Leveraging Pretrained Models for Multimodal Medical Image Interpretation: An Exhaustive Experimental Analysis

Temitayo Matthew Fagbola[1] and Igwebuike Success[2]

[1, 2] Centre of Excellence for Data Science, Artificial Intelligence and Modelling, University of Hull, Kingston Upon Hull, United Kingdom.

* Corresponding author
Email: temitayo-matthew.fagbola@hull.ac.uk (TMF)

[1, 2] The authors contributed equally to this work.

Leveraging Pretrained Models for Multimodal Medical Image Interpretation: An Exhaustive Experimental Analysis

Abstract

Artificial intelligence (AI) in radiology, particularly pretrained machine learning models, holds promise for overcoming image interpretation complexities and improving diagnostic accuracy. Although extensive research highlights their potential, challenges remain in adapting these models for generalizability across diverse medical image modalities, such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and X-rays. Most importantly, limited generalizability across image modalities hinders their real-world application in diverse medical settings. This study addresses this gap by investigating the effectiveness of pretrained models in interpreting diverse medical images. We evaluated ten state-of-the-art convolutional neural network (CNN) models, including ConvNeXtBase, EfficientNetB7, VGG architectures (VGG16, VGG19), and InceptionResNetV2, for their ability to classify multimodal medical images from brain MRI, kidney CT, and chest X-ray (CXR) scans. Our evaluation reveals VGG16's superior generalizability across diverse modalities, achieving accuracies of 96% for brain MRI, 100% for kidney CT, and 95% for CXR. Conversely, EfficientNetB7 excelled in brain MRI with 96% accuracy but showed limited generalizability to kidney CT (56% accuracy) and CXR (33% accuracy), suggesting its potential specialization for MRI tasks. Future research should enhance the generalizability of pretrained models across diverse medical image modalities. This includes exploring hybrid models, advanced training techniques, and utilizing larger, more diverse datasets. Integrating multimodal information, such as combining imaging data with patient history, can further improve diagnostic accuracy. These efforts are crucial for deploying robust AI systems in real-world medical settings, ultimately improving patient outcomes.

Keywords: Pretrained models; medical image synthesis; image modalities; image interpretation; domain adaptation; imaging diagnostics; multimodal image classification; domain generalisation, transfer learning, multi-label classification.

1. Introduction

In modern healthcare, medical imaging is an indispensable cornerstone, providing profound insights into the intricate structures and potential anomalies within the human body. The successful interpretation of medical images across varied modalities—such as CT, MRI, and X-rays—traditionally falls under the expertise of experienced radiologists. However, this interpretative process is multifaceted, marked by complexities inherent in analyzing medical images, from detecting subtle cues to providing comprehensive clinical evaluations due to increased image analysis demands (Pesapane et al., 2018; Balabanova et al., 2005). Each imaging modality has unique strengths and limitations, adding intricacies to its analysis. For instance, while CT scans offer detailed information, X-rays, especially chest X-rays (CXRs), are more cost-effective, expose patients to less radiation, and are more accessible (Power et al., 2016). These factors make CXRs particularly practical in resource-limited settings. Additionally, MRI can sometimes be a more suitable alternative for specific medical images, showcasing its potential to complement or substitute other modalities. In some cases, combining multiple imaging modalities improves accuracy and outcomes,

recognizing the limitations of relying solely on a single modality (Puderbach et al., 2007; Abhisheka et al., 2023). This convergence exemplifies the evolving complexity of medical image interpretation.

Moreover, the increasing demand for radiological investigations is juxtaposed against a diminishing number of radiologists, creating an imbalance that underscores the urgent need for innovative solutions to augment diagnostic capacities. In this context, the evolving role of artificial intelligence (AI) in healthcare has emerged as a beacon of promise (Biswas et al., 2019). AI applications, from image analysis to diagnostic assistance and predictive modelling, hold the potential to significantly enhance outcomes across diverse domains, including patient care (Ben-Israel et al., 2020). Particularly noteworthy is the concept of pre-trained models within AI. Pre-trained models represent a paradigm shift in machine learning, wherein models are initially trained on large and diverse datasets to learn generalized features. This learned knowledge is subsequently repurposed, fine-tuned, and adapted to specific tasks or domains.

In medical image analysis, leveraging these pre-trained models offers a spectrum of potential benefits (Litjens et al., 2017). These models have showcased remarkable performance across various medical image modalities, excelling in both binary and multiclass classification tasks. Models like ResNet, VGG, and DenseNet, initially trained on large-scale natural image datasets, have demonstrated transferability and robustness in classifying medical images. Their ability to extract hierarchical features underscores their adaptability and effectiveness. Notably, these models have achieved high accuracies, sensitivities, and specificities in diagnosing conditions like tumors, fractures, and abnormalities, thus proving their efficacy in diverse medical imaging scenarios. However, continual validation across varied datasets and clinical scenarios is pivotal to ensure consistent and reliable performance across different modalities and classification tasks. The widespread use and effectiveness of pre-trained models in analyzing medical images have sparked interest in understanding how well these models can adapt to different scenarios. In response to this interest, we have developed an evaluation framework designed to assess the performance and adaptability of pre-trained models—including some proposed in previous research studies—particularly in the context of interpreting medical images.

The contribution of this study is as follows:

i. We evaluate the diagnostic accuracy of ten pretrained models across diverse medical imaging modalities, using three datasets containing MRI scans, CT scans, and X-rays. This assessment provides insights into the models' ability to generalize and adapt to different types of medical images.
ii. We explore two distinct classification scenarios to enhance the effectiveness of multi-label diagnosis in medical images. First, we implement a four-category classification approach for two datasets, each corresponding to a different modality. Second, we employ a three-category classification approach, resulting in more accurate and efficient diagnoses in medical settings.
iii. We evaluate the robustness and reliability of the pretrained models, ensuring consistent performance across diverse datasets and imaging modalities. This consistency is crucial for dependable clinical decision-making.
iv. We identify and address biases present in pretrained models, ensuring unbiased and equitable diagnoses across multiple imaging types. This contributes to fairness in medical imaging.

The rest of this paper is structurally organized as follows: Section 2 presents the relevant background and related works; Section 3 elaborates on the materials and methods employed, encompassing aspects such as data acquisition, preprocessing, algorithm selection, and evaluation metrics. Section 4 presents the outcomes of the study, unveiling the results obtained from the proposed framework. Following this, Section

5 engages in discussion, analysing and contextualizing the findings within the broader scope of research. Finally, Section 6 offers a conclusion, summarizing the key insights gleaned from the study.

2. Related Works

Over the years, researchers have shown an increased interest in leveraging machine learning algorithms for medical imaging analysis, particularly in diagnosing and assessing treatment responses. In this context, a notable study Abirami A. (2023) introduces EfficientNetB7 as a pre-trained deep learning model aimed at enhancing accuracy while reducing complexity. Utilizing a dataset comprising 3674 MRI images, the research focuses on evaluating the model's efficacy in classifying tumors as glioma, meningioma, pituitary tumors, or non-tumorous. Leveraging transfer learning, the strategy implemented with EfficientNetB7 yielded a remarkable 98.4% accuracy. However, it's important to note that this high accuracy was achieved without evaluation against other datasets or modalities, raising questions about the model's generalizability beyond the specific dataset used in the study. The study by Abdelaziz Ismael et al. (2020a) introduces an improved deep learning model designed for brain tumor classification using MRI images. The model was tested on a benchmarked dataset comprising 3064 MRI images across three tumor types. It achieved a remarkable accuracy of 99%, surpassing previous benchmarks on the same dataset.

Ramadhan & Baykara (2022) introduced an updated VGG16-CNN model of reduced parameters from approximately 138 million parameters to around 40 million parameters for multiple classifications of COVID-19, two of which classification comprised of three classes: COVID-19, normal, and pneumonia, and binary classification of COVID-19 and normal class. The research utilized three datasets: Db1 containing 21,165 images, Db2 with 5,226 images, and Db3 comprising 6,432 images. Results were promising, with high accuracy achieved across the datasets. Specifically, the model attained 99% accuracy for triple classification and 100% for binary classification on Db3, 98% and 99% on Db2, and 96% and 92% on Db1 for triple and binary classifications, respectively. In another study, a novel model named DarkCovidNet Ozturk et al. (2020) is introduced for the automatic detection of COVID-19 using raw chest X-ray images. The model is specifically designed for accurate diagnostics in both binary classification (COVID vs. No-Findings) and multi-class classification (COVID vs. No-Findings vs. Pneumonia) scenarios. The development phase utilized a dataset comprising 1125 images, categorized into 125 COVID-19 positive, 500 Pneumonia, and 500 No-Findings cases.

Remarkably, the model achieved an accuracy of 98.08% for binary classification and 87.02% for multi-class classification. The DarkNet model employed in the study served as a classifier for the you only look once (YOLO) real-time object detection system. Limitations of these studies, such as the relatively small dataset and the need to expand their research by exploring larger datasets and other medical image types, are acknowledged, they also address the imperative of evaluating competitive models across varied medical images, highlighting their focus on ensuring the model's robustness and accuracy—a pivotal aspect that aligns with the overarching aim and contribution of this study. Leveraging pretrained models may herald advancements in diagnostic accuracy and efficiency, hinting at a transformative paradigm for medical imaging interpretation and augmenting clinical decision-making processes. Table 1 showcases the instrumental role of pretrained models in advancing technological capabilities, as demonstrated by several related studies across diverse domains.

Table 1. Related Works in the application of pretrained model for medical image classification.

Methodology	Number of Dataset & Modality	Classification	Model
Narin et al. (2021)	CXR datasets (3)	Multi – 3 classes	ResNet50
Kassania et al. (2021)	CXR & CT images	Binary	DenseNet121 + Bagging Tree classifier
Monowar et al. (2020)	CXR	Multi – 3 classes	Xception
Xue et al. (2023)	CXR & CT images	Multi – 3 classes	Ensemble: VGG16, DenseNet, ResNet50, ResNet102
Alshmrani et al. (2023)	CXR	Multi – 6 classes	VGG19 + CNN
Alshmrani et al. (2023)	CXR	Binary & Multi – 3 classes	Modified ResNet50
Raza et al. (2023)	CT images	Multi – 3 classes	Lung-EffNet
Humayun et al. (2022)	CT images	Multi – 3 classes	VGG16, VGG19, & Xception
Zhou et al. (2023)	CT images	Multi – 3 classes	Ensemble
Ibrahim et al. (2023)	CT images (2)	Binary	COV-CAF
Jangam et al. (2022)	CXR (2) & CT images (3)	Binary	Stacked Ensemble
Abdelaziz Ismael et al. (2020)	MRI	Multi – 3 classes	Residual Networks
Nayak et al. (2020)	MRI datasets (2)	Multi – 5 classes	Custom CNN

3. Materials and Method

A systematic selection process identified a cohort of leading pretrained models alongside benchmarked architectures for comparative evaluation. These models encompass a spectrum of ten Convolutional Neural Network (CNN) architectures, including DenseNet201, NasNetMobile, NasNetLarge, Xception, ResNet50, EfficientNetB7, VGG16, VGG19, InceptionResNetV2, ConvNeXtBase and four benchmarked models namely; VGG19+CNN, Modified ResNet50, Custom CNN and Ensemble stacked models. The adaptability of these fourteen models across diverse medical imaging modalities was rigorously assessed for their capacity to interpret CT scans, MRIs, and X-rays effectively. The evaluation framework employed a diverse set of metrics, gauging model performance across specific medical imaging tasks, emphasizing diagnostic accuracy, precision, and sensitivity. In Figure 1, the workflow for the experimental analyses is presented.

3.1 Datasets

1. The Brain MRI dataset (Nickparvar, 2021), used for this study encompasses 7023 images sourced from publicly available datasets like figshare, Br35H, and SARTAJ. These images are categorized into glioma, meningioma, no tumor, and pituitary classes. The dataset was split into train and test set with the 80:20 ratio and a separate set of images was used for validation. Various researchers have utilized this dataset for medical image classification tasks. For example, Özkaraca et al. (2023) developed a new deep learning model for MR image classification, leveraging strengths of

DenseNet, VGG16, and basic CNNs. Similarly, Islam *et al*. (2023) explores deep transfer learning architectures for brain tumour diagnosis using MRI.

2. The Kidney CT Scan dataset includes 12,446 images representing tumor, cyst, normal, and stone classes. The dataset underwent preprocessing that involved identifying and removing duplicates using a robust hashing method, resulting in a refined set of 11,929 images. These images were split into ratio of 70:20:10 for training, testing, and validation. The original 512x512x3 pixel images were resized to 224x224x3 pixels to balance model accuracy and computational complexity for efficient model training.

3. The Chest X-ray Dataset for analysis consisted of samples from nine different sources, including augmented images to compensate for the limited availability of a single extensive dataset. It comprises three primary classes: covid, pneumonia, and normal cases, totaling 6,939 samples. This diverse dataset enabled a comprehensive analysis of chest X-ray instances. The dataset underwent preprocessing, transitioning into a structured dataframe. It was then divided into 80% for the training set and 20% for the testing set. The images were standardized to a 224x224x3 dimension using rescaling and resizing techniques. This resizing aimed to enhance the model's focus by reducing irrelevant information assimilation from larger images, promoting a more pertinent learning process. Table 2 illustrates the breakdown of image proportions utilized for model evaluation.

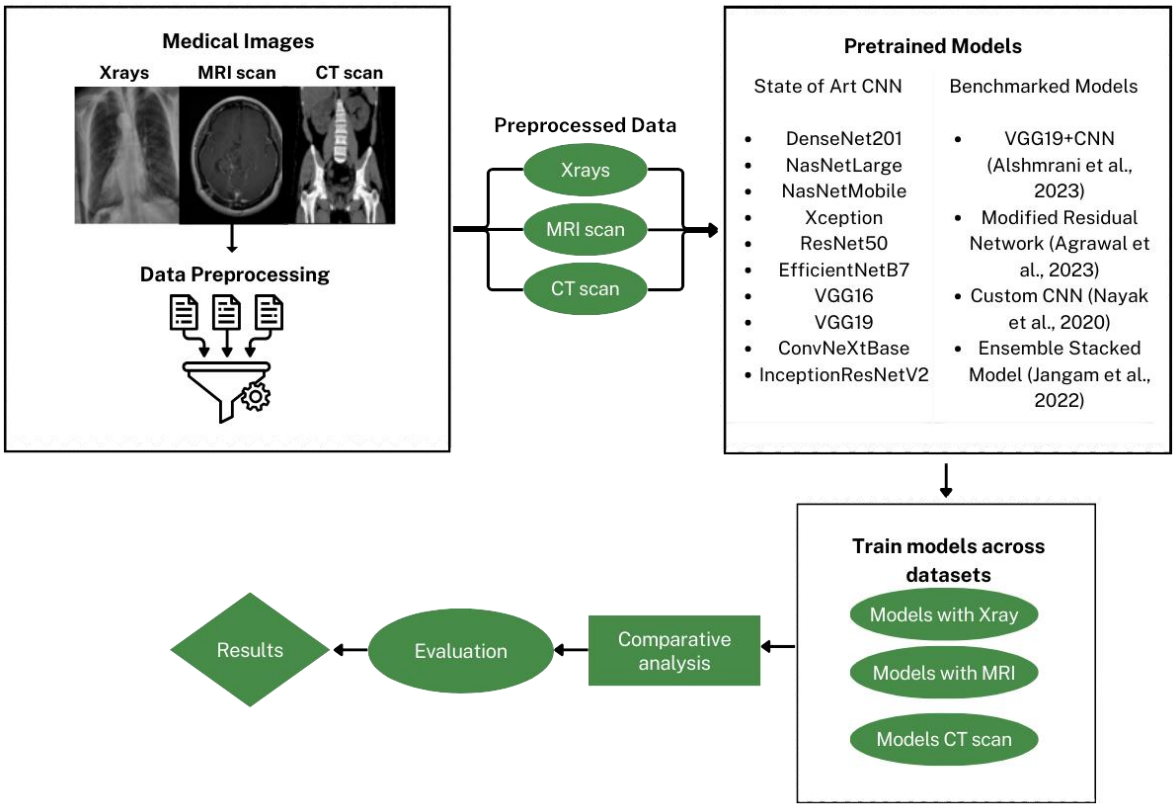
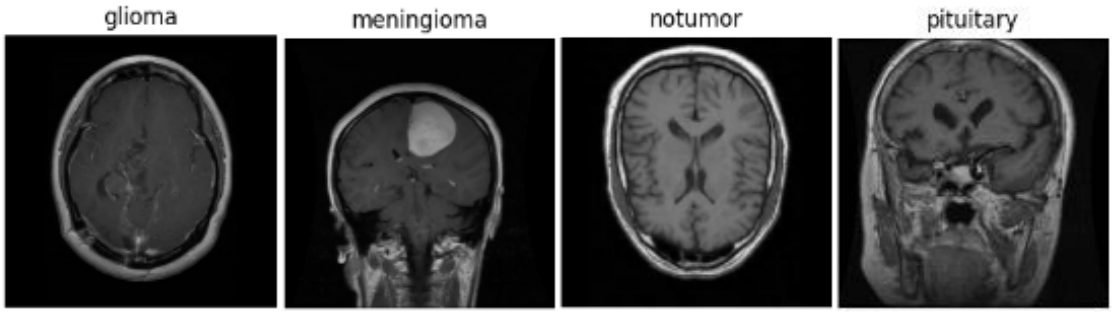


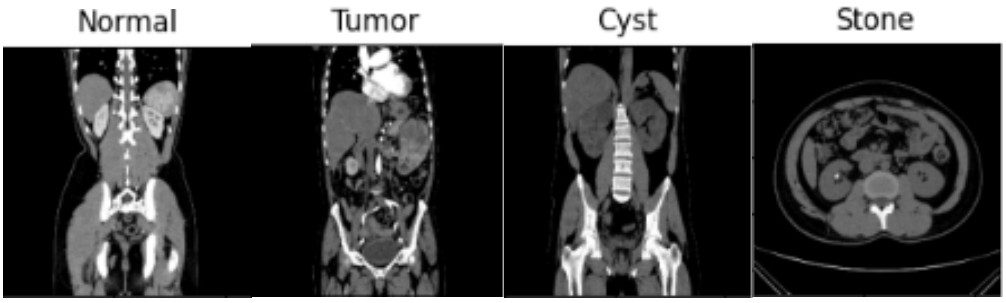
Figure 1: Overall Architecture of the Proposed Framework

Table 2. Proportional Distribution of Images for Model Evaluation

Datasets	Train set	Test set	Validation set
Brain MRI	4570	1142	1311
Kidney CT	8352	2385	1192
Chest X-ray	5518	1384	-



(a)



(b)

(c)

Figure 2: Dataset Snippet Arranged by Modalities a, b, and c -Brain MRI, Kidney CT, and CXR Images respectively

2.2 Algorithm Background

1. **DenseNet201:** DenseNet-201, (Huang et al., 2022), a specific variant of the DenseNet architecture. DenseNet-201 is renowned for its unique structure, characterized by dense connectivity patterns

among layers. In a DenseNet, each layer receives inputs from all preceding layers, promoting extensive feature reuse and propagation. This dense connectivity enhances gradient flow during training, mitigating issues like vanishing gradients and enabling more efficient learning. Due to its densely connected layers, DenseNet-201 can effectively capture intricate dependencies between features within the data.

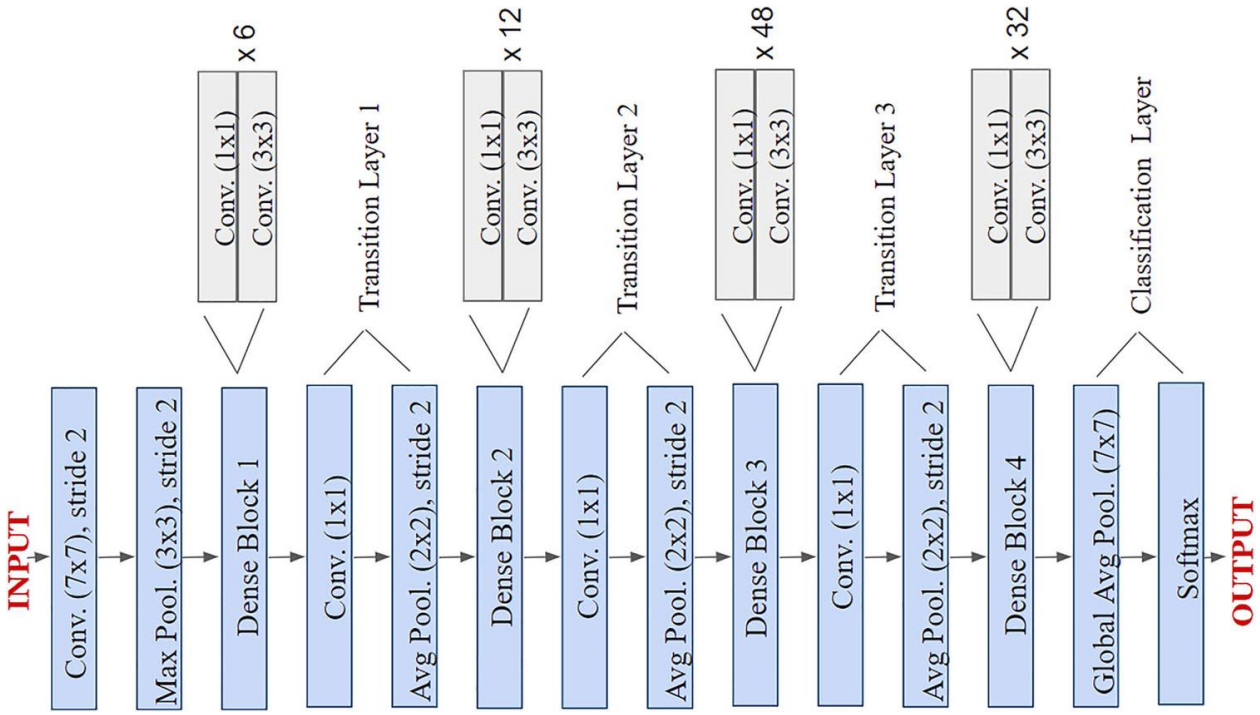


Figure 3: The Architecture of DenseNet201 (Chahar et al., 2020)**.**

2. **NasNetLarge and NasNetMobile:** NasNet (Neural Architecture Search Network), models are designed using neural architecture search techniques, resulting in architectures optimized for performance and efficiency. the creation of a novel search space, referred to as the "NASNet search space," refers to a predefined set of possible neural network architectures or architectural components that are explored during the process of neural architecture search (NAS) , (Zoph et al., 2018).

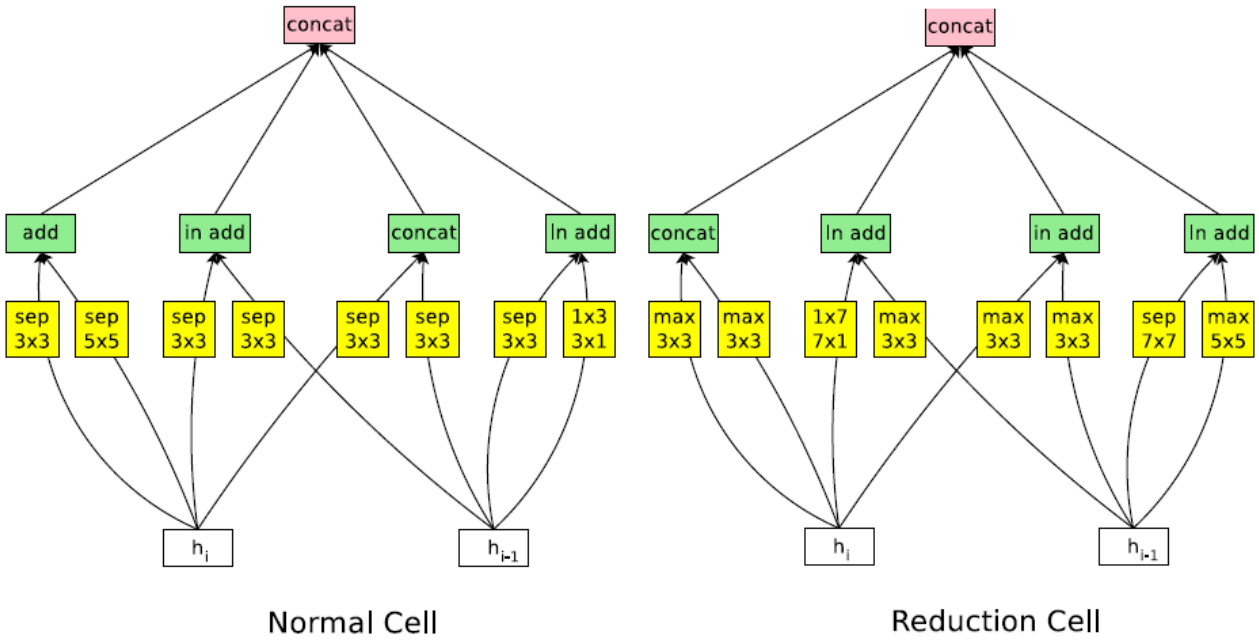


Figure 4: The Architecture of NasNet (Tsang, 2021)**.**

3. **Xception:** Xception derived from the combination of "Extreme Inception" signifies an extreme iteration of the Inception architecture, a well-known convolutional neural network design employing "Inception modules" for feature extraction. In Xception, these Inception modules are replaced with "depthwise separable convolutions," a variant of convolutions that separates spatial and channel-wise information processing. This alteration in convolutional operation distinguishes Xception from Inception. The Xception architecture is characterized by its 36 convolutional layers, serving as the foundational feature extraction component of the network (Chollet, 2017a).

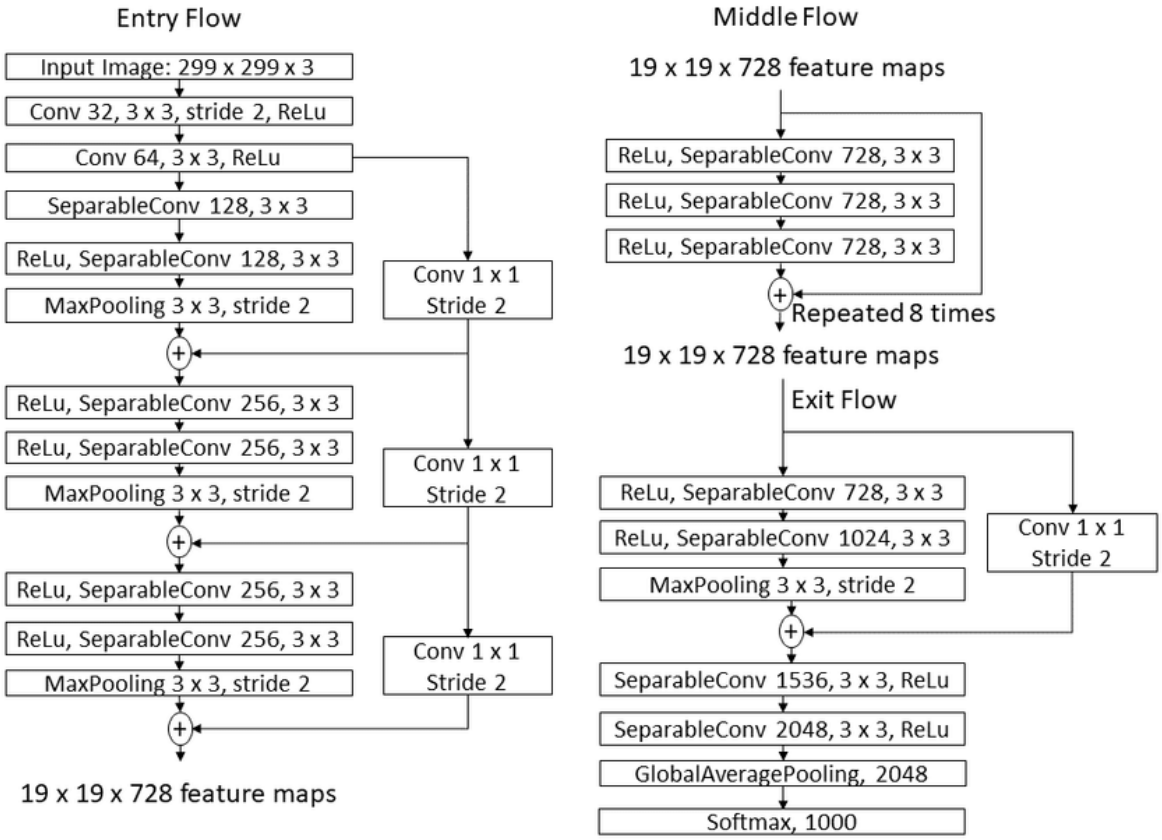


Figure 5: The Architecture of Xception (Chollet, 2017b)**.**

4. **Vgg16 & Vgg19:** VGG (Visual Geometry Group) networks are characterized by their simple and uniform architecture, comprising multiple convolutional layers followed by max-pooling and fully connected layers. The authors (Simonyan & Zisserman, 2015) evaluated convolutional neural networks (CNNs) of increasing depth, up to 16–19 weight layers, using a specific architecture with small (3x3) convolution filters, which positively impacts classification accuracy. VGG16 and VGG19 differ in depth, with VGG19 having more layers. These models are chosen for their simplicity, ease of interpretation, and strong performance.

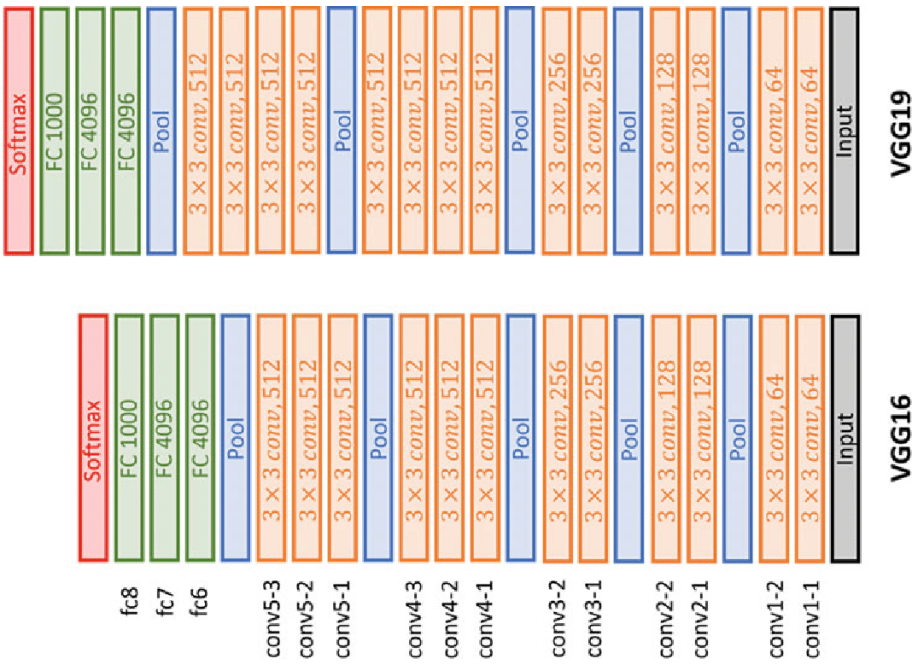


Figure 6: The Architecture of VggNet (El-Rahiem & Hammad, 2021)**.**

5. **EfficientNetB7:** EfficientNet employs a compound scaling method to balance model depth, width, and resolution, resulting in highly efficient yet powerful architectures. EfficientNetB7 is the largest variant in the EfficientNet family, offering superior performance on various computer vision tasks while maintaining high efficiency. Thanks to this compound scaling method, a mobile-sized EfficientNet model can be scaled up very effectively, achieving state-of-the-art accuracy with significantly fewer parameters (Tan & Le, 2019).

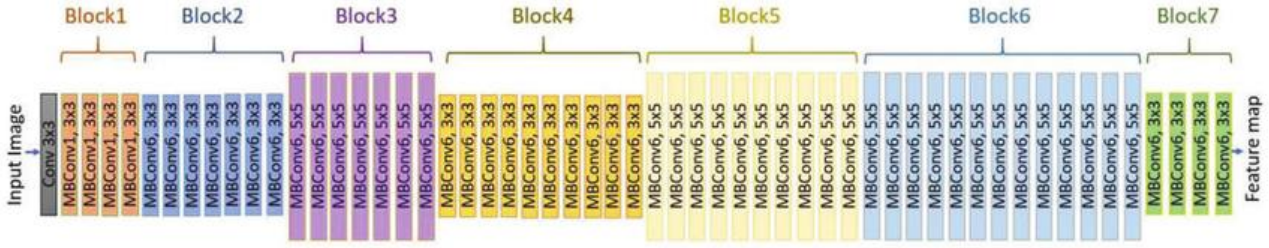


Figure 7: EfficientNetB7 Model Architecture (Tan & Le, 2019)**.**

6. **InceptionResNetV2:** InceptionResNetV2, a combination of Inception architectures with residual connections, leverages the benefits of both approaches. It is a costlier hybrid Inception version but despite the increased computational cost, the model demonstrates better accuracy or effectiveness in recognizing and classifying objects in images compared to previous versions or other architectures (Szegedy et al., 2017).

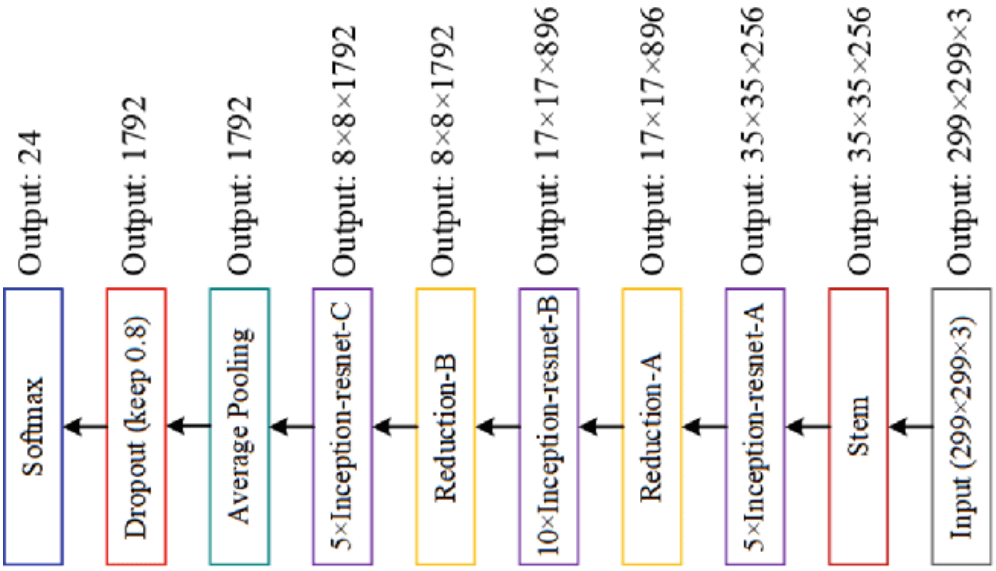


Figure 8: The Architecture of InceptionResNetV2 (Gao et al., 2020)**.**

7. **Resnet50:** ResNet (Residual Network) introduced residual connections to address the vanishing gradient problem in deep neural networks. The ResNet architecture consists of multiple "blocks" of layers, each containing two convolutional layers followed by batch normalization and a ReLU activation function. ResNet50 is a specific variant with 50 layers, each 2-layer block in the 34-layer net was replaced with this 3-layer bottleneck block, resulting in a 50-layer ResNet, striking a balance between depth and computational complexity (He et al., 2015).

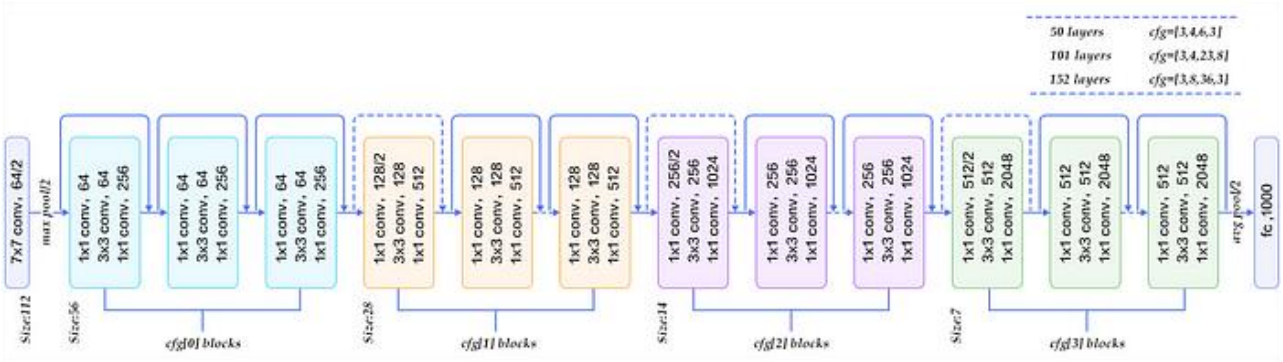


Figure 9: The Architecture of Resnet50 (Rastogi, 2022)**.**

8. **ConvNeXtBase:** ConvNeXtBase employs grouped convolutions to enhance feature extraction while reducing computational cost. ConvNeXts were constructed using standard ConvNet modules, and they perform well in comparison to Transformers across various metrics such as accuracy, scalability, and robustness on major benchmarks. Despite being based on traditional ConvNet modules, ConvNeXts are competitive with Transformers, which represent another type of neural network architecture (Liu et al., 2022).

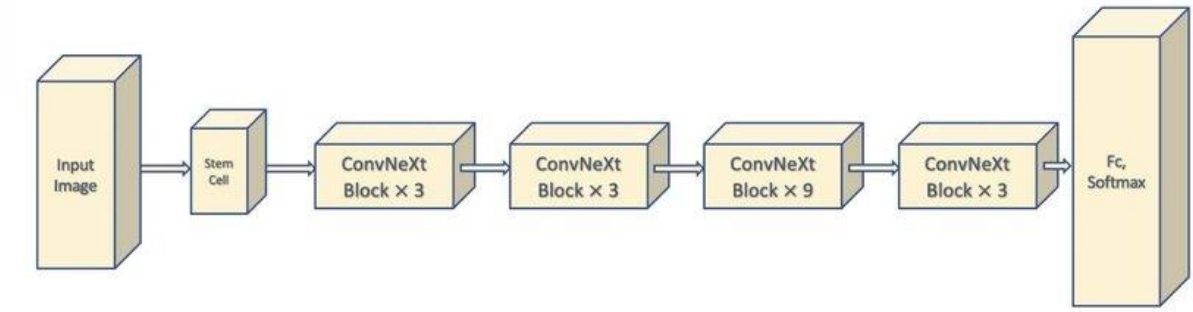


Figure 10: The CovNeXt Architecture (Yengec-Tasdemir et al., 2022)**.**

2.3 Experiments

Computational capabilities -the experiment was carried out by utilizing Google Colab's V100 GPU configuration, boasting 40GB of RAM. Additionally, the Dell machine used is equipped with a Core i9 – 12900 CPU with 32GB of RAM and powered by an NVIDIA GeForce RTX 3090 GPU.

For implementation, the state-of-the-art CNN models underwent initial training on the ImageNet dataset. Following image preprocessing, all layers of these pretrained models were frozen. To ensure uniformity and minimize bias among the models, no extra deep neural networks or dense layers were introduced, ensuring methodological consistency and fairness in model assessment, the input shape was meticulously set to match the size expected by most CNN pretrained models. For the optimization process, we opted for the Adam optimizer, a widely acclaimed choice known for its adaptive learning rate properties. This optimizer has been extensively used in various domains and is considered a standard choice in the machine learning community, also, adaptive nature of the learning rate in Adam ensures robust convergence and efficient training (Wang et al., 2022). In determining the appropriate duration for training, our selection of the number of epochs was guided by the need to witness the stabilization of model performance. By allowing sufficient epochs, we aimed to capture the convergence behaviour and ascertain the models' adaptability. In the choice of batch size, we adhered to a parameter commonly employed in the literature. This careful selection ensures a balanced trade-off between stochastic updates and computational efficiency, aligning our approach with established practices for training deep neural networks. Table 3 provides an overview of the hyperparameters employed throughout the training process of the State of the Art pretrained models.

Table 3. Hyperparameters Used for State of Art CNN Models and Respective Values.

	Input Shape	**Epochs**	**Batch Size**	**Optimizer**	**Learning Rate**	**Loss Function**	**Activation Function**
MRI	224x224	25	32	Adam	0.001	Sparse categorical crossentropy	Softmax
CT	224x224	25	32	Adam	0.001	Sparse categorical crossentropy	Softmax

CXR	224x224	25	32	Adam	0.001	Categorical crossentropy	Softmax

The hyperparameters of the benchmarked pretrained models from existing work were used for the training and evaluation process for the comparative analysis, as depicted in Table 4. This approach enables us to assess the performance of our proposed method against existing benchmarks under standardized conditions

Table 4. Hyperparameters Used for Benchmarked Models according to their specific architectures

	Input Shape	**Epochs**	**Batch Size**	**Optimizer**	**Learning Rate**	**Loss Function**	**Activation Function**
VGG19+CNN Alshmrani et al. (2023)	224x224	5000	32	Adam	0.000009	Sparse categorical crossentropy	Softmax
MODIFIED RESNET Agrawal et al. (2023)	224x224	25	32	Adam	0.001	Sparse categorical crossentropy	Softmax
CUSTOM CNN Nayak et al. (2020)	224x224	10	25	SGD	0.0001	Categorical crossentropy	Softmax
ENSEMBLE STACKED MODEL Jangam et al. (2022)	224x224	100	16	Adam	0.0001	Categorical crossentropy	Softmax

2.4 Evaluation Metrics

The metrics used for the evaluation of the models in this study are presented as follows:

TP, TN, FP, FN is True Positive, True Negative, False Positive, False Negative respectively.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + TN + FP + FN}}$$

Accuracy measures how correct the model is in predicting and classifying both the positive and negative instance in the dataset.

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FP}}$$

Precision specifically measures how accurate the model's capability to precisely identify the true positive instances among all the positive predictions it made.

$$\boldsymbol{Recall\ (Sensitivity)} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FN}}$$

Recall, also known as sensitivity, measures how many positive instances the model accurately predicts in the dataset.

$$F1\ score = 2 \times \frac{precision \times recall}{precision + recall}$$

F1 Score is the harmonic mean and balance of precision and recall metrics. It takes into account both false positives (precision) and false negatives (recall), providing a balanced assessment of the model's performance.

3.0 Results

This section presents the evaluation of selected models across diverse modalities, employing consistent training hyperparameters for reproducibility (outlined in Table 3). Accuracy, Precision, Recall, and F1-Score were computed for each class and averaged across the dataset, offering a comprehensive performance overview. Detailed in the tables below are the experiment results, comparing the reproducibility and generalizability of state-of-the-art CNN models. Additionally, our performance is discussed in comparison with benchmarked pretrained models done by other researchers, across Brain MRI, Kidney CT, and CXR Imaging Modalities.

Table 4. Classification Result of State of Art Pretrained Models with MRI Image Modality

Models	**BRAIN MRI IMAGES**			
	Acc	**Pre**	**Rec**	**F1 Score**
DenseNet201	92	92	91	91
NasNetMobile	87	86	86	86
NasNetLarge	89	89	89	89
ResNet50	**95**	**95**	**95**	**95**
EfficentNetB7	**96**	**96**	**96**	**96**
VGG19	**95**	**95**	**95**	**95**
VGG16	**96**	**96**	**95**	**96**
Xception	89	90	89	89
InceptionResNetV2	70	71	69	64
ConvNeXtBase	**96**	**96**	**95**	**95**

Table 5. Classification Result of State of Art Pretrained Models with CT Image Modality

Models	**KIDNEY CT IMAGES**			
	Acc	**Pre**	**Rec**	**F1 Score**
DenseNet201	100	100	100	100
NasNetMobile	100	100	100	100

NasNetLarge	100	100	100	100
ResNet50	100	100	100	100
EfficentNetB7	53	69	36	32
VGG19	100	100	100	100
VGG16	100	100	100	100
Xception	100	100	100	100
InceptionResNetV2	100	100	100	100
ConvNeXtBase	100	100	100	100

Table 6. Classification Result of State of Art Pretrained Models with CXR Image Modality

Models	**CXR IMAGES**			
	Acc	**Pre**	**Rec**	**F1 Score**
DenseNet201	**92**	**92**	**92**	**92**
NasNetMobile	89	90	89	89
NasNetLarge	86	87	86	86
ResNet50	83	85	83	83
EfficentNetB7	33	44	33	17
VGG19	90	92	90	90
VGG16	**95**	**95**	**95**	**95**
Xception	**92**	**92**	**92**	**92**
InceptionResNetV2	**92**	**93**	**92**	**92**
ConvNeXtBase	75	82	75	75

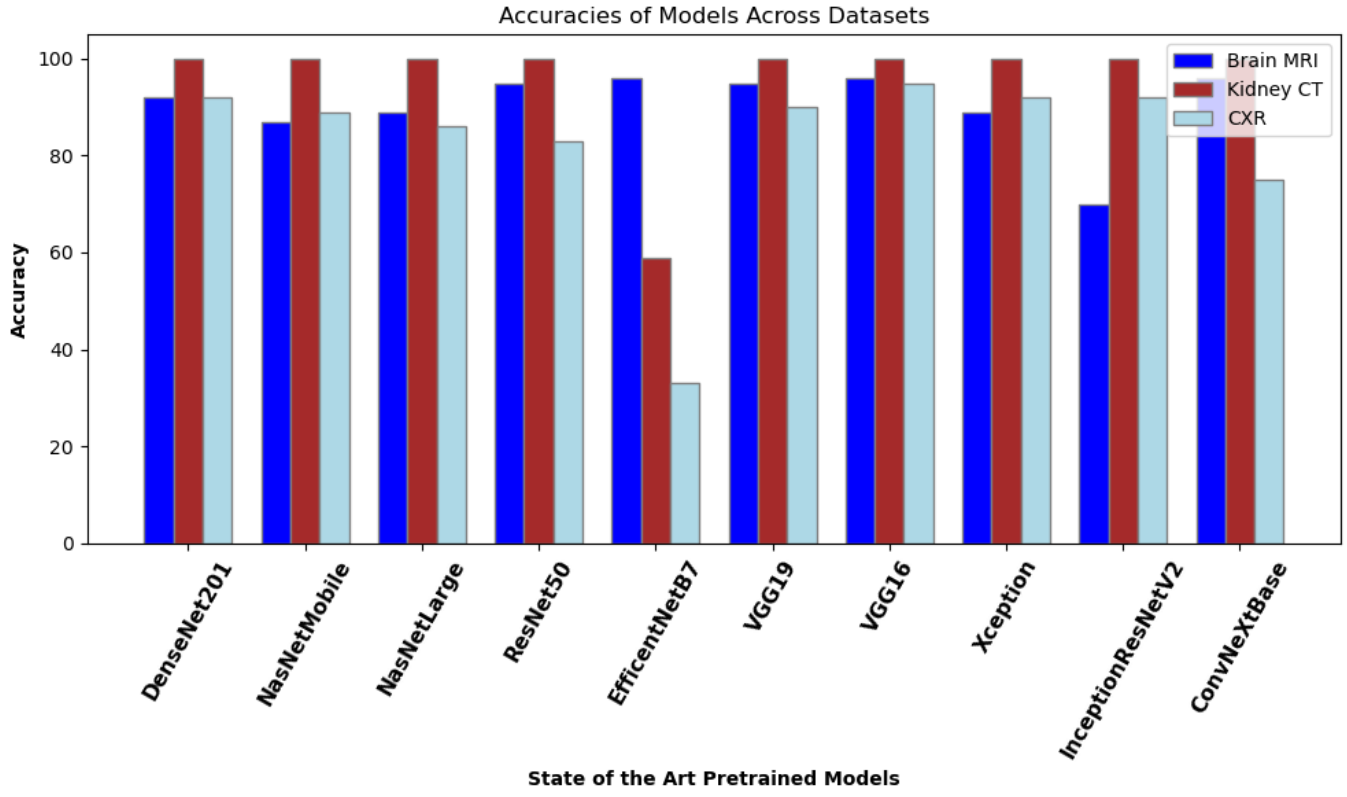
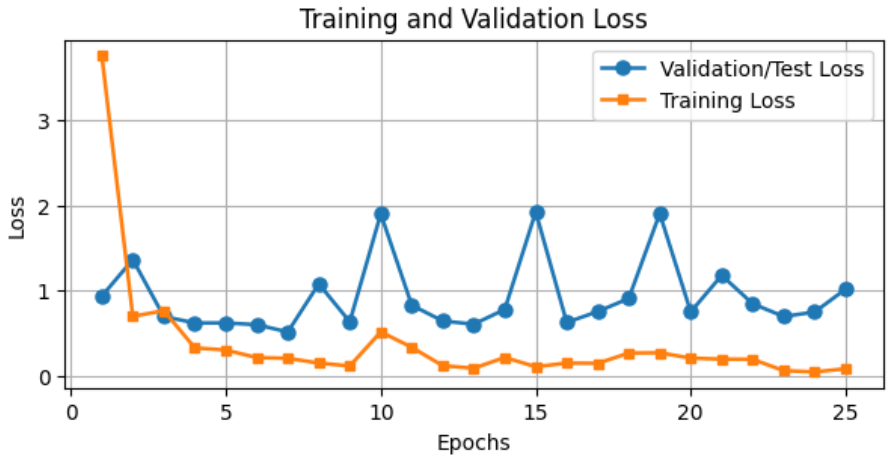


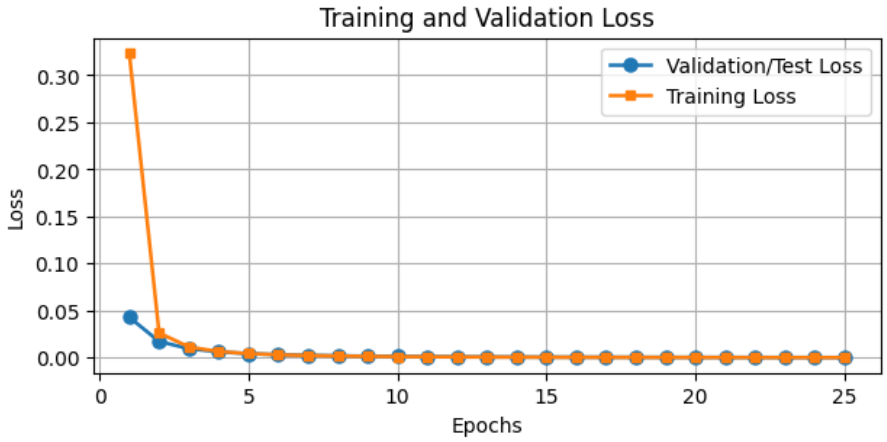
Figure 11: Evaluation Performance of State of Art Pretrained Models on the 3 Imaging Modalities Based Accuracy Metrics

In this experiment, ten state-of-the-art CNN models were utilized for classification across three distinct imaging modalities: brain MRI (categorizing glioma, meningioma, no tumor, and pituitary), kidney CT (identifying cyst, stone, normal, and tumor), and chest X-ray (distinguishing covid, pneumonia, and normal conditions). The results outlined in Tables 4, 5, and 6 present the top-performing models, highlighting their respective modalities. Among these models, VGG16 consistently demonstrated notable performance in identifying the various diseases in the brain, kidney and chest as provided in the dataset across the modalities, achieving 96%, 100%, and 95% accuracy in MRI, CT, and CXR classifications, respectively. While EfficientNetB7 exhibited superior recall in MRI (96% compared to VGG16's 95%), its performance notably declined for CXR image classification as well classifying the kidney diseases with accuracies of 33% and 53% respectively. ConvNeXtBase, InceptionResNetV2, NasNetLarge, Xception exhibited inconsistent performance. On the other hand, models like DenseNet201, ResNet50, and VGG19 consistently showed high performance in MRI, CT, and CXR, although not always ranking as the top-performing models. For instance, DenseNet201 achieved 92%, 100%, and 92%, ResNet50 reached 95%, 100%, and 83%, and VGG19 attained 95%, 100%, and 90% accuracy across the modalities, respectively. Please note that the loss curves obtained for most of the pretrained models evaluated have been included under the appendix section of this paper.

(a)



(b)



(c)

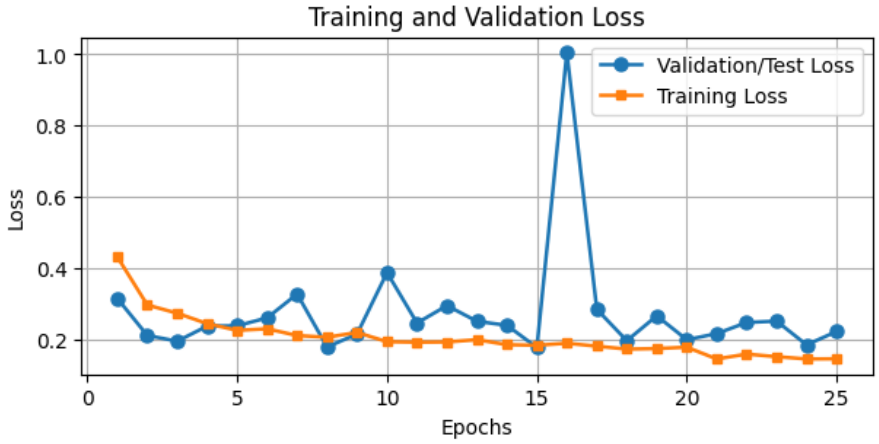
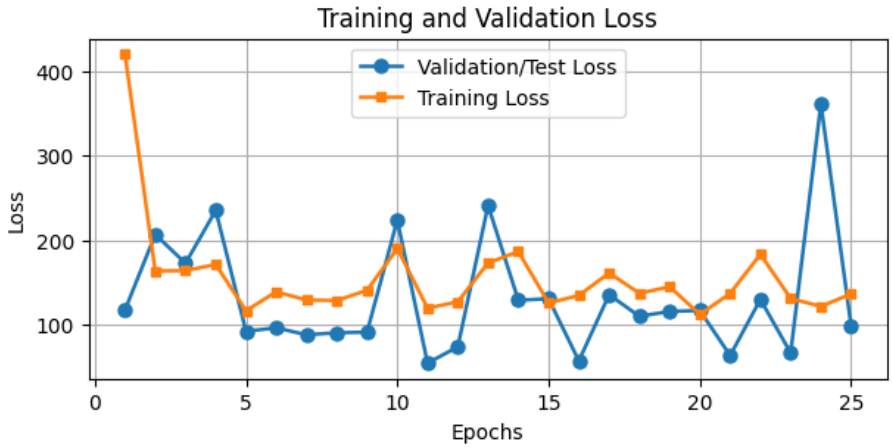
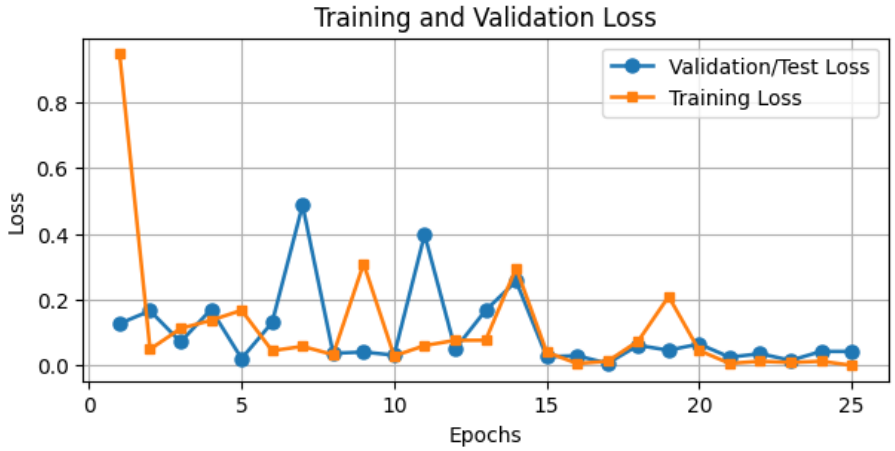


Figure 12: a, b, and c represent the Loss curve of VGG16 in order of MRI, CT scan and CXR modalities

(a)



(b)



(c)

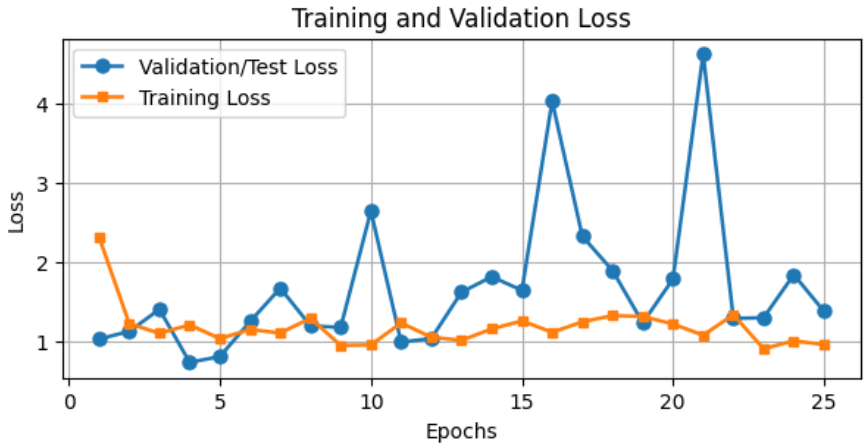


Figure 13: a, b, and c represent the Loss curve of InceptionResNetV2 in order of MRI, CT scan and CXR modalities

Table 7. Classification Result of Evaluating Benchmarked with MRI Image Modality

Method	**Models**	**Classes**	**Modality Used in article**	**Existing Accuracy**	**Evaluation Results**			
					Acc	**Pre**	**Rec**	**F1 Score**
Alshmrani et al. (2023)	VGG19+CNN	6	CXR dataset (1)	98	93	94	93	93
Agrawal et al. (2023)	Modified ResNet50	2 & 3	CXR dataset (2)	99.2, 86.1	96	96	95	96
Nayak et al. (2020)	Custom CNN	5	MRI dataset (2)	100, 97.5	92	92	91	92
Jangam et al. (2022)	Ensemble stacked models	2	CXR dataset (2) & CT dataset (3)	84, 93, 99, 99, 90	94	94	93	94

Table 8. Classification Result of Evaluating Benchmarked with CT Image Modality

Method	**Models**	**Classes**	**Modality Used in article**	**Existing Accuracy**	**Evaluation Results**			
					Acc	**Pre**	**Rec**	**F1 Score**
Alshmrani et al. (2023)	VGG19+CNN	6	CXR dataset (1)	98	100	100	100	100
Agrawal et al. (2023)	Modified ResNet50	2 & 3	CXR dataset (2)	99.2, 86.1	94	94	91	92
Nayak et al. (2020)	Custom CNN	5	MRI dataset (2)	100, 97.5	100	100	100	100

Jangam et al. (2022)	Ensemble stacked models	2	CXR dataset (2) & CT dataset (3)	84, 93, 99, 99, 90	98	97	99	98

Table 9. Classification Result of Evaluating Benchmarked with CXR Image Modality

Method	**Models**	**Classes**	**Modality Used in article**	**Existing Accuracy**	**Evaluation Results**			
					Acc	**Pre**	**Rec**	**F1 Score**
Alshmrani et al. (2023)	VGG19+CNN	6	CXR dataset (1)	98	76	84	76	75
Agrawal et al. (2023)	Modified ResNet50	2 & 3	CXR dataset (2)	99.2, 86.1	75	75	75	75
Nayak et al. (2020)	Custom CNN	5	MRI dataset (2)	100, 97.5	83	85	83	82
Jangam et al. (2022)	Ensemble stacked models	2	CXR dataset (2) & CT dataset (3)	84, 93, 99, 99, 90	94	94	94	94

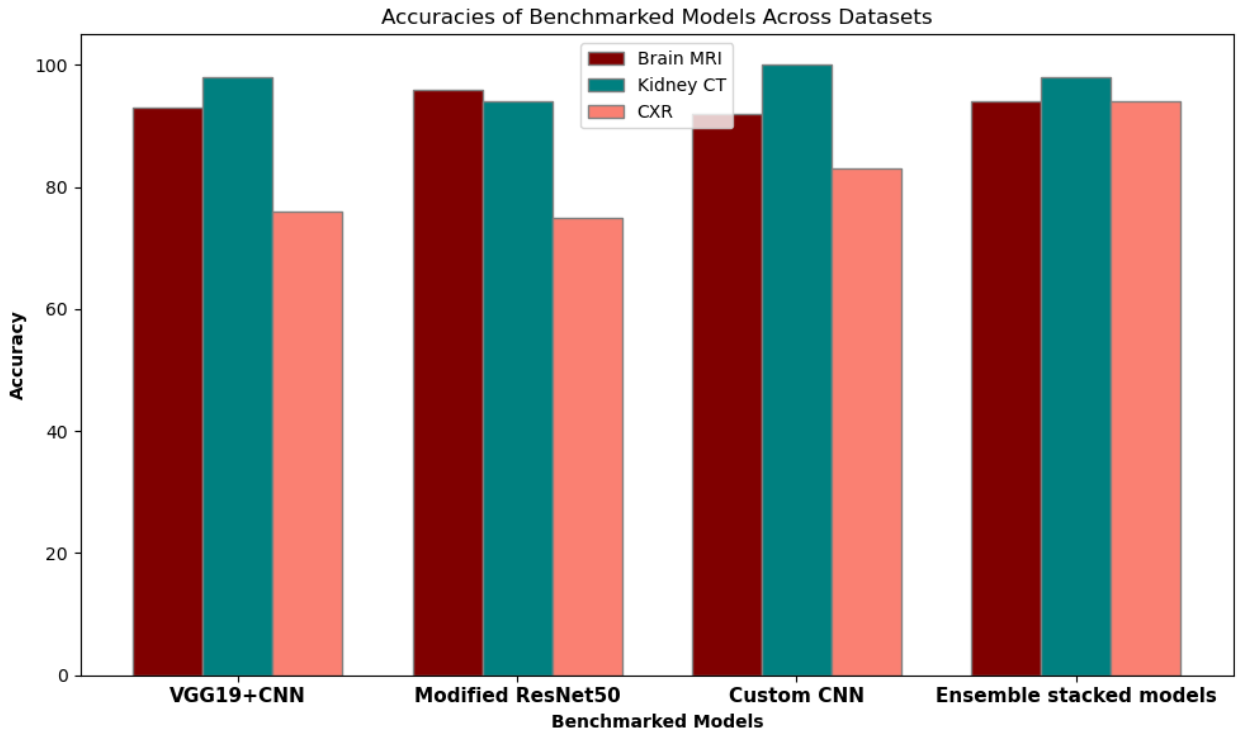


Figure 14: Evaluation Performance of Benchmarked Architecture on the 3 Imaging Modalities Based Accuracy Metrics

The study incorporated model architectures from tables 7, 8, and 9, adapting their distinct hyperparameters to suit the dataset. The VGG19+CNN model, as reported by Alshmrani et al. (2023), exhibited promising performance with an accuracy of 98% on a CXR dataset. Upon evaluation across a broader spectrum of modalities encompassing MRI, CT, and CXR, our study observed accuracies of 93%, 100%, and 76% respectively. The modified ResNet50, initially assessed on two CXR datasets with accuracies of 99.2% and 86.1%, displayed accuracies of 96%, 94%, and 75% across the MRI, CT, and CXR modalities when compared against our study's datasets.

Two MRI datasets were utilized to evaluate the custom CNN, achieving accuracies of 100% and 97.5%. The performance evaluation in our study showcased accuracies of 92%, 100%, and 83% across the MRI, CT, and CXR modalities. The ensemble methods, scrutinized across five datasets—two belonging to CXR and three to CT image modalities—maintained consistent performance with accuracies ranging from 84% to 99% for CXR datasets and from 90% to 99% for CT datasets. In contrast to our study, the ensemble method exhibited analogous results of 94%, 98%, and 94% across the MRI, CT, and CXR modalities. The ensemble method reliably maintained high accuracy across diverse image modalities. This approach utilized a fusion of multiple models, leveraging their individual strengths to collectively achieve a sustained, exceptional level of accuracy of 94%, 98%, and 94% across the diverse spectrum of medical imaging modalities under examination in this study.

4.0 Discussion

The outcome of this study has provided insight into the adaptability of pre-trained models across MRI scans, CT scans, and X-rays, spotlighting VGG16 as a standout performer across Brain MRI, Kidney CT, and CXR datasets, boasting impressive accuracies of 96%, 100%, and 95% respectively. Insights from implementing the ensemble method from existing literature reveals the ensemble method exhibited a sustained exceptional accuracy of 94%, 98%, and 94% for Brain MRI, Kidney CT, and CXR images respectively, showcasing its potential in achieving adaptability across diverse modalities. In contrast to the adaptability of Vgg16, the EfficientNetB7 model demonstrated high performance specifically in Brain MRI classification with 96% accuracy, its performance notably dipped when evaluated on the Kidney CT and CXR images with accuracy of 53% and 33% respectively. The result of efficientNetB7 based on the findings of similar studies by Abirami (2023), where EfficientNetB7 model performed with 98.4% accuracy in classifying four classes of brain MRI diagnosis might suggest the potential specialization for MRI classification specifically. These results build on existing need of ensuring consistent performance across diverse datasets and imaging modalities, crucial for dependable clinical decision-making and mitigating bias present in pretrained models which this study uncovers. While the future of AI in medical image interpretation holds immense potential, certain limitations merit attention. The rapid evolution of AI models may challenge the enduring relevance of findings, emphasizing the need for ongoing assessment and adaptation to emerging methodologies. Additionally, the study's focused approach on specific modalities might inadvertently overlook emerging ones, potentially limiting the broader applicability of findings to a more extensive array of pretrained models not encompassed in the analysis. This study emphasizes the importance of validating pretrained models across various modalities to ensure their robustness and suitability for medical image interpretation within the healthcare sector. These findings serve as a crucial guideline for selecting appropriate models, significantly impacting their successful application in medical imaging tasks.

5.0 Conclusion

The evolution of pretrained models in medical image classification has significantly impacted radiology, offering a reliable means for interpretation. By analyzing these pretrained models for their efficacy across

three key imaging modalities: MRI scans, CT scans, and X-rays, focusing on their adaptability, it can be concluded that pretrained models are robust in reproducing high performance across diverse datasets and imaging modalities. This study underscores the potential widespread application of the VGG16 model in interpreting various medical images for its reproducibility and generalizability, while also emphasizing the specialized use of certain pretrained models for specific imaging task like the EfficientNetB7 which specifically excelled in MRI classification. Additionally, the study highlights that although other high-performing models might not be optimal initially, fine-tuning could yield more favourable outcomes as seen in the benchmarked architectures implemented on. To better understand the implication of these results, future studies should consider the validation of the VGG16 model across a wider range of medical imaging modalities than those explored in this study. Further investigation into optimizing the EfficientNetB7 model for MRI tasks is also recommended. Moreover, validating pretrained models in real-world clinical settings remains pivotal to confirm their robustness and practical applicability. This research offers valuable insights into the potential of certain pretrained models, emphasizing their suitability for distinct imaging tasks, and also leveraging these models effectively in medical imaging, contributing to the ongoing quest for more accurate, adaptable, and widely applicable AI-enabled interpretations in healthcare.

6.0 References

Abdelaziz Ismael, S.A., Mohammed, A. & Hefny, H. (2020a) An enhanced deep learning approach for brain cancer MRI images classification using residual networks. *Artificial Intelligence in Medicine*, 102, 101779. Available online: https://doi.org/10.1016/j.artmed.2019.101779.

Abdelaziz Ismael, S.A., Mohammed, A. & Hefny, H. (2020b) An enhanced deep learning approach for brain cancer MRI images classification using residual networks. *Artificial Intelligence in Medicine*, 102, 101779. Available online: https://doi.org/10.1016/j.artmed.2019.101779.

Abhisheka, B., Biswas, S.K., Purkayastha, B., Das, D. & Escargueil, A. (2023) Recent trend in medical imaging modalities and their applications in disease diagnosis: a review. *Multimedia Tools and Applications* [Preprint]. Available online: https://doi.org/10.1007/s11042-023-17326-1.

Abirami A, S.B. (2023) MRI-based Brain Tumour Classification Using EfficientNetB7 model with transfer learning. *Journal of Survey in Fisheries Sciences*, 10(2S), 1737–1750. Available online: https://doi.org/10.17762/sfs.v10i2S.945.

Agrawal, S., Honnakasturi, V., Nara, M. & Patil, N. (2023) Utilizing Deep Learning Models and Transfer Learning for COVID-19 Detection from X-Ray Images. *SN Computer Science*, 4(4), 326. Available online: https://doi.org/10.1007/s42979-022-01655-3.

Alshmrani, G.M.M., Ni, Q., Jiang, R., Pervaiz, H. & Elshennawy, N.M. (2023) A deep learning architecture for multi-class lung diseases classification using chest X-ray (CXR) images. *Alexandria Engineering Journal*, 64, 923–935. Available online: https://doi.org/10.1016/j.aej.2022.10.053.

Balabanova, Y., Coker, R., Fedorin, I., Zakharova, S., Plavinskij, S., Krukov, N., Atun, R. & Drobniewski, F. (2005) Variability in interpretation of chest radiographs among Russian clinicians and implications for screening programmes: observational study. *BMJ : British Medical Journal*, 331(7513), 379–382.

Ben-Israel, D., Jacobs, W.B., Casha, S., Lang, S., Ryu, W.H.A., de Lotbiniere-Bassett, M. & Cadotte, D.W. (2020) The impact of machine learning on patient care: A systematic review. *Artificial Intelligence in Medicine*, 103, 101785. Available online: https://doi.org/10.1016/j.artmed.2019.101785.

Biswas, M., Kuppili, V., Saba, L., Edla, D.R., Suri, H.S., Cuadrado-Godia, E., Laird, J.R., Marinhoe, R.T., Sanches, J.M., Nicolaides, A. & Suri, J.S. (2019) State-of-the-art review on deep learning in medical imaging. *Frontiers in Bioscience-Landmark*, 24(3), 380–406. Available online: https://doi.org/10.2741/4725.

Chahar, V., Jaiswal, A., Gianchandani, N., Singh, D. & Kaur, M. (2020) Classification of the COVID-19 infected patients using DenseNet201 based deep transfer learning. *Journal of Biomolecular Structure & Dynamics*, 39. Available online: https://doi.org/10.1080/07391102.2020.1788642.

Chollet, F. (2017a) Xception: Deep Learning with Depthwise Separable Convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, 1800–1807. Available online: https://doi.org/10.1109/CVPR.2017.195.

Chollet, F. (2017b) Xception: Deep Learning with Depthwise Separable Convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1800–1807. Available online: https://doi.org/10.1109/CVPR.2017.195.

El-Rahiem, B. & Hammad, M. (2021) A Multi-fusion IoT Authentication System Based on Internal Deep Fusion of ECG Signals. In, 53–79. Available online: https://doi.org/10.1007/978-3-030-85428-7_4.

Gao, Q., Ogenyi, U., Liu, J., Ju, Z. & Liu, H. (2020) A Two-Stream CNN Framework for American Sign Language Recognition Based on Multimodal Data Fusion. In, 107–118. Available online: https://doi.org/10.1007/978-3-030-29933-0_9.

He, K., Zhang, X., Ren, S. & Sun, J. (2015) Deep Residual Learning for Image Recognition. arXiv. Available online: http://arxiv.org/abs/1512.03385 [Accessed 12/02/2024].

Huang, G., Liu, Z., Pleiss, G., Maaten, L. van der & Weinberger, K.Q. (2022) Convolutional Networks with Dense Connectivity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12), 8704–8716. Available online: https://doi.org/10.1109/TPAMI.2019.2918284.

Humayun, M., Sujatha, R., Almuayqil, S.N. & Jhanjhi, N.Z. (2022) A Transfer Learning Approach with a Convolutional Neural Network for the Classification of Lung Carcinoma. *Healthcare*, 10(6), 1058. Available online: https://doi.org/10.3390/healthcare10061058.

Ibrahim, M.R., Youssef, S.M. & Fathalla, K.M. (2023) Abnormality detection and intelligent severity assessment of human chest computed tomography scans using deep learning: a case study on SARS-COV-2 assessment. *Journal of Ambient Intelligence and Humanized Computing*, 14(5), 5665–5688. Available online: https://doi.org/10.1007/s12652-021-03282-x.

Islam, Md.M., Barua, P., Rahman, M., Ahammed, T., Akter, L. & Uddin, J. (2023) Transfer learning architectures with fine-tuning for brain tumor classification using magnetic resonance imaging. *Healthcare Analytics*, 4, 100270. Available online: https://doi.org/10.1016/j.health.2023.100270.

Jangam, E., Barreto, A.A.D. & Annavarapu, C.S.R. (2022) Automatic detection of COVID-19 from chest CT scan and chest X-Rays images using deep learning, transfer learning and stacking. *Applied Intelligence*, 52(2), 2243–2259. Available online: https://doi.org/10.1007/s10489-021-02393-4.

Kassania, S.H., Kassanib, P.H., Wesolowskic, M.J., Schneidera, K.A. & Detersa, R. (2021) Automatic Detection of Coronavirus Disease (COVID-19) in X-ray and CT Images: A Machine Learning Based

Approach. *Biocybernetics and Biomedical Engineering*, 41(3), 867–879. Available online: https://doi.org/10.1016/j.bbe.2021.05.013.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B. & Sánchez, C.I. (2017) A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60–88. Available online: https://doi.org/10.1016/j.media.2017.07.005.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T. & Xie, S. (2022) A ConvNet for the 2020s. In. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11976–11986. Available online: https://openaccess.thecvf.com/content/CVPR2022/html/Liu_A_ConvNet_for_the_2020s_CVPR_2022_paper.html [Accessed 12/02/2024].

Monowar, K.F., Hasan, Md.A.M. & Shin, J. (2020) Lung Opacity Classification With Convolutional Neural Networks Using Chest X-rays. *2020 11th International Conference on Electrical and Computer Engineering (ICECE)*, 169–172. Available online: https://doi.org/10.1109/ICECE51571.2020.9393135.

Narin, A., Kaya, C. & Pamuk, Z. (2021) Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks. *Pattern Analysis and Applications*, 24(3), 1207–1220. Available online: https://doi.org/10.1007/s10044-021-00984-y.

Nayak, D.R., Dash, R. & Majhi, B. (2020) Automated diagnosis of multi-class brain abnormalities using MRI images: A deep convolutional neural network based method. *Pattern Recognition Letters*, 138, 385–391. Available online: https://doi.org/10.1016/j.patrec.2020.04.018.

Nickparvar, M. (2021) Brain tumor MRI dataset. Available online: https://doi.org/10.34740/kaggle/dsv/2645886.

Özkaraca, O., Bağrıaçık, O.İ., Gürüler, H., Khan, F., Hussain, J., Khan, J. & Laila, U. e (2023) Multiple Brain Tumor Classification with Dense CNN Architecture Using Brain MRI Images. *Life*, 13(2), 349. Available online: https://doi.org/10.3390/life13020349.

Ozturk, T., Talo, M., Yildirim, E.A., Baloglu, U.B., Yildirim, O. & Rajendra Acharya, U. (2020) Automated detection of COVID-19 cases using deep neural networks with X-ray images. *Computers in Biology and Medicine*, 121, 103792. Available online: https://doi.org/10.1016/j.compbiomed.2020.103792.

Pesapane, F., Codari, M. & Sardanelli, F. (2018) Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. *European Radiology Experimental*, 2(1), 35. Available online: https://doi.org/10.1186/s41747-018-0061-6.

Power, S.P., Moloney, F., Twomey, M., James, K., O'Connor, O.J. & Maher, M.M. (2016) Computed tomography and patient risk: Facts, perceptions and uncertainties. *World Journal of Radiology*, 8(12), 902–915. Available online: https://doi.org/10.4329/wjr.v8.i12.902.

Puderbach, M., Eichinger, M., Haeselbarth, J., Ley, S., Kopp-Schneider, A., Tuengerthal, S., Schmaehl, A., Fink, C., Plathow, C., Wiebel, M., Demirakca, S., Müller, F.-M. & Kauczor, H.-U. (2007) Assessment of Morphological MRI for Pulmonary Changes in Cystic Fibrosis (CF) Patients: Comparison to Thin-Section CT and Chest X-ray. *Investigative Radiology*, 42(10), 715. Available online: https://doi.org/10.1097/RLI.0b013e318074fd81.

Ramadhan, A.A. & Baykara, M. (2022) A Novel Approach to Detect COVID-19: Enhanced Deep Learning Models with Convolutional Neural Networks. *Applied Sciences*, 12(18), 9325. Available online: https://doi.org/10.3390/app12189325.

Rastogi, A. (2022) *ResNet50 Medium*. Available online: https://blog.devgenius.io/resnet50-6b42934db431 [Accessed 20/02/2024].

Raza, R., Zulfiqar, F., Khan, M.O., Arif, M., Alvi, A., Iftikhar, M.A. & Alam, T. (2023) Lung-EffNet: Lung cancer classification using EfficientNet from CT-scan images. *Engineering Applications of Artificial Intelligence*, 126, 106902. Available online: https://doi.org/10.1016/j.engappai.2023.106902.

Simonyan, K. & Zisserman, A. (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv. Available online: https://doi.org/10.48550/arXiv.1409.1556.

Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. (2017) Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1). Available online: https://doi.org/10.1609/aaai.v31i1.11231.

Tan, M. & Le, Q. (2019) EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*. *International Conference on Machine Learning*, PMLR, 6105–6114. Available online: https://proceedings.mlr.press/v97/tan19a.html [Accessed 12/02/2024].

Tsang, S.-H. (2021) Review: NASNet — Neural Architecture Search Network (Image Classification). *Medium*, 27 July. Available online: https://sh-tsang.medium.com/review-nasnet-neural-architecture-search-network-image-classification-23139ea0425d [Accessed 20/02/2024].

Wang, B., Zhang, Y., Zhang, H., Meng, Q., Ma, Z.-M., Liu, T.-Y. & Chen, W. (2022) Provable Adaptivity in Adam. arXiv. Available online: https://doi.org/10.48550/arXiv.2208.09900.

Xue, X., Chinnaperumal, S., Abdulsahib, G.M., Manyam, R.R., Marappan, R., Raju, S.K. & Khalaf, O.I. (2023) Design and Analysis of a Deep Learning Ensemble Framework Model for the Detection of COVID-19 and Pneumonia Using Large-Scale CT Scan and X-ray Image Datasets. *Bioengineering*, 10(3), 363. Available online: https://doi.org/10.3390/bioengineering10030363.

Yengec-Tasdemir, S., Akay, E., Dogan, S. & Yilmaz, B. (2022) *Classification of Colorectal Polyps from Histopathological Images using Ensemble of ConvNeXt Variants*. Available online: https://doi.org/10.21203/rs.3.rs-1791422/v1.

Zhou, J., Hu, B., Feng, W., Zhang, Z., Fu, X., Shao, H., Wang, H., Jin, L., Ai, S. & Ji, Y. (2023) An ensemble deep learning model for risk stratification of invasive lung adenocarcinoma using thin-slice CT. *Npj Digital Medicine*, 6(1), 1–12. Available online: https://doi.org/10.1038/s41746-023-00866-z.

Zoph, B., Vasudevan, V., Shlens, J. & Le, Q.V. (2018) Learning Transferable Architectures for Scalable Image Recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT: IEEE, 8697–8710. Available online: https://doi.org/10.1109/CVPR.2018.00907.

7.0 Appendix

Loss curves and classification report of the pretrained models in this study

(a)

	precision	recall	f1-score	support
glioma	0.97	0.95	0.96	300
meningioma	0.94	0.88	0.91	306
notumor	0.95	1.00	0.97	405
pituitary	0.97	0.99	0.98	300
accuracy			0.96	1311
macro avg	0.96	0.95	0.96	1311
weighted avg	0.96	0.96	0.96	1311

(b)

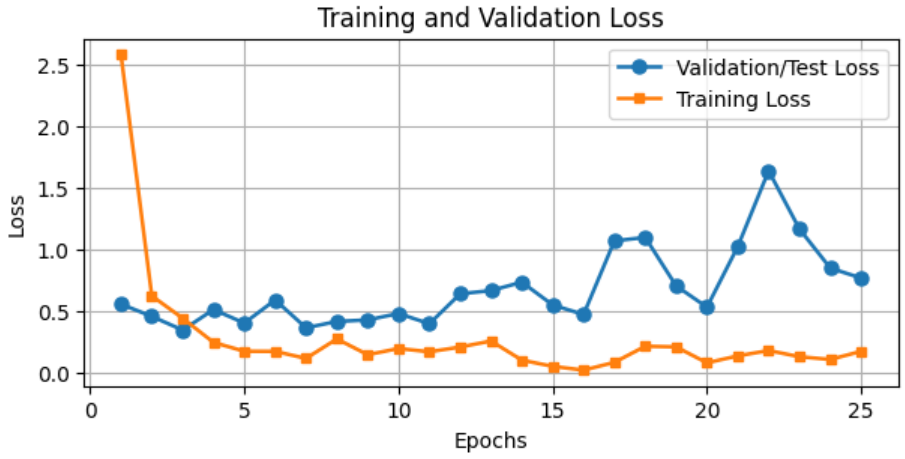
	precision	recall	f1-score	support
Cyst	1.00	1.00	1.00	649
Normal	1.00	1.00	1.00	999
stone	1.00	1.00	1.00	254
tumor	1.00	1.00	1.00	483
accuracy			1.00	2385
macro avg	1.00	1.00	1.00	2385
weighted avg	1.00	1.00	1.00	2385

(c)

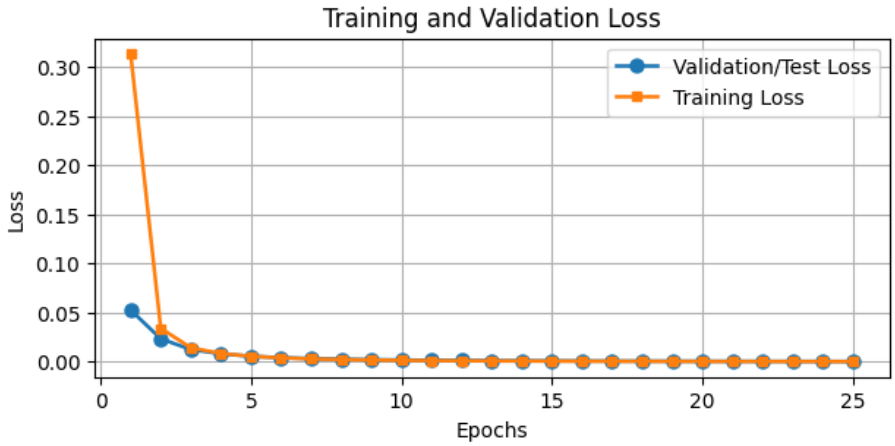
	precision	recall	f1-score	support
0	0.98	0.98	0.98	459
1	0.89	0.98	0.93	463
2	0.99	0.88	0.93	462
accuracy			0.95	1384
macro avg	0.95	0.95	0.95	1384
weighted avg	0.95	0.95	0.95	1384

Figure 6: a, b, and c represent the classification report of VGG16 in order of MRI, CT scan and CXR modalities.

(a)



(b)



(c)

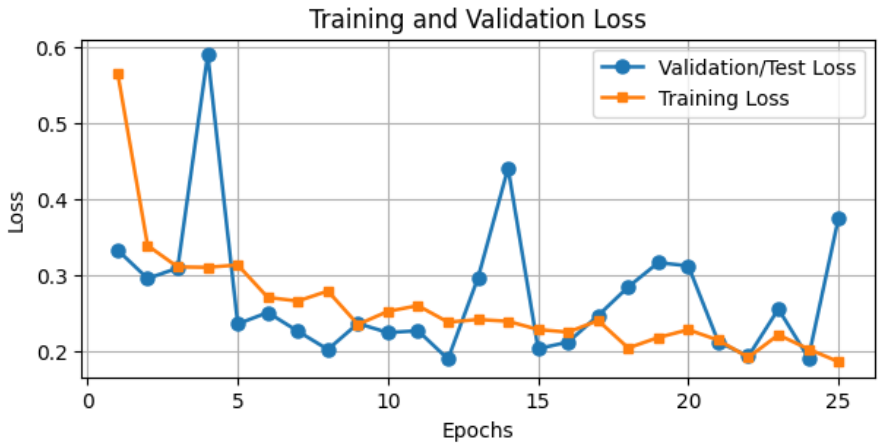


Figure 11: a, b, and c represent the Loss curve of VGG19 in order of MRI, CT scan and CXR modalities

(a)

	precision	recall	f1-score	support
glioma	0.95	0.93	0.94	300
meningioma	0.91	0.89	0.90	306
notumor	0.98	0.99	0.99	405
pituitary	0.96	0.99	0.97	300
accuracy			0.95	1311
macro avg	0.95	0.95	0.95	1311
weighted avg	0.95	0.95	0.95	1311

(b)

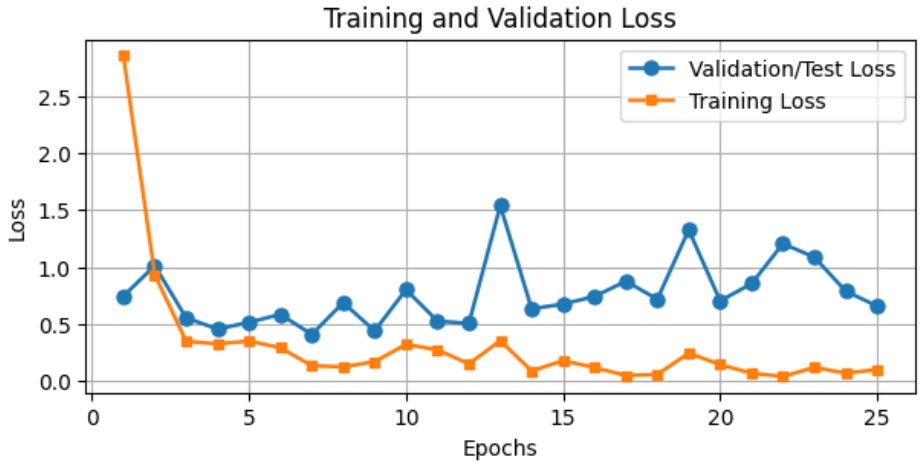
	precision	recall	f1-score	support
Cyst	1.00	1.00	1.00	649
Normal	1.00	1.00	1.00	999
stone	1.00	1.00	1.00	254
tumor	1.00	1.00	1.00	483
accuracy			1.00	2385
macro avg	1.00	1.00	1.00	2385
weighted avg	1.00	1.00	1.00	2385

(c)

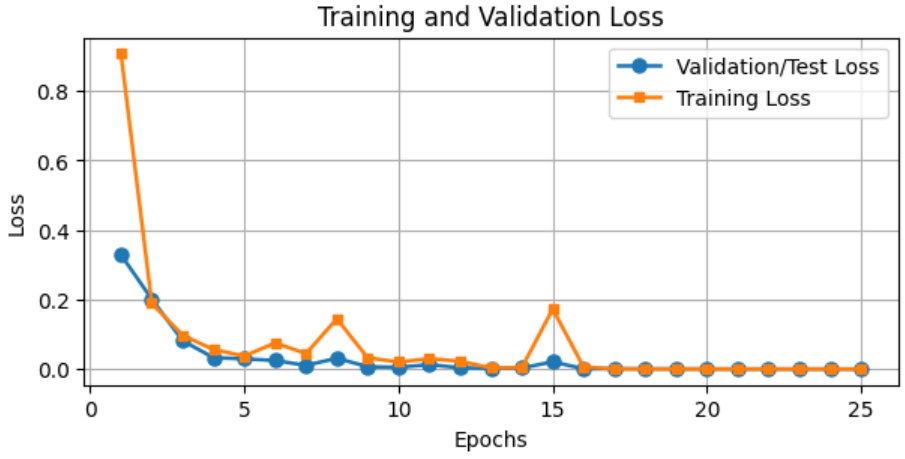
	precision	recall	f1-score	support
0	0.97	0.94	0.95	459
1	0.79	0.99	0.88	463
2	0.99	0.77	0.87	462
accuracy			0.90	1384
macro avg	0.92	0.90	0.90	1384
weighted avg	0.92	0.90	0.90	1384

Figure 12: a, b, and c represent the classification report of VGG19 in order of MRI, CT scan and CXR modalities.

(a)



(b)



(c)

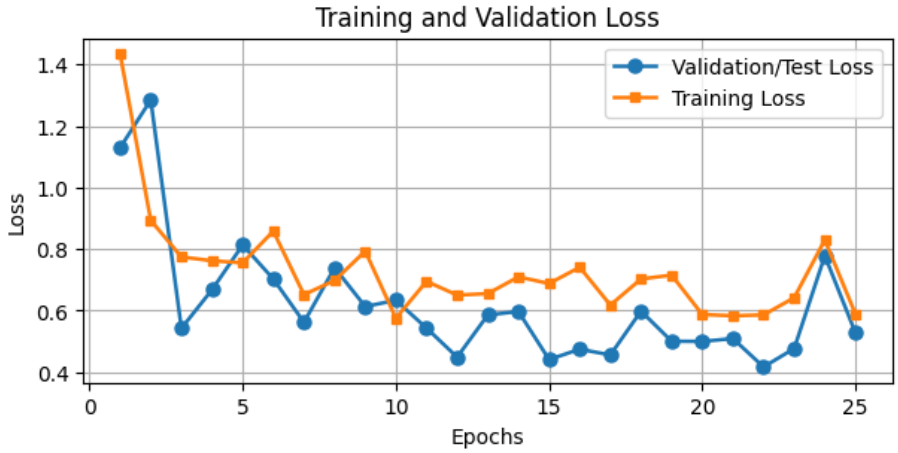


Figure 13: a, b, and c represent the Loss curve of ResNet50 in order of MRI, CT scan and CXR modalities.

(a)

	precision	recall	f1-score	support
glioma	0.98	0.86	0.91	300
meningioma	0.86	0.95	0.90	306
notumor	0.99	0.99	0.99	405
pituitary	0.99	0.99	0.99	300
accuracy			0.95	1311
macro avg	0.95	0.95	0.95	1311
weighted avg	0.95	0.95	0.95	1311

(b)

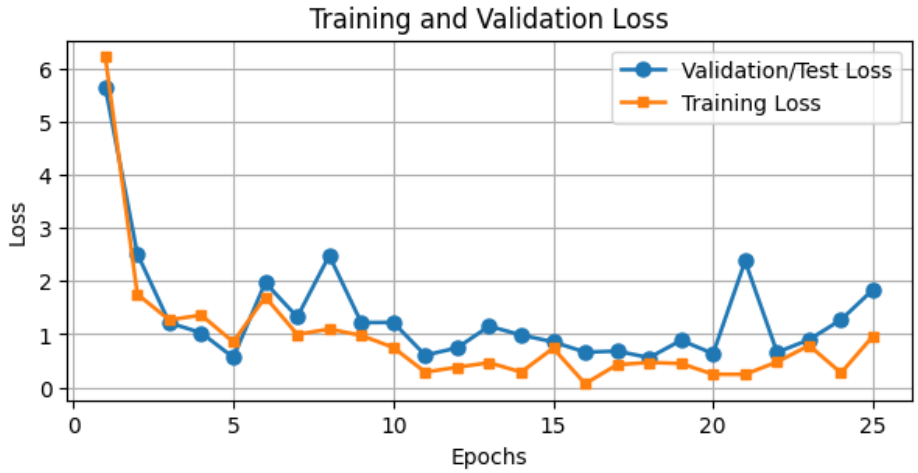
	precision	recall	f1-score	support
Cyst	1.00	1.00	1.00	649
Normal	1.00	1.00	1.00	999
stone	1.00	1.00	1.00	254
tumor	1.00	1.00	1.00	483
accuracy			1.00	2385
macro avg	1.00	1.00	1.00	2385
weighted avg	1.00	1.00	1.00	2385

(c)

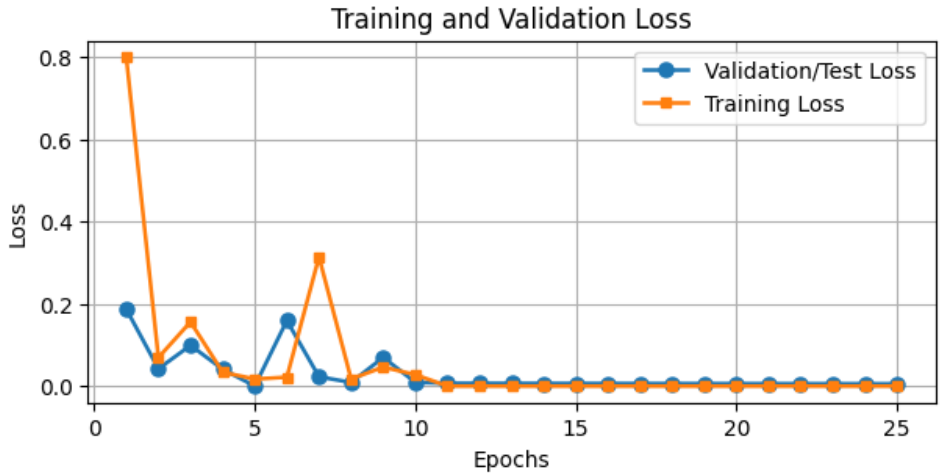
	precision	recall	f1-score	support
0	0.82	0.86	0.84	459
1	0.74	0.89	0.81	463
2	0.99	0.74	0.84	462
accuracy			0.83	1384
macro avg	0.85	0.83	0.83	1384
weighted avg	0.85	0.83	0.83	1384

Figure 14: a, b, and c represent the classification report of ResNet50 in order of MRI, CT scan and CXR modalities.

(a)



(b)



(c)

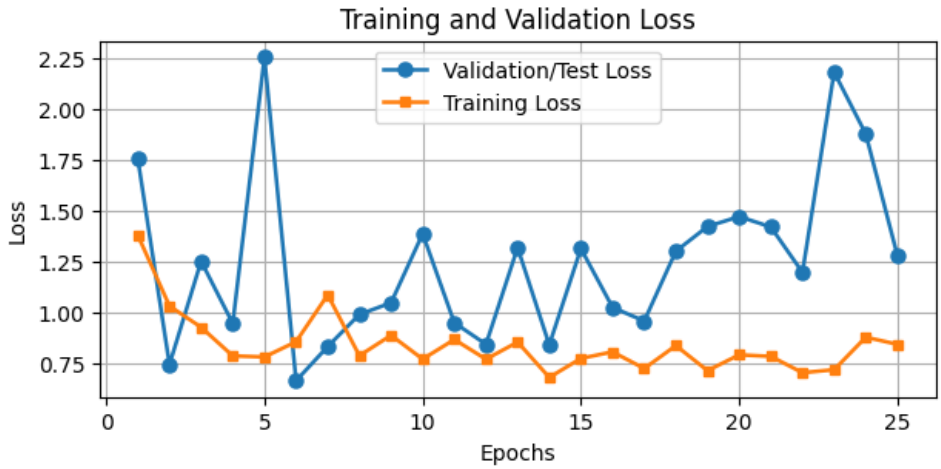


Figure 19: a, b, and c represent the Loss curve of DenseNet201 in order of MRI, CT scan and CXR modalities

(a)

	precision	recall	f1-score	support
glioma	0.89	0.93	0.91	300
meningioma	0.92	0.74	0.82	306
notumor	0.91	1.00	0.95	405
pituitary	0.96	0.98	0.97	300
accuracy			0.92	1311
macro avg	0.92	0.91	0.91	1311
weighted avg	0.92	0.92	0.92	1311

(b)

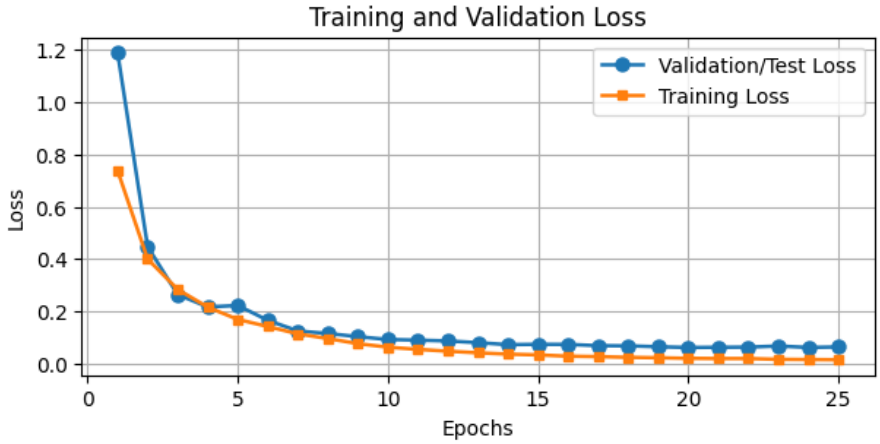
	precision	recall	f1-score	support
Cyst	1.00	1.00	1.00	649
Normal	1.00	1.00	1.00	999
stone	1.00	1.00	1.00	254
tumor	1.00	1.00	1.00	483
accuracy			1.00	2385
macro avg	1.00	1.00	1.00	2385
weighted avg	1.00	1.00	1.00	2385

(c)

	precision	recall	f1-score	support
0	0.95	0.97	0.96	459
1	0.89	0.91	0.90	463
2	0.93	0.89	0.91	462
accuracy			0.92	1384
macro avg	0.92	0.92	0.92	1384
weighted avg	0.92	0.92	0.92	1384

Figure 20: a, b, and c represent the classification report of DenseNet201 in order of MRI, CT scan and CXR modalities.

(a)



(b)

(c)

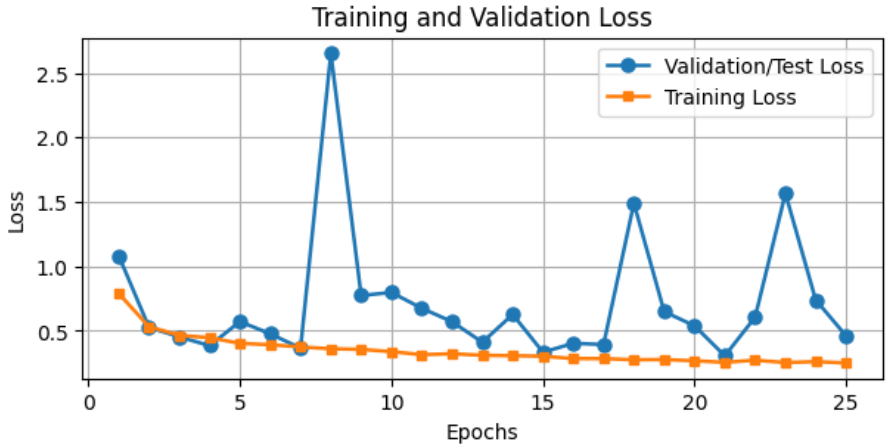


Figure 25: a, b, and c represent the Loss curve of CustomCNN in order of MRI, CT scan and CXR modalities.

(a)

	precision	recall	f1-score	support
glioma	0.95	0.79	0.86	300
meningioma	0.82	0.87	0.84	306
notumor	0.95	1.00	0.98	405
pituitary	0.95	0.98	0.96	300
accuracy			0.92	1311
macro avg	0.92	0.91	0.91	1311
weighted avg	0.92	0.92	0.92	1311

(b)

	precision	recall	f1-score	support
Cyst	1.00	1.00	1.00	649
Normal	1.00	1.00	1.00	999
stone	1.00	1.00	1.00	254
tumor	1.00	1.00	1.00	483
accuracy			1.00	2385
macro avg	1.00	1.00	1.00	2385
weighted avg	1.00	1.00	1.00	2385

(c)

	precision	recall	f1-score	support
0	0.98	0.66	0.79	459
1	0.77	0.90	0.83	463
2	0.79	0.92	0.85	462
accuracy			0.83	1384
macro avg	0.85	0.83	0.82	1384
weighted avg	0.85	0.83	0.82	1384

Figure 26: a, b, and c represent the classification report of Custom CNN in order of MRI, CT scan, and CXR modalities.

(a)

	precision	recall	f1-score	support
glioma	0.99	0.98	0.98	288
meningioma	0.96	0.97	0.96	255
notumor	0.99	0.99	0.99	321
pituitary	0.99	0.99	0.99	278
accuracy			0.98	1142
macro avg	0.98	0.98	0.98	1142
weighted avg	0.98	0.98	0.98	1142

(b)

	precision	recall	f1-score	support
Cyst	1.00	0.94	0.97	649
Normal	1.00	1.00	1.00	999
stone	0.87	1.00	0.93	254
tumor	1.00	1.00	1.00	483
accuracy			0.98	2385
macro avg	0.97	0.99	0.98	2385
weighted avg	0.99	0.98	0.98	2385

(c)

	precision	recall	f1-score	support
0	0.95	0.98	0.97	459
1	0.92	0.92	0.92	463
2	0.95	0.92	0.94	462
accuracy			0.94	1384
macro avg	0.94	0.94	0.94	1384
weighted avg	0.94	0.94	0.94	1384

Figure 29: a, b, and c represent the classification report of Ensemble stacked Models in order of MRI, CT scan and CXR modalities.