

Improving Brain Tumor Classification in MRI Through Hybrid Data Preprocessing and Random Forests

DJAFRI Laouni. Author, Laboratory of Informatics and Mathematics (LIM), University of Tiare, louni.djafri@univ-tiaret.dz

GAFOUR Yacine. Author, Laboratory of Informatics and Mathematics (LIM), University of Tiare, yacine.gafour@univ-tiaret.dz

Mohammed Hoceini, Research Laboratory in Artificial Intelligence and System (LRIAS) , University of Tiare, Mohammed.hoceini@univ-tiaret.dz

HORRI Moura, Research Laboratory in Artificial Intelligence and System (LRIAS) , University of Tiare, Mourad.Horri@univ-tiaret.dz

SILARBI Ibrahim, , Research Laboratory in Artificial Intelligence and System (LRIAS) , University of Tiare, ibrahim.silarbi@univ-tiaret.dz

Abstract

Brain tumor classification using Magnetic Resonance Imaging (MRI) plays a crucial role in medical diagnosis and treatment planning. However, the automatic analysis of MRI data remains challenging due to noise, variability, and class imbalance. In this work, two approaches for binary brain tumor classification are investigated. The first approach is based on classical machine learning using a Random Forest classifier combined with a hybrid data preprocessing pipeline that integrates oversampling and data cleaning techniques. The second approach relies on a deep learning model based on a Convolutional Neural Network (CNN). Both approaches are evaluated under identical experimental conditions. The experimental results demonstrate that the machine learning approach combined with effective data preprocessing achieves superior performance compared to the CNN model, particularly in terms of recall and F1-score, which are critical metrics for medical diagnosis.

Index Terms— Brain tumor classification, Random forest, Smote, Data analyses

1. INTRODUCTION

Medical data plays a central role in modern healthcare, particularly in oncology, where medical imaging such as Magnetic Resonance Imaging (MRI) is widely used for diagnosis, treatment planning, and clinical decision support [1]. MRI provides detailed information about soft tissues, making it a valuable tool for detecting pathological conditions such as brain tumors. However, the automatic analysis of MRI data remains

challenging due to noise, variability, and the inherent complexity of medical images.

Medical data processing typically involves several stages, including data acquisition, preprocessing, feature extraction, and classification [2]. In MRI-based applications, raw images can either be transformed into descriptive datasets through feature extraction techniques or processed directly using deep learning models. These representations are then used as input for classification systems, whose performance strongly depends on the quality and distribution of the training data.

A major challenge in medical datasets is class imbalance, where pathological cases are significantly less represented than normal cases. This situation is common in oncology; as abnormal cases are naturally rarer than healthy ones. When trained on imbalanced data, both machine learning and deep learning models tend to favor the majority class, resulting in high overall accuracy but poor detection of minority (pathological) cases [3]. This limitation is particularly critical in medical applications, where missing tumor cases can have serious clinical consequences.

To address class imbalance, several data-level preprocessing techniques have been proposed. Oversampling methods, such as the Synthetic Minority Over-sampling Technique (SMOTE), aim to increase the number of minority class samples by generating synthetic data points [4]. SMOTE has been

shown to improve minority class representation and enhance classification performance for classical machine learning models. However, applying SMOTE alone may introduce overlapping samples and noise, which can negatively affect class separability [5].

To overcome these limitations, oversampling is often combined with data cleaning techniques. Methods such as ALL-KNN remove noisy and misclassified samples, while One-Sided Selection (OSS) focuses on eliminating ambiguous samples near class boundaries [6]. Combining SMOTE with ALL-KNN and OSS helps reduce overlap and improve data quality.

In parallel, Convolutional Neural Networks (CNNs) have been widely adopted for brain tumor classification from MRI images due to their ability to automatically learn discriminative spatial features. Several studies have demonstrated the effectiveness of CNN-based models for brain tumor detection and classification without handcrafted feature extraction. However, CNN-based approaches often require large and well-balanced datasets and may suffer from reduced sensitivity when trained on imbalanced medical data.

In this paper, a complete preprocessing pipeline is adopted to improve classification performance under class imbalance. The proposed approach is evaluated using a classical machine learning classifier and is compared with a CNN-based deep learning model to assess the impact of data preprocessing and model selection on binary brain tumor classification.

2. RELATED WORK

The issue of class imbalance has been widely studied in the field of machine learning and deep learning, particularly in medical and biomedical applications. Several studies have shown that classifiers trained on imbalanced datasets tend to achieve high accuracy while exhibiting low recall and F1-score for minority classes, regardless of the underlying learning model [3].

SMOTE, introduced by Chawla et al. [4], is one of the most popular oversampling techniques for addressing class imbalance. By generating synthetic samples through interpolation between minority class instances, SMOTE improves the representation of minority classes and has been successfully applied in various medical classification tasks. However, subsequent studies reported that SMOTE may increase class overlap and introduce noisy samples, especially in complex datasets [5].

To address these issues, hybrid preprocessing approaches combining oversampling and data cleaning techniques have been proposed. Batista et al.

demonstrated that combining oversampling with undersampling techniques can significantly improve classification performance on imbalanced datasets [5]. Similarly, Tomek introduced methods such as OSS to remove borderline and ambiguous samples, thereby improving class separability [6]. ALL-KNN has also been widely used to eliminate noisy samples based on nearest neighbor rules [7].

CNN-based deep learning models have been extensively explored for brain tumor classification using MRI data. Pereira *et al.* proposed a deep CNN architecture for brain tumor segmentation and classification, achieving high accuracy on MRI datasets [8]. Similarly, Havaei *et al.* introduced a CNN-based approach that integrates both local and contextual information, significantly improving tumor detection performance [9]. Despite their strong results, several studies have reported that CNN performance is highly dependent on dataset size, class balance, and data

diversity, often requiring additional data augmentation or preprocessing strategies [2], [3].

Consequently, classical machine learning approaches combined with effective data preprocessing pipelines remain competitive and attractive solutions, particularly when datasets are small or imbalanced. Comparing such approaches with CNN-based models provides valuable insight into the trade-offs between data-centric and model-centric strategies in brain tumor classification. This work contributes to this comparison by evaluating a complete preprocessing pipeline and contrasting its performance with a CNN-based approach for binary brain tumor detection.

3. PROPOSED METHOD

In the first approach, a classical machine learning pipeline is adopted for binary brain tumor classification. MRI images are first transformed into numerical feature vectors through feature extraction. A hybrid preprocessing strategy is then applied to handle class imbalance and data noise by combining oversampling and data cleaning steps. The resulting balanced dataset is finally used to train a Random Forest classifier, which performs the classification task.

In the second approach, a deep learning pipeline based on a Convolutional Neural Network combined with an Artificial Neural Network classifier is employed. MRI images are directly provided as input to the CNN, which automatically learns relevant feature representations. These features are subsequently classified by the ANN to produce the final binary output corresponding to tumor or non-tumor classes.

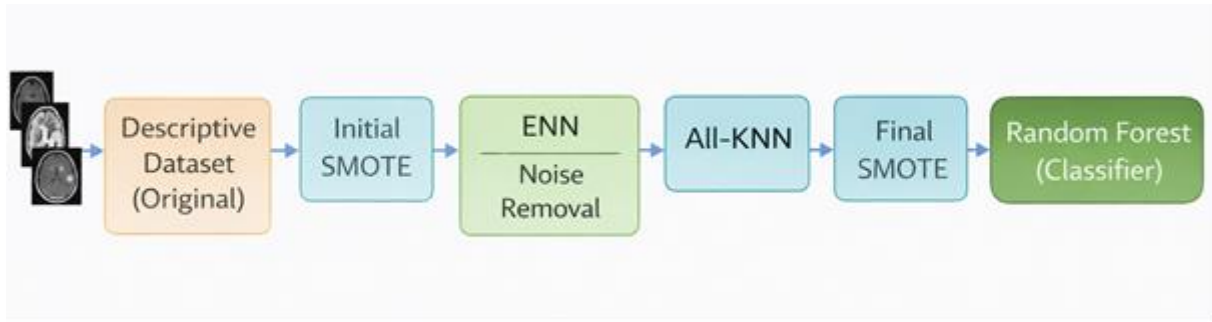


Figure 1. Proposed data preprocessing and augmentation pipeline for MRI-based classification.

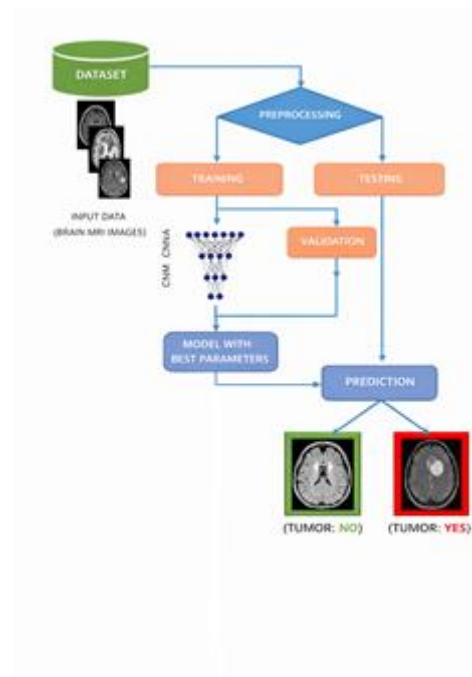


Figure 2. Overall framework of the MRI-based classification system.

4. EXPERIMENTATION AND RESULTS

All experiments were conducted on the same dataset to ensure a fair comparison between the two approaches. The dataset consists of MRI brain images labeled into two classes, tumor and non-tumor, and exhibits a significant class imbalance due to the medical nature of the data, with tumor cases being underrepresented. For the machine learning approach, MRI images were first converted into a descriptive numerical dataset through feature extraction, as classical machine learning algorithms cannot directly process raw image data. The resulting dataset was then used as input for the Random Forest classifier. To address the class imbalance problem, a hybrid preprocessing strategy combining successive

oversampling and data cleaning steps was applied to generate a final dataset that is both balanced and cleaned, thereby improving the reliability of the learning process. In parallel, a CNN model was trained on the same classification task, learning features automatically from the data and serving as a deep learning baseline for comparison. Both models were evaluated using identical training and testing splits and the same evaluation metrics.

4.1 EVALUATION METRICS

To provide a comprehensive evaluation of classification performance, especially in the context of imbalanced medical data, the following metrics were used:

- Accuracy

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

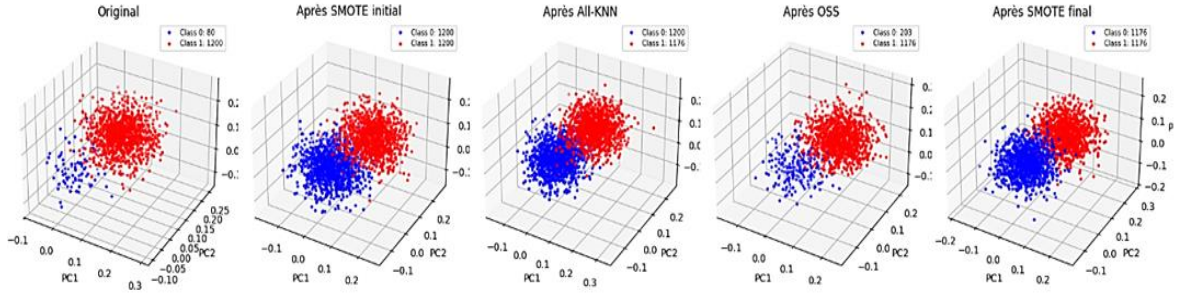


Figure 3. Visualization of dataset evolution through the different preprocessing steps.

- Precision

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

- Recall (Sensitivity)

$$Recall = \frac{TP}{TP + FN}$$

- F1-score

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

These metrics are widely used in medical image analysis and provide meaningful insight into diagnostic performance. To better understand the impact of the proposed preprocessing pipeline, a visual analysis of the data distribution was performed.

Figure 3 illustrates the progressive evolution of the dataset distribution through the different preprocessing steps. In the original dataset, a strong class imbalance and a significant overlap between classes are observed, which explains the difficulty of class separation. After the application of the initial SMOTE, the minority class is reinforced; however, some overlap is introduced. The subsequent cleaning steps using All-KNN and OSS remove noisy and ambiguous samples, improving class separability. Finally, the SMOTE final step produces a balanced and well-structured dataset, which visually confirms the improvement in data quality and supports the observed performance gains.

4.2 RESULTS OF THE RANDOM FOREST MODEL WITH HYBRID PREPROCESSING

The Random Forest classifier was trained on the final dataset obtained after applying the hybrid preprocessing strategy. This final dataset is balanced and free from noisy and overlapping samples, which contributes to improved class separability.

The experimental results show that the proposed model achieved optimal performance across all evaluation metrics, including accuracy, precision, recall, F1-score, and AUC. In particular, the model demonstrates excellent sensitivity, which is crucial for detecting tumor cases in medical diagnosis.

4.3 COMPARATIVE RESULTS WITH CNN

To evaluate the effectiveness of the proposed approach, a comparative analysis was conducted with a CNN-based model. The quantitative results of this comparison are presented in **Figure 4**.

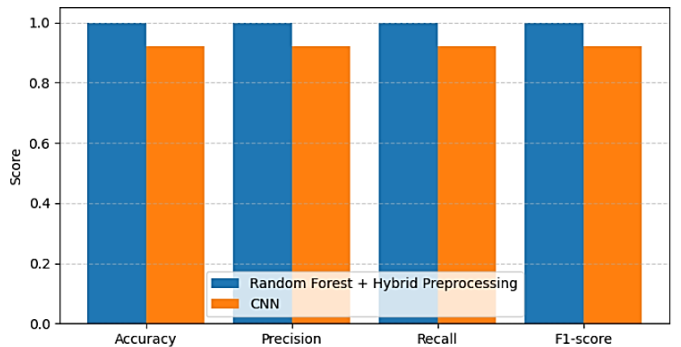


Figure 4. Performance comparison between the proposed Random Forest-based approach and the CNN model.

Figure 4 provides a graphical comparison of the classification performance between the proposed

approach and the CNN model. While the CNN achieves good overall performance, the Random Forest combined with the hybrid preprocessing strategy shows superior results across all evaluation metrics. In particular, the improvement in recall and F1-score highlights a better detection of tumor cases, which is essential in medical diagnosis. These results emphasize the importance of data preprocessing and show that data-centric approaches can outperform deep learning models when data availability is limited.

The CNN model achieved an overall accuracy of **92%** on the test set. While it showed high precision for tumor detection (**0.96**), its recall for the tumor class (**0.87**) was lower, indicating that some tumor cases were missed.

Overall, the results indicate that the Random Forest model combined with the hybrid preprocessing strategy outperformed the CNN model across all evaluation metrics, particularly in terms of recall and F1-score, which are critical for medical diagnosis.

4.4 SUMMARY OF EXPERIMENTAL FINDINGS

The experimental evaluation confirms that the proposed preprocessing and classification strategy significantly improves performance for binary brain tumor detection. The results demonstrate that, under the given experimental conditions, the Random Forest approach with hybrid preprocessing provides superior performance compared to the CNN model.

5 CONCLUSION

This paper addressed the problem of automatic brain tumor classification from MRI data, with particular emphasis on the challenges of noise, intra-class variability, and class imbalance that commonly affect medical imaging datasets. Two different strategies were investigated: a classical machine learning pipeline based on a Random Forest classifier combined with a hybrid data preprocessing scheme, and a deep learning approach based on a Convolutional Neural Network.

The experimental results show that the proposed hybrid preprocessing strategy, which combines oversampling and data cleaning, leads to a clear improvement in classification performance. When integrated with the Random Forest model, this pipeline achieved superior results across all evaluation metrics, with especially strong gains in recall and F1-score, which are critical for reducing false negatives in medical diagnosis. These outcomes underscore the importance of data quality and class balance for building reliable decision-support systems.

By contrast, although the CNN-based approach achieved competitive overall accuracy, it obtained lower recall and F1-score under the same experimental conditions, indicating that a non-negligible number of tumor cases were missed. This suggests that deep learning models in this setting are more sensitive to dataset size, balance, and variability, and may require larger, more diverse, and better-curated MRI collections to reach their full potential.

Overall, the findings of this study support the view that data-centric strategies, such as hybrid preprocessing, can be as effective as or even more effective than more complex model-centric solutions when working with imbalanced medical datasets. Future work will focus on extending the proposed pipeline to multiclass tumor classification, evaluating it on larger and more heterogeneous MRI datasets, and exploring hybrid frameworks that combine the strengths of classical machine learning and deep learning to further improve robustness and clinical applicability.

REFERENCES

- [1] L. O. Hall, A. B. Gepp, and D. J. Cook, "Class Imbalance in Medical Data Classification," *IEEE Engineering in Medicine and Biology Magazine*, vol. 30, no. 6, pp. 36–45, 2011.
- [2] G. Litjens et al., "A Survey on Deep Learning in Medical Image Analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.
- [3] H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.
- [4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
- [5] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20–29, 2004.
- [6] I. Tomek, "Two Modifications of CNN," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 6, no. 11, pp. 769–772, 1976.
- [7] D. L. Wilson, "Asymptotic Properties of Nearest Neighbor Rules Using Edited Data," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 2, no. 3, pp. 408–421, 1972.

[8] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1240–1251, 2016.

[9] M. Havaei et al., "Brain Tumor Segmentation with Deep Neural Networks," *Medical Image Analysis*, vol. 35, pp. 18–31, 2017.