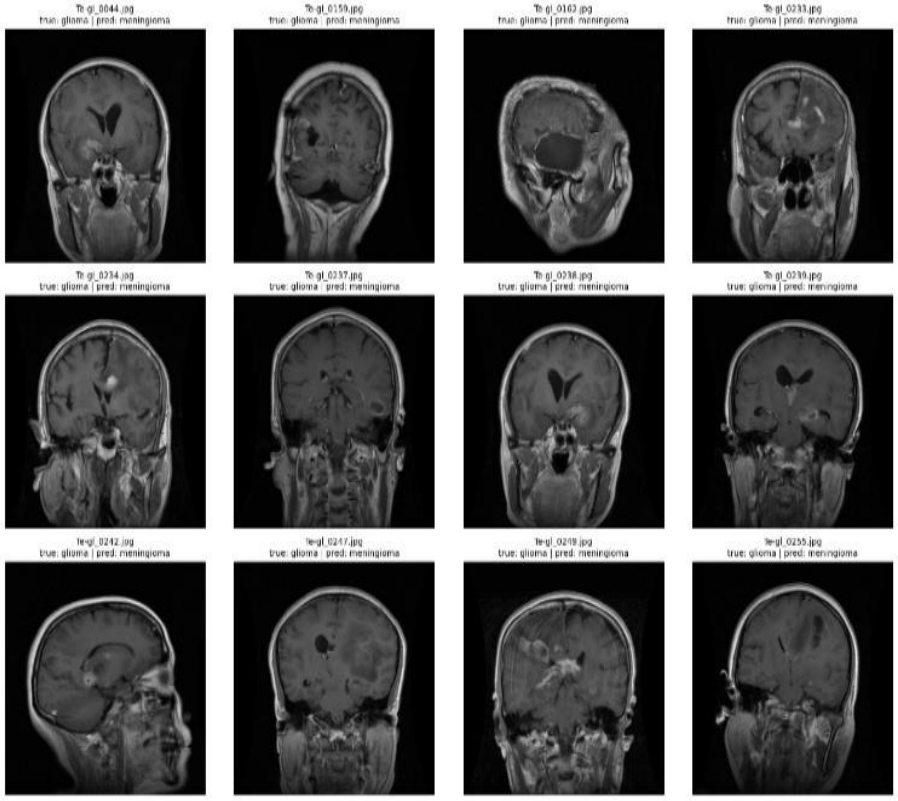


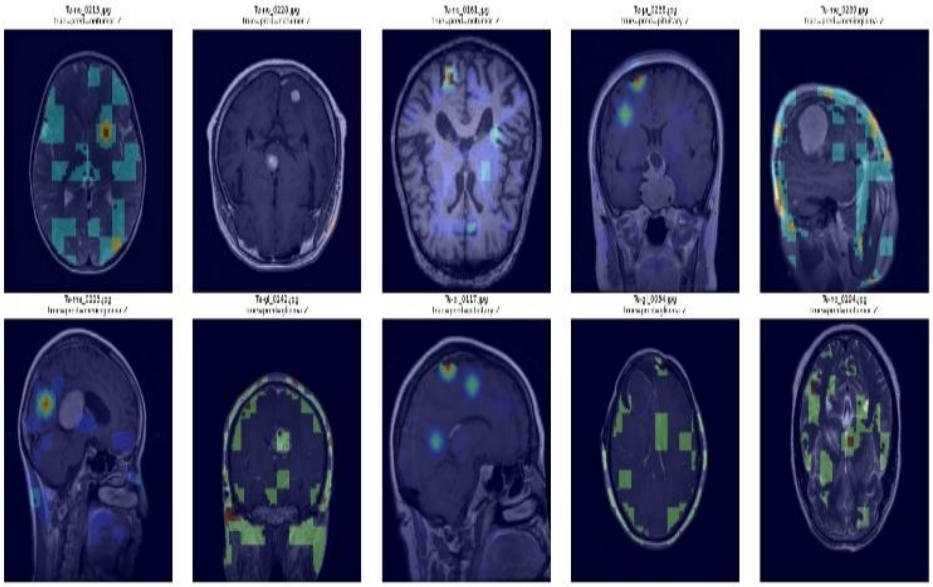
Fig. 5 Ablation analysis on tumour classification accuracy. Baseline ViT (no DWT, Dense Softmax Output) achieves 94.1% accuracy. Adding DWT improves accuracy to 97%. Using the KELM classifier instead of the dense output increases performance to 98%, highlighting the synergy between wavelet features and KELM heads

6.5. Explainability: Qualitative Results

Representative attention-based explanations are in Figure 6. In panel (a), Grad-CAM heatmaps are superimposed on misclassified examples, showing that failure cases are often extremely low-contrast tumours. In panel (b), ViT attention rollout maps are added on top of photos that are correctly classified, demonstrating that the model looks strong at the tumour cores and peri-tumoral regions and ignores healthy tissue. These visualizations help build confidence in the model's decision, as they demonstrate that predictions are based on medically meaningful features.



(a) Misclassified examples with grad-CAM overlays



(b) Correct classifications with attention rollout

Fig. 6(a) Grad-CAM heatmaps on misclassified images show spread or misplaced attention, indicating model uncertainty, and (b)ViT attention rollout overlays on correctly classified images concentrate on the tumour core and peri-tumoral regions, conforming to clinical expectations.

6.6. Error Analysis

Misclassifications are analyzed in order to identify common modes of failure. The confusion matrix in Figure 7 reveals that the errors are found mostly between similar tumour types (e.g., pituitary vs meningioma) rather than dissimilar classes (tumour vs. no tumour). Figure 6(a) gives an example of typical misclassification: images with very low contrast or a typical tumour appearance are responsible for erroneous predictions.

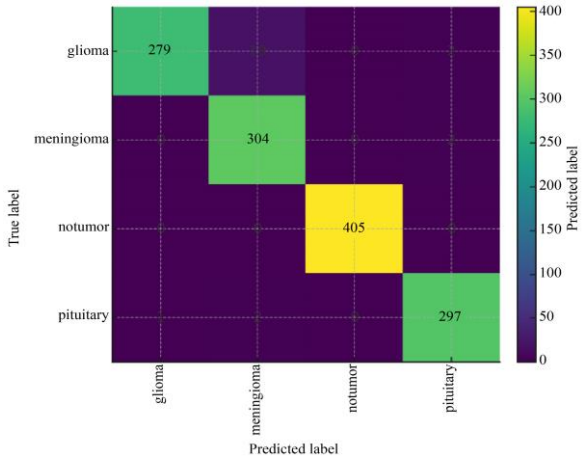


Fig. 7 Confusion matrix for the four tumour classes. The majority of errors occur between similar tumour types, while dissimilar classes (tumour vs. no tumour) are rarely confused. Values indicate the number of test samples per cell.

Attention maps are diffuse in these cases, denoted by the uncertainty. Such cases can be flagged for closer human review. Overall, the errors of the model are understandable and reflect radiologists' difficulties.

7. Future Work

The use of unified DWT+ViT+KELM presents several research opportunities in the future. First of all, three-dimensional versions of the ViT could be explored to process the entire volumetric MRI, rather than representative slices. Recent attempts in applying Swin transformers to volumetric data have suggested that 3D patching can be used to capture more robust contextual information about anatomy.

Second, multi-modal fusion, respectively combining MRI with perfusion, diffusion, PET, or clinical biomarkers, may allow for better classification and also better survival prediction.

Third, federated learning (privacy-preserving training on multi-institutional data) and privacy-preserving learning might enable this to be done, training using multi-institutional datasets, without having to share the raw patient data.

Fourth, an ordinal regression formulation of survival could supersede the discrete categories employed here; this would allow levels of prognostic estimate with the potential to be finer grained. Finally, the unified architecture could be extended for other tasks such as tumour segmentation, treatment response prediction, or cognitive impairment assessment in AD. By adding appropriate heads, the backbone can be used as a general feature extractor for neuroimaging.

7.1. Limitations & Bias

Despite promising results, this study has several constraints that must be recognized. First, the tumor classification experiments are conducted on 2D MRI slices, which may not fully capture three-dimensional tumor morphology. While this choice enables larger sample sizes

and reduced computational cost, future work should thoroughly investigate 3D transformer architectures. Second, survival prediction is evaluated on the BraTS 2020 dataset, which contains a finite quantity of annotated cases, potentially introducing dataset-specific bias and limiting generalizability across institutions.

Additionally, the use of publicly available datasets may reflect inherent acquisition and demographic biases that are not representative of real-world clinical populations. Although data augmentation and cross-validation mitigate overfitting, potential multi-center validation is necessary to verify robustness. Finally, while explainability methodologies, including Grad-CAM and attention mechanism rollout, provide qualitative insights, they do not guarantee causal interpretability, and clinical decision-making should not rely solely on model explanations.

8. Conclusion

A single architecture is presented here that achieves both tumor classification and survival prediction on multi-institutional MRI data. The model is used to simultaneously perform four-class tumour classification and three-category survival prediction using a combination of wavelet-based pre-processing, transformer feature extraction, and closed-form KELM classifiers. On the FigshareImages and BraTS2020 data sets, the method obtained 98.02% and 94.67% classification and survival prediction accuracy, respectively. Breaking the classical CNN and transformer baselines and closing in on the recently reported fine-tuned ViT results. Ablation studies showed that both DWT and KELM are important components. Interpretability using Grad-CAM and attention rollout showed the focus of the model on tumour areas and clinically relevant structures. This cohesive and computable approach holds promise for integrated neuroimaging analysis. Future research will involve volumetric transformers, multi-modal fusion, and challenges of inter-institutional federated training.

Informed Consent

The datasets utilized, publicly available and totally anonymized (FigshareImages and BraTS2020), are used in this study. The original data providers addressed the issue of informed consent and ethical approvals.

Acknowledgment

The work proposed here was done without any special external funding. Some of this manuscript has appeared in earlier versions of the Proceedings of ISPCC in 2025 and ISACC in somewhat more primitive form; here, the manuscript builds an elaborate elaboration upon those ideas and includes experiments and discussion in its entirety. The authors credit the BraTS organizers with providing BraTS2013-2020 datasets and the brain MRI contributors of Figshare. The authors also acknowledge the assistance of the computing facilities of the institutions we belong to and the valuable discussions with colleagues.

Author Contributions

The conceptualization process, design of methods, experiments, and writing of the manuscript were led by A.Gupta. Anita assisted in supervising, reviewing the design of the experiment, and revising the manuscript. M. Gupta gives clinical commentary, interpretation, and critique of the paper.

References

[1] Ze Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 9992-10002, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[2] Sérgio Pereira et al., "Brain Tumor Segmentation using Convolutional Neural Networks in MRI Images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1240-1251, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[3] Alexey Dosovitskiy et al., "An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale," *International Conference on Learning Representations*, 2021. [Google Scholar] [Publisher Link]

[4] Jieneng Chen et al., "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," *arXiv Preprint*, pp. 1-13, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[5] Ali Hatamizadeh et al., "UNETR: Transformers for 3D Medical Image Segmentation," *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, pp. 1748-1758, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[6] Finale Doshi-Velez, and Been Kim, "Towards a Rigorous Science of Interpretable Machine Learning," *arXiv Preprint*, pp. 1-13, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[7] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, "Why Should I Trust You?' Explaining the Predictions of Any Classifier," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA, pp. 1135-1144, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[8] Konstantinos Kamnitsas et al., "Efficient Multi-Scale 3D CNN with Fully Connected CRF for Accurate Brain Lesion Segmentation," *Medical Image Analysis*, vol. 36, pp. 61-78, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[9] Abdullah Almuhaimeed et al., "Brain Tumor Classification using GAN-Augmented Data with Autoencoders and Swin Transformers," *Frontiers in Medicine*, vol. 12, pp. 1-26, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[10] K. Chandraprabha, L. Ganesan, and K. Baskaran, "A Novel Approach for the Detection of Brain Tumor and Its Classification via End-to-

End Vision Transformer-CNN Architecture," *Frontiers in Oncology*, vol. 15, pp. 1-18, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[11] Clifford R. Jack "NIA-AA Research Framework: Toward a Biological Definition of Alzheimer's Disease," *Alzheimer's & Dementia*, vol. 14, no. 4, pp. 535-562, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[12] Chunfeng Lian et al., "Multi-Task Weakly-Supervised Attention Network for Dementia Status Estimation with Structural MRI," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 8, pp. 4056-4068, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[13] Emerald U. Henry, Onyeka Emebob, and Conrad Asotie Omonhinmin, "Vision Transformers in Medical Imaging: A Review," *arXiv Preprint*, pp. 1-31, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[14] Hugo J. W. L. Aerts et al., "Decoding Tumour Phenotype by Noninvasive Imaging using a Quantitative Radiomics Approach," *Nature Communications*, vol. 5, pp. 1-9, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[15] Jiangwei Lao et al., "A Deep Learning-Based Radiomics for Prediction of Survival in Glioblastoma," *Scientific Reports*, vol. 7, pp. 1-8, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[16] Spyridon Bakas et al., "Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BraTS Challenge," *arXiv Preprint*, pp. 1-49, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[17] Ujjwal Baid et al., "The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification," *arXiv Preprint*, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[18] Ujjwal Baid et al., "The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification,"*arXiv Preprint*, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[19] Bohua Deng et al., "An Overview of Extreme Learning Machines," *2019 4th International Conference on Control, Robotics and Cybernetics (CRC)*, Tokyo, Japan, pp. 189-195, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[20] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew, "Extreme Learning Machine: Theory and Applications," *Neurocomputing*, vol. 70, no. 1-3, pp. 489-501, 2006. [CrossRef] [Google Scholar] [Publisher Link]

[21] Guang-Bin Huang et al., "Extreme Learning Machine for Regression and Multiclass Classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 2, pp. 513-529, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[22] Ramprasaath R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336-359, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[23] Aditya Chattopadhay et al., "Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks," *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Tahoe, NV, USA, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[24] Samira Abnar, and Willem Zuidema, "Quantifying Attention Flow in Transformers," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190-4197, 2020. [Google Scholar] [Publisher Link]

[25] Scott M Lundberg, and Su-In Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems*, vol. 30, pp. 1-10, 2017. [Google Scholar] [Publisher Link]

[26] Mukund Sundararajan, Ankur Taly, and Qiqi Yan, "Axiomatic Attribution for Deep Networks," *Proceedings of the 34th International Conference on Machine Learning*, Sydney NSW Australia, vol. 70, pp. 3319-3328, 2017. [Google Scholar] [Publisher Link]

[27] Jost Tobias Springenberg et al., "Striving for Simplicity: The All Convolutional Net," *arXiv Preprint*, pp. 1-14, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[28] Nicholas J. Tustison et al., "N4ITK: Improved N3 Bias Correction," *IEEE Transactions on Medical Imaging*, vol. 29, no. 6, pp. 1310-1320, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[29] Stephen M. Smith, "Fast Robust Automated Brain Extraction," *Human Brain Mapping*, vol. 17, no. 3, pp. 143-155, 2002. [CrossRef] [Google Scholar] [Publisher Link]

[30] Joost J.M. van Griethuysen et al., "Computational Radiomics System to Decode the Radiographic Phenotype," *Cancer Research*, vol. 77, no. 21, pp. e104-e107, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[31] Ashish Vaswani et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998-6008, 2017. [Google Scholar] [Publisher Link]

[32] MONAI, Medical Open Network for Artificial Intelligence, 2020. [Online]. Available: https://project-monai.github.io/

[33] C. Kishor Kumar Reddy et al., "A Fine-Tuned Vision Transformer based Enhanced Multi-Class Brain Tumor Classification using MRI Scan Imagery," *Frontiers in Oncology*, vol. 14, pp. 1-23, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[34] Min Hao et al., "Survival Prediction in Gliomas based on MRI Radiomics Combined with Clinical Factors and Molecular Biomarkers," *PeerJ*, vol. 13, pp. 1-23, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[35] Ahmad Chaddad et al., "A Radiomic Model for Gliomas Grade and Patient Survival Prediction," *Bioengineering*, vol. 12, no. 5, pp. 1-19, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[36] Roya Poursaeed, Mohsen Mohammadzadeh, and Ali Asghar Safaei, "Survival Prediction of Glioblastoma Patients using Machine Learning and Deep Learning: A Systematic Review," *BMC Cancer*, vol. 24, no. 1, pp. 1-36, 2024. [CrossRef] [Google Scholar] [Publisher

Link]
[37] Palani Thanaraj Krishnan et al., "Enhancing Brain Tumor Detection in MRI with a Rotation Invariant Vision Transformer," *Frontiers in Neuroinformatics*, vol. 18, pp. 1-13, 2024. [CrossRef] [Google Scholar] [Publisher Link]
[38] Fatma E. AlTahhan et al., "Refined Automatic Brain Tumor Classification using Hybrid Convolutional Neural Networks for MRI Scans," *Diagnostics*, vol. 13, no. 5, pp. 1-16, 2023. [CrossRef] [Google Scholar] [Publisher Link]
[39] Bjoern H. Menze et al., "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993-2024, 2015. [CrossRef] [Google Scholar] [Publisher Link]
[40] Navoneel Chakrabarty, Kaggle Contributor, Brain MRI Images for Brain Tumour Classification, 2020. [Online]. Available: https://www.kaggle.com/datasets/navoneel/brain-mri-images-for-brain-tumor-detection
[41] Diederik P. Kingma, and Jimmy Ba, "Adam: A Method for Stochastic Optimization," *arXiv Preprint*, pp. 1-15, 2017. [CrossRef] [Google Scholar] [Publisher Link]
[42] Ilya Loshchilov, and Frank Hutter, "Decoupled Weight Decay Regularization," *arXiv Preprint*, pp. 1-19, 2019. [CrossRef] [Google Scholar] [Publisher Link]
[43] Ilya Loshchilov, and Frank Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts," *arXiv Preprint*, pp. 1-16, 2017. [CrossRef] [Google Scholar] [Publisher Link]
[44] Adam Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," *Advances in Neural Information Processing Systems*, vol. 32, 2019. [Google Scholar] [Publisher Link]
[45] Lutz Prechelt, *Early Stopping-But When?*, Neural Networks: Tricks of the Trade, Springer, Berlin, Heidelberg, pp. 55-69, 1998. [CrossRef] [Google Scholar] [Publisher Link]