Differential Diagnosis of Brain Diseases Using Ensemble Learning and Explainable AI

Nighat Bibi[1][a], Kathleen M. Curran[2][b] and Jane Courtney[1][c]
[1]*Technological University Dublin, School of Electrical and Electronic Engineering, Dublin, Ireland*
[2]*University College Dublin, School of Medicine, UCD Belfield, Dublin, Ireland*
{*nighat.bibi, jane.courtney*}*@tudublin.ie, kathleen.curran@ucd.ie*

Keywords: Brain Tumour, Multiple Sclerosis, Glioma, Meningioma, Pituitary, Ensemble Learning, Explainability, Interpretability, XAI, Grad-Cam.

Abstract: The differential diagnosis of brain diseases by magnetic resonance imaging (MRI) is a crucial step in the diagnostic process, and deep learning (DL) has the potential to significantly improve the accuracy and efficiency of these diagnoses. This study focuses on creating an ensemble learning (EL) model that utilizes the ResNet50, DenseNet121, and EfficientNetB1 architectures to concurrently and accurately classify various brain conditions from MRI images. The proposed ensemble learning model identifies a range of brain disorders that encompass different types of brain tumours, as well as multiple sclerosis. The proposed model trained on two open source datasets, consisting of MRI images of glioma, meningioma, pituitary tumours, non-tumour and multiple sclerosis. Central to this research is the integration of gradient-weighted class activation mapping (Grad-CAM) for model interpretability, aligning with the growing emphasis on explainable AI (XAI) in medical imaging. The application of Grad-CAM improves the transparency of the decision-making process of the model, which is vital for clinical acceptance and trust in AI-assisted diagnostic tools. The EL model achieved an impressive 99.84% accuracy in classifying these various brain conditions, demonstrating its potential as a versatile and effective tool for differential diagnosis in neuroimaging. The model's ability to distinguish between multiple brain diseases underscores its significant potential in the field of medical imaging. Additionally, Grad-CAM visualizations provide deeper insights into the neural network's reasoning, contributing to a more transparent and interpretable AI-driven diagnostic process in neuroimaging.

1 INTRODUCTION

Early and accurate diagnosis of brain conditions is essential for effective treatment. Magnetic resonance imaging (MRI) plays a crucial role in this process (Anantharajan et al., 2024; Zebari et al., 2024), but interpreting these images requires specialized knowledge and can be time-consuming (Segato et al., 2020). Recently, deep learning (DL) algorithms have become powerful tools for medical image analysis, improving diagnostic accuracy and efficiency (Arbabshirani et al., 2018; Lopatina et al., 2020). DL, a subset of Artificial intelligence (AI), trains neural networks on large datasets for tasks like classification, detection, and segmentation. DL models have shown promise in identifying brain pathologies, including tumours and neurodegenerative diseases (Segato et al., 2020). MS, a severe autoimmune disease affecting the central nervous system, is common in many countries and often leads to non-traumatic neurological impairment in young people (Dobson and Giovannoni, 2019; Browne et al., 2014). MS diagnosis relies on McDonald's criteria (Thompson et al., 2018), which require clinical and imaging evidence, necessitating clinical expertise (Lopatina et al., 2020). Several studies have demonstrated the potential of machine learning and deep learning in MS diagnosis (Macin et al., 2022; Reddy et al., 2023; Ekmekyapar and Taşçı, 2023; Bibi et al., 2024a). Brain tumours are also require precise identification for effective management (Nair et al., 2023). MRI modalities like T1w, T2w, and FLAIR provide essential contrasts for lesion identification, but interpreting these images is complex. Computer-aided diagnosis (CAD) systems aid in tumour detection (Esmaeili et al., 2021). DL models, including CNNs, have achieved high accu-

[a] https://orcid.org/0000-0002-3586-7363
[b] https://orcid.org/0000-0003-0095-9337
[c] https://orcid.org/0000-0002-9175-7855

Bibi, N., Curran, K. M. and Courtney, J.
Differential Diagnosis of Brain Diseases Using Ensemble Learning and Explainable AI.
DOI: 10.5220/0013325800003912
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 2: VISAPP*, pages 958-964
ISBN: 978-989-758-728-3; ISSN: 2184-4321
Proceedings Copyright © 2025 by SCITEPRESS – Science and Technology Publications, Lda.

racy in tumour detection, segmentation, and classification. Techniques like CAM, SHAP, and LIME improve model interpretability, aiding medical professionals (Zhang et al., 2022; Bibi et al.,).

Despite progress, challenges remain in applying DL to diverse brain conditions. The diagnosis of overlapping diseases, such as MS and brain tumours, can be challenging, and AI aids in this process (Shafi et al., 2021). Most studies focus on specific diseases, which limits their scope. Yousaf et al. (Yousaf et al., 2023) proposed a CNN method to detect tumours and ischemic stroke with high accuracy. Talo et al. (Talo et al., 2019) used pre-trained CNN models to classify various brain diseases, with ResNet-50 achieving the highest accuracy. Shafi et al. (Shafi et al., 2021) proposed an SVM-based ensemble classifier for tumour and MS lesion classification.

These studies underscore the need for models to diagnose multiple brain diseases simultaneously. The black-box nature of DL models poses a challenge in clinical settings, where understanding the rationale behind a diagnosis is crucial. Incorporating explainability into DL models is necessary for transparency and trust. Our proposed study uses an ensemble learning model to classify multiple brain conditions from MRI images, including various tumours and MS, which goes beyond the typical single-disease focus of previous studies. Using Grad-CAM for interpretability, it enhances transparency and trust in the diagnostic process, providing insights into the model's decision-making. The proposed model demonstrates high accuracy (99.84%) in diverse datasets. Furthermore, our future work will further validate the Grad-CAM visual evaluations with clinical insights, ensuring the practical applicability of the model in different clinical settings.

2 MATERIALS AND METHODS

2.1 Dataset Description

The foundation of this study lies in the combination of two publicly available datasets, which are instrumental in the classification of brain conditions through MRI imaging: the Brain Tumour MRI Dataset (Nickparvar, 2021) and the Multiple Sclerosis dataset Muslim et al. (Muslim et al., 2022). The datasets provide a diverse range of images, crucial for training and validating the deep learning model. The distribution of images across different classes in the dataset is shown in Table 1, and representative samples from both datasets are illustrated in Figure 1.

2.1.1 Brain Tumour MRI Dataset

The Brain tumour MRI dataset (Nickparvar, 2021) comprehensively aggregates data from three sources: figshare (Cheng, 2017), the SARTAJ dataset (Bhuvaji et al., 2020) and Br35H: Brain tumour Detection 2020. This dataset is integral to the study and contains a total of 7023 human brain MRI images. These images are classified into four categories: Glioma, meningioma, no-tumour, and pituitary. According to Nickparva (Nickparvar, 2021) during the dataset preparation phase, a detailed examination revealed misclassification issues with Glioma images from the SARTAJ dataset. This discovery was supported by a comparative analysis of similar research work and performance evaluations of various models. To rectify this, the Glioma images of the SARTAJ dataset were replaced with accurately classified images from figshare, ensuring the integrity and reliability of the dataset (Nickparvar, 2021).

2.1.2 Multiple Sclerosis Dataset

The Multiple Sclerosis dataset is derived from a comprehensive study by Muslim et al. (Muslim et al., 2022). This dataset is crucial for the study's objective of extending its classification capabilities beyond tumours. Includes T1-weighted, T2-weighted, and Flair magnetic resonance images of 60 MS patients. T1-weighted images selected from this dataset because the tumour dataset contains T1-weighted images.

Table 1: Image distribution across training and validation and testing datasets for each class.

Class	Training	Validation	Testing
MS	1210	150	150
Glioma	1321	150	150
Meningioma	1339	153	153
Control	1595	202	203
Pituitary	1457	150	150

2.2 Methods

The Methodology of the study involves in dataset's preprocessing, model selection, training, and the approach to interpreting the model's results.

2.2.1 Preprocessing

A critical preprocessing step, particularly for the MS dataset, involved selecting 2D slices from T1-weighted magnetic resonance images containing MS lesions by comparing them with segmented binary masks created by three radiologists and a neurologist.

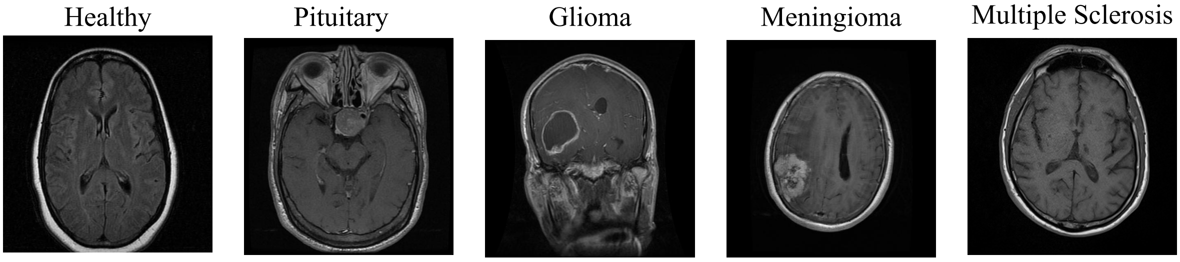


Figure 1: Representative images for each class in the dataset.

The dataset was divided into training, validation, and testing sets, following the distribution demonstrated in Table 1, with approximately 80% allocated for training, 10% for validation, and 10% for testing. All images from the Brain Tumour and MS datasets were normalized using the mean and standard deviation of the ImageNet dataset and resized to a uniform dimension of 240x240 pixels to ensure consistency across the datasets and optimize processing conditions.

2.2.2 Model Training

The training process involved selecting state-of-the-art convolutional neural network architectures known for their proficiency in image classification tasks. DenseNet121 (Huang et al., 2017), EfficientNetB1 (Tan and Le, 2019), and ResNet50 (He et al., 2015) were chosen due to their proven success in medical imaging. These models were pre-trained on the ImageNet dataset to leverage transfer learning, enabling faster convergence and enhanced performance.

The models were trained using the Adam optimizer, with categorical cross-entropy as the loss function. Training was conducted with a batch size of 32 for up to 50 epochs. Early stopping, based on validation loss, was implemented to prevent overfitting. An ensemble model was created by averaging the predictions of the three architectures, effectively combining their strengths to enhance classification accuracy. Training and testing were performed on a MacBook M2, utilizing its computational capabilities for efficient processing of the models and datasets.

2.2.3 Model Explainability

Understanding the decision-making process of the model is crucial to validate its effectiveness in clinical applications. Gradient-weighted class activation mapping (Grad-CAM) used for each individual model to achieve this. Grad-CAM generates heatmaps highlighting the regions in MRI images that significantly influenced the model's classification decisions, providing visual explanations of the model's reasoning.

Grad-CAM works by using the gradients flowing into the final convolutional layer to produce a localization map of important regions in the image. This technique ensures that the model focuses on relevant anatomical structures, such as tumour boundaries and internal lesions, which are critical for accurate diagnosis. For multiple sclerosis, it highlights the lesions essential for correct identification.

These visualizations can help validate the model's predictions against clinical expectations, ensuring that the AI's decision-making process aligns with medical knowledge. By examining the heatmaps, researchers and clinicians can detect potential model biases or errors, indicating areas needing further training or adjustment.

3 RESULTS AND DISCUSSION

3.1 Model Performance

The findings of this study demonstrate the high accuracy of classifying various brain diseases using magnetic resonance images from diverse datasets. These results represent significant advancements over existing studies, as shown in Table 2, which compares methodologies, accuracy, and the incorporation of explainable AI across related works. The proposed models ResNet50, DenseNet121, and EfficientNetB1 achieved accuracies of 99.13%, 99.38%, and 99.13%, respectively. When combined in an ensemble approach, the classification accuracy improved to 99.84%, surpassing the performance of earlier studies, such as Shafi et al. (Shafi et al., 2021), which reported an accuracy of 97.957% for classifying multiple sclerosis and tumours. Additionally, integrating Grad-CAM enhanced transparency by providing visual explanations of the model's predictions, building trust for clinical applications.

The confusion matrices shown in Figures 2a, 2b, and 2c illustrate the classification performance of ResNet50, DenseNet121, and EfficientNetB1, respectively. Diagonal elements represent correctly classified instances, while off-diagonal elements indicate

Table 2: Comparison of the proposed model with existing studies.

Author	Classification	Methodology	Accuracy	XAI
(Maqsood et al., 2022)	3 Tumour types	CNN	97.47%, 98.92%	Yes
(Gaur et al., 2022)	3 Tumour types, Healthy	CNN	94.64%	Yes
(Agrawal et al., 2024)	3 Tumour types	CNN-based classifier	96.4%	No
(Shafi et al., 2021)	3 Tumour types, MS	Ensemble classifier	97.957%	No
Proposed Approach	**3 Tumour types, MS, and Healthy**	**Ensemble Learning**	**99.84%**	**Yes**

misclassifications. ResNet50 and EfficientNetB1 performed comparably, achieving high accuracy across all classes. DenseNet121 demonstrated superior performance with fewer misclassifications in overlapping classes, likely due to its deeper architecture, which is adept at capturing finer features. These observations emphasize the complementary strengths of the individual models, which are effectively leveraged in the ensemble to further reduce errors.

The high accuracy achieved by the EL model in this study is encouraging for the application of deep learning models in healthcare, particularly in radiology and neurology. The ability to accurately and efficiently diagnose various brain conditions using MRI scans can significantly assist clinicians in making informed decisions and providing timely treatment. However, the performance of the model in a controlled study may differ when applied in clinical practice, where data variability and patient demographics are more diverse. The study's reliance on specific datasets raises questions about the model's performance across different imaging technologies and patient groups.

3.2 Grad-CAM Visualizations

Grad-CAM visualizations, shown in Figure 3, provide critical insights into the model's decision-making process by highlighting the regions within MRI scans that most significantly influenced its classifications. For tumour cases, the visualizations typically emphasize the boundaries and internal structures, while for multiple sclerosis, they focus on lesion areas that are essential for accurate diagnosis.

While Grad-CAM provides valuable interpretability, the visualizations also reveal certain areas for improvement in the models. For DenseNet121, the heatmaps occasionally highlight regions outside the brain, such as the skull or surrounding areas, rather than the critical brain structures. This misfocus requires refinement to ensure that the model consistently concentrates on relevant anatomical regions for more reliable predictions.

A similar issue is observed in the case of multiple sclerosis, where the heatmaps sometimes emphasize the skull or external regions instead of focusing solely on the lesion areas. This unintended focus may hinder the model's ability to reliably diagnose MS and should be addressed in future iterations to enhance accuracy.

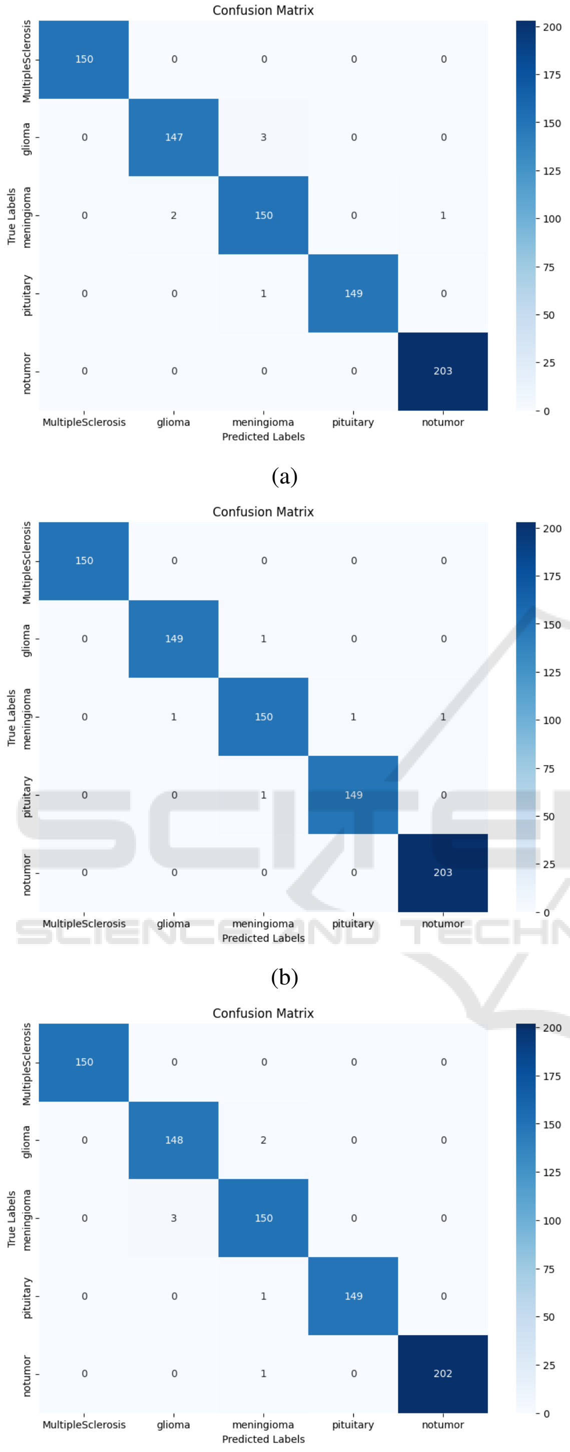
In the case of pituitary tumours, the heatmaps predominantly highlight the tumour boundaries, with less focus on the internal structures of the tumour. While boundary emphasis is important for delineation, a more balanced focus on both the boundaries and the interior regions of the tumour could enhance the model's diagnostic reliability.

By producing these heatmaps that localize important regions in each image, Grad-CAM bridges the gap between deep learning outputs and clinical understanding. These visualizations enable clinicians to validate the AI model's focus areas against established diagnostic criteria, ensuring alignment with medical expertise.

Additionally, Grad-CAM supports the identification of model biases or inaccuracies (Bibi et al., 2024b), such as the aforementioned issues with DenseNet121, MS and pituitary tumour visualizations, highlighting areas that need further refinement. Addressing these limitations is crucial for improving model robustness and ensuring reliable application in clinical workflows. Continued advancements in explainable AI, coupled with thorough validation in real-world settings, are essential to foster widespread adoption and clinical confidence.

4 CONCLUSION

The Ensemble learning model used to classify various brain conditions from MRI scans with a high accuracy of 99.84%. This approach is broader than typical single-disease studies. It successfully identifies different brain tumours, such as glioma, meningioma, and pituitary tumours, and it can also differentiate between brain images of non tumour and multiple sclerosis. A key aspect of the research is the use of Grad-CAM to visualize the MRI areas that influence the predictions of the model, which validated its accuracy. Although promising, the study notes limitations



(a)

(b)

(c)

Figure 2: Confusion matrices for (a) ResNet50, (b) DenseNet121, and (c) EfficientNetB1. Each matrix represents the classification performance of the respective model on the test set.

such as data dependence and potential variability in different clinical settings. The findings point to the potential of deep learning models such as EL to accurately and efficiently diagnose brain conditions, suggesting further tests in various clinical settings.

Future work will involve expanding model training and testing on larger, more diverse datasets, including a wider range of patient demographics and imaging techniques. We will also validate the Grad-CAM results with clinical insights to ensure the model's reliability and applicability in real-world clinical scenarios.

ACKNOWLEDGEMENTS

This research is funded by the Science Foundation Ireland Center for Research Training in Machine Learning (18/CRT/6183).

REFERENCES

Agrawal, T., Choudhary, P., Shankar, A., Singh, P., and Diwakar, M. (2024). Multifenet: Multi-scale feature scaling in deep neural network for the brain tumour classification in mri images. *International Journal of Imaging Systems and Technology*, 34(1):e22956.

Anantharajan, S., Gunasekaran, S., Subramanian, T., and Venkatesh, R. (2024). Mri brain tumor detection using deep learning and machine learning approaches. *Measurement: Sensors*, page 101026.

Arbabshirani, M. R., Fornwalt, B. K., Mongelluzzo, G. J., Suever, J. D., Geise, B. D., Patel, A. A., and Moore, G. J. (2018). Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. *NPJ digital medicine*, 1(1):9.

Bhuvaji, S., Kadam, A., Bhumkar, P., Dedge, S., and Kanchan, S. (2020). Brain tumor classification (mri).

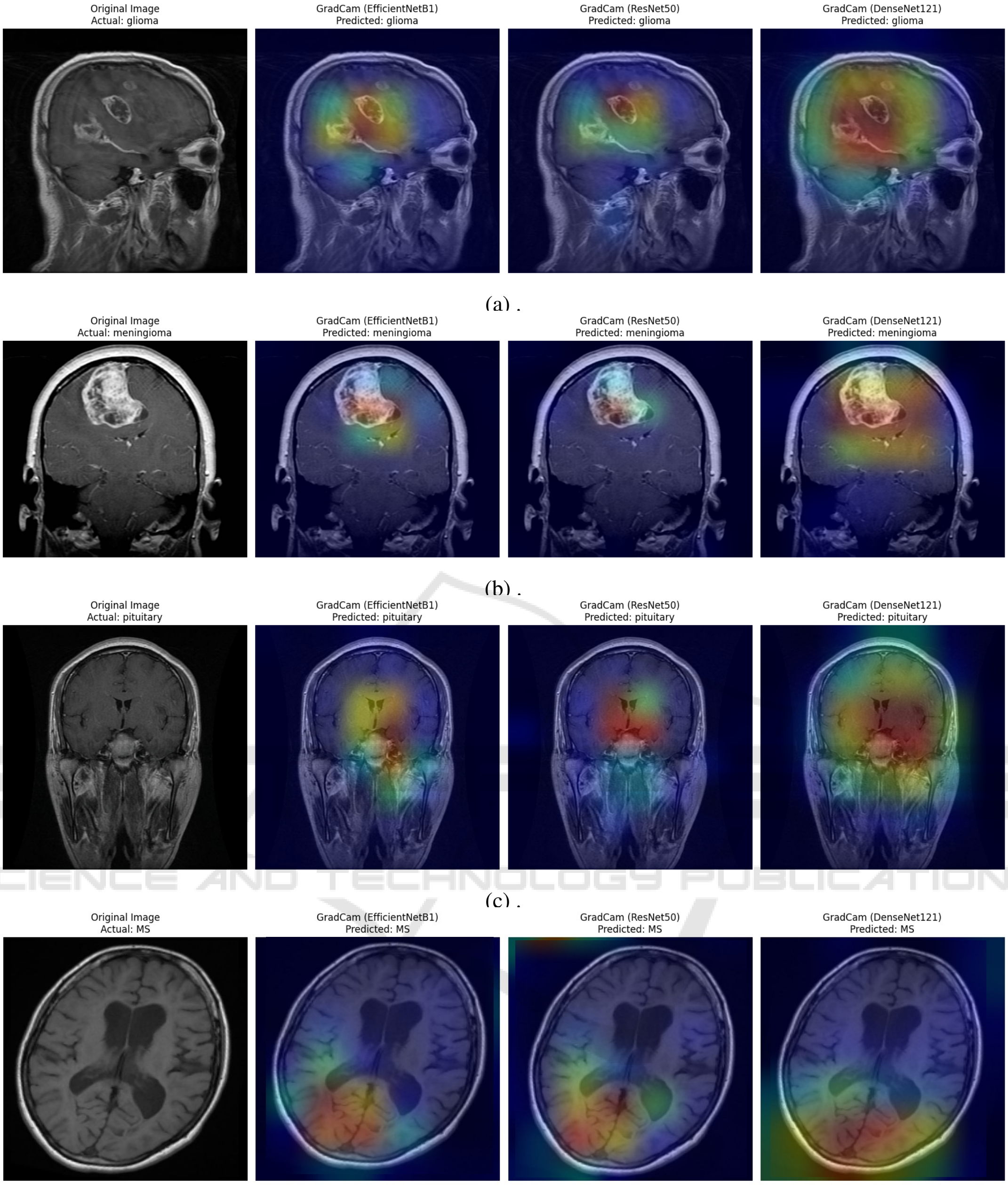
Bibi, N., Courtney, J., and Curran, K. M. (2024a). Enhancing multiple sclerosis diagnosis with explainable ai. In *IET Conference Proceedings CP887*, volume 2024, pages 218–225. IET.

Bibi, N., Courtney, J., and Curran, K. M. (2024b). Multiple sclerosis diagnosis with deep learning and explainable ai.

Bibi, N., Courtney, J., and McGuinness, K. Enhancing brain disease diagnosis with xai: A review of recent studies. *ACM Transactions on Computing for Healthcare*.

Browne, P., Chandraratna, D., Angood, C., Tremlett, H., Baker, C., Taylor, B. V., and Thompson, A. J. (2014). Atlas of multiple sclerosis 2013: a growing global problem with widespread inequity. *Neurology*, 83(11):1022–1024.

Cheng, J. (2017). Brain tumor dataset. *figshare. Dataset*, 1512427(5).



(a) .

(b) .

(c) .

(d) .

Figure 3: Grad-CAM visualizations for (a) Glioma, (b) Meningioma, (c) Pituitary, and (d) Multiple Sclerosis. Each subfigure displays Grad-CAM results from ResNet50, DenseNet121, and EfficientNetB1 for the respective disease.

Dobson, R. and Giovannoni, G. (2019). Multiple sclerosis–a review. *European journal of neurology*, 26(1):27–40.

Ekmekyapar, T. and Taşcı, B. (2023). Exemplar mobilenetv2-based artificial intelligence for robust and accurate diagnosis of multiple sclerosis. *Diagnostics*, 13(19):3030.

Esmaeili, M., Vettukattil, R., Banitalebi, H., Krogh, N. R., and Geitung, J. T. (2021). Explainable artificial intelligence for human-machine interaction in brain tumor localization. *Journal of Personalized Medicine*, 11(11):1213.

Gaur, L., Bhandari, M., Razdan, T., Mallik, S., and Zhao, Z. (2022). Explanation-driven deep learning model for prediction of brain tumour status using mri image data. *Frontiers in Genetics*, page 448.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. corr abs/1512.03385 (2015).

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger,

K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Lopatina, A., Ropele, S., Sibgatulin, R., Reichenbach, J. R., and Güllmar, D. (2020). Investigation of deep-learning-driven identification of multiple sclerosis patients based on susceptibility-weighted images using relevance analysis. *Frontiers in neuroscience*, 14:609468.

Macin, G., Tasci, B., Tasci, I., Faust, O., Barua, P. D., Dogan, S., Tuncer, T., Tan, R.-S., and Acharya, U. R. (2022). An accurate multiple sclerosis detection model based on exemplar multiple parameters local phase quantization: Exmplpq. *Applied Sciences*, 12(10):4920.

Maqsood, S., Damaševičius, R., and Maskeliūnas, R. (2022). Multi-modal brain tumor detection using deep neural network and multiclass svm. *Medicina*, 58(8):1090.

Muslim, A. M., Mashohor, S., Al Gawwam, G., Mahmud, R., binti Hanafi, M., Alnuaimi, O., Josephine, R., and Almutairi, A. D. (2022). Brain mri dataset of multiple sclerosis with consensus manual lesion segmentation and patient meta information. *Data in Brief*, 42:108139.

Nair, P. C., Gupta, D., Devi, B. I., and Kanjirangat, V. (2023). Building an explainable diagnostic classification model for brain tumor using discharge summaries. *Procedia Computer Science*, 218:2058–2070.

Nickparvar, M. (2021). Brain tumor mri dataset.

Reddy, A. C., Akhila, C., Mehdi, M. J., et al. (2023). Exploring multifaceted cnn architectures for enhanced detection of multiple sclerosis. In *2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN)*, pages 483–489. IEEE.

Segato, A., Marzullo, A., Calimeri, F., and De Momi, E. (2020). Artificial intelligence for brain diseases: A systematic review. *APL bioengineering*, 4(4):041503.

Shafi, A., Rahman, M. B., Anwar, T., Halder, R. S., and Kays, H. E. (2021). Classification of brain tumors and auto-immune disease using ensemble learning. *Informatics in Medicine Unlocked*, 24:100608.

Talo, M., Yildirim, O., Baloglu, U. B., Aydin, G., and Acharya, U. R. (2019). Convolutional neural networks for multi-class brain disease detection using mri images. *Computerized Medical Imaging and Graphics*, 78:101673.

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Thompson, A. J., Banwell, B. L., Barkhof, F., Carroll, W. M., Coetzee, T., Comi, G., Correale, J., Fazekas, F., Filippi, M., Freedman, M. S., et al. (2018). Diagnosis of multiple sclerosis: 2017 revisions of the mcdonald criteria. *The Lancet Neurology*, 17(2):162–173.

Yousaf, F., Iqbal, S., Fatima, N., Kousar, T., and Rahim, M. S. M. (2023). Multi-class disease detection using deep learning and human brain medical imaging. *Biomedical Signal Processing and Control*, 85:104875.

Zebari, N. A., Mohammed, C. N., Zebari, D. A., Mohammed, M. A., Zeebaree, D. Q., Marhoon, H. A., Abdulkareem, K. H., Kadry, S., Viriyasitavat, W., Nedoma, J., et al. (2024). A deep learning fusion model for accurate classification of brain tumours in magnetic resonance images. *CAAI Transactions on Intelligence Technology*.

Zhang, Y., Weng, Y., and Lund, J. (2022). Applications of explainable artificial intelligence in diagnosis and surgery. *Diagnostics*, 12(2):237.