International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(6s) 2026 pp. 243-258
https://doi.org/10.70917/ijcisim-2026-2882
Article

Self-Supervised Deep Learning Framework for Secure Image Classification under Adversarial Attack Environments

Rashmi Abhijeet Jadhav[1], Devyani Swapnil Jadhav[2]

[1]School of Engineering and Technology, Sanjivani University, Kopargaon, Maharashtra, India
rashmijadhav3@gmail.com
[2]School of Engineering and Technology, Sanjivani University, Kopargaon, Maharashtra, India
bhamare.devyani29@gmail.com

Abstract: Ensuring accurate detection and classification of brain tumors in magnetic resonance imaging (MRI) is essential for early diagnosis and effective treatment. Conventional methods often suffer from limited robustness, inadequate feature representation, and reduced generalization, particularly under adversarial perturbations. To develop a robust and efficient model for brain tumor detection and classification that performs reliably under both normal and adversarial conditions. Validation is performed on two MRI datasets comprising 7,200 and 253 images. Images undergo preprocessing steps, including resizing, Z-score normalization, and Gaussian noise reduction, to enhance image quality and reduce false positives. Feature extraction is carried out using a pretrained VGG16 network to capture discriminative spatial and hierarchical patterns. The proposed model, Improved Squirrel Search Algorithm–Attention based Capsule Network (ISSA-Att-CapsNet), integrates deep learning and optimization techniques for robust tumor classification. The Attention-based Capsule Network (Att-CapsNet) emphasizes significant tumor regions and captures spatial relationships between features, while the Improved Squirrel Search Algorithm (ISSA) optimizes feature weighting and hyperparameters, enhancing stability and performance. To assess robustness, adversarial attacks including Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Universal Adversarial Perturbations (UAP) are incorporated, where FGSM generates single-step gradient-based perturbations, PGD produces iterative perturbations, and UAP creates universal perturbations affecting multiple images. Implementation is performed in Python 3.9 using TensorFlow/Keras. Experimental results show high performance, with accuracy, precision, recall, and F1-score ranging from 98.0% to 99.0% under normal conditions and 96.5% to 97.8% under adversarial attacks, demonstrating the model's effectiveness, robustness, and clinical potential for automated brain tumor detection and classification.

Keywords: Self-Supervised Learning, Adversarial Attacks, Secure Image Classification, Deep Learning, Robustness, Universal Adversarial Perturbation.

1. Introduction

Medical imaging devices like MRI is important in the modern healthcare system as they allow viewing of the internal organs and structures of the body without making any invasions and form the basis of accurate diagnosis of the disease [1]. With the recent developments in DL and ML, medical image analysis has significantly improved [2]. Brain tumors, among other neurological complications, are some of the most life-threatening complications. Brain tumors may be harmless or cancerous, and the most aggressive form of brain tumors is the gliomas. Early and precise diagnosis is essential since classification and grading of tumors determine the therapy plan and prognosis of a patient [3]. Because of its superior soft-tissue contrast resolution and ability to see finer details of cerebral structures, MRI is widely used in the detection of tumors in the brain. However, MRI scans can only be analyzed manually, which is a laborious task and requires the expertise of an individual, thereby underscoring the need for automated diagnostic systems based on artificial intelligence [4]. The computer-aided diagnostic systems based on deep learning have dramatically enhanced the efficiency and the accuracy of the brain tumor classification using MRI images. These technologies enable early identification of anomalies, which is vital for effective clinical therapy and patient care [5].

Copyright: © 2026 by the authors

Deep learning models have achieved remarkable results but at the same time, adversarial attacks are a serious challenge to this technology. Adversarial perturbations are little adjustments to the input images that are often invisible but may make models make inaccurate predictions. Medical imaging requires such manipulations that may lead to misdiagnosis which is very dangerous to patients [6]. It has been demonstrated that deep learning models are vulnerable to image classification by several methods of adversarial attacks, including FGSM and PGD. These attacks take advantage of the natural weaknesses in neural networks hence significantly compromising the performance of model [7]. It is against this background that secure and safe deep learning network development to classify medical images has become a significant research goal. It is necessary to improve the quality of diagnostic systems that work in the adversarial environment with the help of defensive techniques and effective learning procedures [8]. More advanced learning models are being investigated to enhance model stability and generalization such as self-supervised learning. To make such medical image classification systems more robust and resistant to adversarial attacks, the system may be trained to make informed decisions with the help of self-supervised architectures [9]. Health and treatment depends on the early and accurate diagnosis of the disease. MRI enables accurate and safe diagnosis of brain tumours with ease, whereas computer-aided analysis enhances classification and effective planning of treatment as well as reliable evaluation of tumour progression [10].

This research aims to come up with a flexible and reliable model to automatically detect and classify brain tumours in MRI images to guarantee reliable performance under both normal and adversarial conditions with constraints regarding limited robustness, inadequate feature representation, and isopimality in generalization. The key contributions are as follows:

- Improved Squirrel Search Algorithm VGG16 Attention -Capsule Network (ISSA-Att-CapsNet) can successfully extracted features and classified brain tumours in MRI images effectively.
- Att -CapsNet focuses on salient tumour regions and takes care of the spatial relations, thus, boosting the hierarchical representation of tumour features.
- The ISSA is a more advanced search algorithm that also optimizes the parameter of the model, which is the weights of the features and thus enhances the stability of the model and the performance of the classification.

The research is organized as follows: Section 2 is the literature review on the topic of CNNs, capsule networks, adversarial robustness; Section 3 introduces the proposed ISSA-Att-CapsNet model that includes preprocessing of the dataset, the extraction of features with VGG16, attention-based capsule networks, and the ISSA optimisation; Section 4 explores the experimental setup, evaluation metrics, classification performance, adversarial attack resistance and the comparison with baseline models; Section 5 presents the key research findings, limitations, and future research.

2. Related Work

A brain tumor was detected and classified with improved accuracy and reduced computational complexity [11]. Gray-Level Co-occurrence Matrix (GLCM) was used to extract features from MRI images after Otsu binarization preprocessing. Using a hybrid Deep Convolutional Neural Network (DCNN) with Covid-19 optimization and Residual Network with 152 Layers (ResNet-152) transfer learning, 97.57% accuracy with a low error rate was attained. A number of studies have used deep learning models that are based on transfer learning to automatically identify and classify brain tumors using MRI images [12]. Specifically, CNN transfer-learning models that include ResNet-50, Inception-V3 and VGG-16 were evaluated, and VGG-16 proved to be the most successful regarding training and validation accuracy on the detection of brain tumors. An automated method in brain tumorclassification was created using an MRI scan [13]. GLCM was used to obtain features in MRI images following adaptive filtering and segmentation using Improved K -Means Clustering (IKMC). A Recurrent Convolutional Neural Network (RCNN) was used to classify the Kaggle data set with an accuracy of 95.17%.

The accuracy of brain MRI image categorization was increased further by improving previous transfer-learning algorithms [14]. On Kaggle MRI data, a deep-learning network comprised of CNN, VGG-16, and DenseNet was trained using an 80/20 split of data and 10-fold cross validation. Classification accuracy improved considerably. Improve the speed and accuracy of brain tumour identification using MRI scans [15]. It employed a two-module strategy: (1) Image Enhancement Technique with neural networks, independent component analysis and adaptive Wiener filtering (2) Support Vector Machines (SVM) for tumour separation and prediction, and. Outperforming current approaches in efficiency and precision, it achieved 0.989 accuracy. Despite the complexity and variability of tumor characteristics, accurately categorize brain tumor MRI pictures as benign or malignant [16]. It used segmentation (Otsu's thresholding), feature extraction, classification using SVM, and noise reduction. It outperformed earlier methods with a classification accuracy of 97.9% on 24 MRI scans.

The goal was to create a powerful and fast DL framework for accurate and reliable brain MRI classification. A Multi-Expert Fusion approach combined three models with Taguchi optimisation, denoising, adversarial training, and fuzzy logic to produce a lightweight architecture with 738,100 parameters, high classification accuracy, improved PSNR (58.09 dB), and increased robustness to FGSM/PGD attacks. To assess DL models' resilience and susceptibility to adversarial attacks in medical image processing [18]. FGSM and PGD attacks were used to evaluate four distinct deep learning models on complete and partial images of varied sizes from three medical datasets. The models displayed varying degrees of susceptibility, with the amount and location of adversarial perturbations having a significant impact on accuracy and resilience. The gradient-free activation networks were compared to MLP and ResNet18 in [19] using MRI and histopathology images. The proposed model was better than the two baselines, as more distortion had to be surpassed to reach 88.89 % accuracy on hostile cases.

2.1 Research Gap

The primary weakness of the studies mentioned above is their reliance on accuracy as the sole performance indicator, while omitting other important metrics. The Ability to resist adverse conditions or adversarial examples has not been studied and may undermine reliability [11]. The use of various expert models and special optimization methods increases the complexity of implementation. As a result, they can cause worse performance on unexplored or fine diverse MRI datasets and require great expertise to be fine-tuned [17]. Also, the proposed models might not be able to cope with the complex or new adversarial approaches, and their performance in various medical imaging modalities would fluctuate. The calculations required to perform training in sign-activation networks also provide an additional challenge to scalability [19]. To overcome these constraints, computational needs may be minimized by adopting light, efficient architectures. Noisy, low-resolution, or unseen MRI data can be strengthened by means of data augmentation and adversarial training. Automated hyper parameter optimization and modular design of models can be used to reduce the complexity of implementation, and transfer learning as well as efficient inference techniques can be used to improve model efficiency and scalability.

3. Methodology

The proposed ISSA–Att-CapsNet model classifies brain tumors from MRI images by first preprocessing data with Resizing, Gaussian filtering, and Z-score normalization. Features are extracted using VGG16, while the Att-CapsNet captures spatial relationships for accurate classification. ISSA optimizes feature weighting and model parameters, and robustness is evaluated under FGSM, PGD, and UAP adversarial attacks. Figure 1 illustrates the ISSA-Att-CapsNet pipeline for brain tumour detection and adversarial robustness.

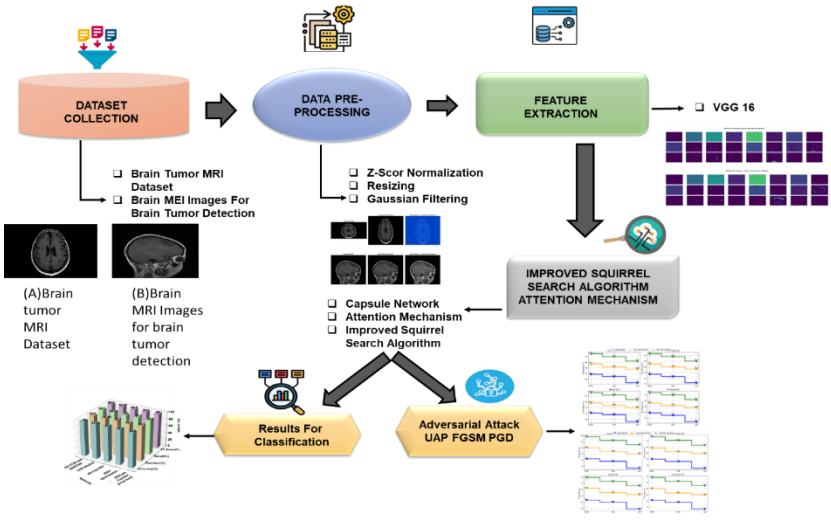


Figure 1. ISSA-Att-CapsNet model for tumour MRI categorisation and under adversarial attacks

3.1 Dataset Collection

Accurate and reliable brain tumor image classification is achieved using deep learning with robust feature extraction and adversarial-resilient models.

- **Brain Tumor MRI Dataset (BTD)** – This dataset includes 7,200 MRI images divided into four categories: glioma, meningioma, no tumour and pituitary tumour. The images have been separated into training and testing where the classes are equally divided and offer a stable source of development and testing of automated brain tumour classification models. (https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset).
- **Brain MRI Images for Brain Tumor Detection (BITD):** This data comprise of 253 MRI images, which were coded to differentiate tumour-affected brain tissues and normal ones. The dataset involves the variations in MRI sequences and image quality; the models can be generalized to different clinical conditions. Representatives are tumour-affected and normal brain scans, which can be used to check classification and adversarial robustness (https://www.kaggle.com/datasets/navoneel/brain-mri-images-for-brain-tumor-detection). Figure 2 shows representative MRI scans from the (A) BTD and (B) BITD used.

(A) Brain Tumor MRI Dataset (B)Brain MRI Images for Brain Tumor Detection

Figure 2. Sample MRI Images from the (A) BTD and (B) BITD

3.2 Data Preprocessing

To improve the extraction of features and the performance of the model, the images of the brain MRI were resized to 224 x 224 pixels, followed by a denoising with Gaussian filtering and normalization with Z-score standardization.

3.2.1 Z-Score Normalization for Brain MRI Image and Metadata

Normalization as one of the preprocessing processes is critical to dependable classification of brain tumours. Normalization scales pixel intensities and makes the models more efficient besides learning strong features that are resistant to adversarial perturbations. It also minimizes the variability in the input data hence improving the consistency and quality of the model.. In particular, z-score normalization standardizes each pixel value v_i of an image channel E to have a mean of 0 and standard deviation of 1, mapping it into a normalized range suitable for DL models. This can be expressed as Eq. (1).

$$u' = \frac{u_j - F_j}{Std(F)} \quad (1)$$

Where $std(F)$ is the standard deviation attribute F, u_j represents the value to be normalized in the attribute, F_j is the mean value of the attribute, and u' is the result of the normalization value.

3.2.2 Image Resizing for Brain MRI Image Data

MRI images were resized to 224×224 pixels to ensure model compatibility, preserve key tumor features, reduce computation, and improve feature extraction, training stability, and generalization for robust brain tumor classification under adversarial attacks.

3.2.3 Noise Reduction Using Gaussian Filtering for Brain MRI Images

A Gaussian filter was applied to reduce noise in brain MRI images while preserving key tumor features, improving image quality, and enabling more reliable feature extraction for robust classification. The Gaussian Filter is defined as Eq. (2).

$$G(x,y) = \frac{1}{2\pi\sigma^2} exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \quad (2)$$

In Equation 1, $G(x, y)$ is the Gaussian weight at pixel location (x, y), x and y are the horizontal and vertical offsets from the center of the kernel, σ represents the standard deviation controlling the smoothing intensity, and $\exp(.)$ is the exponential function that reduces the weight of distant pixels. Figure 3 (a-b) illustrates the preprocessing steps applied to MRI images from both the (A) BTD and (B) BITD, including resizing, Z-score normalization, and Gaussian filtering.

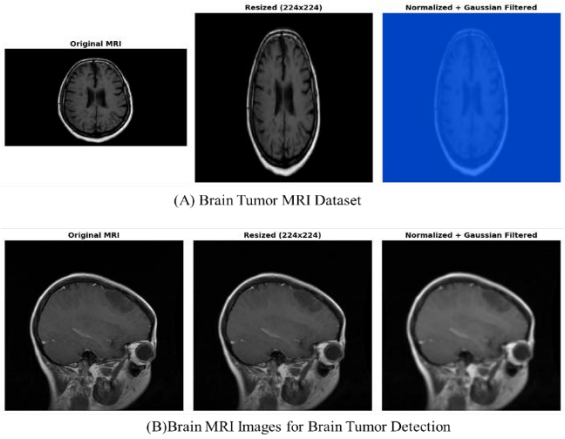


Figure 3. Pre-processing stages of brain MRI images from the (A) BTD and (B) BITD, including resizing to 224 × 224 pixels, Z-score normalization, and Gaussian filtering

3.3 Feature Representation via VGG16

The approach uses VGG16 as a feature extractor, removing its fully connected layers and adding custom classification layers. The convolutional layers capture generalized features thus enhancing model resilience and robustness to adversarial attacks and reducing reliance on labelled data. Brain MRI features extracted via the pretrained VGG16 model are illustrated in Figure 4.

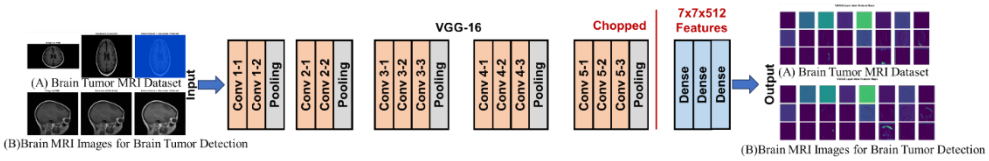


Figure 4. Feature extraction from brain MRI images using VGG16 pretraining with final fully connected layers

Input Layers: In this layer, the image data is passed to subsequent layers where meaningful and discriminative features are extracted, forming the foundation for the network to resist adversarial perturbations and make reliable predictions.

Convolutional: To extract key characteristics from images, the convolutional layer analyses the input through many filters. These retrieved characteristics assist the network in learning discriminative representations, increasing its resilience and reliability against adversarial attacks. These features are used in calculating the suitability of each test procedure as Eq. (3)

$$Output = \frac{m+2o-e}{t} + 1 \tag{3}$$

Where $output$ represents the final computed result of the expression, t is the strides, e is the kernel filter length or kernel feature height, o is the padding, and m is the input length or height.

Pooling: The pooling layer reduces spatial dimensions while preserving key features, enhancing efficiency and adversarial robustness. Max Pooling selects the highest values, and Global Average Pooling averages values to retain essential information without extra parameters. The MaxPool output is computed as shown in Eq. (4).

$$Outputmaxpool = \frac{m-e}{t} + 1 \quad (4)$$

Fully connected layer: The fully linked layer transforms extracted information into classes, enabling secure and reliable image classification even under adversarial attacks. The use of this layer is to unite all nodes into one dimension, Eq. (5)

$$Y_i = \sum_{i=1}^{d} x_{j,i}^{S} \; W_j + a_i \quad (5)$$

Where Y_i is the output value of the network, a_i is the bias of the network, W_j is the input result from feature extraction, $x_{j,i}$ represents the network weight value whose size is $j \times i$, i is the number of target classes, j is the number of input features, d is the total number of input features and s represents the preprocessing step applied to the input. Figure 5 (a-b) illustrates the layer-wise feature maps extracted by VGG16 from brain MRI images of both the (A) BTD and (B) BITD.

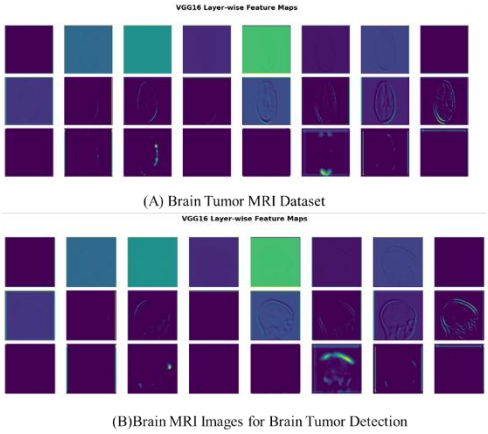


Figure 5. VGG16 layer-wise feature map visualizations for (A) BTD and (B) BITD

3.4 Proposed ISSA-Att-CapsNet Model for Robust Brain Tumor Classification and Adversarial Resilience

The proposed ISSA-Att-CapsNet model classifies brain tumours from MRI images using a layer-wise design that combines the Capsule Network, Attention Mechanism, and Improved Squirrel Search Algorithm (ISSA).First, VGG16 feature maps are fed into the Capsule Network, where convolutional capsules encode local features, followed by fully connected capsules that aggregate them into higher-level hierarchical representations, preserving spatial and part-whole relationships for robust class separation. The attention module works in parallel with trunk and mask branches, producing refined feature maps by prioritizing the clinically important parts of the tumor and downplaying the irrelevant ones. This increases feature discriminability as well as adversarial robustness. These enriched features are then used in a multi-class model for tumor classification using fully connected layers. The ISSA optimization layer optimizes hyperparameters and feature weights dynamically by using adaptive predator probability, cloud-based position updates, dimension-wise search, and fitness-based selection, which guarantees hyperparameters of the network to be highly tuned.

3.4.1 Capsule Network for Hierarchical Feature Learning

CapsNet enhances adversarial robustness by capturing spatial and hierarchical relationships. Its convolutional and capsule layers extract local features and encode their presence and orientation, improving classification accuracy and reliability. As shown in Figure 6 (A-B), the Capsule Network is formed of a complex three-layer deep network.

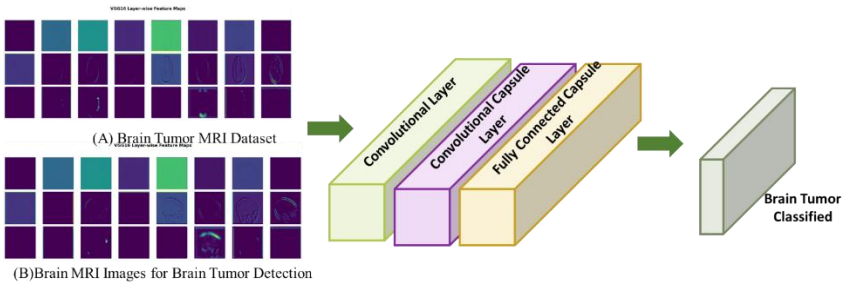


Figure 6. Proposed Capsule Network (CapsNet) Architecture for Brain Tumor Classification using VGG16 Feature Extraction

To calculate the result capsule, apply a nonlinear limiting function since the size of the result vector indicates the chance that it exists, Eq. (6)

$$u_i = \frac{||t_i||^2}{1+||t_i||^2} \frac{t_i}{||t_i||} \tag{6}$$

Where u_i is the vector output of capsule i following nonlinear squashing, $and\ t_i$ is its total input. With this purpose, small vectors are restricted to practically zero length, while large vectors are scaled to a length lower than one, Eq. (7)

$$t_i = \sum_j D_{ji} \hat{v}_{i|j} \tag{7}$$

A weight matrix is multiplied by a capsule's output v_j to get the input t_i. The coupling coefficient D_{ji} in Eq. (8) is determined using an iterative dynamic routing method.

$$D_{ji} = \frac{\exp(a_{ji})}{\sum_l \exp(a_{jl})} \tag{8}$$

Where $a_{ji}\ and\ a_{jl}$ show the log prior probability for two connected capsules. D_{ji} represents the normalized attention weight or probability value for the i^{th}element in the j^{th}group, obtained using the softmax function.

Each lower-level capsule sends predictions through dynamic routing to higher-level capsules, using transformation matrices to encode class-specific features and capture hierarchical relationships. Therefore, the overall length of the vector is representative of the probability level. For each capsule, the margin loss K_l can be written as follows, Eq. (9)

$$K_l = S_l \max(0, n^+ - ||u_l||)^2 + \lambda(1 - S_l)\max(0, ||u_l|| - n^-)^2 \tag{9}$$

Where $S_l = \begin{cases} 1, if class k is active \\ 0, otherwise \end{cases}, \lambda = 0.5, n^+ = 0.9\ , and\ n^- = 0.1$ Where λ is the Down-weighting parameter to reduce loss for absent classes, $\max(0,.)$ is the ensures only positive margin violations contribute to the loss, n^- is the lower margin threshold for negative class, n^+ is the upper margin threshold for positive class, $\|u_l\|$ length (magnitude) of the capsule output vector, representing class probability, u_l is the output vector of the l^{th}capsule, S_l ground truth label indicator (1 if class $l$is present, 0 otherwise) and K_l is the loss value for the l^{th} class(margin loss in capsule network).

The capsule-based hierarchical feature learning allows the model to collect locations and particular tumor properties, boosting classification accuracy and resilience to adversarial perturbations in brain MRI images.

3.4.2 Attention Mechanism for Feature Prioritization

The attention mechanism enhances robustness by highlighting important features and suppressing irrelevant ones, improving classification accuracy and reliability under adversarial attacks. The attention module is divided into

two branches: the trunk branch, which creates features E_o, and the mask branch, which combines LBP features to generate attention maps E_n. As indicated in Eq. (10), the element-wise product is then applied to attention maps E_n and feature maps E_o to produce refined feature maps E_n.

$$E_{refine} = E_o \otimes E_n \tag{10}$$

Where E_{refine} is the refined feature representation after enhancement, E_o is the original feature map or extracted feature representation, E_n is the attention map or weighted feature applied to enhance important regions, and $\otimes$ is the element-wise multiplication used to refine the original features. When the input of a final layer in the filter section is e_n, the attention mappings E_n are created as Eq. (11).

$$E_n = Sigmoid\ (X \copyright e_n + a) \tag{11}$$

The convolution layer's values and tilt are represented by X and a, respectively, with © representing the convolution operation and sigmoid representing the sigmoid function. The Sigmoid activation function gives (0, 1) probability ratings, enabling the network to discriminate the significance of distinct factors. The attention mechanism emphasizes the most relevant tumor regions in brain MRI images, enhancing feature representation and improving classification accuracy and robustness against adversarial attacks.

3.4.3 Improved Squirrel Search Algorithm for Hyperparameter Optimization

The ISSA enhances the Attention-Enhanced Capsule Network's parameters for reliable and secure image classification against adversarial attacks. It effectively looks for ideal values, enhancing accuracy, and durability against adversarial perturbations, and feature weighting. This section introduces four techniques to increase the squirrel search optimization algorithm's searching capabilities.

An Adaptive Predator Probability Strategy: The ISSA uses the predator presence probability O_{dp} to balance exploration and exploitation. Early in the search, a higher O_{dp} encourages diverse exploration, while later iterations reduce O_{dp} to focus on optimal parameter tuning, improving network performance and resilience against adversarial perturbations, which actively alter as a function of the repetition amount, as follows: Eq. (12)

$$O_{dp} = \left(O_{dpmax} - O_{dpmin}\right) \times (1 - Iter|Iter_{max})^{10} + O_{dpmin} \tag{12}$$

Where O_{dpmax} and O_{dpmin} are the greatest and least predator occurrence risks, respectively.

Random Position Generation for Flying Squirrels Using a Cloud Generator: When $Q_1, Q_2, Q_3 < Q_{dp}$, flying squirrels glide at random to new sites, with fluctuation in direction and speed representing the unpredictability and looseness of their feeding. It helps the algorithm explore the search space effectively and improve parameter tuning for robust and adversarial-resilient image classification. Eq. (13) - (15) describe the position update rules of the improved Squirrel Search Algorithm.

$$FS_{bs}^{new} = \begin{cases} FS_{bs}^{old} + c_h H_d \left(FS_{gs}^{old} - FS_{bs}^{old}\right), if\ Q_1 \geq O_{dp} \\ Dw\left(FS_{bs}^{old},\ F_m G_f\right),\ Otherwise \end{cases} \tag{13}$$

$$FS_{ms}^{new} = \begin{cases} FS_{bs}^{old} + c_h H_d \left(FS_{bs}^{old} - FS_{ms}^{old}\right), if\ Q_2 \geq O_{dp} \\ Dw\left(FS_{ms}^{old},\ F_m G_f\right),\ Otherwise \end{cases} \tag{14}$$

$$FS_{ms}^{new} = \begin{cases} FS_{ms}^{old} + c_h H_d \left(FS_{gs}^{old} - FS_{ms}^{old}\right), if\ Q_3 \geq O_{dp} \\ Dw\left(FS_{ms}^{old},\ F_m G_f\right),\ Otherwise \end{cases} \tag{15}$$

Where $F_m G_f$ is the feature map used in the cloud-based random update, $Dw\left(FS_{bs}^{old},\ F_m G_f\right)$ is the random position generated using the normal cloud model for diversity, O_{dp} is the predator presence probability threshold, H_d

is the glide distance factor, C_h is the glide coefficient controlling movement step size, FS_{gs}^{old} position of the gliding squirrel, FS_{bs}^{old} is the previous position of the best squirrel and FS_{bs}^{new} is the updated position of the best squirrel in the current position.

A Strategy for Selecting Successive Positions: During optimization, new positions replace previous ones only if they improve performance, ensuring optimal parameters and enhancing robustness and accuracy against adversarial attacks. It can be mathematically stated by Eq. (16).

$$FS_j = \begin{cases} FS_j^{new}, if\ e_j^{new} < e_j^{old} \\ FS_j^{old}, Otherwise \end{cases} \quad (16)$$

Where FS_j is the recently altered spot of the j^{th} flying squirrel, FS_j^{new} is the newly generated candidate position for the j^{th} squirrel, FS_j^{old} is the previous of the j^{th} squirrel , e_j^{new} is a crucial value for the new location, measuring quality and e_j^{old} is the fitness value at the old position.

Improve the intensity level of search: The basic SSA updates all dimensions of a solution at the same time, which might produce interference between dimensions. To improve parameter optimization for robust and secure image classification, each dimension is updated individually: the best solution is identified, and a new solution is generated by changing one dimension at a time, its fitness is compared with the original, and the better value is retained. This enhances the fine-tuning of network parameters and improves resilience against adversarial attacks. The newly generated solution is produced by Eq. (17).

$$FS_{best,i}^{new} = Dw\left(FS_{best,i}^{new}, F_m G_f\right), i = 1,2, \dots, m \quad (17)$$

Where $FS_{best,i}^{new}$ updated value of the i^{th} dimension of the best squirrel's position, $FS_{best,i}^{old}$ previous value of the i^{th}dimension of the best squirrel's position, $Dw(.)$ is the cloud-based random generator function used to refine or diversify the position, $i = 1,2, \dots, m$ is the index of dimensions in the solution vector, with $m$being the total number of dimensions.

The proposed ISSA-Att-CapsNet integrates VGG16, Attention-based Capsule Network, and ISSA for hyperparameter optimization, enabling accurate and robust brain tumor classification from MRI images, including resilience to FGSM, PGD, and UAP attacks. The ISSA-Att-CapsNet model, uses ISSA to optimize the hyperparameters of the Att-CapsNet for accurate and robust brain tumor classification.

3.5 Adversarial Attack Techniques for Robustness Evaluation

The proposed model's robustness is evaluated using FGSM, PGD, and UAP attacks, demonstrating its ability to maintain accurate and reliable brain tumor classification even under adversarial perturbations.

3.5.1 FGSM – Fast Gradient Sign Method

The ISSA–Att-CapsNet model is evaluated under FGSM attacks, which add small gradient-based perturbations to MRI images to test and improve its robustness and reliability in brain tumor classification. FGSM is a gradient-based approach that transforms a source image X corresponding to the loss function's gradient L relative to the input. Eq. (18) produces the adversarial image X_{adv}.

$$X_{adv} = X + \epsilon . sign(\nabla_X \mathcal{L}(\theta, X, y)) \quad (18)$$

Where $\nabla_X \mathcal{L}(\theta, X, y)$ shows the gradient of loss L with respect to X, X is the dataset's initial input image, ϵ is a little positive scalar that regulates the perturbation's magnitude, which establishes the strength of the adversarial attack, θ are the model parameters, and X_{adv} is the adversarial image that results from a little disturbance. The $sign(.)$ ensures the perturbation maximally increases the loss, fooling the classifier while keeping changes small and y is the true label.

3.5.2 Projected Gradient Descent (PGD)

Projected Gradient Descent (PGD) is used to evaluate the robustness of the proposed ISSA–Att-CapsNet by generating stronger adversarial MRI images through iterative perturbations. As a powerful white-box attack, PGD helps assess and improve the model's resistance to multi-step adversarial perturbations in brain tumor classification. PGD generates adversarial images by iteratively updating the input while keeping the perturbation within a bounded region ϵ. Mathematically, the adversarial image at iteration $t+1$ is given in Eq. (19)

$$X_{adv}^{t+1} = Proj_{X+S}\left(X_{adv}^{t} + \alpha\,.\, sign\left(\nabla_X \mathcal{L}(\theta, X_{adv}^{t}, y)\right)\right) \tag{19}$$

Where y is the true label, θ are the model parameters, X_{adv}^{t} is the adversarial image at iteration t,, $\nabla_X \mathcal{L}(\theta, X_{adv}^{t}, y)$ is the gradient of the loss relative to the current adversarial input, $Proj_{X+S}(.)$ Ensures the perturbed image stays within the allowed perturbation range S around the original input X , X_{adv}^{t+1} is the updated adversarial image at iteration $t+1$ and α is the step size controlling perturbation per iteration.

3.5.3 Universal Adversarial Perturbations (UAPs)

UAPs aim to find a single perturbation that can fool the model on most inputs, thereby testing the model's vulnerability to general attacks. In the proposed model, defending against UAPs ensures robustness against broad and general adversarial threats, not just image-specific attacks. Mathematically, the goal is to find a perturbation v such that Eq. (20)

$$min\|v\|_p \; s.t. P_{X\sim D}\left(f_\theta(X+v) \neq f_\theta(X)\right) \geq 1-\delta, \|v\|_p \leq \epsilon \tag{20}$$

Where v is the adversarial perturbation, $min\|v\|_p$ is its p-norm magnitude, X is an input from the Dataset D, f_θ is the model with parameters θ, $P_{X\sim D}(.)$ is the probability over the dataset, $1-\delta$ specifies the required success rate of the attack, and ϵ bounds the maximum allowed perturbation to keep it imperceptible. The ISSA-Att-CapsNet model is evaluated under FGSM, PGD, and UAP attacks to ensure that it maintains accurate brain tumor classification against single-step, multi-step, and universal adversarial perturbations.

4. Result

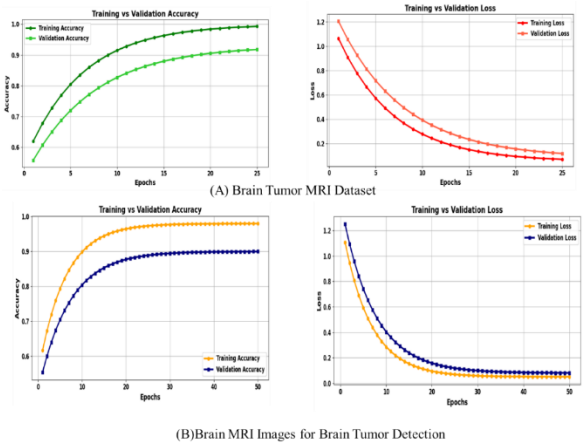
The BTD and BITD datasets' MRI images were used to train and assess the ISSA, AAT-CapsNet, and ISSA-Att-CapsNet models (70% training, 30% testing) using Python-TensorFlow/Keras on a CUDA-enabled GPU. Key steps such as preprocessing, VGG16 feature extraction, Att-CapsNet classification, and ISSA optimization were tested, showing superior performance and robustness compared to baseline models in both standard classification and adversarial attack scenarios. Table 1 summarizes the experimental setup that was implemented.

Table 1. Experimental Setup and System Configuration

Component	Configuration
Programming Language	Python 3.9
Framework	TensorFlow / Keras
Libraries	OpenCV, NumPy, Scikit-learn
Operating System	Windows 11 (64-bit)
Processor	Intel Core i7 / i9
RAM	16–32 GB
GPU	NVIDIA RTX 30 Series

Environment	Anaconda / Jupyter Notebook

The suggested model of classifying brain tumors was trained and tested effectively with the help of a Python-based deep learning model and was used on a high-performance computer with the support of a graphics card. Figure 7 depicts both training loss and success curves of the (A) BTD and (B) BITD, demonstrating that both hybrid models have converged and are functioning well.

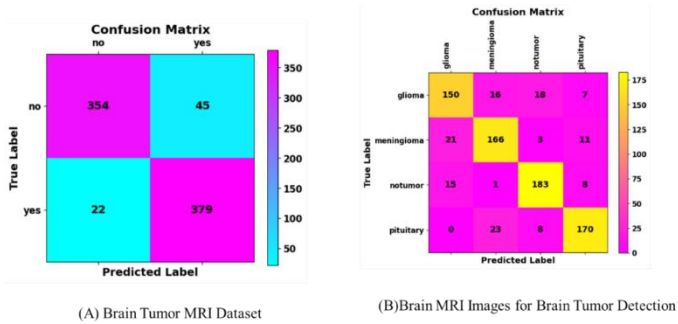


(A) Brain Tumor MRI Dataset

(B)Brain MRI Images for Brain Tumor Detection

Figure 7. Reliability and Losses Rates for (A) BTD and (B) BITD

The suggested a combination was tested on both datasets, displaying continuous training convergence and properly categorizing tumour types, including glioma, meningioma, pituitary, and no-tumor, indicating high generalization across both datasets. Figure 8 also shows the confusion matrices of the developed hybrid model, and this indicates the classification performance of the (A) BTD and (B) BITD.



(A) Brain Tumor MRI Dataset

(B)Brain MRI Images for Brain Tumor Detection

Figure 8. Confusion matrices showing classification performance on the (A) BTD and (B) BITD

The ISSA-Att-CapsNet confusion matrices show strong performance: for binary classification, 354 “no” and 379 “yes” were correct with 45 false positives and 22 false negatives; for multi-class, correct predictions were 150 glioma, 166 meningioma, 183 no tumor, and 170 pituitary, with minimal misclassifications (21 meningioma as glioma, 23 pituitary as meningioma), demonstrating reliable tumor type differentiation.

4.1 Evaluation Metrics

The hybrid ISSA-Att-CapsNet model accurately classified brain tumour pictures with high precision, F1-score, accuracy, specificity, recall, and steady learning, establishing baselines in efficacy and resilience versus adversarial attacks.

Accuracy: The proportion of the correct number of brain MRI images that were correctly classified among all the predictions. It gives an overall measure of the model in terms of differentiating between normal and abnormal images, which is resistant to adversarial attacks (Eq. 21).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (21)$$

Where FP = False Positives, TP =True Positives, TN = True Negatives and FN =False Negatives.

Precision: Measures the proportion of real good projected data (tumor pictures) among all projected positive samples. False positives are reduced when it is more exact (Eq. 22).

$$Precision = \frac{TP}{TP+FP} \times 100 \quad (22)$$

Recall: The recall is the fraction of accurately anticipated positive samples among positive samples. There is a high recall rate, which suggests that the majority of tumor cases are correctly detected (Eq. 23).

$$Recall = \frac{TP}{TP+FN} \times 100 \quad (23)$$

F1-Score: The harmonic mean of accuracy and recall, or F1, shows the balance of false positives and false negatives. Such a statistic is critical for medical imaging reliability (Eq. 24).

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (24)$$

Specificity: Of all the actual negative samples, what proportion of them are correctly identified (negative brain image) rather than false alarms (the last and most common type of error)? It indicates the capability of this model to prevent false alarms (Eq. 25).

$$\text{Specificity} = \frac{TN}{TN+FP} \times 100 \quad (25)$$

Testing Time (s): The mean time of categorization of one MRI image. The reduced time of testing implies an increased efficacy, which is a necessity for use in clinical practice.

4.2 Comparative Analysis

The suggested hybrid model achieved excellent performance, robustness, and reliable feature representation in both standard and adversarial brain tumor classification on the BTD and BITD datasets, with consistent improvement across all metrics due to effective feature extraction, attention, and optimization.

4.2.1 Performance on Standard Brain MRI Classification

The ISSA-Att-CapsNet model was tested on normal images of tests in both BTD and BITD. Att-CapsNet with ISSA optimization retrieved discriminating features with an excellent F1-score, precision, accuracy, and recall that demonstrated consistent and trustworthy brain tumor image categorization. Table 2 summarizes the hybrid ISSA-Att-CapsNet performance, showing improvements in all evaluation metrics on the Brain Tumor MRI datasets.

Table 2. Performance Comparison of ISSA-Att-CapsNet for Brain Tumor Image categorization

Methods	Datasets	Accuracy	Precision	Recall	F1-Score	Specificity	Testing Time(s)
RG-FCM With HHOCNN [20]	BTD	98%	97.9%	98.7%	98.8%	98.3%	-
CNN Models [21]	BITD	97.84%	97.85%	97.85%	97.90%	-	0.83
Att-CapsNet	BTD and BITD	98%	98%	97%	97.5%	98.2%	0.70
ISSA		98.2%	98.1%	97.5%	97.7%	98.4%	0.60
ISSA-Att-CapsNet [Proposed]		98.5%	98.3%	98.9%	99%	98.9%	0.55

RG-FCM with HHOCNN reached 98% accuracy on BTD, while CNN models (VGG16, ResNet50, InceptionV3, MobileNetV2, VGG19) achieved up to 97.84% on BITD. The proposed Att-CapsNet with ISSA optimization improved performance on both datasets, with the hybrid ISSA-Att-CapsNet achieving the best results: 98.9% specificity, 98.9% recall, 98.3% precision, 99% F1-score, 98.5% accuracy and 0.55 s testing time. The comparative performance of the proposed ISSA-Att-CapsNet for brain tumor image classification is illustrated in Figure 9.

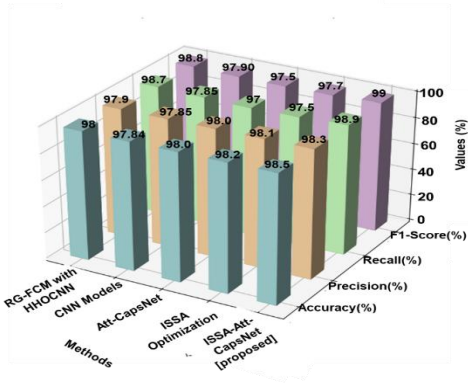


Figure 9. Image Classification Performance Comparison of ISSA-Att-CapsNet with Existing Models

The proposed ISSA-Att-CapsNet outperformed RG-FCM with HHOCNN, traditional CNNs, Att-CapsNet, and ISSA optimization, achieving , 98.9% recall, 98.3% precisio, 99% F1-score and 98.5% accuracy demonstrating enhanced robustness and precise brain tumor classification.

4.2.2 Performance under Adversarial Attacks

Hybrid ISSA-Att-CapsNet was tested using adversarial perturbed brain MRI images using FGSM, PGD, and UAP attacks. The model was robust and reliable, as it was able to withstand the perturbations and still made correct classification in very difficult adversarial settings by combining Att-CapsNet and ISSA Optimization. Table 3 shows performance under adversarial attacks on the BTD.

Table 3. Performance metrics under adversarial attacks on BTD using Att-CapsNet, ISSA, and ISSA-Att-CapsNet

Attack Type	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
FGSM	Att-CapsNet	94.2	94.0	94.3	94.1
	ISSA	96.0	95.8	96.2	96.0
	ISSA-Att-CapsNet	97.8	97.7	97.9	97.8
PGD	Att-CapsNet	93.8	93.6	94.0	93.8
	ISSA	95.4	95.2	95.5	95.3
	ISSA-Att-CapsNet	97.2	97.1	97.3	97.2
UAP	Att-CapsNet	92.7	92.5	92.9	92.7
	ISSA	95.0	94.9	95.2	95.0
	ISSA-Att-CapsNet	96.5	96.3	96.6	96.5

The hybrid ISSA-Att-CapsNet showed strong robustness on the BTD under adversarial attacks, achieving FGSM: 97.8% accuracy, PGD: 97.2%, and UAP: 96.5%, with consistent precision, recall, and F1-scores, demonstrating stable classification via Att-CapsNet and enhanced feature weighting through ISSA optimization. The

performance of Att-CapsNet, ISSA Optimization, and ISSA-Att-CapsNet under FGSM, PGD, and UAP attacks on the BTD is shown in Figure 10.

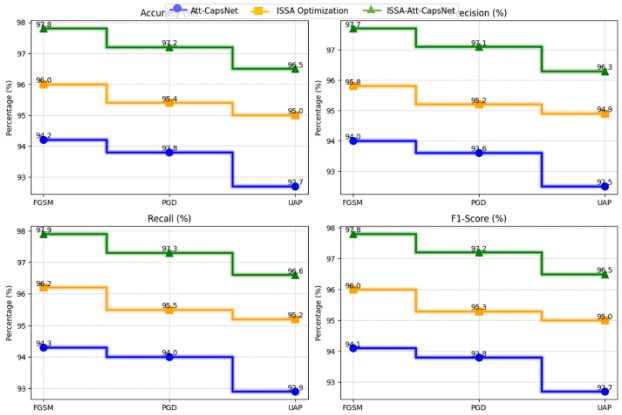


Figure 10. Robustness Comparison under Adversarial Attacks on BTD

On the BTD, the suggested ISSA -Att -CapsNet achieves the best results in terms of all the metrics. It has 97.8% accuracy, 97.7% precision, 97.9% recall, and 97.8% F1-score with FGSM, 97.2% accuracy, 97.1% precision, 97.3% recall, and 97.2% F1-score with PGD, and 96.3%, 96.2%, 96.6%, and 96.5% with UAP, and it consistently. Table 4 shows performance under adversarial attacks on the BITD.

Table 4. Performance metrics under adversarial attacks on BITD using Att-CapsNet**,** ISSA, and ISSA-Att-CapsNet

Attack Type	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
FGSM	Att-CapsNet	93.9	93.7	94.1	93.9
	ISSA Optimization	95.7	95.5	95.8	95.6
	ISSA-Att-CapsNet	97.5	97.3	97.6	97.4
PGD	Att-CapsNet	93.5	93.3	93.8	93.5
	ISSA Optimization	95.2	95.0	95.4	95.2
	ISSA-Att-CapsNet	97.0	96.8	97.2	97.0
UAP	Att-CapsNet	92.4	92.2	92.6	92.4
	ISSA Optimization	94.8	94.6	95.0	94.8
	ISSA-Att-CapsNet	96.2	96.0	96.4	96.2

On the BITD, ISSA-Att-CapsNet outperformed Att-CapsNet and ISSA alone under adversarial attacks, achieving FGSM: 97.5% accuracy, PGD: 97.0%, and UAP: 96.2%, with consistently high precision, recall, and F1-scores, demonstrating robust and reliable classification. Adversarial robustness of Att-CapsNet, ISSA Optimization, and ISSA-Att-CapsNet on BITD is illustrated in Figure 11**.**

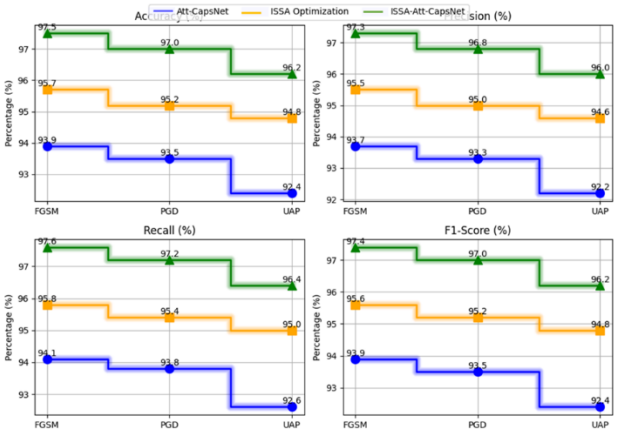


Figure 11. Adversarial Attack Performance of ISSA-Att-CapsNet and Baseline Models on BITD

On the BITD, the hybrid ISSA-Att-CapsNet shows the strongest resistance to adversarial attacks. It achieves 97.5% Accuracy, 97.3% Recall,97.4% F1-Score and Precision, 97.6%. under FGSM, 97.0%, 96.8%, 97.2%, and 97.0% under PGD, and 96.2%, 96.0%, 96.4%, and 96.2% under UAP, consistently outperforming Att-CapsNet and ISSA Optimization and demonstrating robust classification under all attack scenarios.

4.3 Discussion

The suggested method intends to boost its precision of identifying tumors from MRI images by employing better deep learning algorithms. This highlights the need for robust medical image classification, as standard CNNs show limited adversarial resilience, generalization, and advanced feature representation [20]. The method mainly uses standard CNNs, evaluates only normal conditions, and lacks analysis of robustness, generalization, and computational efficiency, limiting its reliability in diverse and adversarial scenarios [21]. The proposed ISSA-Att-CapsNet enhances the feature representation and the classification accuracy as it involves the combination of optimization and attention-based capsule networks. Testing the model on BTD and BITD datasets improves its resilience against adversarial attacks as well as its generalization. Moreover, optimized features weighting is a coding of complex spatial relationships in MRI images, which results in higher and more consistent brain tumour classification in both normal and adversarial case. The ISSA -Att-CapsNet model is highly accurate and robust but can face unexpected adversarial attacks and low computational capacities, future research needs to focus on efficiency, generalization, segmentation, multimodal imaging incorporation, and investigate lightweight architectures to implement it in the clinic more quickly.

5. Conclusion

The suggested ISSA-Att-CapsNet model demonstrates the high efficiency of the brain tumour detection and classification with the help of MRI images of both BTD and BITD. With the combination of attention-based capsule networks and ISSA optimisation, the model is effective in extracting features of complex objects and at the same time achieving high accuracy even in adversarial cases. The accuracy, precision, recall and F1-score of both datasets are displayed as 98.5, 98.3, 98.9, and 98.9 respectively and the specificity is 99.0. The model was very robust under FGSM, PGD and UAP attacks with an accuracy of between 97.8 96.5% on BTD and 97.5 96.2 on BITD respectively as the other metrics showed similar stability and thus validity of the model and clinical use..

References

1. Tsai, M.J., Lin, P.Y. and Lee, M.E., 2023. Adversarial attacks on medical image classification. Cancers, 15(17), p.4228. https://doi.org/10.3390/cancers15174228
2. Yinusa, A. and Faezipour, M., 2025. A multi-layered defense against adversarial attacks in brain tumor classification using ensemble adversarial training and feature squeezing. Scientific Reports, 15(1), p.16804.
3. Mathivanan, S.K., Srinivasan, S., Koti, M.S., Kushwah, V.S., Joseph, R.B., and Shah, M.A., 2025. A secure hybrid deep learning framework for brain tumor detection and classification. Journal of Big Data, 12(1), p.72.
4. Jeong, S.W., Cho, H.H., Lee, S., and Park, H., 2022. Robust multimodal fusion network using adversarial learning for brain tumor grading. Computer Methods and Programs in Biomedicine, 226, p.107165. https://doi.org/10.1016/j.cmpb.2022.107165

5. Ahmad, B., Sun, J., You, Q., Palade, V., and Mao, Z., 2022. Brain tumor classification using a combination of variational autoencoders and generative adversarial networks. Biomedicines, 10(2), p.223. https://doi.org/10.3390/biomedicines10020223
6. Vasan, D. and Hammoudeh, M., 2024. Enhancing resilience against adversarial attacks in medical imaging using advanced feature transformation training. Current Opinion in Biomedical Engineering, 32, p.100561. https://doi.org/10.1016/j.cobme.2024.100561
7. Maliamanis, T.V., Apostolidis, K.D., and Papakostas, G.A., 2022. How resilient are deep learning models in medical image analysis? The case of the moment-based adversarial attack (Mb-AdA). Biomedicines, 10(10), p.2545. https://doi.org/10.3390/biomedicines10102545
8. Jiang, S., Wu, Z., Yang, H., Xiang, K., Ding, W. and Chen, Z.S., 2024. A prior knowledge-guided distributionally robust optimization-based adversarial training strategy for medical image classification. Information Sciences, 673, p.120705. https://doi.org/10.1016/j.ins.2024.120705
9. Apostolidis, K.D. and Papakostas, G.A., 2022. Digital watermarking as an adversarial attack on medical image analysis with deep learning. Journal of Imaging, 8(6), p.155. https://doi.org/10.3390/jimaging8060155
10. Veeramuthu, A., Meenakshi, S., Mathivanan, G., Kotecha, K., Saini, J.R., Vijayakumar, V., and Subramaniyaswamy, V., 2022. MRI brain tumor image classification using a combined feature and image-based classifier. Frontiers in Psychology, 13, p.848784. https://doi.org/10.3389/fpsyg.2022.848784
11. Kumar, K.A., Prasad, A.Y., and Metan, J., 2022. A hybrid deep CNN-Cov-19-Res-Net Transfer learning architecture for an enhanced Brain tumor Detection and Classification scheme in medical image processing. Biomedical Signal Processing and Control, 76, p.103631. https://doi.org/10.1016/j.bspc.2022.103631
12. Srinivas, C., KS, N.P., Zakariah, M., Alothaibi, Y.A., Shaukat, K., Partibane, B., and Awal, H., 2022. Deep transfer learning approaches in performance analysis of brain tumor classification using MRI images. Journal of Healthcare Engineering, 2022(1), p.3264367. https://doi.org/10.1155/2022/3264367Digital Object Identifier (DOI)
13. Vankdothu, R. and Hameed, M.A., 2022. Brain tumor MRI images identification and classification based on the recurrent convolutional neural network. Measurement: Sensors, 24, p.100412. https://doi.org/10.1016/j.measen.2022.100412
14. Özkaraca, O., Bağrıaçık, O.İ., Gürüler, H., Khan, F., Hussain, J., Khan, J. and Laila, U.E., 2023. Multiple brain tumor classification with dense CNN architecture using brain MRI images. Life, 13(2), p.349. https://doi.org/10.3390/life13020349
15. Asiri, A.A., Soomro, T.A., Shah, A.A., Pogrebna, G., Irfan, M., and Alqahtani, S., 2024. Optimized brain tumor detection: a dual-module approach for MRI image enhancement and tumor classification. IEEE Access, 12, pp.42868-42887. 10.1109/ACCESS.2024.3379136
16. Alemu, B.S., Feisso, S., Mohammed, E.A., and Salau, A.O., 2023. Magnetic resonance imaging-based brain tumor image classification performance enhancement. Scientific African, 22, p.e01963. https://doi.org/10.1016/j.sciaf.2023.e01963
17. Singh, A., Singh, M.P., and Singh, A.K., 2025. Robust multi-expert deep learning framework for brain MRI classification with Taguchi optimization. Neurocomputing, 650, p.130824. https://doi.org/10.1016/j.neucom.2025.130824
18. Pal, S., Rahman, S., Beheshti, M., Habib, A., Jadidi, Z. and Karmakar, C., 2024. The impact of simultaneous adversarial attacks on the robustness of medical image analysis. IEEE Access, 12, pp.66478-66494. 10.1109/ACCESS.2024.3396566
19. Yang, Y., Shih, F.Y. and Roshan, U., 2022. Defense against adversarial attacks based on stochastic descent sign activation networks on medical images. International Journal of Pattern Recognition and Artificial Intelligence, 36(03), p.2254005. https://doi.org/10.1142/S0218001422540052
20. Kurdi, S.Z., Ali, M.H., Jaber, M.M., Saba, T., Rehman, A., and Damaševičius, R., 2023. Brain tumor classification using meta-heuristic optimized convolutional neural networks. Journal of Personalized Medicine, 13(2), p.181. https://doi.org/10.3390/jpm13020181
21. Rasheed, Z., Ma, Y.K., Ullah, I., Ghadi, Y.Y., Khan, M.Z., Khan, M.A., Abdusalomov, A., Alqahtani, F., and Shehata, A.M., 2023. Brain tumor classification from MRI using image enhancement and convolutional neural network techniques. Brain Sciences, 13(9), p.1320. https://doi.org/10.3390/brainsci13091320