big data and cognitive computing

MDPI

Article

Brain Tumor Detection Using Magnetic Resonance Imaging and Convolutional Neural Networks

Rafael Martínez-Del-Río-Ortega [1], Javier Civit-Masot [1,2,3], Francisco Luna-Perejón [1,2,3,4,*] and Manuel Domínguez-Morales [1,2,3,4]

[1] E.T.S. Ingeniería Informática, Universidad de Sevilla, Avda. Reina Mercedes s/n, 41012 Seville, Spain; lmsaavedra@us.es (R.M.-D.-R.-O.); mjavier@us.es (J.C.-M.); mjdominguez@us.es (M.D.-M.)
[2] Robotics and Technology of Computers Research Group (TEP-108), Architecture and Computer Technology Department, E.T.S. Ingeniería Informática, Universidad de Sevilla, Avda. Reina Mercedes s/n, 41012 Seville, Spain
[3] Escuela Politécnica Superior (EPS), Universidad de Sevilla, 41011 Seville, Spain
[4] Computer Engineering Research Institute (I3US), E.T.S. Ingeniería Informática, Universidad de Sevilla, Avda. Reina Mercedes s/n, 41012 Seville, Spain
* Correspondence: fluna1@us.es

Abstract: Early and precise detection of brain tumors is critical for improving clinical outcomes and patient quality of life. This research focused on developing an image classifier using convolutional neural networks (CNN) to detect brain tumors in magnetic resonance imaging (MRI). Brain tumors are a significant cause of morbidity and mortality worldwide, with approximately 300,000 new cases diagnosed annually. Magnetic resonance imaging (MRI) offers excellent spatial resolution and soft tissue contrast, making it indispensable for identifying brain abnormalities. However, accurate interpretation of MRI scans remains challenging, due to human subjectivity and variability in tumor appearance. This study employed CNNs, which have demonstrated exceptional performance in medical image analysis, to address these challenges. Various CNN architectures were implemented and evaluated to optimize brain tumor detection. The best model achieved an accuracy of 97.5%, sensitivity of 99.2%, and binary accuracy of 98.2%, surpassing previous studies. These results underscore the potential of deep learning techniques in clinical applications, significantly enhancing diagnostic accuracy and reliability.

Keywords: brain tumors; MRI; convolutional neural networks; deep learning; image classification; medical imaging

Citation: Martínez-Del-Río-Ortega, R.; Civit-Masot, J.; Luna-Perejón, F.; Domínguez-Morales, M. Brain Tumor Detection Using Magnetic Resonance Imaging and Convolutional Neural Networks. *Big Data Cogn. Comput.* **2024**, *8*, 123. https://doi.org/10.3390/bdcc8090123

Academic Editor: Yoichi Hayashi

Received: 19 August 2024
Revised: 8 September 2024
Accepted: 19 September 2024
Published: 21 September 2024

1. Introduction

The early and precise detection of brain tumors is crucial for improving clinical outcomes and quality of life for patients. Brain tumors represent a significant cause of morbidity and mortality worldwide. According to the World Health Organization (WHO), approximately 300,000 new cases of brain tumors are diagnosed annually, making them one of the leading causes of cancer-related death in children and young adults [1]. The survival rate for brain tumor patients heavily depends on early detection and the effectiveness of the subsequent treatment.

Magnetic resonance imaging (MRI) is an advanced imaging technique that provides excellent spatial resolution and soft tissue contrast, making it an essential tool for identifying brain abnormalities. MRI can non-invasively detect various types of brain tumors and other neurological disorders. However, interpreting MRI scans is a complex task that requires the expertise of highly trained radiologists. This task is prone to errors, due to variability in tumor appearance and to human limitations, such as fatigue and subjectivity [2]. Additionally, in regions with a shortage of specialists, access to accurate and timely diagnosis can be limited, delaying treatment initiation and adversely affecting patient outcomes.

Convolutional neural networks (CNNs) have emerged as a powerful technique in the field of deep learning for image classification. CNNs can learn relevant features directly from image data, eliminating the need for manual feature engineering and enabling higher classification accuracy [3]. These networks have demonstrated outstanding performance in image recognition tasks across various domains, including the detection of pathologies in medical images [4]. In medical imaging, CNNs have shown promise in tasks such as tumor detection, segmentation, and classification, providing a reliable second opinion, helping prioritize urgent cases, and allowing physicians to focus on more complex and critical tasks [5].

The aim of this work was to develop an image classifier based on CNNs for detecting brain tumors in MRI scans, with the goal of supporting medical diagnosis and improving patient outcomes. Using a public MRI dataset, the project involved preprocessing the data to ensure image quality and consistency, designing a specific CNN architecture for brain tumor detection. A grid search method was employed to optimize hyperparameters, ensuring the best-possible model performance.

The remainder of the paper is structured as follows. In the next section, Materials and Methods, the dataset used, the preprocessing steps performed, and the proposed models, along with the hyperparameters considered for performance analysis, are described. Following this, in the Results sections, the metrics obtained are presented and the best models are identified. Subsequently, in the Discussion section, a comparison with recent works addressing the binary problem of brain tumor identification is included, and the limitations, challenges, and future work are outlined. Finally, the conclusions are presented.

2. Materials and Methods

This section provides an overview of the publicly available dataset, which underwent preprocessing for training and testing the classification system. It also describes the classifiers developed and the evaluation metrics used. Figure 1 summarizes the methodology used to achieve these tasks:

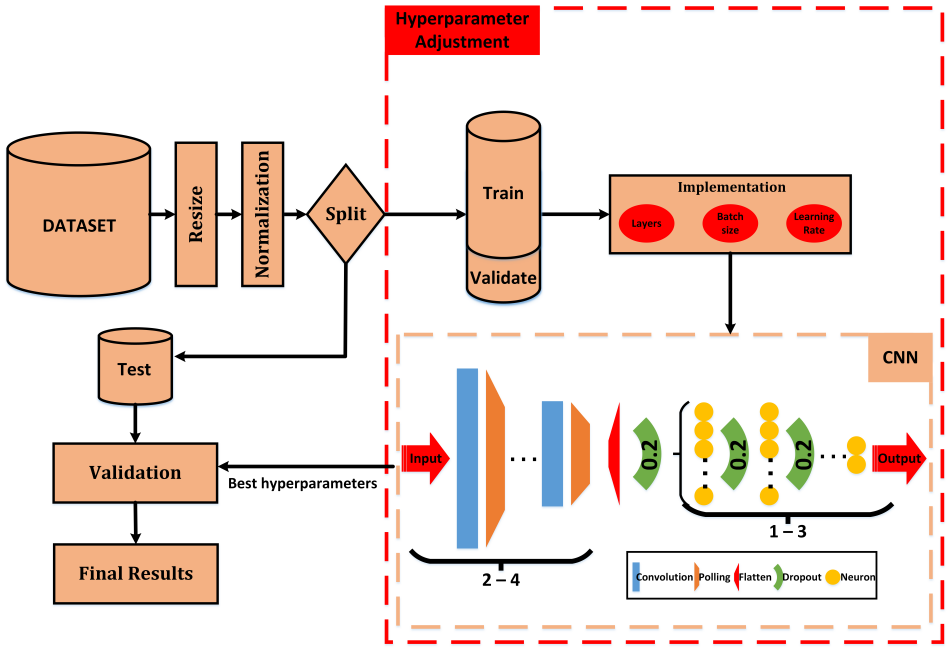

Figure 1. Workflow overview. The dataset is preprocessed (resizing and normalization), split for training and testing, and used to train a convolutional neural network (CNN). Hyperparameters like layers, batch size, and learning rate are optimized during training. The best model is validated and tested to produce the final results.

2.1. Dataset

There are various contributions in the literature providing publicly available datasets for brain tumor identification, each offering different advantages and limitations. One such dataset is the Br35H dataset (https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection, accessed on September 2023), which contains a well-balanced collection of 1500 tumoral and 1500 non-tumoral images. This balance makes it suitable for binary classification tasks; however, its limitations include the high contrast of the images and the fact that all are captured exclusively on the axial plane, which may restrict the model's ability to generalize across different anatomical perspectives. Similarly, the BraTS dataset, widely used for brain tumor segmentation competitions since 2014 [6], compiles over 1000 tumoral images but lacks non-tumoral samples, making it more appropriate for segmentation tasks than binary classification. Another dataset, the IXI dataset (https://brain-development.org/ixi-dataset/, accessed on September 2023), though useful for various MRI analyses, contains only 600 images, which poses a challenge for deep learning models that typically require larger datasets for effective training.

The dataset used in this study is the publicly available Preet Viradiya Brain Tumor Dataset (https://www.kaggle.com/datasets/preetviradiya/brian-tumor-dataset, accessed on September 2023), consisting of brain scans categorized into healthy and tumor-affected groups. This dataset contains a substantial number of images and is well-balanced, with 55% of the images depicting tumors and 45% representing healthy brain scans. Additionally, it includes a diverse range of anatomical planes and image shifts, which is highly advantageous for deep learning models. This diversity helps mitigate the risk of overfitting and reduces the likelihood of the model becoming overly specialized to specific features or patterns. The inclusion of varied imaging perspectives ensures that the model can generalize more effectively to different scenarios, ultimately enhancing its robustness and performance when applied to new, unseen data.

Figure 2 displays representative samples from each class, illustrating not only the variability and characteristics inherent in both the healthy and tumor-affected brain scans but also the impact of this diversity. The presence of different anatomical planes and shifts plays a critical role in improving the model's capacity to adapt and perform reliably across a wide range of clinical cases.

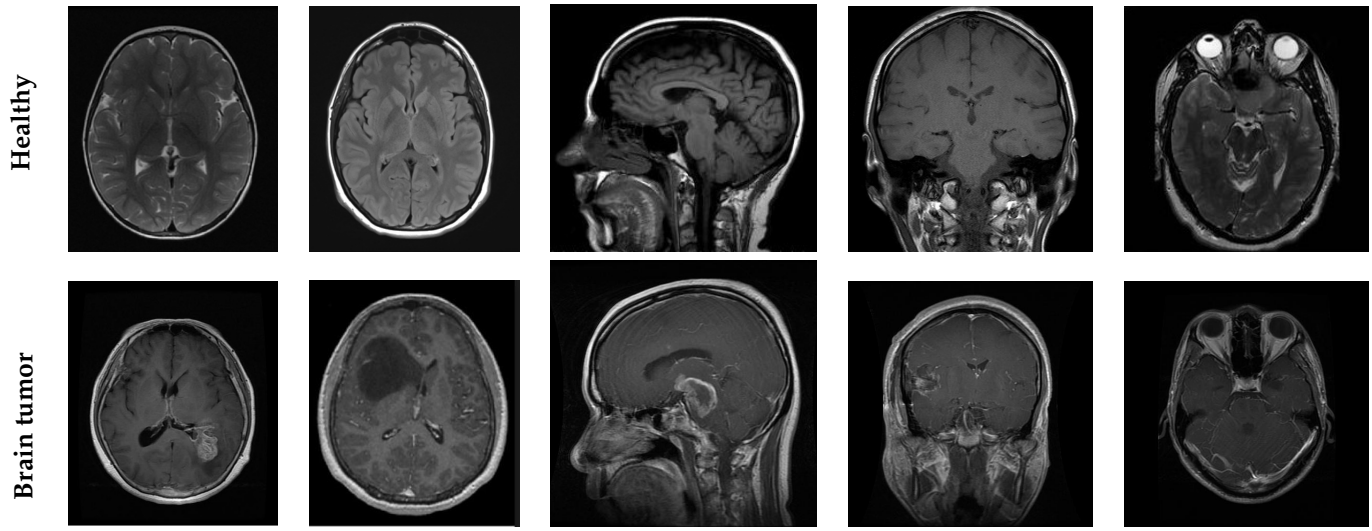

Figure 2. Example MRI images from the dataset used in this study. The top row shows MRI scans of healthy brain tissue, while the bottom row depicts scans with visible brain tumors. Each row includes different views, such as axial, coronal, and sagittal planes, demonstrating the anatomical differences between healthy and pathological cases.

The dataset was split into training, validation, and testing sets to ensure robust model evaluation. The training set included 70% of the data, the validation set 20%, and the testing set 10%. The proportion of samples within each class was maintained consistently across all three subsets, thereby ensuring that the class distribution remained balanced and representative in each subset. The specific distribution of samples within these subsets can be consulted in Table 1:

Table 1. Dataset used in this work and subsets division.

Class	Train (70%)	Validation (20%)	Test (10%)	Total
Healthy	1461	418	208	2087
Brain tumor	1759	502	252	2513
Total	3220	920	460	4600

2.2. Data Preprocessing

Preprocessing steps are essential to enhancing the quality and consistency of MRI images, as highlighted by Akkus et al. [7]. Several key techniques were implemented to prepare the data for effective analysis. Initially, all the images were uniformly scaled to a resolution of 256 × 256 pixels. Furthermore, images exhibiting artifacts within the cranial region were excluded from further analysis, to maintain data integrity. Following this, a min–max global normalization technique was applied, to adjust the pixel intensity values to a common scale ranging between 0 and 1, thereby ensuring uniformity across all images. The smallest pixel value in the dataset was transformed to 0, the largest to 1, and the remaining values were proportionally scaled in between. This step was crucial for mitigating the effects of varying acquisition protocols and for facilitating consistent model performance. Therefore, by applying these global normalization parameters across the entire dataset, the intensity levels remained consistent, even for images acquired from different MRI sources or settings. This step played a critical role in improving uniformity and minimizing biases due to brightness or contrast variations across the scans. Moreover, segmentation techniques were employed to isolate regions of interest (ROIs), specifically focusing on brain tissues while excluding non-relevant parts of the image. This approach enabled the analysis to concentrate on the most critical areas, reducing the influence of extraneous structures and thereby enhancing the accuracy of subsequent tasks, such as classification.

Images exhibiting artifacts within the cranial region were also addressed during preprocessing. Initially, manual inspection was conducted, identifying a total of 20 images with artifacts. Since these artifacts were located outside the regions of interest (ROIs), it was possible to retain these images for training and evaluation by cropping the affected areas. In terms of segmentation, the dataset had already undergone preprocessing to remove parts of the background unrelated to the skull. However, an additional step of edge cropping was performed to further focus on the regions of interest, particularly centering the relevant brain areas. This involved eliminating portions of the neck or jaw in images where they appeared, which further refined the dataset. This extra refinement step not only enhanced the model's ability to focus on the most pertinent areas but also reduced irrelevant background noise, thereby improving the accuracy and efficiency of the training process.

2.3. Model Architecture

A convolutional neural network (CNN) was designed specifically for the task of detecting brain tumors. The architecture consisted of the following:

- Input Layer: Accepting preprocessed MRI images.
- Convolutional Layers: Extracting spatial features through filters applied to the input images.
- Pooling Layers: Reducing the dimensionality of the feature maps while retaining important information.

- Fully Connected Layers: Performing high-level reasoning based on the extracted features.
- Output Layer: Providing binary classification outputs (healthy or tumor).

2.4. Implementation

The development environment used Jupyter Notebook with Python 3.3. TensorFlow and Keras built and trained the CNN. The study was carried out with a grid search for hyperparameter tuning. Next, the optimal models identified through the grid search were subsequently validated using the test subset. For this validation, we considered the average values of three key metrics: accuracy, precision, and sensitivity (recall).

A grid search is a systematic method for exploring hyperparameter space by evaluating all possible combinations of predefined hyperparameter values. While this approach can be computationally intensive, it was chosen due to its thoroughness in identifying the optimal configuration. Given the manageable size of our hyperparameter space, we were able to conduct this search efficiently using available computational resources. Additionally, the use of the grid search allowed us to ensure that every potential combination was tested, providing comprehensive coverage of the hyperparameter space to achieve the best results. The hyperparameters included were learning rate, batch size, number of convolutional and max pooling layers, and batch size. Various combinations of these hyperparameters were tested to identify the best-performing models configurations. The hyperparameters and their values are detailed in Table 2. The selection of values for the grid search was guided both by our expertise and by insights gained from previous studies in diagnostic support through image analysis [8–11]. These studies, which have focused on optimizing similar machine learning models in healthcare applications, provided a foundation for determining appropriate hyperparameter ranges.

Each model was trained using the defined hyperparameter combinations, with the training process involving multiple iterations and adjustments to optimize performance. The grid search was implemented through exhaustive loops to explore all possible combinations of hyperparameters, with the results systematically recorded for comparative analysis. To prevent overfitting, early stopping was applied using a patience parameter of 5 iterations. Specifically, if the target metric on the validation subset did not improve after 5 consecutive iterations, the training was halted, and the model with the best performance within those iterations was selected. Model checkpointing was also employed to ensure that the best-performing model was saved throughout the training process. These measures helped ensure that the models were not overtrained, facilitating the identification of the optimal hyperparameter configurations.

Table 2. Hyperparameters and their values used for grid search.

Hyperparameter	Values
Learning rate	0.1, 0.01, 0.001
Batch size	10, 20, 30
Number of convolutional and max pooling layers	2, 3, 4
Number of dense layers	1, 2, 3

2.5. Evaluation Metrics

The models' performances were evaluated using several metrics:

- Accuracy: The percentage of correctly classified images out of the total images.
- Sensitivity (Recall): The ability of the model to correctly identify positive cases (tumor).
- Specificity: The ability of the model to correctly identify negative cases (no tumor).
- Precision: The proportion of positive identifications that were actually correct.
- F1 Score: The harmonic mean of precision and recall, providing a single metric that balanced both concerns.

These metrics provided a comprehensive assessment of the model's classification capabilities, ensuring that the model not only performed well on average but also effectively distinguished between tumor and non-tumor cases. Accordingly, the high-level metrics are presented in the following equations:

$$Accuracy = \sum_{c} \frac{TP_c + TN_c}{TP_c + FP_c + TN_c + FN_c}, c \in classes \quad (1)$$

$$Specificity = \sum_{c} \frac{TN_c}{TN_c + FP_c}, c \in classes \quad (2)$$

$$Precision = \sum_{c} \frac{TP_c}{TP_c + FP_c}, c \in classes \quad (3)$$

$$Sensitivity = \sum_{c} \frac{TP_c}{TP_c + FN_c}, c \in classes \quad (4)$$

$$F1_{score} = 2 \times \frac{precision \times sensitivity}{precision + sensitivity}. \quad (5)$$

3. Results

The results will be presented systematically, beginning with the grid search results for the validation and testing subsets. Following this, the top five models from all training iterations will be discussed, and detailed performance metrics for these models will be provided.

3.1. Grid Search

During this phase, a total of 81 training sessions were conducted, encompassing all possible combinations of the hyperparameters outlined in the previous section. The results of the validation subset, recorded during the training process, are summarized in Figure 3. Furthermore, the results of the testing subset, evaluated after training, are shown in Figure 4. Supplementary Material Tables S1 and S2 contains detailed tabulated results for further reference.

From a global perspective, it is evident that the models with lower learning rates (e.g., 0.0001) and moderate batch sizes (e.g., 20) tended to achieve a more favorable balance between accuracy and generalization. This characteristic is particularly significant in the context of brain tumor detection, where a model's ability to generalize effectively to new and unseen data is crucial for its clinical applicability.

For instance, a model trained with a learning rate of 0.0001 and a batch size of 20 demonstrated a validation loss of 0.35 and a validation accuracy of 90.5%. This specific combination of parameters not only facilitated improved convergence during the training process but also contributed to the model's stability, enabling more gradual and precise weight updates. Such behavior suggests enhanced generalization to unseen data, which is essential for accurate tumor detection across varying imaging conditions and diverse patient populations.

In contrast, models utilizing higher learning rates (e.g., 0.001) exhibited rapid convergence; however, they often displayed a significant disparity between training accuracy and validation accuracy, indicative of overfitting. Overfitting occurs when a model becomes overly attuned to the training data, capturing noise and idiosyncrasies specific to that dataset rather than learning patterns that are broadly generalizable. Consequently, these models tend to perform poorly when applied to new datasets.

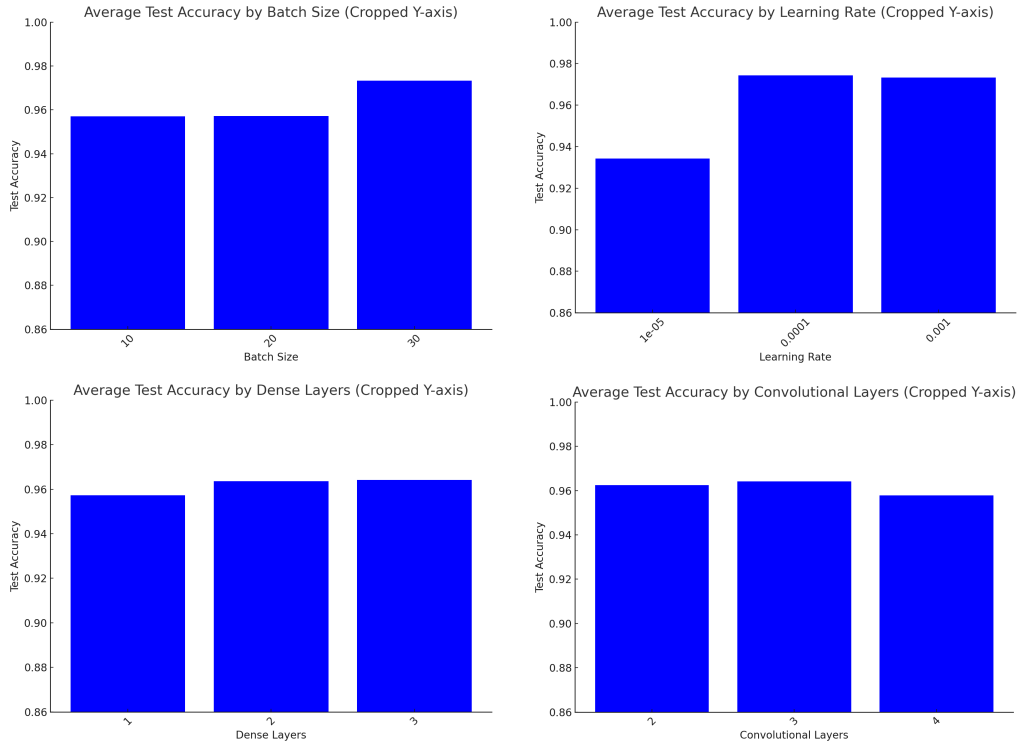

Figure 3. Average validation accuracy results for different hyperparameter settings. These bar plots illustrate the impact of hyperparameter choices on the model's validation accuracy. The top-left plot shows the effect of batch size. The top-right plot presents the influence of learning rate. The bottom-left plot shows the effect of dense layers, while the bottom-right plot reflects the impact of convolutional layers.

The stability observed in models with lower learning rates and moderate batch sizes can be attributed to the fact that these settings allowed the model to make finer adjustments during the training process. A moderate batch size, such as 20, provides an optimal balance between the frequency of weight updates and the stability of those updates. This is particularly critical in deep convolutional neural networks, where abrupt adjustments can lead to destabilization during training and prevent the model from achieving a global optimum.

The top-performing models achieved accuracies exceeding 90%, demonstrating exceptional performance in the task of image classification for brain tumor detection. These results are particularly significant in the clinical setting, where precision and accuracy are critical for ensuring diagnostic confidence. For instance, a model configured with a batch size of 20, a learning rate of 0.0001, three convolutional layers, and two dense layers achieved a precision of 97% and an accuracy of 98%. These metrics underscore the model's ability to consistently make correct predictions, accurately identifying both true positives and true negatives when applied to the test dataset.

This model's configuration enabled it to capture complex and detailed features within the MRI images, which is essential for accurately distinguishing between healthy and tumor tissues. A precision of 97% indicates that the model correctly identified 97% of the positive cases, minimizing the margin of error and reducing the likelihood of false positives. Furthermore, an accuracy of 98% signifies that the model effectively maintained a strong balance in correctly identifying both healthy and tumor states. This is crucial for avoiding misdiagnoses, which could otherwise lead to inappropriate treatments or delays in medical intervention. Such high performance metrics are vital in ensuring that the model can be reliably integrated into clinical practice, where accurate and timely diagnosis is paramount.

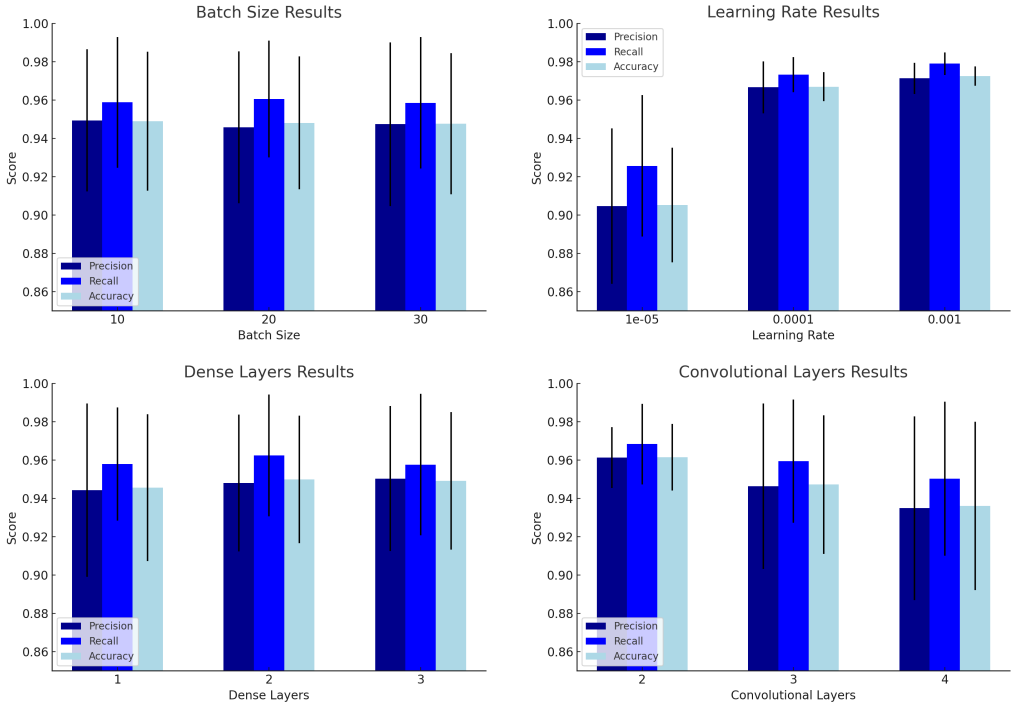

Figure 4. Average metrics results for different hyperparameter settings. These bar plots illustrate the impact of hyperparameter choices on the model's test accuracy, precision, and recall. The top-left plot shows the effect of batch size. The top-right plot presents the influence of learning rate. The bottom-left plot shows the effect of dense layers, while the bottom-right plot reflects the impact of convolutional layers.

Moreover, the precision and sensitivity metrics provided a more nuanced understanding of the models' performances. For instance, a model configured with a batch size of 10, a learning rate of 0.001, four convolutional layers, and two dense layers achieved a precision of 96% and a sensitivity of 98%. The 96% precision indicates the model's capability of accurately classifying images as positive, thereby minimizing the occurrence of false positives. On the other hand, the sensitivity of 98% demonstrates the model's effectiveness in correctly identifying images that actually contained tumors, thus reducing the likelihood of false negatives. This balance between precision and sensitivity is critical, as it ensures that the model is not only accurate in its positive predictions but also reliable in detecting all tumor cases.

This model, with its specific configuration, successfully achieved a balance between identifying positive cases and minimizing false positives. This balance is particularly significant in medical applications, where false positives can lead to unnecessary invasive procedures and increased patient anxiety, while false negatives may result in the failure to provide timely treatment for patients with brain tumors. The model's ability to maintain high sensitivity ensures that the vast majority of tumor cases are detected, which is essential for effective disease management and treatment.

3.2. Searching for the Best Models

The models with the best accuracy scores stood out for their ability to consistently make correct predictions. Table 3 shows the results of the models with the highest accuracy.

These models demonstrated consistent performance, with accuracies exceeding 95%, underscoring their robustness and reliability in classifying magnetic resonance images for brain tumor detection. These high accuracy rates are particularly significant, as they indicate the models' ability to effectively differentiate between images of healthy brains and those with tumors, thereby minimizing both false positives and false negatives.

Table 3. List of the five models with the highest accuracy. The columns represent the batch size (BS), learning rate (LR), number of convolutional layers (nCL), number of dense layers (nDL), precision, recall, and accuracy.

BS	LR	nCL	nDL	Precision	Recall	Accuracy
30	0.001	3	1	0.97510	0.99156	0.98214
10	0.001	2	3	0.97983	0.98380	0.97991
30	0.001	2	3	0.98393	0.98000	0.97991
10	0.0001	2	3	0.97983	0.97983	0.97767
20	0.001	3	2	0.97580	0.98373	0.97767

Additionally, these models generally employed moderate batch sizes and medium learning rates, which suggests an optimal balance between training efficiency and model stability. Moderate batch sizes, such as 20 or 30, enabled efficient training by processing a sufficient number of samples in each iteration to develop a robust representation of the data, while also minimizing the introduction of excessive noise. This approach helped stabilize model weight updates, allowing for gradual and controlled adjustments that prevent abrupt oscillations that could destabilize the training process.

Medium learning rates, such as 0.0001 or 0.001, also played a crucial role in the performance of these models. A medium learning rate ensures that the model makes meaningful progress during each training iteration without making overly large adjustments that could lead to overfitting. This balance is essential for maintaining both the accuracy and generalizability of the model. The use of these learning rates allowed the models to effectively adapt to the training data, thereby enhancing their ability to generalize to new test data.

The models with the best precision scores excelled in minimizing false positives. Table 4 presents the five models with the highest precision.

Table 4. List of the five models that achieved the best precision results.

BS	LR	nCL	nDL	Precision	Recall	Accuracy
10	0.001	3	3	0.98804	0.96875	0.97544
30	0.0001	3	1	0.98770	0.96787	0.97544
20	0.001	4	1	0.98412	0.97637	0.97767
30	0.001	2	3	0.98393	0.98000	0.97991
20	0.0001	2	1	0.98031	0.98031	0.97767

The model configured with three convolutional layers and three dense layers achieved an impressive accuracy of 98.8%, underscoring its remarkable efficiency in correctly identifying brain tumor cases. This configuration enabled the model to extract intricate and complex features from the MRI images, enhancing its ability to distinguish between healthy and tumor tissues. The three convolutional layers were instrumental in capturing varying levels of image features, ranging from basic edges and textures in the initial layers to more complex and tumor-specific patterns in the deeper layers. The subsequent three dense layers consolidated this information, enabling the model to make precise final classification decisions.

In addition to their high accuracy, these models employed configurations that prioritized high specificity, a critical factor in clinical applications. High specificity indicates a low false-positive rate, ensuring that the model accurately identifies images that do not contain tumors. In a clinical context, this is particularly important, as false positives can lead to misdiagnosis, causing unnecessary anxiety for patients and potentially resulting in unwarranted invasive procedures.

The models' ability to minimize false positives enhances confidence in their predictions, ensuring that only cases with a high probability of tumor presence are flagged for further diagnosis and treatment.

Collectively, these models demonstrate that well-optimized configurations can achieve exceptional levels of accuracy and specificity, which are crucial for clinical applications. These models are particularly valuable in the context of brain tumor detection, where each classification decision carries significant implications for patient health. The high specificity of these models ensures that patients are not subjected to unnecessary procedures due to false positives, while their high accuracy ensures that true-positive cases are identified and treated promptly.

The models with the highest sensitivity scores were exceptionally effective in identifying true positives. Table 5 presents the top five models based on this metric, highlighting their effectiveness in accurately detecting tumor cases:

Table 5. List of the five models that achieved the best recall results.

BS	LR	nCL	nDL	Precision	Recall	Accuracy
30	0.001	3	1	0.97510	0.99156	0.98214
20	0.001	2	2	0.96551	0.98823	0.97321
10	0.001	4	3	0.95703	0.98790	0.96875
20	0.0001	3	3	0.92217	0.98750	0.94866
10	0.0001	2	1	0.96311	0.98739	0.97321

The model with three convolutional layers and one dense layer achieved an impressive sensitivity of 99%, demonstrating its exceptional ability to detect the majority of positive cases of brain tumors in MRI images. This specific configuration enabled the model to extract a vast array of detailed and complex features from the images, capturing both fine patterns and more abstract structures that were indicative of the presence of tumors. The convolutional layers were instrumental in identifying and amplifying different levels of image features, ranging from edges and textures to more complex structures, while the dense layer effectively integrated this information to make accurate and well-informed decisions in the final classification.

The high sensitivity of this model indicates its remarkable effectiveness in detecting true positives, meaning it accurately identifies images that actually contain tumors. This is particularly crucial in the clinical setting, where missing a tumor can have severe consequences for a patient's health. The ability to detect nearly all positive cases ensures that patients with brain tumors receive the correct diagnosis and timely treatment, significantly improving the chances of early intervention and potentially saving lives.

Importantly, these models often employed settings that favored positive detection, even at the expense of a higher number of false positives. This strategy of prioritizing sensitivity over accuracy may be advantageous in the medical field, where erring on the side of caution is preferable. In other words, it is more acceptable to have false positives that can be ruled out through additional testing than to miss a positive case, which could result in a critical omission of treatment. The priority of these models is to ensure that no case of brain tumor goes undetected, which is vital for patient safety and well-being.

The emphasis on high sensitivity also aligns with an early-detection strategy, where identifying potential brain tumor cases as early as possible can significantly impact treatment outcomes. The ability of these models to detect tumors at early stages increases the likelihood of early intervention and expands treatment options, which is crucial for improving the long-term prognosis of patients.

The models that achieved the best combination of precision, sensitivity, and accuracy represented an optimal balance across all evaluated metrics. Table 6 below presents the results of the five models with the highest average percentages across these three key metrics, highlighting their overall performance and reliability in clinical applications.

From a comprehensive perspective, the model configured with a batch size of 30, a learning rate of 0.001, three convolutional layers, and one dense layer emerged as the top performer, achieving an average score of 98.3%. This model demonstrated a precision of 97.5%, a sensitivity of 99.2%, and an accuracy of 98.2%. This specific configuration

enabled the model to capture detailed and complex image features, enhancing its ability to distinguish between healthy and tumor tissues. High sensitivity ensures that nearly all tumor cases are detected, while high accuracy minimizes false positives—an essential consideration in clinical settings, to prevent misdiagnosis and avoid unnecessary procedures.

Table 6. Five models with the best combination of accuray, precision and recall.

BS	LR	nCL	nDL	Precision	Recall	Accuracy	Metrics Average
30	0.001	3	1	0.97510	0.99156	0.98214	0.98292
30	0.001	2	3	0.98393	0.98000	0.97991	0.98128
10	0.001	2	3	0.97983	0.98380	0.97991	0.98118
20	0.0001	2	1	0.98031	0.98031	0.97767	0.97943
20	0.001	4	1	0.98412	0.97637	0.97767	0.97939

The second-best model, configured with a batch size of 30, a learning rate of 0.001, two convolutional layers, and three dense layers, achieved an average score of 98.1%. This model also exhibited high levels of precision (98.4%) and sensitivity (98.0%), reflecting its ability to accurately identify both positive and negative cases. The combination of two convolutional layers and three dense layers provided an optimal balance between feature extraction and precise classification, resulting in excellent generalization to unseen data.

The third model, configured with a batch size of 10, a learning rate of 0.001, two convolutional layers, and three dense layers, attained an average score of 98.1%, with a precision of 97.9% and a sensitivity of 98.4%. Despite the smaller batch size, this configuration maintained high generalizability and accuracy in brain tumor detection. The smaller batch size allowed for more frequent weight updates, which may have contributed to more stable and faster convergence during training.

The fourth model, with a batch size of 20, a learning rate of 0.0001, two convolutional layers, and one dense layer, obtained an average score of 97.9%. This model demonstrated a precision of 98.0% and a sensitivity of 98.0%, indicating a well-balanced ability to detect positive cases while minimizing false positives. The lower learning rate allowed for finer adjustments of the weights during training, which likely contributed to the model's stability and accuracy.

The fifth model, configured with a batch size of 20, a learning rate of 0.001, four convolutional layers, and one dense layer, also achieved an average score of 97.9%. This model displayed a precision of 98.4% and a sensitivity of 97.6%, reflecting its ability to handle complex features through a deeper network architecture. The four convolutional layers facilitated extensive feature extraction, while the final dense layer consolidated this information for accurate classification decisions.

The confusion matrices depicted in Figure 5 provide a detailed visualization of the performance of these five best models, allowing for an examination of true positives, true negatives, false positives, and false negatives. These matrices were instrumental in understanding the specific strengths and weaknesses of each model in terms of their classification capabilities.

The first model demonstrated 233 true negatives, 211 true positives, two false negatives, and two false positives, reflecting its excellent ability to minimize errors while maintaining high accuracy and sensitivity. The second model showed 224 true negatives, 218 true positives, three false negatives, and three false positives, also exhibiting a strong balance between accuracy and sensitivity.

The third model achieved 231 true negatives and 214 true positives, with only one false positive and two false negatives, standing out for its notably low false positive rate. The fourth model recorded 219 true negatives, 223 true positives, two false negatives, and four false positives, indicating its strong performance in tumor detection.

Finally, the fifth model demonstrated 219 true negatives and 224 true positives, with three false negatives and two false positives, reflecting its high accuracy and sensitivity.

After a detailed analysis of the training results and model evaluations, including a thorough examination of the confusion matrices, it can be concluded that the model configured with a batch size of 30, a learning rate of 0.001, three convolutional layers, and one dense layer was the most optimal final model. This model excelled in minimizing errors—both false positives and false negatives—while maintaining an optimal balance across all key metrics.

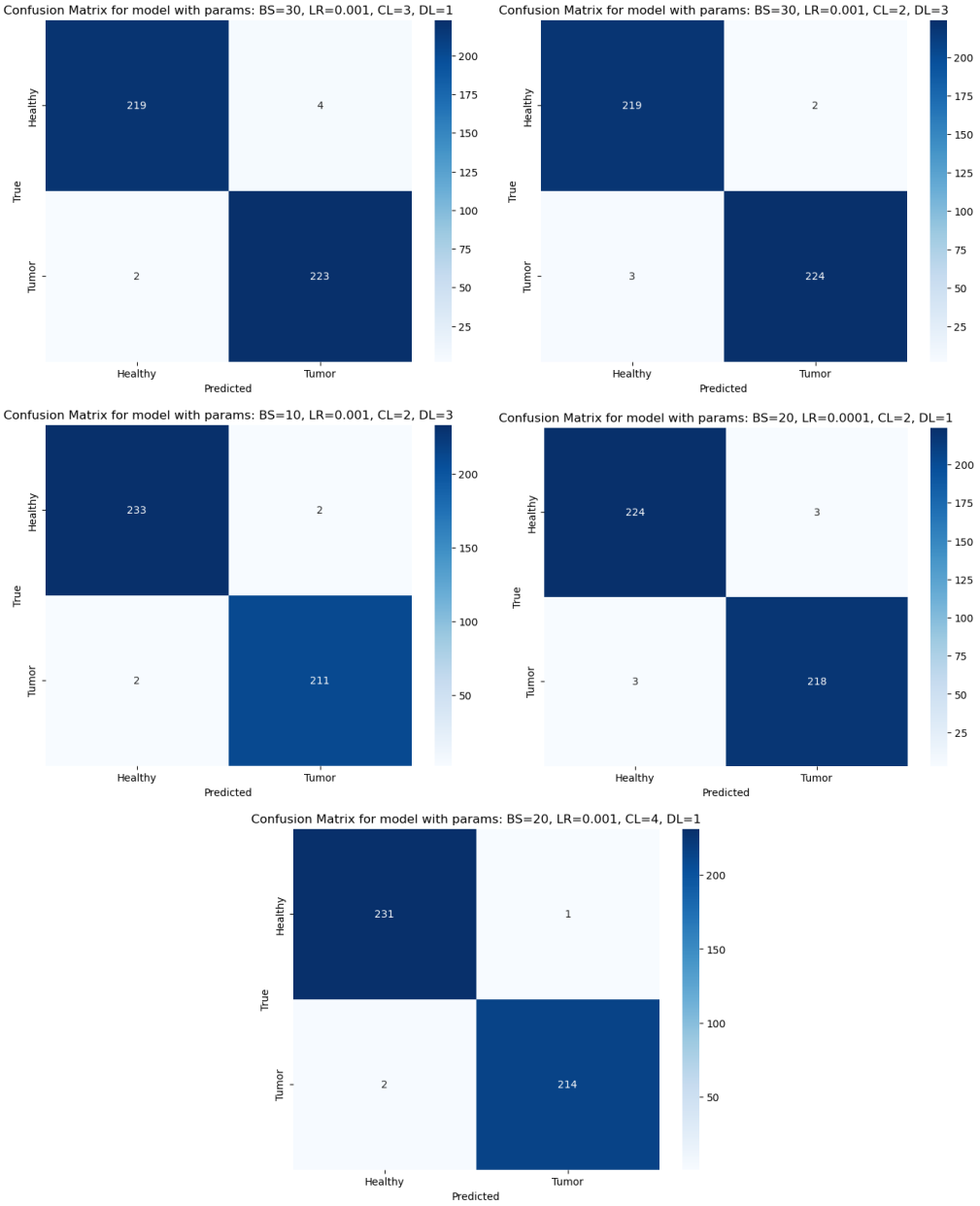

Figure 5. Confusion matrices of the five best-performing models based on their average scores. This figure displays the confusion matrices for the five top models, each showing the true-positive, false-positive, true-negative, and false-negative counts for healthy and tumor predictions. The darker shades in the diagonal indicate higher numbers of correct predictions.

The model's high accuracy of 97.5% and sensitivity of 99.2% highlight its ability to accurately identify both positive and negative cases, which is crucial for clinical applications where diagnostic precision is essential. Furthermore, the model's low false-positive rate minimizes the risk of patients undergoing unnecessary procedures due to incorrect diag-

noses. The combination of performance metrics, including precision, sensitivity, and binary accuracy, underscores the model's robustness and reliability for brain tumor detection.

To further validate its performance, the model was tested on an additional external dataset, specifically the Brain MRI Scan Images from the RSNA-MICCAI Brain Tumor competition (https://www.rsna.org/rsnai/ai-image-challenge/brain-tumor-ai-challenge-2021), consisting of 227 images. The model achieved accuracy greater than 90%; while slightly lower than the results obtained with the main test subset, this still reinforced the model's robustness and generalizability in practical, real-world scenarios. The slight performance decrease suggests some dependence on the primary training dataset, but the results remain high to consider a strong generalizability.

3.3. Execution Times

The study was conducted using the resources freely provided by Google Colab, specifically utilizing a T4 GPU for training. For the best-performing model, which consisted of three convolutional layers and one dense layer, we observed an increase in efficiency as the batch size grew. The average processing time was 3.5 ms per image for batches of 10 images, 2.9 ms for batches of 20, and 2.5 ms for batches of 30. However, when performing inference on a single image, the time increased to 10 ms per image. This indicates that in real-world clinical settings, batch processing would be a more efficient use of available resources, particularly when using a GPU.

We also conducted inference tests using a CPU, specifically an Intel Xeon Broadwell series processor (2.20 GHz) with two to four cores available. In this case, batch size did not have a significant effect on processing times, and the average inference time was approximately 350 ms.

4. Discussion

4.1. Clinical Implications

The integration of convolutional neural networks (CNNs) into MRI analysis holds the potential to significantly improve the diagnostic process by providing accurate and rapid second opinions. This could help prioritize urgent cases and allow radiologists to focus on more complex diagnoses, ultimately enhancing patient care and outcomes. The proposed CNN-based approach demonstrated high precision (97.5%), sensitivity (99.2%), and accuracy (98.2%) on the main test subset. Additionally, when evaluated on an external dataset, the model maintained an accuracy greater than 90%, underscoring its robustness and adaptability to new data. While the slight decrease in performance compared to the primary dataset suggests some dependence on the original data, the model remains a highly effective tool for real-world applications. The consistent performance across different datasets further reinforces its potential as a reliable screening tool, particularly in large-scale diagnostic efforts.

These systems are not intended to replace healthcare professionals but rather to serve as a support tool in assisting with large-scale population screening. The primary clinical benefit lies in the significant reduction in response times for professional medical reports. Moreover, such tools could expand access to diagnostic tests in remote or underserved areas where healthcare facilities are limited. In these cases, individuals would only be referred to hospitals if the system detects potential signs of malignancy, improving both efficiency and resource allocation in healthcare.

In addition to performance metrics, computational resources and inference times are critical factors in determining the model's feasibility in real-world clinical settings. The study revealed that batch processing, particularly when using a GPU, can significantly reduce image inference time. In practical environments, processing multiple images simultaneously could maximize resource efficiency, with batch sizes of 30 images yielding the shortest processing time of 2.5 milliseconds per image on a GPU. This speed makes the CNN-based approach viable in resource-constrained yet high-volume healthcare systems, where rapid image analysis is essential.

In settings where only CPU resources are available, the model's performance remains acceptable, although with slower inference times of approximately 350 milliseconds per image. While this is a more pronounced delay compared to GPU-based processing, it is still within a reasonable range for use in diagnostic workflows. Thus, in healthcare systems with limited access to advanced computational infrastructure, CPU-based inference offers a cost-effective alternative, ensuring the model's utility in a wide range of clinical environments.

4.2. Comparison with Similar Studies

In this section, a comprehensive comparison is made between the results obtained in this study and the results reported in previous relevant work. This comparison allows us to contextualize the performance of the models developed and justify the differences in the results obtained.

Several studies have explored the application of CNNs for brain tumor detection in MRI images. A comparison of the previous works to this work is presented in Table 7:

Table 7. Comparison with previous works.

Work	Year	Classifier	Accuracy	Recall
Zahoor et al. [12]	2024	Res-BRNet (CNN)	98.22%	98.11%
Mandle et al. [13]	2022	VGG-19 CNN	99.83%	97.82–98.87%
Zeineldin et al. [14]	2022	Multimodal CNN	94.2%	92.5%
Xue et al. [15]	2022	Multimodal CNN	93.0%	91.0%
This Work	2024	Optimized CNN	98.2%	99.2%

Both Zahoor et al. [12] and Mandle et al. [13] (2022) presented deep learning approaches that achieved competitive results in brain tumor classification using MRI images. Zahoor et al. proposed the Res-BRNet, a novel architecture that combined regional and residual blocks, achieving accuracy of 98.22% and recall of 98.11%. Similarly, Mandle et al.'s model, based on the VGG-19 architecture, reported impressive accuracy of 99.83%, with recall rates of 97.82% to 98.87% for different tumor types. These results are very close and demonstrate the effectiveness of both models in this task. However, a key distinction between these approaches and the optimized CNN proposed in this work is the complexity of the architectures. Both of the abovementioned studies considered deep, multi-layered models that, while powerful, demanded substantial computational resources, due to their depth and the number of operations required for feature extraction and classification. In contrast, the optimized CNN in this work achieved comparable or superior performance with a less complex architecture, reducing the computational burden. This makes the model more suitable for environments where computational resources are limited, offering a practical balance between accuracy and efficiency.

In the work by Zeineldin et al. [14] the authors achieved accuracy of 94.2% and sensitivity of 92.5% in the detection of brain tumors. They used a multimodal convolutional neural network integrating different MRI imaging modalities. In comparison, the best model of the present project achieved accuracy of 97.5% and sensitivity of 99.2%, significantly surpassing the results of Zeineldin et al. This difference can be attributed to the simplicity and efficiency of the optimized architecture, as well as the careful selection and tuning of hyperparameters. In addition, the use of advanced preprocessing techniques in the present study may have contributed to improving the quality of the input images, which, in turn, improved the performance of the model.

Xue et al. [15] reported average accuracy of 93.0% and sensitivity of 91.0% in several studies. The studies encompassed multiple approaches, including deep convolutional networks and multimodal machine learning techniques. The results obtained in this study, with accuracy of 97.5% and sensitivity of 99.2%, are superior. This may be due to the specificity of the dataset used and the advanced preprocessing and optimization techniques applied. The use of clean and well-scaled images allowed the models to better learn the relevant features for tumor detection.

The models developed in this study exhibited excellent performance, in terms of both accuracy and sensitivity. Their superiority in these critical metrics, along with the ability to generalize effectively to external datasets and maintain reasonable execution times suitable for healthcare systems, highlights the effectiveness of the proposed models. This suggests their strong potential for real-world clinical implementation, where they could provide precise and dependable diagnoses, ultimately contributing to improved treatment outcomes for brain tumor patients.

4.3. Challenges and Limitations

One of the main challenges encountered was dealing with the heterogeneity of tumor appearances and the quality of MRI images. Advanced preprocessing techniques, including data augmentation, were essential to addressing these issues and improving model robustness. Additionally, the grid search method for hyperparameter tuning was computationally intensive but necessary, to achieve optimal model performance.

Despite the advantages of the proposed system, several challenges must be addressed for successful clinical deployment. One major challenge is integrating the system into existing workflows, which rely on diverse imaging systems and software platforms. Compatibility with medical imaging software, PACS, and RIS is crucial, requiring collaboration with IT teams and vendors to ensure seamless adoption through standardized interfaces or custom APIs. Another challenge is obtaining regulatory approval, as the system must meet the strict safety and efficacy standards set by regulatory bodies like the FDA and EMA. This process involves extensive validation and clinical trials, which can be costly and time-consuming. The evolving regulatory frameworks for AI in healthcare add further complexity. Human factors must also be considered, as radiologists and healthcare professionals need proper training to use the tool effectively. The system should complement their expertise, and resistance to new technologies or over-reliance on AI could hinder adoption. Ensuring user-friendly interfaces and clear explanations of the AI's decision-making process is essential for acceptance.

Additionally, there are limitations within the dataset used in this study that could introduce bias. Specifically, the dataset presents an imbalance in the proportion of images captured from different anatomical planes, with a higher percentage taken from the axial plane compared to other views. This imbalance may affect the model's performance by leading it to favor certain perspectives over others, potentially reducing its generalizability. Furthermore, the dataset lacks demographic data, meaning that critical factors, such as age and gender distribution, might be unbalanced. Without this information, it cannot be determinate whether the model's predictions are equally effective across different age groups or genders, which is a key consideration in medical applications. Addressing these potential biases would require further refinement of the dataset or the addition of supplemental data to ensure more equitable model training and evaluation across different population subsets and imaging orientations.

4.4. Future Work

Future research should focus on expanding the dataset to include a more diverse range of MRI images and exploring more advanced network architectures, such as combining CNNs with recurrent neural networks (RNNs) for temporal analysis of sequential imaging data. Additionally, incorporating data augmentation techniques and integrating other clinical information, such as patient demographics and clinical history, could further enhance the model's performance and applicability. Investigating the use of ensemble methods, where multiple models are combined to improve prediction accuracy and robustness, could also be beneficial.

Another promising direction is the application of explainable AI (xAI) techniques to interpret and visualize the decision-making process of convolutional neural networks (CNNs). Techniques like Grad-CAM and saliency maps can highlight the areas of medical images influencing the model's predictions, offering valuable insights to clinicians. This

transparency enables healthcare professionals to verify the relevance of the model's focus, increasing trust and aiding in clinical decision making. Studies have demonstrated that xAI can improve model adoption in healthcare by enhancing interpretability and identifying potential biases or errors [16].

5. Conclusions

This study highlights the effectiveness of a well-structured CNN in accurately detecting brain tumors from MRI scans, with performance that matches or exceeds those reported in similar studies, both in terms of accuracy and execution times. By employing advanced preprocessing methods, comprehensive hyperparameter optimization, and rigorous evaluation metrics, our model offers a scalable, efficient, and precise solution for brain tumor detection, enhancing diagnostic workflows. Its potential to facilitate large-scale population screening, especially in remote or underserved regions, while maintaining high accuracy, underscores its practicality in contemporary healthcare settings. Future improvements, such as incorporating more diverse datasets, adopting advanced architectures, and integrating explainable AI (xAI) techniques to provide interpretable results, could further enhance the abilities of CNNs in medical imaging.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/bdcc8090123/s1, Table S1: Results obtained in the Grid Search for the training and validation subsets; Table S2. Detailed results obtained after the Grid Search for the testing subset.

Author Contributions: Conceptualization: J.C.-M. and M.D.-M.; methodology: F.L.-P. and M.D.-M.; software: R.M.-D.-R.-O.; validation: J.C.-M. and F.L.-P.; formal analysis: M.D.-M.; investigation and resources: J.C.-M. and M.D.-M.; data curation: R.M.-D.-R.-O.; writing—original draft preparation: F.L.-P. and M.D.-M.; writing—review and editing: R.M.-D.-R.-O.; visualization: F.L.-P.; supervision: J.C.-M. and M.D.-M.; project administration: M.D.-M.; funding acquisition: M.D.-M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data and code associated with this study are available upon request by contacting the corresponding author.

Acknowledgments: We want to thank the research group "TEP108—Robotics and Computer Technology" from University of Seville (Spain).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Osborn, A.; Louis, D.; Poussaint, T.; Linscott, L.; Salzman, K. The 2021 World Health Organization classification of tumors of the central nervous system: What neuroradiologists need to know. *Am. J. Neuroradiol.* **2022**, *43*, 928–937. [CrossRef] [PubMed]
2. Bi, W.L.; Hosny, A.; Schabath, M.B.; Giger, M.L.; Birkbak, N.J.; Mehrtash, A.; Allison, T.; Arnaout, O.; Abbosh, C.; Dunn, I.F.; et al. Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA Cancer J. Clin.* **2019**, *69*, 127–157. [CrossRef] [PubMed]
3. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
4. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 84–90. [CrossRef]
5. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef] [PubMed]
6. Menze, B.H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; Burren, Y.; Porz, N.; Slotboom, J.; Wiest, R.; et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* **2014**, *34*, 1993–2024. [CrossRef] [PubMed]
7. Akkus, Z.; Galimzianova, A.; Hoogi, A.; Rubin, D.L.; Erickson, B.J. Deep learning for brain MRI segmentation: State of the art and future directions. *J. Digit. Imaging* **2017**, *30*, 449–459. [CrossRef] [PubMed]
8. Roth, H.R.; Lu, L.; Liu, J.; Yao, J.; Seff, A.; Cherry, K.; Kim, L.; Summers, R.M. Efficient false positive reduction in computer-aided detection using convolutional neural networks and random view aggregation. In *Deep Learning and Convolutional Neural Networks for Medical Image Computing: Precision Medicine, High Performance and Large-Scale Datasets*; Springer: Cham, Switzerland, 2017; pp. 35–48.

9. Muñoz-Saavedra, L.; Civit-Masot, J.; Luna-Perejón, F.; Domínguez-Morales, M.; Civit, A. Does two-class training extract real features? a COVID-19 case study. *Appl. Sci.* **2021**, *11*, 1424. [CrossRef]
10. Karthik, R.; Menaka, R.; Kathiresan, G.; Anirudh, M.; Nagharjun, M. Gaussian dropout based stacked ensemble CNN for classification of breast tumor in ultrasound images. *IRBM* **2022**, *43*, 715–733. [CrossRef]
11. Gago-Fabero, Á.; Muñoz-Saavedra, L.; Civit-Masot, J.; Luna-Perejón, F.; Rodríguez Corral, J.M.; Domínguez-Morales, M. Diagnosis Aid System for Colorectal Cancer Using Low Computational Cost Deep Learning Architectures. *Electronics* **2024**, *13*, 2248. [CrossRef]
12. Zahoor, M.M.; Khan, S.H.; Alahmadi, T.J.; Alsahfi, T.; Mazroa, A.S.A.; Sakr, H.A.; Alqahtani, S.; Albanyan, A.; Alshemaimri, B.K. Brain tumor MRI classification using a novel deep residual and regional CNN. *Biomedicines* **2024**, *12*, 1395. [CrossRef] [PubMed]
13. Mandle, A.K.; Sahu, S.P.; Gupta, G.P. CNN-based deep learning technique for the brain tumor identification and classification in MRI images. *Int. J. Softw. Sci. Comput. Intell. (IJSSCI)* **2022**, *14*, 1–20. [CrossRef]
14. Zeineldin, R.A.; Karar, M.E.; Burgert, O.; Mathis-Ullrich, F. Multimodal CNN networks for brain tumor segmentation in MRI: A BraTS 2022 challenge solution. In Proceedings of the International MICCAI Brainlesion Workshop, Singapore, 18 September 2022; Springer: Berlin, Germany, 2022; pp. 127–137.
15. Xue, J.; Yao, Y.; Teng, Y. Multi-modal Tumor Segmentation Methods Based on Deep Learning: A Narrative Review. In *Quantitative Imaging in Medicine and Surgery*; AME Publishing Company: Hong Kong, China, 2022.
16. Pawar, U.; O'Shea, D.; Rea, S.; O'Reilly, R. Incorporating Explainable Artificial Intelligence (XAI) to aid the Understanding of Machine Learning in the Healthcare Domain. In Proceedings of the AICS, Dublin, Ireland, 7–8 December 2020; pp. 169–180.