Advanced CNN and Explainable AI Based Architecture for Interpretable Brain MRI Analysis

Shuvashis Sarker
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
shuvashisofficial@gmail.com

Shamim Rahim Refat
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
n.a.refat2000@gmail.com

Faika Fairuj Preotee
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
faikafairuj2001@gmail.com

Tashreef Muhammad
Department of Computer Science and Engineering
Southeast University
Dhaka, Bangladesh
tashreef.muhammad@seu.edu.bd

ABSTRACT

Convolutional Neural Networks (CNNs) serve as a foundational component in the domain of Computer Vision (CV). In order to enhance the Interpretability of CNN models, a critical aspect for clinical adoption, this study incorporates Explainable AI (XAI) methodologies. Through applying CNNs and XAI to a dataset comprising 5285 Brain MRI Images, a classification accuracy of 86% was achieved. The LIME framework was employed to generate localized explanations, thereby augmenting the model's transparency and facilitating a deeper understanding of its decision-making process. This research explores the potential of synergistically integrating deep learning and XAI to foster the development of more reliable and comprehensible medical image analysis systems. Such systems hold the promise of improving diagnostic accuracy and clinical decision-making by providing healthcare professionals with transparent and explainable insights into the model's predictions, ultimately leading to more informed and effective patient care.

CCS CONCEPTS

• **Computing methodologies** → **Object recognition**; **Object detection**.

KEYWORDS

Convolutional Neural Network (CNN), Explainable AI (XAI), Local Interpretable Model-agnostic Explanations (LIME), Magnetic Resonance Imaging (MRI), Brain Disease, T1-weighted

ACM Reference Format:
Shuvashis Sarker, Shamim Rahim Refat, Faika Fairuj Preotee, and Tashreef Muhammad. 2024. Advanced CNN and Explainable AI Based Architecture for Interpretable Brain MRI Analysis. In *3rd International Conference on Computing Advancements (ICCA 2024), October 17–18, 2024, Dhaka, Bangladesh.* ACM, Dhaka, Bangladesh, 8 pages. https://doi.org/10.1145/3723178.3723220

ICCA 2024, October 17–18, 2024, Dhaka, Bangladesh

ACM ISBN 979-8-4007-1382-8/2024/10…$15.00
https://doi.org/10.1145/3723178.3723220

1 INTRODUCTION

Medical imaging is a valuable resource for detailed internal body examinations. In recent years, the developments in Artificial Intelligence (AI) and Machine Learning (ML) have significantly transformed the healthcare sector. These technologies are being used to optimize tasks like analyzing patient records, providing personalized healthcare and improving medical image analysis. AI and ML has a profound impact across various healthcare sectors, especially in the analysis of Magnetic Resonance Imaging (MRI). These techniques have achieved many meaningful results [6] that have been proved by different conducted studies.

Medical imaging process named Magnetic Resonance Imaging (MRI) is used in radiology to make clear pictures of the body's organs and cells. MRI does not use ionising radiation like X-rays and CT scans. Strong magnetic fields, radio waves and a computer are used instead to make cross-sectional images of the body that are clear and full of information.

The brain is the command center of the human body which is mainly responsible for coordinating all bodily functions and processes. This extraordinarily complex organ is composed of the cerebellum, cerebrum and brain stem. All together, they make the central nervous system. The brain keeps changing and adjusting through a process called neuroplasticity. It helps to reform itself in the presence of new experiences, learning new lessosn or the fear of getting hurt [14, 26].

Brain diseases are a broad term for a number of conditions that affect the brain and its processes, including movement, memory, balance and such. Alzheimer's diseases, traumatic injuries, vascular diseases, infections and tumors are some of the different types of these illnesses. In reality, Magnetic Resonance Imaging (MRI) helps experts to find brain diseases by giving them clear pictures of soft tissues. This makes it easier to understand and make more accurate diagnosis throughout a wide range of brain conditions and problems. Complex MRI data can be analyzed much more quickly and correctly with AI algorithms, especially those that are based on deep learning. These algorithms or techniques can identify patterns

or leads to different problems that might just get skipped by the human eye easily in the brain. Also, Explainable AI (XAI) can make the procedure even better by allowing people to see and understand how AI algorithms make decisions. XAI tells experts how and why certain patterns and changes were found in the MRI data, so that they can believe the AI's suggestions and better understand what they mean.

2 RELATED WORK

Several new publications about how MRI can classify different types of brain related images were studied prior to conducting the experiments in this study. In this area of research Computer Vision, Image Processing, Artificial Intelligence and Machine Learning were used simultaneously for achieving ground breaking results.

One such approach implemented by Yusuf Brima et al. [6] where authors utilized the Deep Residual Convolutional Neural Network (ResNet50) architecture for multi-class classification of brain tumors in MRI images. They used three dataset, among which there is a novel dataset collected from National Institute of Neuroscience & Hospitals, Bangladesh. From this dataset they achieved 86.40% accuracy. Another one was the Harvard Whole Brain Atlas Dataset [31], 93.80% accuracy was gained using this dataset and the third dataset was the Biomedical School of Engineering brain tumor dataset [9], which gained 97.05% accuracy.

Authored by Sushama Ghodke et al. [12], a research which was published in recent days, had detected various kinds of brain diseases from MRI images using CNN and VGG19 architecture models. In this study a clinical database was used. There were two distinct approaches to categorizing brain MRIs in this system. Brain MRIs were classified as either Normal or Abnormal using the first approach and as having No Tumor, Glioma or Meningioma using the second way. Utilizing the CNN method, the first strategy was able to achieve an accuracy of 87.5% after 50 iterations. The second technique, which yielded an accuracy of 89.6% and 94.1% respectively, by using the CNN and VGG19 algorithms.

Through a mechanized process, Samia Ferdous Mou et al. [22] developed a capability of diagnosis accurate disease by extracting essential feature from complex image data using MRI. They employed the EfficientNetB0 model, a CNN model optimized for classifying brain diseases. Two datasets - one was Brain Tumor Dataset from Kaggle [7] which includes images classified into tumor and healthy categories, another Alzheimer Parkinson dataset [24] was used, which includes images classified into Alzheimer, Parkinson, and Control Parkinson categories in this regard. The accuracy for Brain Tumor Dataset was 96.1%. Besides, in Alzheimer Parkinson Dataset, it achieved 96.2%. By combining Brain Tumor Dataset (for multi-class classification), 95.4% accuracy was achieved.

Potentially improving the efficiency and precision of medical diagnosis and treatment planning, S Basil Xavier et al.[33] in their recent paper imposed an approach using MobileNet and ResNet deep convolutional neural network architectures, utilizing transfer learning to classify brain diseases and estimate the age of patients from MRI scans. To train the deep learning model, a large number of MRI scans was gathered, together with age estimations and the related disease labels like Normal, Mild Cognitive Impairment and Alzheimer's disease. The paper examined high accuracy achievements in disease classification and age estimation tasks, though specific numeric results were not provided in the context.

Dimitrios Amanatidis et al. [3] proposed basic CNN for classifying Brain MRI images to diagnose autoimmune diseases, particularly focusing on multiple sclerosis. Privilege of using CNNs was their capability to effectively handle spatial hierarchies in image data, which enabled accurate segmentation and classification of brain MRI images for diagnosing diseases like multiple sclerosis. CNN model achieved a remarkable accuracy of nearly 100%, indicating highly effective classification of the MRI images.

Harish Rohan Kambhampaty et al. [15] also worked on a Sequential Deep Convolutional Neural Network (DCNN) in adults using Brain MRI scans. The study combined data from two sources: the Alzheimer Disease Neuroimaging Initiative (ADNI) and a healthy brains dataset from Kaggle[13]. The model achieved a maximum accuracy of 93.57% with Batch Config-3b (6000 including CatB).

U. Sakthi et al. [27] used a Deep Convolutional Neural Network (DCNN) to identify brain tumour MRI images more accurately than SVM, KNN and standard CNNs by extracting more features. A dataset from Kaggle with 5712 pictures including types like Pituitary, Glioma and Meningioma and a testing set of 1310 images were used to train and test the DCNN thoroughly. The paper discussed a lot about how accurate the DCNN was, but it didn't give a number. There were hints of problems like the need for a lot of data and problems with generalisation across various MRI setups, but they were not fully explained.

A method of using a dual-input CNN architecture utilizing Explainable AI (XAI) tools, including LIME and "SHapley Additive exPlanations (SHAP)" by Loveleen Gaur [11] to enhance classification accuracy for brain tumors using MRI images. The dataset used MRI images which are annotated into three categories: Glioma, Meningioma and Pituitary Tumors, sourced from publicly available data on Kaggle. Out of 2870 total images, 2296 images of distinct types were used as training sets and the remaining as test sets. The CNN model achieved overall 85.37% test accuracy.

Another researcher Hamza Ahmed Shad [28] used a pipeline of three deep learning models: ResNet-50, VGG16 and Inception V3, for predicting stages of Alzheimer's disease using MRI images. For proper classification they also used XAI(LIME). This study used a hybrid dataset from Kaggle, Alzheimer's disease from T1 weighted MRI images and it was labeled as non demented, very mild demented, mild demented and moderate demented. These models had their ability to provide high categorical accuracy in early detection of Alzheimer's disease stages, which was crucial for timely intervention. The models achieved categorical accuracies of 86.82% for VGG16, 82.56% for ResNet50 and 82.04% for Inception V3.

Recent studies used lightweight models such as MobileNetV2 and EfficientNet for diagnosing brain tumour and achieved better efficiency. Majeed et al.[19] improved on this with MobileNetV3, achieved 91% accuracy overall, 91% on the NINS 2022 dataset [5], and 94% on the SBE-SMU dataset[9], outperformed other methods.

3 METHODOLOGY

3.1 Dataset

The "Brain MRI Dataset" which has used in our study is a collection of 5285 Brain MRI Images that are publicly available in Figshare [5]. This rich dataset is put together with joint efforts from the Computer Science and Engineering Department at the University of Dhaka and the National Institute of Neuroscience (NINS) in Bangladesh. The images are T1-weighted and contrast-enhance, providing detailed views of the brain across 37 different categories. Figure 1 shows a bar plot that visually represents the dataset consisting of 37 different classes, effectively showing the quantity of MRI images for each class and Figure 2 highlights some specific classes from the dataset, offering a closer look at the variety of brain scans of the dataset.

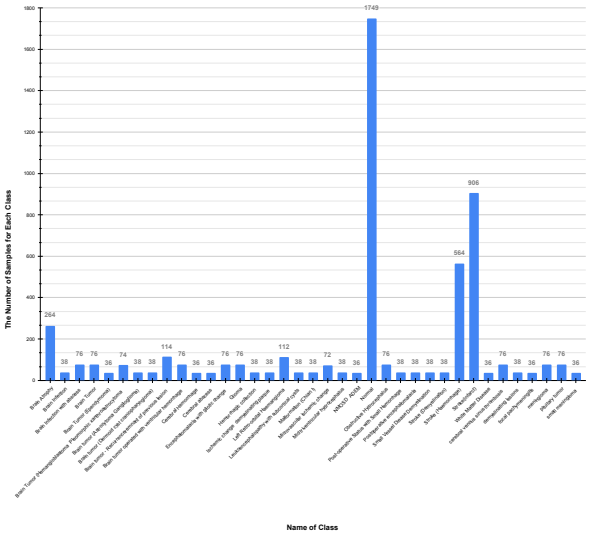


Figure 1: A Visual Representation of the Dataset

3.2 Data Preprocessing

The initial step is involved choosing brain MRI images from the "Brain MRI Dataset" for model training and evaluation. To standardize processing by CNN models, the images are resized to a consistent dimension of 256 × 256 pixels, despite their original varying sizes. The dataset is then split into training, testing and validation sets, following an 80:10:10 division and using a batch size of 16 images. The training set is used to train the model, the validation set assists in fine-tuning hyperparameters and monitoring performance, and the testing set evaluates the model's ability to generalize. These preparatory actions including image resizing, dataset splitting and setting batch sizes, prepare the "Brain MRI dataset" for effective training and evaluation using CNN.

3.3 CNN Architecture

Convolutional Neural Networks or CNNs or ConvNets, are a subclass of neural networks that are particularly good at processing

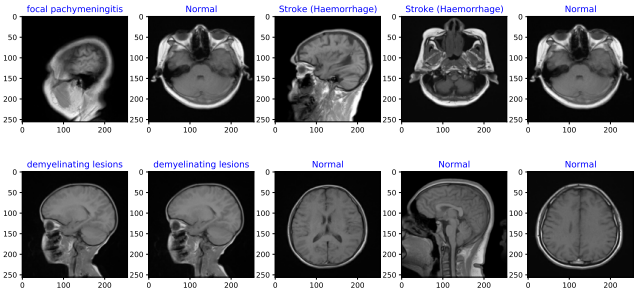


Figure 2: Sample Images of the Brain MRI Dataset

input with a topology resembling a grid, like images. Neurons in the brains of humans and animals which is served as the model for CNNs architecture.

Convolutional, Pooling and Fully Connected Layers are the three layers that make up a CNN [29]. Figure 3 shows the architecture of the basic CNN model. The primary component of a CNN structure, the convolutional layer is made up of the computational burden on the network. This layer does a dot product between two matrices: one matrix is called the kernel, contains the learnable parameters and the second matrix contains parts of the receptive fields, which may be any type of grid-like input, such as an image.

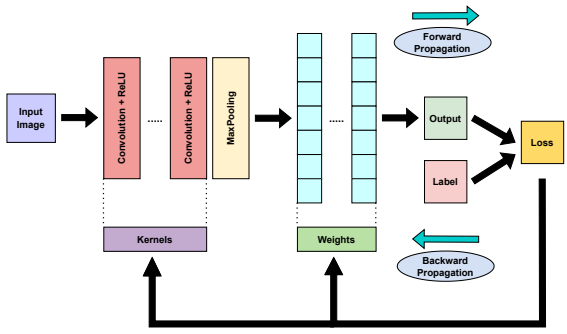


Figure 3: A flow diagram of CNN Architecture

While the kernel layer is deeper than the image data, it is still smaller than the picture. For instance, an RGB image with three channels (Red, Green and Blue) has a depth of three. The kernel layer depth is designed to cover all three channels.The kernel layer moves across a certain number of pixels or an interval of pixels called the stride, to receive data segments from the receptive field or picture. The output of the ReLU function [16] is

$$f(x) = \max(0, x) \tag{1}$$

Convolution is a linear layer, but real-world data like pictures are rarely linear. Activation functions [21] are employed to impart nonlinearity to our convolution. Although there are other activation functions as well, such that Sigmoid, Softmax, tanh etc. ReLU often performs better in terms of performance.

By downsampling the data, the pooling layer [20] minimizes the number of parameters when the data is too big. MaxPooling2D is

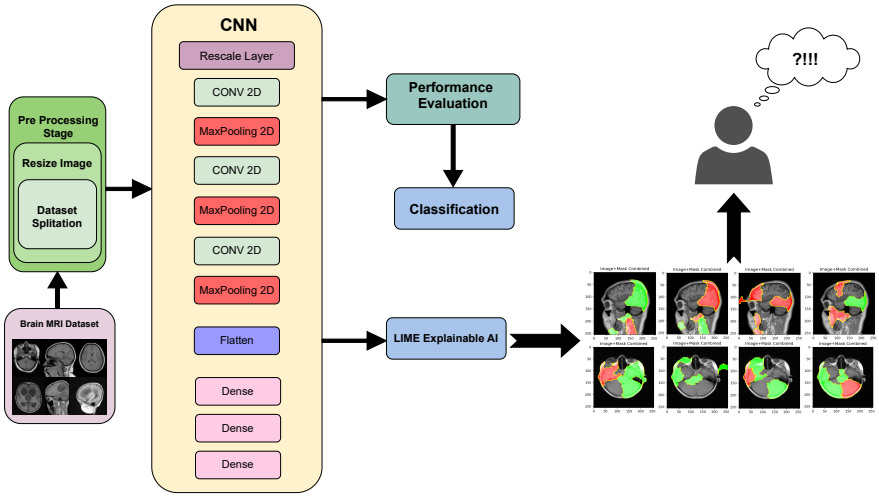


Figure 4: A Flow Diagram of Our Proposed Model

the pooling layer that employes in CNN model. By cutting the 2D spatial data along their spatial dimensions, or height and breadth, in accordance with a set window which is known as the pool size and shifting over, akin to in the convolutional layer following a defined Stride [8], MaxPooling2D downsamples the data.

Through forward propagation on a training dataset, the performance of the model with specific kernels and weights is determined using a loss function. Learnable parameters, or kernels and weights, are then updated based on the loss value through back propagation using the gradient descent optimization algorithm.

In the majority of current radiomics [18] research, manual feature extraction methods like texture analysis are used first and then traditional machine learning classifiers like random forests and support vector machines [17]. There exist several distinctions between these techniques and CNN. Firstly, handcrafted feature extraction is not necessary with CNN. Second, human expertise are not always needed to segregate tumors or organs in CNN structures. Third CNN has millions of learnable parameters to estimate, it is far more data-hungry and computationally expensive, necessitating the use of graphics processing units (GPU) for model training.

3.4 Proposed Model

Figure 4 indicates the flow diagram of our proposed model using a sequential framework made for sorting brain MRI images into groups that can help find and study brain related conditions. It initiates with a rescaling layer to normalize image pixel values. Then it has three convolutional (Conv2D) layers and each one is paired with a MaxPooling2D layer that pulls out and downsample features. The Conv2D layers have depths that change gradually, starting with 32 filters, growing to 128 filters and then shrinking back down to 64 filters. This makes the model better at capturing different picture characteristics. The data is flattened after convolutional processing so that it is ready for the fully linked step. In this step, there is a dense network with a dropout layer placed after each dense layer to keep from being overfitted.

The network is made up of two dense layers, each with 1024 and 512 neurons that represent a different classification category. Additionally, a dense layer is used following Softmax activation function with 37 neurons according to the 37 categories in the dataset. The model has an amazing 59.6 million parameters, and all of them can be trainable. This means that it is a very complex model that can learn and classify features in very particular manners.

4 RESULT ANALYSIS

The methodologies are executed on a computer equipped with an Intel Core i5 13400F, NVIDIA GeForce RTX 3060 12GB and 16GB DDR4 RAM. Keras [10], a TensorFlow [1] API, which are served as the foundation for implementing the model.

The suggested method's performance has assessed using precision, recall, f1-score and accuracy once classification completed using CNN. In our proposed model, "SparseCategoricalCrossentropy" is the selected loss function, worked well for classification jobs with numerous categories where the labels are integers. The network weights has been adjusted using the Adam optimizer, which has a learning rate of 1e-4 that has determined how frequently the weights should change during training. Last but not least, the accuracy of the model, which has quantified the percentage accurately.

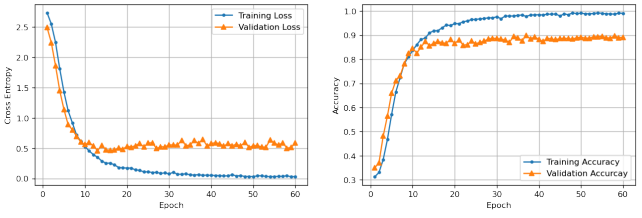


Figure 5: Training and Validation Loss and Accuracy Over Numerous Iterations.

The Figure 5 represents two graphs which demonstrate how training and validation loss and accuracy have changed over a

Table 1: Detailed Performance Metrics for Various Medical Conditions

Condition	Precision	Recall	F1-score
Brain Atrophy	0.89	0.64	0.74
Brain Infection	0.99	0.99	0.99
Brain Infection with abscess	0.80	0.50	0.62
Brain Tumor	0.98	0.64	0.77
Brain tumor (Astrocytoma Ganglioglioma)	1.00	0.80	0.89
Brain tumor (Dermoid cyst craniopharyngioma)	0.80	0.80	0.80
Brain Tumor (Ependymoma)	0.98	0.67	0.80
Brain Tumor (Hemangioblastoma Pleomorphic xanthroastrocytoma metastasis)	1.00	0.83	0.91
Brain tumor - Recurrence remnant of previous lesion	0.75	0.64	0.69
Brain tumor operated with ventricular hemorrhage	0.99	0.99	0.99
Cerebral abscess	0.98	0.75	0.85
Cerebral Hemorrhage	0.99	1.00	0.99
Cerebral venous sinus thrombosis	0.98	0.67	0.80
Demyelinating lesions	0.98	0.60	0.74
Encephalomalacia with gliotic change	0.98	0.50	0.66
Focal pachymeningitis	1.00	0.75	0.86
Glioma	0.92	1.00	0.96
Hemorrhagic collection	0.99	1.00	0.99
Ischemic change demyelinating plaque	0.98	0.60	0.74
Left Retro-orbital Haemangioma	0.86	0.75	0.80
Leukoencephalopathy with subcortical cysts	0.50	1.00	0.67
Malformation (Chiari I)	0.99	1.00	0.99
Meningioma	0.80	0.95	0.87
Microvascular ischemic change	0.99	0.99	0.99
Mid triventricular hydrocephalus	1.00	0.67	0.80
NMOSD ADEM	0.98	0.67	0.80
Normal	1.00	1.00	1.00
Obstructive Hydrocephalus	0.57	0.80	0.67
Pituitary tumor	0.83	0.83	0.83
Post-operative Status with Small Hemorrhage	0.90	0.90	0.90
Postoperative encephalomalacia	1.00	0.80	0.89
Small meningioma	0.99	1.00	0.99
Small Vessel Disease Demyelination	0.80	0.57	0.67
Stroke (Demyelination)	0.99	1.00	0.99
Stroke (Haemorrhage)	1.00	0.75	0.86
Stroke (infarct)	1.00	0.86	0.92
White Matter Disease	0.99	1.00	0.99

Table 2: Comparison of Different Approaches

Authors	Approach	Accuracy	Use of XAI
Brima et al.[6]	Transfer Learning of ResNet50	0.87	NO
	Transfer Learning of ResNet50 with Augmented Data	0.81	
	Fine-tuning ResNet50 with Augmented Data	0.84	
Ahmed et al.[19]	Transfer Learning of MobileNetV3 with Augmented Data	0.92	NO
Proposed model	**CNN**	**0.86**	**YES (LIME)**

number of epochs. As the training goes on, a clear pattern has appeared, the loss goes down and then stays unchanged, which shows that the model has been learnt as predicted. At the same time, the accuracy goes up and then stays the same, which shows that the model can recall what it has learned. The validation measures are comparable but there are a few minor modifications. This shows how well the model might work with data it hasn't seen before. The

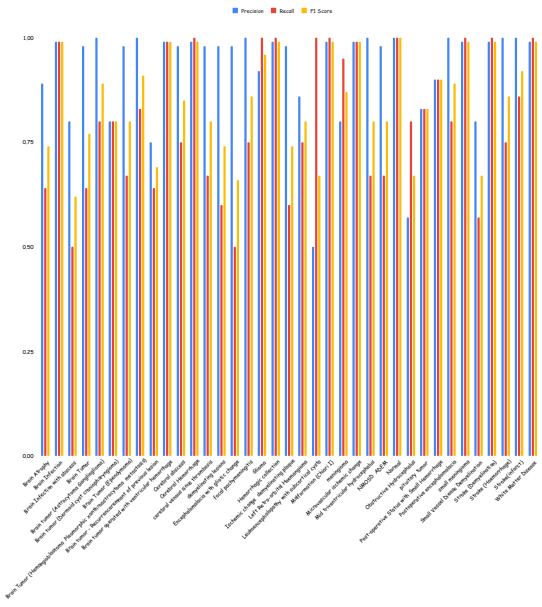


Figure 6: Performance Evaluation Metrics for Various Medical Conditions.

Figure 6 that follows compares the performance evaluation metrics (precision, recall and f1-score) visually, making it easy to observe the differences between classes right away. This visualization helps illustrate which classes the model can guess more confidently and accurately and which ones may need additional training.

A good indication of a model's predictive power is its test accuracy. Our proposed model does well because it is about 86% accurate

on the test. Although precision, recall and f1-score, three specific performance metrics that vary greatly depending on the situation generally display high values. According to the data, the accuracy varies between 84% to 86% after five independent runs, with 86% accuracy being the most common result. The fact that, the accuracy rate is increasing over time indicates that the model is robust in its predictions and that it functions effectively even after many tests. Here is a performance metrics of some category in the Table 1 Precision, Recall and F1-score offer a detailed view of the model effectiveness for each specific condition.

Table 2 showcases different model training approaches, focusing on accuracy and the use of Explainable AI (XAI). Our proposed approach, which uses a CNN with LIME for interpretability, achieves an accuracy of 86% emphasizing the importance of model explainability in addition to performance.

5 EXPLAINABLE AI

Many recent machine learning and deep learning research methods investigate results in complex ways which are hard to understand. Often, researchers and programmers use some pre-existing models and struggle to explain how the models arrive at their decisions, classifications or predictions. As many doctors might not trust AI's help in medicine, if they don't understand why it makes certain choices [4]. Recently, several methods which is known as XAI have been developed to make deep learning models more understandable. These XAI methods are divided into two types: those that are specific to a particular model and those that can be used with any model, according to a survey by Adadi and Berrada [2]. For instance, researchers use a technique where they give input data and watch how the output from deep learning models changes, to understand which parts of the data influence the model's decisions [23].

5.1 Lime

Ribeiro et al. [25] came up with LIME (Local Interpretable Model-agnostic Explanations). This method can explain how any classifier or regressor works, no matter how complicated it is. LIME illuminates the rationale behind predictions by locally approximating any given model to a comprehensible counterpart.[30] This elucidation is achieved through two pivotal attributes: *Interpretable representation* and *Local fidelity*. These elements ensure that LIME not only presents a clear connection between input features and outputs but also assures that the explanations are trustworthy around the specific data point under consideration. The term *Model-agnostic* signifies LIME's capability to regard the original model as a *Black-box*, thereby maintaining versatility across various data forms such as images, text and tabular data, providing explanations in textual, numerical or visual formats. For instance, in image classification, LIME may transform a complex image tensor into a binary vector representing the presence or absence of certain pixel patches to aid in elucidation of the predictive decision-making process. The mathematical representation of the LIME process is denoted by:

$$\xi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{2}$$

where ξ denotes the model to be explained, G is the class of interpretable models that LIME can utilize, L is the loss function quantifying the discrepancy between the predictions of the black-box model ξ and the interpretable model g, π_x weights the proximity of instances and $\Omega(g)$ imposes a penalty for complexity to maintain the interpretability of the model g [23, 32].

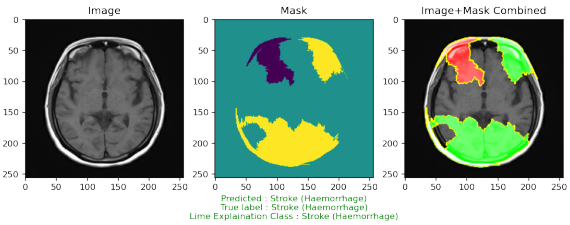
In our work, we have implemented LIME for multi-class classification to explain the predictions of our proposed model trained on brain MRI images. The goal is to understand how the models are making predictions for different brain conditions. Figure 7 shows a sample brain MRI images of the actual class, which is predicted correctly by the model. The green regions in the image represent the positive class (the ground truth class) that has the highest probability according to the model's prediction.

The combination of four images shows in Figure 7 the visualisation clearly shows how the model makes decisions in Brain MRI image classification. In Figure 7a, the model predicts a Stroke (Haemorrhage), marking important areas in green to show a high chance of this condition as well as areas in red to indicate a lower chance of other classes. In Figure 7b, the Normal condition is displayed. The green highlights shows that the model is positive that there will be no major medical findings. This is in direct opposition to the red areas, which are thought to have a lower chance of changing the prediction. Figure 7c shows a Cerebral Abscess. The green areas show the infection sites that are important for this finding, while the red areas show areas that are not as important. Lastly, Figure 7d shows a Small Meningioma. The green areas show how important the small meningiomas are to the model's ability to predict, while the red areas show other, less important areas. This colour coding across all images not only helps us figure out where the model is focused, but it also makes it more likely that the model will correctly rank the most important parts of each medical scenario, which leads to accurate and reliable predictions.

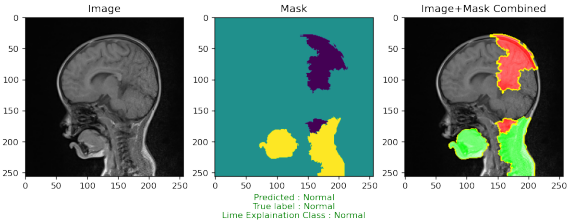
Figure 8 shows multiple images that show the differences between actual classes and the model's predictions, along with LIME explanations. Each image shows in green the parts that the model considers as the most important for making predictions, which causes mistakes in classification. For instance, Figure 8a shows a Hemorrhagic Collection that is wrongly classified as a Stroke (Infarct), by the model which highlights the areas it gets inappropriate. Figure 8b shows a Normal MRI that is incorrectly classified as a Brain Tumour (Hemangioblastoma Pleomorphic xanthroastrocytoma metastasis). The green areas show where the misunderstanding occurred. In Figure 8c, a Brain Tumour - Recurrenceremnant of Previous Lesion is incorrectly classified as Brain Atrophy. Green areas show where the model gets it incorrect and gives the wrong characteristics. Finally, Figure 8d shows Microvascular Ischemic Change that is wrongly classified as Normal. The green colour again shows that important characteristics that should have been considered are not taken into consideration. The red areas in every image represent to other classes. This visual difference helps explain why and how the model makes these mistakes. This color-coded visualisation helps you get a better sense of how the model usually works and where it could be improved.

6 LIMITATIONS AND FUTURE WORK

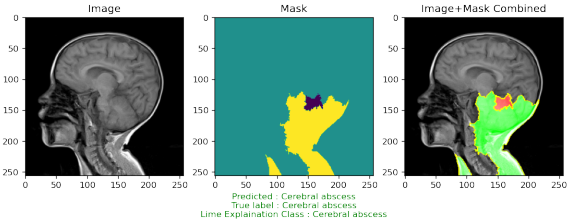
In our research, we have noticed a notable limitation in the form of dataset imbalance. We have employed one particular deep learning



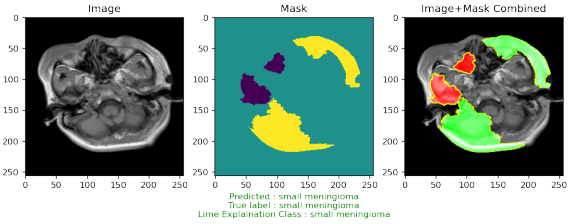
(a) Actual: Stroke (Haemorrhage)
Predicted: Stroke (Haemorrhage)
Lime: Stroke (Haemorrhage)



(b) Actual: Normal
Predicted: Normal
Lime: Normal

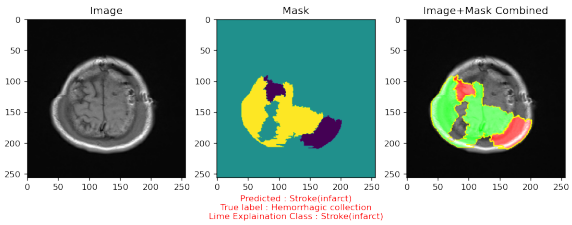


(c) Actual: Cerebral Abscess
Predicted: Cerebral Abscess
Lime: Cerebral Abscess

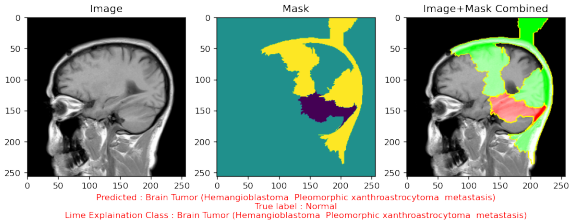


(d) Actual: Small Meningioma
Predicted: Small Meningioma
Lime: Small Meningioma

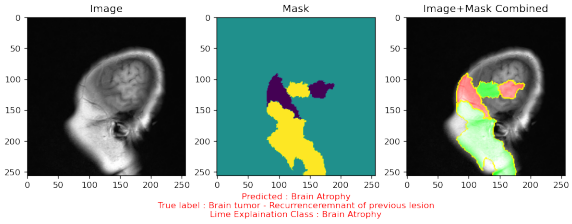
Figure 7: Some Sample Images of Actual Class Predicted Correctly by the Model



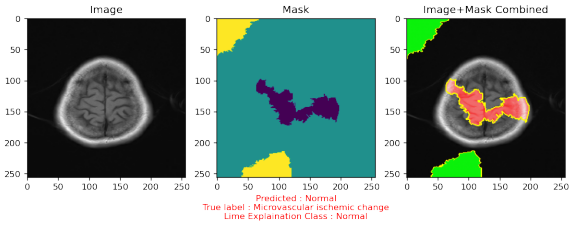
(a) Actual: Hemorrhagic (Collection)
Predicted: Stroke(Infarct)
Lime: Stroke(Infarct)



(b) Actual: Normal
Predicted: Brain Tumor (Hemangioblastoma Pleomorphic xanthroastrocytoma metastasis)
Lime: Brain Tumor (Hemangioblastoma Pleomorphic xanthroastrocytoma metastasis)



(c) Actual: Brain tumor - Recurrenceremnant of Previous Lesion
Predicted: Brain Atrophy
Lime: Brain Atrophy



(d) Actual: Microvascular Ischemic Change
Predicted: Normal
Lime: Normal

Figure 8: Some Sample Images of Actual Class Predicted Incorrectly by the Model

model and utilized LIME to clarify the outcomes, enhancing comprehension of the machine's decision-making mechanism. However, the study may have been improved by using numerous models and utilizing a range of Explainable AI methodologies to strengthen the reliability and comprehensibility of our findings.

In our future work, we intend to utilize transfer learning models to take advantage of their established capabilities and also investigate the creation of a hybrid model in order to potentially enhance accuracy and performance. In order to rectify the existing imbalance in the dataset, our primary objective will be to generate a dataset that is balanced. Furthermore, given the information utilized is specific to Bangladesh and has been publicly released, our objective is to gather and employ additional local datasets to strengthen the

research and ensure that the conclusions are more representative and relevant to the local people.

7 CONCLUSION

In conclusion, our method use a CNN model on a localized dataset from NINS, Bangladesh, which is taught us a lot about how to identify brain MRI images. Utilizing LIME to explain the model's predictions across all 37 classes, we've provided insight on what is commonly thought of as a "Black-box" by providing reasons for each class prediction. This clear way not only builds trust in our model's diagnostic abilities, but it also makes it more credible for both patients and healthcare professionals. Diseases have been successfully identified and LIME's answers are easy to understand. This shows that trustworthy AI-driven diagnostic tools could help and even change the health sector.

ACKNOWLEDGMENTS

The contents, including the tables, graphs, images and analytical thoughts presented in this paper, are fully original. AI tools(ChatGPT-4) have been utilized for the formal organization and presentation of the text. The core ideas, methodoogies, findings and explainations are author's thoughtful contribution and effort.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.

[2] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.

[3] Dimitrios Amanatidis, Georgios Chatzisavvas, and Michael Dossis. 2022. Brain MRI based diagnosis of autoimmune diseases using deep learning. In *2022 7th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*. IEEE, IEEE, Piraeus, Greece, 1–5.

[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82 – 115. https://doi.org/10.1016/j.inffus.2019.12.012

[5] Yusuf Brima, Mossadek Hossain Kamal Tushar, Upama Kabir, and Tariqul Islam. 2021. Brain MRI Dataset. https://doi.org/10.6084/m9.figshare.14778750.v2.

[6] Yusuf Brima, Mosaddek Hossain Kamal Tushar, Upama Kabir, and Tariqul Islam. 2021. Deep Transfer Learning for Brain Magnetic Resonance Image Multi-class Classification. *Dhaka University Journal of Applied Science and Engineering* 6, 2 (2021), 14–29.

[7] Navoneel Chakrabarty. 2019. Brain MRI Images for Brain Tumor Detection. https://www.kaggle.com/datasets/navoneel/brain-mri-images-for-brain-tumor-detection. Accessed on April 22, 2024.

[8] Ting-Hao Chen. 2017. What is "stride" in Convolutional Neural Network? https://medium.com/machine-learning-algorithms/what-is-stride-in-convolutional-neural-network-e3b4ae9baedb. accessed on 22 April 2024.

[9] Jun Cheng. 2017. Brain Tumor Dataset. https://doi.org/10.6084/m9.figshare.1512427.v5 Figshare.

[10] François Chollet et al. 2015. Keras. https://keras.io.

[11] Loveleen Gaur, Mohan Bhandari, Tanvi Razdan, Saurav Mallik, and Zhongming Zhao. 2022. Explanation-driven deep learning model for prediction of brain tumour status using MRI image data. *Frontiers in genetics* 13 (2022), 822666.

[12] Sushama Ghodke and Sunita Nandgave. 2023. Brain MRI Classification Using Convolutional Neural Networks and VGG19: A Deep Learning Approach for Accurate Brain Disease Diagnosis. In *2023 International Conference on Integration of Computational Intelligent System (ICICIS)*. IEEE, IEEE, Bangalore, India, 1–6.

[13] Ahmed Hamada. 2021. Br35H :: Brain Tumor Detection 2020. https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection Accessed on April 22, 2024.

[14] Tonya Hines. 2018. Anatomy of the Brain. https://mayfieldclinic.com/pe-anatbrain.htm. accessed on 22 April 2024.

[15] Harish Rohan Kambhampaty, Battula Naga Datha Saikiran, Y Vijayalata, and Ashlin Deepa. 2022. A Novel Deep-Learning Based Classification Of Alzheimer's Disease In Adults. In *2022 IEEE Delhi Section Conference (DELCON)*. IEEE, IEEE, New Delhi, India, 1–6.

[16] Serkan Kızılırmak. 2023. Rectified Linear Unit (ReLU) Function: Understanding the Basics. https://medium.com/@serkankizilirmak/rectified-linear-unit-relu-function-in-machine-learning-understanding-the-basics-3770bb31c2a8. accessed on 22 April 2024.

[17] Philippe Lambin, Emmanuel Rios-Velazquez, Ralph Leijenaar, Sara Carvalho, Ruud GPM Van Stiphout, Patrick Granton, Catharina ML Zegers, Robert Gillies, Ronald Boellard, André Dekker, et al. 2012. Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer* 48, 4 (2012), 441–446.

[18] Kristofef Linton-Reid. 2021. What is Radiomics? https://towardsdatascience.com/what-is-radiomics-d9fb790a58c7. accessed on 22 April 2024.

[19] Ahmed Firas Majeed, Pedram Salehpour, Leili Farzinvash, and Saeid Pashazadeh. 2024. Multi-class Brain Lesion Classification using Deep Transfer Learning with MobileNetV3. *IEEE Access* TBA, TBA (2024), TBA.

[20] Nikita Malviya. 2023. Convolution & Pooling Layers. https://medium.com/@nikitamalviya/convolution-pooling-f8e797898cf9. accessed on 22 April 2024.

[21] Mayank Mishra. 2020. Convolutional Neural Networks, Explained. https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939. accessed on 22 April 2024.

[22] Samia Ferdous Mou and SM Abdur Razzak. 2023. Brain Disease Classification from MRI Scans Using EfficientNetB0 Feature Extraction. In *2023 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*. IEEE, IEEE, Dhaka, Bangladesh, 336–340.

[23] Jinzhao Qian, Hailong Li, Junqi Wang, and Lili He. 2023. Recent advances in explainable artificial intelligence for magnetic resonance imaging. *Diagnostics* 13, 9 (2023), 1571.

[24] Gokul Ramasamy. 2022. alzheimer_parkinson_disease. https://www.kaggle.com/datasets/gokulramasamy/alzheimer-parkinson-disease Accessed on April 22, 2024.

[25] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, San Francisco, CA, USA, 1135–1144.

[26] Anand I Rughani. 2015. Brain Anatomy. https://emedicine.medscape.com/article/1898830-overview?form=fpf. accessed on 22 April 2024.

[27] U Sakthi, K Thangaraj, A Tamizhselvi, and MK Kirubakaran. 2023. Deep Convolutional Neural Network Framework for Brain Tumor Classification using MRI Images. In *2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS)*. IEEE, IEEE, Coimbatore, India, 548–553.

[28] Hamza Ahmed Shad, Quazi Ashikur Rahman, Nashita Binte Asad, Atif Zawad Bakshi, SM Faiaz Mursalin, Md Tanzim Reza, and Mohammad Zavid Parvez. 2021. Exploring Alzheimer's disease prediction with XAI in various neural network models. In *TENCON 2021-2021 IEEE Region 10 Conference (TENCON)*. IEEE, IEEE, Auckland, New Zealand, 720–725.

[29] SAGAR SHARMA. 2017. Activation Functions in Neural Networks. https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6. accessed on 22 April 2024.

[30] Md Tanvir Rouf Shawon, GM Shibli, Farzad Ahmed, and Sajib Kumar Saha Joy. 2023. Explainable cost-sensitive deep neural networks for brain tumor detection from brain MRI images considering data imbalance. *arXiv preprint arXiv:2308.00608* (2023).

[31] D Summers. 2003. Harvard Whole Brain Atlas: www.med.harvard.edu/AANLIB/home.html. *Journal of Neurology, Neurosurgery & Psychiatry* 74, 3 (2003), 288–288. https://doi.org/10.1136/jnnp.74.3.288 arXiv:https://jnnp.bmj.com/content/74/3/288.full.pdf

[32] Viswan Vimbi, Noushath Shaffi, and Mufti Mahmud. 2024. Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection. *Brain Informatics* 11, 1 (2024), 10.

[33] S Basil Xavier, Eddula Sai Manoj, Bande Rohith, K Abhilash, et al. 2023. Brain Disease Classification along with Age Estimation from MRI. In *2023 4th International Conference for Emerging Technology (INCET)*. IEEE, IEEE, Belgaum, India, 1–4.