Reliable Central Nervous System Tumor Differentiation on MRI Images with Deep Neural Networks and Conformal Prediction

Luis Balderas[1,*], María Moreno de Castro[1], Miguel Lastra[2], Jose P. Martínez[3,4], Francisco J. Pérez[3], Antonio Laínez[3,4], Antonio Arauzo-Azofra[5] and José M. Benítez[1,4]

[1]*School of Sciences, Technology and Engineering's Doctorates. Department of Computer Science and Artificial Intelligence, DiCITS, DaSCI, iMUDS, University of Granada, 18071, Granada, Spain*

[2]*Department of Software Engineering, DiCITS, DaSCI, iMUDS, University of Granada, 18071, Granada, Spain*

[3]*Radiodiagnosis Service, Hospital Universitario "Virgen de las Nieves", Granada, Spain*

[4]*ibs.Granada: Instituto Biosanitario de Granada, Granada, Spain*

[5]*Rural Engineering Department, DiCITS Lab, University of Córdoba, 14005, Córdoba, Spain*

Abstract

Central nervous system tumors, particularly gliomas, rank among the top 10 causes of cancer-related deaths worldwide. Thus, precise differentiation of these tumors is crucial for effective treatment, which can reduce patient suffering and lower mortality rates. We propose a deep learning-based technique for glioma differentiation using MRI imaging, which incorporates two novel approaches. First, we represent perfusion sequences as time series, which serve as inputs for a Deep Neural Network (DNN) classifier. This classifier is trained to distinguish the sequences between low-grade glioma (LGG) and high-grade glioma (HGG). Second, we employ Conformal Prediction to calibrate the results, ensuring they include the true category with a 90% probability. This approach has been rigorously tested to evaluate its performance. Our experimental results demonstrate that our deep neural network not only provides accurate but also ensures trustworthy predictions.

1. Introduction

Central nervous system (CNS) tumors have traditionally been regarded as a lethal form of cancer, ranking among the top 10 causes of deaths globally. The GLOBOCAN project using nationwide data from 184 countries estimated the annual incidence of malignant-only CNS tumors to be 3.4 per 100,000 individuals [1]. Within the variety of tumor types, gliomas are the most frequent primary intracerebral tumors in adults. Despite their high fatality rates, the prognosis of CNS tumors has considerably improved over the last decades, possibly because of the prompt detection, the optimization of treatment protocols including the introduction of temozolomide and the advances in neurosurgical procedures.

Accurate preoperative differentiation of primary CNS tumors is crucial because the treatment strategies differ substantially (e.g. stereotactic biopsy + chemotherapy vs. gross total resection + chemotherapy in lymphoma vs. glioblastoma). Magnetic Resonance Imaging (MRI) is the most important non-invasive technique in the diagnosis and differentiation of CNS tumors. In this context, MRI-perfusion is an imaging technique frequently used to assess the vascularization of brain lesions in a minimally invasive way. There are different approaches for the study of perfusion and lesions of CNS.

Machine learning has been employed in numerous medical problems as part of expert decision support systems. However, in most instances, these systems do not provide any information regarding the confidence of their predictions. In this paper we propose an innovative approach that treats perfusion curves as time series data and apply Deep Neural Networks (DNN) and Conformal Predictors (CP) to assess the type of glioma a patient suffers, along with measures of confidence.

EXPLIMED - First Workshop on Explainable Artificial Intelligence for the medical domain - 19-20 October 2024, Santiago de Compostela, Spain

[*] Corresponding author.

0000-0002-3845-8848 (L. Balderas); 0000-0003-0440-0864 (M. Moreno de Castro); 0000-0002-7278-2668 (M. Lastra); 0000-0002-0663-2856 (A. Laínez); 0000-0002-2486-5792 (A. Arauzo-Azofra); 0000-0002-2346-0793 (J. M. Benítez)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

In essence, this involves framing the differentiation problem as a classification task using the perfusion curve as input. This requires finding a suitable representation for the series and then identifying a technique that delivers high classification performance. On top of the classifier, a conformal predictor is built to quantify the level of uncertainty inherent in the model's predictions using specific metrics. Our research has resulted in an effective automated procedure to assist radiologists in their work, contributing to the quantification of uncertainty at the patient level, with applications in personalized and precision medicine.

The structure of the article is as follows: In Section 2, we present state-of-the-art methods that use deep learning and conformal prediction for glioma differentiation, including medical context for a better understanding of the problem. In Section 3 we detail Conformal Predictors. In Section 4, we present our proposal for glioma differentiation. In Section 5, we include the empirical study and discussion of the results. Finally, Section 6 contains the conclusions of the work.

2. Related work

2.1. Medical contextualization

As noted in the introduction, gliomas are the most frequent primary intracerebral tumors in adults. The age-adjusted annual incidence of histologic verified glioma has been reported to be 7.3 cases per 100,000 person-years [2]. Most patients with gliomas have a fatal prognosis, and the disease has considerable impact on patients and their families' physical, psychological, and social status. A recent study on the epidemiology of gliomas found that high-grade gliomas —HGG, grades 3 and 4 according to World Health Organization (WHO)— were present in 85% and low-grade glioma (LGG, WHO 1 and 3 grades) in 15% of the cases, with 5-year overall survival of 82, 54, 22 and 3% for grade 1, 2, 3, and 4, respectively [3].

Accurate preoperative differentiation of primary CNS tumors is essential because the treatment strategies differ substantially. MRI is the most important non-invasive technique in the diagnosis of CNS tumors. In particular, MRI-perfusion is frequently used to assess the vascularization of brain lesions. There are different approaches for the study of perfusion and lesions of CNS. The most widely used MRI-perfusion approach in clinical practice is T2 gradient echo or dynamic susceptibility contrast (DSC) due to its high sensitivity and specificity to differentiate lesions and its short acquisition time. DSC perfusion allows the study of changes in the signal intensity of the tissues derived from the passage of an intravenous gadolinium bolus administered in a peripheral vein. The passage of the contrast bolus leads to specific imaging patterns in different tissues and tumors, with changes in their signal intensity proportional to the amount of gadolinium that passes through or is deposited in them. Changes in signal intensity over time allow a curve to be obtained, from which hemodynamic parameters are inferred.

At present, pathological anatomy techniques are essential in the diagnosis and genetic and histological differentiation of gliomas. However, in the presence of lesions such as metastases, an adequate diagnostic imaging can avoid a brain biopsy. The procedure of stereotactic brain biopsy entails a considerable risk of complications (e.g., haemorrhage, infection, wound breakdown), which can be as high as 15.3% in tumors located in midline areas, while 28.8% of post-procedural intracranial haemorrhages have been reported [4].

A recent Cochrane review [5] reported limited available evidence on the ability of perfusion MRI to distinguish HGG from LGG, which precludes reliable estimation of the performance of DSC MRI perfusion-derived parameters (e.g., relative cerebral blood volume or rCBV) for determining the tumor grade, specifically in untreated solid and non-enhancing LGG vs. HGG. MRI-perfusion has also been advocated as a useful tool to differentiate between gliomas and other tumors of the CNS, such as lymphomas or metastases, but with variable diagnostic yields depending on the publications. Furthermore, it demonstrated to be useful in differentiating between tumor recurrence of gliomas and post-treatment changes such as pseudo progression or radionecrosis. However, the available evidence is restricted to isolated non-clinically validated studies.

Most of these publications have focused on a classical analysis of MRI perfusion data, with special emphasis on the quantification of rCBV (an estimate of tumor perfusion), which is not quantifiable in absolute terms (thus requiring comparison with areas of normal cerebral parenchyma). A number or publications that have applied Artificial Intelligence algorithms to the analysis of brain perfusion can be found. They are discussed in section 2.2.

2.2. State-of-art Artificial Intelligence methods

In this section, we present the most relevant articles in the state of the art related to the topic. To the best of our knowledge, there is no technique in the literature that utilizes the perfusion curve as input for a time series classification model in low-grade gliomas (LGG) and high-grade gliomas (HGG). Moreover, there are numerous approaches in the literature that apply CP to quantify the uncertainty of deep learning models ([6], [7]). In fact, the combination of deep learning and CP is especially useful in biomedical problems. For example, in skin lesion classification ([8]), for the diagnosis and grading of prostate biopsies ([9]), for rating breast density in mammography ([10]), for grading the severity of spinal stenosis in lumbar spine MRI ([11]); among other applications ([12], [13], [14]), we have not encountered in the literature any work that combines DNN and CP methods in the differentiation of central nervous system tumors.

However, the non-invasive differentiation of gliomas through the application of machine learning, specifically distinguishing between LGG and HGG gliomas, has been extensively investigated in recent years. For instance, in [15], a substantial number of radiological features were extracted from MRI sequences, including T1-weighted, T1-weighted contrast-enhanced, T2-weighted, and FLAIR, across a total of 285 cases (210 HGG, 75 LGG). These features were used to train three classifiers (logistic regression, random forest, and support vector machine) to determine the glioma type, achieving an average AUC of 0.9030 for test cohorts.

In [16], a deep multi-scale 3D Convolutional Neural Network (CNN) architecture was proposed to categorize gliomas into LGG and HGG using volumetric T1 contrast-enhanced MRI sequences, achieving an accuracy of 96.49%.

Recently, some studies advocate mixed approaches incorporating molecular profiling for the differentiation of LGG and HGG ([17]). For instance, in the work by authors in [18], clinical and laboratory data were integrated to create a tool for predicting the molecular status (ATRX, IDH1/2, MGMT, and 1p19q co-deletion), distinguishing between low-grade and high-grade gliomas. The system achieved an AUC of 0.885 for this specific learning task.

Finally, other studies such as [19], [20], [21], and [22] utilized cases of high-grade or low-grade gliomas to conduct specific studies on various features, but did not present tools for classification between these two grades of gliomas.

3. Conformal predictors

In spite of their generally accurate performance, many machine learning models are known to estimate poorly prediction probabilities. To deliver reliable estimates and avoid over or under-confidence predictions, prediction uncertainty must be quantified. This decision aligns with the demand of trustworthy Artificial Intelligence, which, as expressed in recent European Parliament AI regulations [23], includes as a requirement the models ability to quantify and communicate their confidence in their outputs ([24], [6]) with special focus on high-risk applications as AI-supported medicine [25].

Performance metrics, such as accuracy or F1-score, cannot tell about the confidence or uncertainty of the model's predictions. Even for a model that is right most of the time, not all the cases are equally easy to classify. Some of them might even be doubtful (e.g., they belong to a different distribution or their kind was mostly under represented, overlooked, or misclassified). For example, we can think about patients images from a different hospital, taken from a different MRI scanner, or labeled by another team. The model may provide unnoticed overly confident predictions because machine learning algorithms do not include a built-in warning mechanism to prevent well-informed predictions from looking the same as wild guesses [26] and refers to the alignment between predicted probabilities directly provided by the model and relative frequencies in the actual data. Well-calibrated models are those where the predicted probabilities match the empirical probabilities. Thus, it is about calculating *"the probability that the predicted probability is right"*. For example, given an MRI scan, if the model predicts class A with 80% probability, then class A should occur approximately 80 times out of 100 predictions. More formally, a well-calibrated classifier satisfies the following formula [27]:

$$P(\text{actual class is A}|\text{predicted probability of A is } p) \approx p \tag{1}$$

To provide confidence to classification results, calibration models should be applied. Two of the most popular calibration methods are Platt's Scaling and Isotonic Regression. Platt's Scaling [28] is a very popular method that maps each prediction to its empirical frequency by passing it through a sigmoid. The method is therefore parametric and presupposes normally distributed and heteroscedastic per-class scores, a notably limiting assumption [29]. On the other hand, Isotonic Regression [30] is more general because the map is performed with an isotonic (monotonically increasing) function. It assumes that the classifier perfectly ranks objects in the test set, essentially implying an ROC AUC of 1 and it is not recommended for small sets [29]. A detailed explanation of how both calibration methods can be found in [31].

One step ahead of calibration is conformalization. To conformalize means that each single-point prediction is not only calibrated but also provided with a prediction interval, with statistical guarantees of including the true value at patient-level ([27], [32], [29]). Conformal Prediction (CP) methods conformalize the prediction. CP methods, which are distribution-free and lightweight, have shown some successful application in health-related domains, mostly for cancer prediction ([33], [34], [35], [32],[36], [34], [37], [38], [39]).

Venn-ABERS methods ([40], [27]) are members of the CP family [41], [42], [43]). They share the skill to evaluate how different is an unseen instance from the training dataset. These methods were created to work on top of binary classifiers (although new implementations work for multi-class problems too [44]). They conformalize the predicted probability to match the actual frequency of its class, while delivering in the process the upper (p_1) and lower (p_0) bounds of the probability interval that includes the true class (a property known as coverage or validity) with statistical guarantee. This prediction interval quantifies the uncertainty of predictions at instance level: the larger the interval, the lower the confidence the model has in that prediction. Venn-ABERS are an adaptation of Isotonic Regression (where it is applied twice, to fit the probability of each class) and a special case of Venn Predictors ([45], [46] ,[29]), from which they inherit guaranteed validity in the form of well-calibrated probability prediction.

There are two possible implementations of the Venn-ABERS algorithm depending on whether or not an independent calibration set is required separate from the training set. On one hand, an inductive approach, or Inductive Venn-ABERS (I-V-A), where the calibration is trained over a hold-on split of the training set, thus the conformalization is model-agnostic [46]. On the other hand, Cross Venn-ABERS (C-V-A) where the splits are performed via k-fold, thus there is not need to reserve data just for calibration as in the inductive case and we use all available data to train the classifier but the classifier must be retrained each time we predict and conformalize the probabilitiy of a new patient [33].

4. Our proposal

This work presents a novel methodology for the non-invasive differentiation of a specific family of CNS tumors, specifically distinguishing between high-grade (HGG) and low-grade gliomas (LGG). Patient data are 3D MRI scans focused on the T2 sequence. The T2 sequence is designed to measure the evolution of tissue signal intensity as the contrast bolus passes through. From these changes in signal intensity in the image, once processed, we construct the perfusion curve, which serves as input for our machine learning model. Thus, we transform an image processing problem into a time series classification one, which we address by combining two impactful technologies: on one hand, a deep

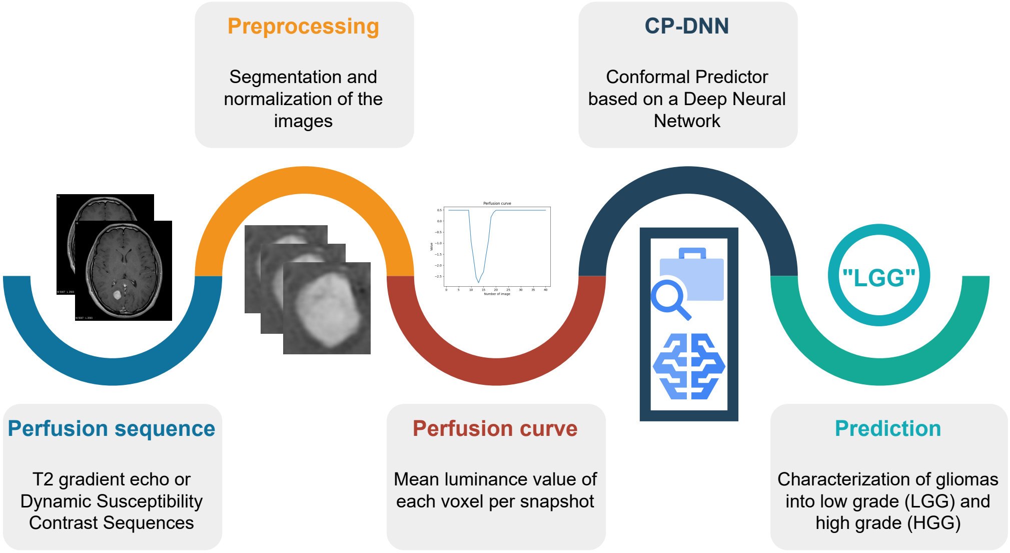
Preprocessing
Segmentation and normalization of the images
CP-DNN
Conformal Predictor based on a Deep Neural Network
Perfusion curve
"LGG"
Perfusion sequence
T2 gradient echo or Dynamic Susceptibility Contrast Sequences
Perfusion curve
Mean luminance value of each voxel per snapshot
Prediction
Characterization of gliomas into low grade (LGG) and high grade (HGG)

Figure 1: Complete processing pipeline for glioma differentiation between low-grade and high-grade. The process initiates with the reception of a new perfusion sequence. After preprocessing the images, we extract the perfusion curve by calculating the mean luminance of the voxels at different time points. Subsequently, a CP-DNN is trained to extract the final prediction and quantify its uncertainty.

neural network capable of extracting patterns with high precision; and on the other hand, conformal predictors for quantifying the uncertainty of predictions. Consequently, we develop a decision support system with significant generalization capacity, providing valuable and calibrated information to the radiologist.

The presented pipeline is divided into three phases: preprocessing of the sequence, generation of the perfusion curve, and finally, glioma differentiation with a machine learning model, specifically, a conformal predictor based on a deep neural network classifier. Below, we provide a detailed description of each phase. Figure 1 show the complete pipeline.

4.1. Preprocessing the perfusion MRI sequence

The differentiation process begins with the arrival of a new perfusion sequence. The first phase, related to preprocessing, involves normalizing the image in both size and grayscale levels using the Statistical Parametric Mapping technique [47]. Additionally, it is crucial that the image is segmented. Despite numerous initiatives aiming to develop deep learning models for brain segmentation in MRI (as illustrated in [48] and other reviews), there is, to the best of our knowledge, no automatic technique for segmenting an image in the perfusion sequence. Therefore, in our approach, cases have been segmented by radiologists using the 3D Slicer tool (https://www.slicer.org, [49]).

4.2. Generating the perfusion curve

Once the perfusion sequence is normalized and segmented, we generate the curve. The perfusion sequence consists of a series of snapshots taken over a specified period. This procedure generates n 3D images of the brain and, specifically, the glioma. By processing each of these n images, calculating the mean luminance value of each voxel within the image in each snapshot, we generate a time series with as many points as snapshots. This way, a time series is generated out of the perfusion sequence, the perfusion curve. After that, the time series is normalized by subtracting the mean of the points and dividing by their standard deviation.

4.3. Characterizing the Glioma with a CP based on DNN

The differentiation of gliomas into low grade (LGG) and high grade (HGG) requires a machine learning system capable of extracting relevant patterns to distinguish between the two tumor types. Consequently, we constructed a deep dense neural network that takes a perfusion curve as input and classifies it into LGG and HGG.

This is an unbalanced classification problem due to the prevalence of the HGG class in the data set over the LGG class. Thus, to evaluate the DNN performance, we calculate the accuracy and F1-score.

To enhance the trustworthiness of the classifier outputs, its probabilities must be calibrated. We calibrate the predictions of the model applying four independent calibration techniques: Platt's Scaling (PS), Isotonic Regression (IR), Cross Venn-ABERS (C-V-A), and Inductive Venn-ABERS (I-V-A).

The performance of the calibrators is evaluated with classification and calibration metrics. In this way, physicians have access not only to artificial intelligence-based tools with a high level of accuracy but also to the ability to assess the confidence that these models have in each of the predictions associated with the patients, lending credibility to those predictions in which the model has sufficient confidence. To achieve this, we employ a Conformal Predictor with a specified confidence level, η, (e.g. 90%) on top of the DNN. Thus, the radiologist will know, with each prediction, whether the model has confidence exceeding η in determining whether the tumor is a low-grade or high-grade glioma, what would increase his or her trust on the differentiation model performance.

To the best of our knowledge, there is no state-of-the-art method that addresses the problem of glioma differentiation by applying explainability and trustworthiness methods, using perfusion sequences as time series and applying DNN and CP in this binary classification context.

5. Empirical study

To demonstrate the ability of our method to differentiate between high-grade and low-grade gliomas satisfactorily, we have conducted a rigorous experimentation. It is detailed in this section along with the results and their analysis.

5.1. Dataset

We have compiled a cohort of 58 patients from the radiology department of the Virgen de las Nieves Hospital in Granada, Spain, selected for their exceptional quality and homogeneity in the MRI scans taken for their diagnosis and treatment.

Due to the rather small size of the dataset, we have designed a data augmentation strategy to effectively train our machine learning models. For the training partition, we generate perfusion curves not only at the level of the entire volume but also across the different 2D slices that compose it along the Z-axis. All slices of the same sequence inherit the label, either LGG or HGG, from the original sequence. This extends the training set to exceed 700 curves. Besides the anticipated imbalanced character of the problem arises. The distribution of instances between the two classes is 70% (HGG) and 30% (LGG). To address this issue, the models need to be adjusted to weigh the prediction error, making it more costly to make errors in predicting the LGG class than the HGG class.

5.2. Deep Neural Network: architecture and hyperparameters

Taking into account that the patient cohort is not very extensive, the dense neural network to be used cannot be very large, as we might incur in overfitting issues. After different attempts, an architecture with five layers (input, output, and three hidden layers with 70, 30, and 5 neurons) is considered, which yields competitive results.

For the selection of hyperparameters, a Grid Search Cross Validation is applied to choose the optimization algorithm and the parameters α (strength of the L2 regularization term), initial learning rate, β_1 (exponential decay rate for fist moment vector), β_2 (exponential decay rate for second moment vector), and ϵ (value for numerical stability). The result of this hyperparameter tuning process produced the following value selection:

- Optimization algorithm: Adam
- Initial Learning rate: 0.001
- α : 0.0001
- β_1 : 0.9
- β_2 : 0.9
- ϵ : 0.000000001

5.3. Metrics

Our proposal combines deep learning and conformal prediction. We use classic metrics to evaluate the goodness of the prediction and calibration metrics to assess the level of uncertainty in the predictions. For classification metrics, we use accuracy and F1-Score. For calibration metrics, we use Brier Score and Log Loss. The first one is defined as:

$$BS = \frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2, \tag{2}$$

where p is the prediction probability of occurrence of the event i, and o_i is defined as follows:

$$o_i = \begin{cases} 1 & \text{if event } i \text{ ocurred} \\ 0 & \text{if event } i \text{ not ocurred} \end{cases} \tag{3}$$

Log Loss is defined as

$$L_{log}(y, p) = -(y\log(p) + (1 - y)\log(1 - p)) \tag{4}$$

with $y \in \{0, 1\}$ and a probability estimate $p = Pr(y = 1)$.

For both of them, Brier Score and Log Loss, the smaller the value, the better is the calibration.

5.4. Results and analysis

We calculated the accuracy and F1-score to compare the performance of our novel DNN for time series classification approach with the leading state-of-the-art method for time series classification, which is a 1NN with Dynamic Time Warping ([50], [51]).

We have designed a data strategy to ensure the rigor of the study. In all cases, we guarantee the separation between patients is maintained. Firstly, to confirm the reliability of the results, a Stratified 5-fold Cross Validation is used. This involves generating five datasets where patients are split into 80% for training and 20% for testing each time, maintaining class proportions in each fold.

The values in Tables 1 and 2 represent the mean values from the five runs for each metric on the test set, demonstrating that our DNN approach outperforms the state-of-the-art 1NN-DTW in both performance and calibration. These results confirm previous analyses where Inductive Venn-ABERS and other CP methods outperformed Platt's Scaling and Isotonic Regression [46].

In order to unleash the true power of the DNN+CP (actually, Inductive Venn-ABERS) combination, we conduct another experiment where the dataset is partitioned into training and test sets, with 80% of the patients used for training and 20% for testing, again in a stratified manner. We plot in Figure 2 the predicted and conformalized probabilities for the 12 instances (MRI scans) in the test set. On the X-axis, we can see the 12 predictions made by the model. On the Y-axis, we find the probability expressed by the model, with 0 representing a prediction of the LGG class and 1 representing a prediction of the HGG class. The black dots indicate the actual labels of the test examples. As can be seen, there are 3

Method	Acc	F1-Score	Brier Score	Log Loss
DNN	0.642	0.769	0.289	1.168
DNN + IR	0.733	0.842	0.2	1.185
DNN + PS	0.733	0.844	0.2	0.612
DNN + C-V-A	0.677	0.787	0.249	0.755
DNN + I-V-A	**0.752**	**0.857**	**0.197**	**0.598**

Table 1
Performance metrics include accuracy and F1-score (the higher the better), and calibration metrics include Brier Score and Log Loss (the smaller the better) for the DNN model, which outperforms the state-of-the-art 1NN-DTW model (see Table 2). The last four rows correspond to the implementations of the four calibration methods: IR stands for Isotonic Regression, PS stands for Platt's Scaling, C-V-A stands for Cross Venn-ABERS, and I-V-A stands for Inductive Venn-ABERS. All methods improve model performance, with Inductive Venn-ABERS significantly reducing calibration errors. The best results are highlighted in bold.

Method	Acc	F1-Score	Brier Score	Log Loss
1NN-DTW	0.692	0.769	0.288	10.376
1NN-DTW + IR	0.714	0.813	0.233	0.636
1NN-DTW + PS	0.73	0.844	0.248	0.634
1NN-DTW + C-V-A	0.675	0.786	0.27	0.857
1NN-DTW + I-V-A	**0.746**	**0.848**	**0.201**	**0.62**

Table 2
Similar to Table 1, here are the results for the state-of-the-art 1NN-DTW model. The classifier shows lower performance and calibration metrics compared to our approach. IR stands for Isotonic Regression, PS stands for Platt's Scaling, C-V-A stands for Cross Venn-ABERS, and I-V-A stands for Inductive Venn-ABERS. The best results are highlighted in bold.

cases of LGG and 9 cases of HGG, maintaining the stratification of the original dataset. The larger light blue dot corresponds to the predictions made by the dense neural network without conformalization. When the black dot and the blue dot coincide, it indicates that the neural network correctly predicted the class. Conversely, if they do not coincide, the neural network made an incorrect prediction.

The smaller blue dot (p_0) and the red dot (p_1) represent the probability interval containing each of the calibrated predictions (p), depicted by the orange dot. This orange dot represents the predictions of our DNN+CP model expressed in terms of probability, where a probability $p \geq 0.5$ results in a predicted label of 1 (HGG class). Conversely, if $p < 0.5$, the predicted label is 0 (LGG class). The distance between p_0 and p_1 ($p_1 - p_0$) indicates the uncertainty in the model's decision: the greater the distance between them, the higher the uncertainty in the prediction. This uncertainty is represented by the pink curve.

As observed, the neural network without conformalization makes two errors (the second and the eleventh cases). Thanks to the conformalization of the predictions, the DNN+CP model correctly predicts the second case, as the predicted probability is below 0.5, resulting in an LGG class label. Regarding the eleventh case, the DNN+CP, like the non-conformalized DNN, makes the error of categorizing the case as HGG when it is actually LGG. However, it can be seen that the uncertainty in that prediction is very high, showing confidence of less than 60%. Similarly, predictions with some uncertainty are made in the first (slightly over 60%) and third cases (close to 80%). The rest of predictions are more reliable, showing a confidence ranging from 85% to 99%.

Therefore, the combination of DNN+CP is not only more accurate, as it corrects errors that an uncalibrated neural network would make, but it also provides very valuable information to doctors about the confidence of the predictions, allowing specialists to assess each case with much more trustworthy information.

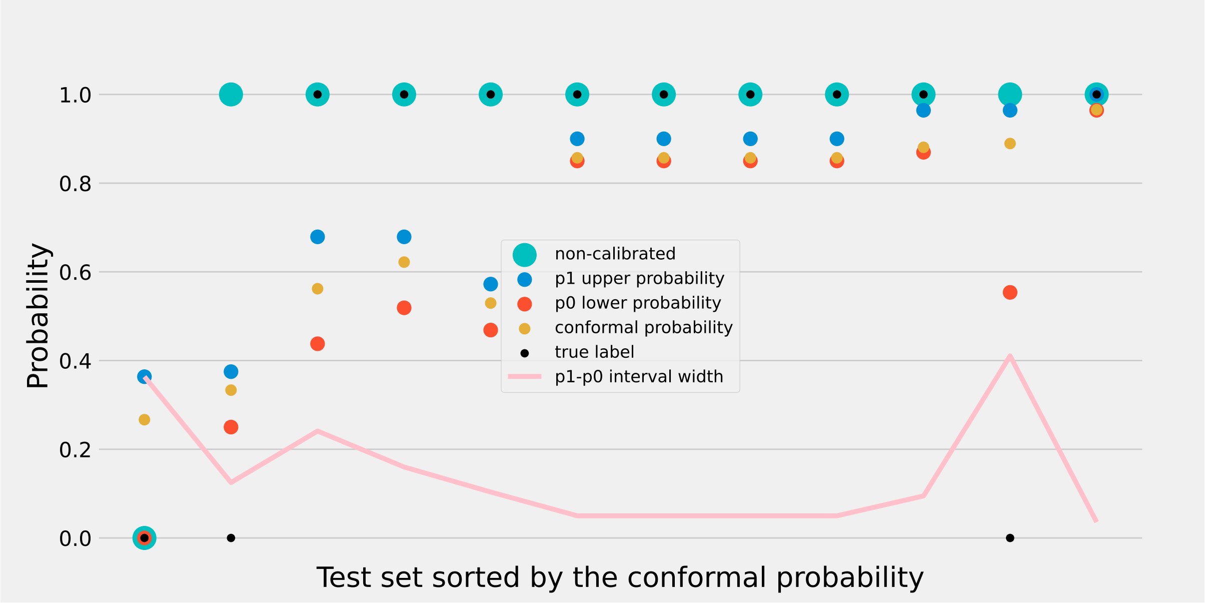

Figure 2: Probability vs conformal probability for MRI scans in the test set. Only thanks to the conformalization, the misclassified patient (second from the left) is correctly classified as a high glioma case. Large intervals show a low confidence in those predictions.

6. Conclusions

In this article, we address a global public health issue, ranked among the top 10 causes of cancer-related mortality worldwide: the detection of central nervous system tumors. Specifically, we focus on differentiating gliomas into their two types: high-grade gliomas (HGG) and low-grade gliomas (LGG). To achieve this, we introduce a novel methodology based on time series classification. By utilizing MRI-perfusion, we transform the image into a time series, called the perfusion curve, which reflects tissue vascularization in a non-invasive manner. To classify these perfusion curves between LGG and HGG, a conformal predictor based on a deep neural network is trained. We have compared our proposal with the state-of-the-art technique in time series classification through rigorous experimentation, based on a dataset of patients from the radiology department of the "Virgen de las Nieves Hospital" in Granada, Spain, obtaining satisfactory results. We demonstrate that our methodology is not only innovative in the way it transforms the image problem into sequences, but also that the combination of deep neural networks and conformal prediction for time series classification generates an ideal tool for radiologists This tool displays exceptional generalization capability and the reduced uncertainty in its predictions, making it useful, reliable and trustworthy. This study contributes to quantifying uncertainty at patient-level with applications in personalized and precision medicine. Besides, by improving the confidence in AI-medical applications, this study also aligns with the requirements of the European Parliament AI regulation to achieve Trustworthy AI.

Acknowledgments

This research has been partially funded by Spanish Ministry of Economy, Industry and Competitiveness (PID2020-118224RB-100 and PID2023-151336OB-I00), con-financed by the European Union (FEDER).

References

[1] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, *et al.*, Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012, Int. J. Cancer

136 (2015) E359e86.

[2] Q. Ostrom, L. Bauchet, F. Davis, *et al.*, The epidemiology of glioma in adults: a "state of the science" review, Neuro Oncology 16 (2014) 894–913.

[3] B. Rasmussen, S. Hansen, R. Larsen, M. Kosteljanetz, H. Schultz, B. Bøgård, *et al.*, Epidemiology of glioma: clinical characteristics, symptoms, and predictors of glioma pationes grade I–IV in the danish neuro-oncology registry, Journal of Neuro-Oncology 135 (2017) 571–549. doi:10.1007/s11060-017-2607-5.

[4] Y. Mizobuchi, K. Nkajima, T. Fujihara, K. Matsuzaki, H. Mure, S. Nagahiro, Y. Takagi, The risk of hemorrhage in steriotactic biopsy for brain tumorus, J. Medical Investigation 66 (2019) 317–318. doi:10.2152/jmi.66.314.

[5] J. Abrigo, D. Fountain, J. Provenzale, E. Law, J. Kwong, M. Hart, W. Tam, Magnetic resonance perfusion for differentiating low-grade frmo high-grade gliomas at first presentation (Review), Cochrane Database of Systematic Reviews (2018). doi:10.1002/14651858.CD011551.pub2.

[6] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, S. Nahavandi, A review of uncertainty quantification in deep learning: Techniques, applications and challenges, Information Fusion 76 (2021) 243–297. URL: https://www.sciencedirect.com/science/article/pii/S1566253521001081. doi:https://doi.org/10.1016/j.inffus.2021.05.008.

[7] H. Karimi, R. Samavi, Quantifying deep learning model uncertainty in conformal prediction, Proceedings of the AAAI Symposium Series 1 (2023) 142–148. URL: http://dx.doi.org/10.1609/aaaiss.v1i1.27492. doi:10.1609/aaaiss.v1i1.27492.

[8] J. Fayyad, S. Alijani, H. Najjaran, Empirical validation of conformal prediction for trustworthy skin lesions classification, Computer Methods and Programs in Biomedicine 253 (2024) 108231. URL: https://www.sciencedirect.com/science/article/pii/S0169260724002268. doi:https://doi.org/10.1016/j.cmpb.2024.108231.

[9] H. Olsson, K. Kartasalo, N. Mulliqi, M. Capuccini, P. Ruusuvuori, H. Samaratunga, B. Delahunt, C. Lindskog, E. A. M. Janssen, A. Blilie, L. Egevad, O. Spjuth, M. Eklund, I. P. I. E. Panel, Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction, Nature Communications 13 (2022) 7761. URL: https://doi.org/10.1038/s41467-022-34945-8. doi:10.1038/s41467-022-34945-8.

[10] C. Lu, K. Chang, P. Singh, J. Kalpathy-Cramer, Three applications of conformal prediction for rating breast density in mammography, 2022. URL: https://arxiv.org/abs/2206.12008. arXiv:2206.12008.

[11] C. Lu, A. N. Angelopoulos, S. Pomerantz, Improving trustworthiness of ai disease severity rating in medical imaging with ordinal conformal prediction sets, in: L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, S. Li (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2022, Springer Nature Switzerland, Cham, 2022, pp. 545–554.

[12] J. Vazquez, J. C. Facelli, Conformal prediction in clinical medical sciences, Journal of Healthcare Informatics Research 6 (2022) 241–252. URL: https://doi.org/10.1007/s41666-021-00113-8. doi:10.1007/s41666-021-00113-8.

[13] C. Lu, A. Lemay, K. Chang, K. Höbel, J. Kalpathy-Cramer, Fair conformal predictors for applications in medical imaging, Proceedings of the AAAI Conference on Artificial Intelligence 36 (2022) 12008–12016. URL: https://ojs.aaai.org/index.php/AAAI/article/view/21459. doi:10.1609/aaai.v36i11.21459.

[14] H. Wieslander, P. J. Harrison, G. Skogberg, S. Jackson, M. Fridén, J. Karlsson, O. Spjuth, C. Wählby, Deep learning with conformal prediction for hierarchical analysis of large-scale whole-slide tissue images, IEEE Journal of Biomedical and Health Informatics 25 (2021) 371–380. doi:10.1109/JBHI.2020.2996300.

[15] H.-H. Cho, S.-H. Lee, J. Kim, H. Park, Classification of the glioma grading using radiomics analysis, PeerJ 6 (2018) e5982.

[16] H. Mzoughi, I. Njeh, A. Wali, M. B. Slima, A. BenHamida, C. Mhiri, K. B. Mahfoudhe, Deep multi-scale 3d convolutional neural network (cnn) for mri gliomas brain tumor classification,

Journal of Digital Imaging 33 (2020) 903–915. URL: https://doi.org/10.1007/s10278-020-00347-9. doi:10.1007/s10278-020-00347-9.

[17] S. C. P. D. E. P. A. I. Agusti Alentorn, Alberto Duran-Peña, S. Kesari, Molecular profiling of gliomas: potential therapeutic implications, Expert Review of Anticancer Therapy 15 (2015) 955–962. URL: https://doi.org/10.1586/14737140.2015.1062368. doi:10.1586/14737140.2015.1062368. arXiv:https://doi.org/10.1586/14737140.2015.1062368.

[18] J. Haubold, R. Hosch, V. Parmar, M. Glas, N. Guberina, O. A. Catalano, D. Pierscianek, K. Wrede, C. Deuschl, M. Forsting, F. Nensa, N. Flaschel, L. Umutlu, Fully automated mr based virtual biopsy of cerebral gliomas, Cancers 13 (2021). URL: https://www.mdpi.com/2072-6694/13/24/6186. doi:10.3390/cancers13246186.

[19] M. Kim, S. Y. Jung, J. E. Park, Y. Jo, S. Y. Park, S. J. Nam, J. H. Kim, H. S. Kim, Diffusion- and perfusion-weighted mri radiomics model may predict isocitrate dehydrogenase (idh) mutation and tumor aggressiveness in diffuse lower grade glioma, European Radiology 30 (2020) 2142–2151. URL: https://doi.org/10.1007/s00330-019-06548-3. doi:10.1007/s00330-019-06548-3.

[20] Y. Ren, X. Zhang, W. Rui, H. Pang, T. Qiu, J. Wang, Q. Xie, T. Jin, H. Zhang, H. Chen, Y. Zhang, H. Lu, Z. Yao, J. Zhang, X. Feng, Noninvasive prediction of idh1 mutation and atrx expression loss in low-grade gliomas using multiparametric mr radiomic features, Journal of Magnetic Resonance Imaging 49 (2019) 808–817. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jmri.26240. doi:https://doi.org/10.1002/jmri.26240. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/jmri.26240.

[21] Z. A. Shboul, J. Chen, K. M. Iftekharuddin, Prediction of molecular mutations in diffuse low-grade gliomas using mr imaging features, Scientific Reports 10 (2020) 3711. URL: https://doi.org/10.1038/s41598-020-60550-0. doi:10.1038/s41598-020-60550-0.

[22] E. Calabrese, J. D. Rudie, A. M. Rauschecker, J. E. Villanueva-Meyer, J. L. Clarke, D. A. Solomon, S. Cha, Combining radiomics and deep convolutional neural network features from preoperative MRI for predicting clinically relevant genetic biomarkers in glioblastoma, Neuro-Oncology Advances 4 (2022) vdac060. URL: https://doi.org/10.1093/noajnl/vdac060. doi:10.1093/noajnl/vdac060. arXiv:https://academic.oup.com/noa/article-pdf/4/1/vdac060/43778051/vdac060.pdf.

[23] AI Act — digital-strategy.ec.europa.eu, https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai, 2024. [Accessed 24-04-2024].

[24] N. Díaz-Rodríguez, J. Del Ser, M. Coeckelbergh, M. López de Prado, E. Herrera-Viedma, F. Herrera, Connecting the dots in trustworthy artificial intelligence: From ai principles, ethics, and key requirements to responsible ai systems and regulation, Information Fusion 99 (2023) 101896. URL: https://www.sciencedirect.com/science/article/pii/S1566253523002129. doi:https://doi.org/10.1016/j.inffus.2023.101896.

[25] CHAI — coalitionforhealthai.org, https://www.coalitionforhealthai.org/, 2024. [Accessed 24-04-2024].

[26] C. Molnar, Introduction to conformal prediction with python, 2023.

[27] V. Vovk, I. Petej, Venn-abers predictors, in: Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence, UAI'14, AUAI Press, Arlington, Virginia, USA, 2014, p. 829–838.

[28] J. Platt, Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, 1999. URL: https://api.semanticscholar.org/CorpusID:56563878.

[29] V. Manokhin, Practical guide to applied conformal prediction in python: Learn and apply the best uncertainty frameworks to your industry applications, 2023.

[30] B. Zadrozny, C. Elkan, Transforming classifier scores into accurate multiclass probability estimates, in: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '02, Association for Computing Machinery, New York, NY, USA, 2002, p. 694–699. URL: https://doi.org/10.1145/775047.775151. doi:10.1145/775047.775151.

[31] A. Niculescu-Mizil, R. Caruana, Predicting good probabilities with supervised learning, in: Proceedings of the 22nd International Conference on Machine Learning, ICML '05, Association for Computing Machinery, New York, NY, USA, 2005, p. 625–632. URL: https://doi.org/10.1145/

1102351.1102430. doi:10.1145/1102351.1102430.

[32] S. Arvidsson, O. Spjuth, L. Carlsson, P. Toccaceli, Prediction of metabolic transformations using cross Venn-ABERS predictors, in: A. Gammerman, V. Vovk, Z. Luo, H. Papadopoulos (Eds.), Proceedings of the Sixth Workshop on Conformal and Probabilistic Prediction and Applications, volume 60 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 118–131. URL: https://proceedings.mlr.press/v60/arvidsson17a.html.

[33] I. Nouretdinov, S. G. Costafreda, A. Gammerman, A. Chervonenkis, V. Vovk, V. Vapnik, C. H. Y. Fu, Machine learning classification with confidence: application of transductive conformal predictors to MRI-based diagnostic and prognostic markers in depression, Neuroimage 56 (2010) 809–813.

[34] D. Devetyarov, I. Nouretdinov, B. Burford, S. Camuzeaux, A. Gentry-Maharaj, A. Tiss, C. Smith, Z. Luo, A. Chervonenkis, R. Hallett, V. Vovk, M. Waterfield, R. Cramer, J. F. Timms, J. Sinclair, U. Menon, I. Jacobs, A. Gammerman, Conformal predictors in early diagnostics of ovarian and breast cancers, Progress in Artificial Intelligence 1 (2012) 245–257. URL: https://doi.org/10.1007/s13748-012-0021-y. doi:10.1007/s13748-012-0021-y.

[35] H. Papadopoulos, A. Gammerman, V. Vovk, Reliable diagnosis of acute abdominal pain with conformal prediction, International journal of engineering intelligent systems for electrical engineering and communications 17 (2009) 127–137. URL: https://api.semanticscholar.org/CorpusID:18515829.

[36] A. Lambrou, H. Papadopoulos, A. Gammerman, Evolutionary conformal prediction for breast cancer diagnosis, in: 2009 9th International Conference on Information Technology and Applications in Biomedicine, 2009, pp. 1–4. doi:10.1109/ITAB.2009.5394447.

[37] L. Alnemer, L. Rajab, I. Aljarah, Conformal prediction technique to predict breast cancer survivability, International journal of advanced science and technology 96 (2016) 1–10. doi:10.14257/ijast.2016.96.01.

[38] A. S. Millar, J. Arnn, S. Himes, J. C. Facelli, Uncertainty in breast cancer risk prediction: A conformal prediction study of race stratification (2024). URL: http://dx.doi.org/10.3233/SHTI231113. doi:10.3233/shti231113.

[39] S. Hernandez-Hernandez, Q. Guo, P. J. Ballester, Conformal prediction of molecule-induced cancer cell growth inhibition challenged by strong distribution shifts, bioRxiv (2024). URL: https://www.biorxiv.org/content/early/2024/03/17/2024.03.15.585269. doi:10.1101/2024.03.15.585269. arXiv:https://www.biorxiv.org/content/early/2024/03/17/2024.03.15.585269.full.pdf.

[40] M. Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, E. Silverman, An Empirical Distribution Function for Sampling with Incomplete Information, The Annals of Mathematical Statistics 26 (1955) 641 – 647. URL: https://doi.org/10.1214/aoms/1177728423. doi:10.1214/aoms/1177728423.

[41] A. Gammerman, V. Vapnik, V. Vovk, Learning by transduction, in: Proceedings of the Fourteenth Conference on Uncertainty in Articial Intelligence, Morgan Kaufmann, 1998, pp. 148–156.

[42] C. Saunders, A. Gammerman, V. Vovk, Transduction with confidence and credibility, in: Sixteenth International Joint Conference on Artificial Intelligence (IJCAI '99) (01/01/99), 1999, pp. 722–726. URL: https://eprints.soton.ac.uk/258961/.

[43] V. Vovk, A. Gammerman, C. Saunders, Machine-learning applications of algorithmic randomness, in: Proceedings of the Sixteenth International Conference on Machine Learning, Morgan Kaufmann, 1999, pp. 444–453.

[44] V. Manokhin, Multi-class probabilistic classification using inductive and cross Venn–Abers predictors 60 (2017) 228–240. URL: https://proceedings.mlr.press/v60/manokhin17a.html.

[45] V. VOVK, Algorithmic learning in a random world, 2023.

[46] T. Pereira, S. Cardoso, M. Guerreiro, A. Mendonça, de, S. C. Madeira, Alzheimer's Disease Neuroimaging Initiative, Targeting the uncertainty of predictions at patient-level using an ensemble of classifiers coupled with calibration methods, Venn-ABERS, and conformal predictors: A case study in AD, J. Biomed. Inform. 101 (2020) 103350.

[47] K. J. Friston, Statistical Parametric Mapping, Springer US, Boston, MA, 2003, pp. 237–250. URL: https://doi.org/10.1007/978-1-4615-1079-6-16. doi:10.1007/978-1-4615-1079-6-16.

[48] R. Ranjbarzadeh, A. Caputo, E. B. Tirkolaee, S. Jafarzadeh Ghoushchi, M. Bendechache, Brain tumor segmentation of mri images: A comprehensive review on the application of artificial intelligence

tools, Computers in Biology and Medicine 152 (2023) 106405. URL: https://www.sciencedirect.com/science/article/pii/S0010482522011131. doi:https://doi.org/10.1016/j.compbiomed.2022.106405.

[49] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka, J. Buatti, S. Aylward, J. V. Miller, S. Pieper, R. Kikinis, 3d slicer as an image computing platform for the quantitative imaging network, Magnetic Resonance Imaging 30 (2012) 1323–1341. URL: https://www.sciencedirect.com/science/article/pii/S0730725X12001816. doi:https://doi.org/10.1016/j.mri.2012.05.001, quantitative Imaging in Cancer.

[50] A. Bagnall, A. Bostrom, J. Large, J. Lines, The great time series classification bake off: An experimental evaluation of recently proposed algorithms. extended version, 2016. URL: https://arxiv.org/abs/1602.01711. arXiv:1602.01711.

[51] B. Dhariyal, T. L. Nguyen, G. Ifrim, Back to basics: A sanity check on modern time series classification algorithms, 2023. URL: https://arxiv.org/abs/2308.07886. arXiv:2308.07886.