Brain Tumor Classification Using MRI Images

Copenhagen Business School

MSc in Business Administration and Data Science

Machine Learning and Deep Learning

Authors:
Francesco Esposito (student ID [redacted])
Stylianos Ilias Skafidas (student ID [redacted])
Mattia Malipiero (student ID [redacted])

Submission Date: 16 May 2025

Number of Pages: 13 pages

Number of Characters: 33112 characters

Abstract

Accurate and early diagnosis is one of the most critical factors in effective brain tumor treatment. This study presents a comparative analysis of four machine learning models developed to support physicians in the classification of brain tumors using MRI scan data. We construct a unified dataset from multiple public sources, applying rigorous preprocessing techniques including normalization, grayscale conversion, and data augmentation. Three distinct modeling methodologies are employed: a baseline Multilayer Perceptron (MLP), two Convolutional Neural Networks (CNNs) of differing complexity, and a Support Vector Machine (SVM), the latter optimized through Bayesian hyperparameter tuning. The models are evaluated using accuracy, precision, recall, and F1 score. Both the optimized SVM and the complex CNN yielded adequately high performances, with the SVM achieving 93.6% accuracy and 93.01 F1 score, and the CNN reaching 90% for both. Beyond reporting results, we discuss limitations, address ethical considerations in a medical context, and propose directions for future research focused on clinical integration and model interpretability. This work contributes to the ongoing effort to evaluate how model complexity and methodological choices affect classification performance in real-world medical contexts.

Keywords: brain tumor, image classification, Magnetic Resonance Imaging (MRI), Convolutional Neural Networks (CNN), Support Vector Machine (SVM)

1. Introduction

Artificial intelligence (AI), and more specifically machine learning (ML), has been met with great enthusiasm in the field of medicine due to its broad range of applications, including drug discovery, disease prediction, and diagnostic support (Rahmani et al., 2021). Within ML, image processing has emerged as a particularly impactful subfield, enabling medical advancements in areas such as deep venous thrombosis detection (Contreras-Luján et al., 2022), cancer screening (Kale et al., 2024), and, notably, brain tumor identification (Alam et al., 2024). In the case of brain tumors, diagnosis and accurate classification rely on the interpretation of magnetic resonance imaging (MRI) scans by radiologists. However, this process is time consuming and prone to human error, as radiologists may provide differing assessments (Alam et al., 2024). To this end, ML-based image processing tools have shown great potential in assisting doctors by significantly reducing diagnosis times and achieving better accuracy than field experts (Forghani, 2020; Chang et al., 2023).

This paper examines the performance of different modeling techniques in labeling MRI scans across the four most common brain tumor categories (Price et al., 2024), which can serve as a primary classification system for doctors at the early stages of a diagnosis. Specifically, we compare four custom models, all trained on the combination of four publicly sourced datasets, with overlapping classes, comprising a total of 13.608 MRI scans.

2. Research Questions & Motivation

This research aims to streamline the diagnostic process for brain tumors by developing and evaluating multiple machine learning models that support doctors in identifying tumor types both rapidly and accurately. The importance of such tools is emphasized by Jabbar et al. (2023), who note that delays in timely tumor diagnosis significantly contribute to the high mortality associated with brain tumors. The aforementioned motivation leads us to the following research question and sub-questions:

How does a diverse set of models perform in classifying brain tumors from MRI scans?

RQ1: What are the practical trade-offs between different model types?

RQ2: How does model complexity impact the model's overall performance?

3. Related Work

The urgency of accurate brain tumor diagnosis has led to a surge of research in the past few years, with many studies proposing a diverse set of machine learning solutions. Ghaffar et al. (2024) built upon the EfficientNet-B1 architecture by applying numerous image preprocessing techniques, along with minor internal modifications to the model's structure. Their enhanced network achieved high classification accuracy across brain tumor types. The dataset they employed consisted of nearly 6,500 MRI images, with class categories closely aligned with those used in our study, making their work directly comparable to ours. Semwal et al. (2025) proposed a hybrid classification model consisting of convolutional neural networks and support vector machines, whose model's hyperparameters were later optimized using particle swarm optimization. Although their model is trained on a relatively small dataset of approximately 3000 MRI images of the same classes we have, they have achieved fairly satisfactory results. Their work offers a useful point of comparison for our study, particularly in examining the effect of dataset size and model complexity. Moreover, Díaz-Pernas et al. (2021) propose a more complex model which contains three CNN branches working in parallel, that operate on different receptive fields, capturing low, mid, and high-level patterns simultaneously, and thus achieving high accuracy. However, their approach requires a lengthy and sophisticated preprocessing stage. In contrast, our models follow a more straightforward design, prioritizing simplicity, and speed. Contrary to the prior highly complex models, Basthikodi et al. (2024**)** introduce a basic and enhanced version of a model grounded on the support vector machine classification algorithm. Their baseline model achieved satisfactory accuracy of 86.5%, whereas they gradually introduced Histogram Oriented Gradients, Local Binary Pattern, and Principal Component Analysis as feature extraction and dimensionality reduction techniques, effectively optimizing their model accuracy. While the overall structure of our SVM-based model is comparable, they do not mention any form of hyperparameter optimization, whereas ours incorporates Bayesian optimization for hyperparameter tuning.

4. Conceptual Framework

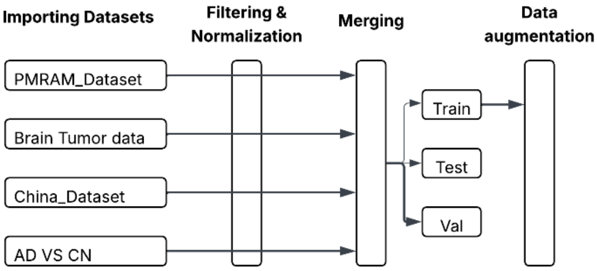


Figure 1: Overview of data preparation.

Our paper employs four distinct datasets to ensure comprehensive representation and enhanced generalizability of results. As with most deep learning tasks involving image classification, an initial exploratory analysis of each dataset is essential (Goodfellow et al., 2016). This exploration helps identify intrinsic characteristics, such as the file format of the datasets and the overall distribution of the classes. Figure 1 provides an overview of the introductory process before the training of the models.

Upon downloading the datasets, we cleaned the datasets to ensure that only readable images proceed further. We then preprocess the images which leaves us with consistent dimensions in terms of image intensity and resolution, bringing the dataset to a uniform condition

Finally, to address our dataset's limited size and existing class imbalances we employ augmentation techniques. This introduces controlled variability into our datasets, promoting model robustness and enhancing its generalizability.

After the augmentation, the datasets are merged and subsequently split into training, validation, and testing sets.

We continue with the creation of four custom models. Firstly, we develop a simple multilayer perceptron model as baseline and proceed with two CNN models which differ in level of complexity. Finally, a standard vector machine model is created and optimized using Bayesian Optimization. For the development of the models, numerous normalization techniques are put into place to avoid overfitting and boost the models' performance and accuracy. The models are then evaluated across different metrics and compared against each other.

These elements together form a conceptual pipeline that links data preparation with model design and performance evaluation. The preprocessing phase serves to standardize input complexity, while augmentation introduces variability to enhance generalization. The inclusion of models with varying levels of complexity (from MLP to CNN to SVM) allows for comparative evaluation across both learning depth and architecture type. This structured approach enables a nuanced understanding of how each component contributes to overall classification performance.

5. Methodology

5.1 Dataset Description

The models presented in our paper were trained and tested on a merged dataset constructed from four publicly available datasets. Each dataset required a tailored loading procedure due to differences in directory structure, file formats, and metadata. An overview of the datasets, including number of classes and image counts, is provided in Table 1.

Name	Classes	Images
(1) PMRAM_Dataset	4	1505
(2) Brain Tumor data	4	7023
(3) China_Dataset	3	3064
(4) AD VS CN	1	2016

Table 1: Datasets overview

The *PMRAM_Dataset* comprises moderate resolution images of MRI classified into four categories: *no_tumor*, *glioma*, *pituitary*, and *meningioma*. All images are square (512x512), with a resolution of 96 DPI and 24-bit color depth. An augmented version of this dataset was available but excluded from our study, as we applied our own augmentation techniques during preprocessing.

The *Brain Tumor data* dataset includes 7023 MRI scans spanning the same four classes. The images have a broad range of different resolutions ranging from 512x512 to 221x228. The dataset was originally split into separate training and testing folders, which were unified into one folder.

The *China_Dataset* contains 3,064 square images, primarily with 512×512 resolution, and a small subset at 256×256. This dataset does not include the *no_tumor* class, which we added as an empty category to maintain class consistency across all datasets. The original files were provided in *.mat* format and were converted to *.jpg* for compatibility with our image processing pipeline.

The *AD_VS_CN Dataset* initially contained 5,526, with resolutions ranging from 256x256 to 160x192. The dataset consists of Alzheimer's patients MRI scans and a control group of healthy brain pictures. We drop the first group's image and utilize healthy MRI images adding the remaining 2016 images to our *no_tumor* class.

Figure 2 depicts some sample images.

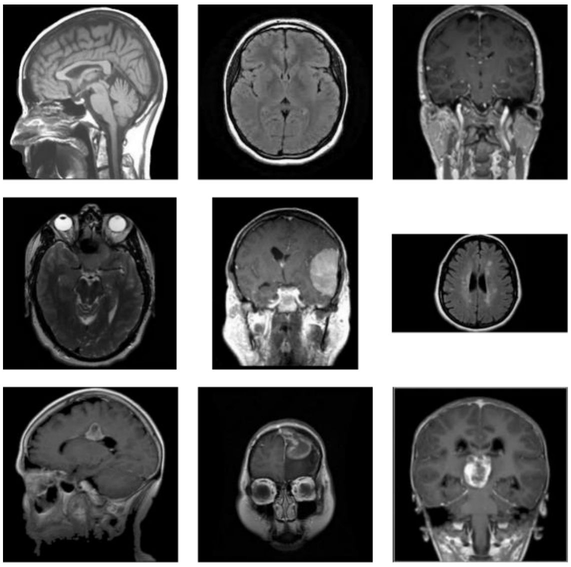


Figure 2: Sample images from the four datasets

While both the *PMRAM_Dataset* and *China_Dataset* exhibit relatively uniform image resolution, the *Brain Tumor data* and *AD_VS_CN* Dataset contains considerable variation in image dimensions. These inconsistencies were addressed during the preprocessing stage, detailed in the next section. Among the four datasets, only the *China_Dataset* includes extensive metadata on image origin and clinical attributes. Although the *PMRAM_Dataset* has some descriptive information, they were disregarded due to inconsistencies with the actual contents of the images.

5.2 Data Preprocessing

The figure below (Fig. 3) displays the data filtering and normalization workflow we followed before merging our datasets.

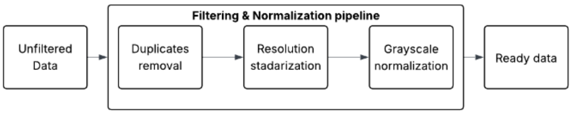


Figure 3: Filtering & Normalization process.

a) Duplicates removal: We commence by ensuring our datasets do not contain duplicate values. For this process we utilize the hash algorithm. This algorithm generates a fixed-length 160-bit hash value from the binary content, allowing us to detect exact duplicates by comparing hash outputs (National Institute of Standards and Technology, 2015). In turn we used the perceptual hash, as seen in Li et al. (2016), to conduct a more thorough check. In total we detected 3882 duplicates images which were removed from the dataset. Table 2 presents the change in numbers after removing duplicates.

Dataset	Total	Unique	Change %
(1) PMRAM_Dataset	1505	726	-51.8%
(2) Brain Tumor data	7023	5259	-25.1%
(3) China_Dataset	3064	1174	-61.7%
(4) AD VS CN	2016	1811	-10.2%

Table 2: Image count after duplicates removal

b) Image resizing: Secondly, we examine the images' dimensions by analyzing their aspect ratios. This step helps identify abnormalities and outliers while supporting the standardization of input sizes for the models. As shown in Figure 4, the original dataset contains a wide variation in aspect ratios. To address this, all images are resized to a fixed resolution of 256×256 pixels, ensuring uniformity across the dataset. Figure 4 illustrates the distribution of aspect ratios both before and after the resizing process.

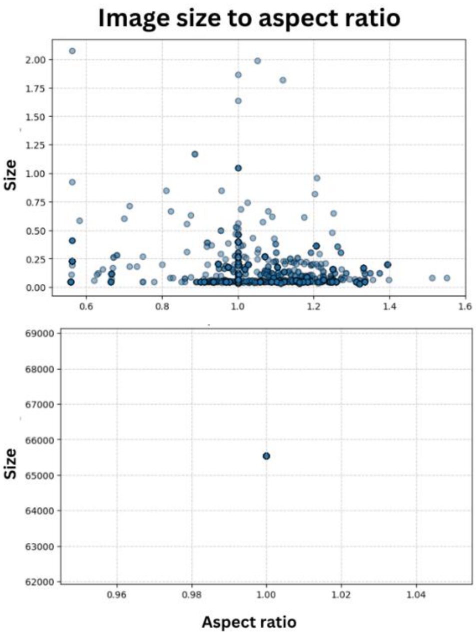


Figure 4: Aspect ratio before and after resizing.

c) *Grayscale conversion:* Lastly, we ensure that our images are in grayscale. Besides uniformity, this process simplifies the data by reducing the image into a single intensity channel, reducing complexity and enhancing model performance through lower memory usage. If all three intensity channels are equal, the image is in grayscale already, and thus we drop the two channels. Otherwise, they are logged and transformed into grayscale. From our data, 56 images were converted from RGB to grayscale.

At this point we have completed the filtering and normalization process. The four datasets are merged and split into Train, Test, and Validation, with percentages 70-15-15% respectively. Table 3 depicts the distribution of images across datasets and classes.

Classes	Train	Test	Validation
glioma	2023	435	433
meningioma	1251	269	268
no_tumor	1869	402	400
pituitary	1249	269	267

Table 3: Images distribution across classes & datasets.

5.3 Data Augmentation

As deep learning models typically require large and diverse datasets, we employed data augmentation to further enhance data volume and variation. This process allows us to synthetically generate new training samples through transformations into existing images, and thus mitigate the risk of overfitting, improving the robustness of our models. These augmentations were designed to introduce plausible scenarios that might be encountered in clinical MRI acquisition without altering the underlying diagnostic information.

We take the *glioma* class, being the largest, and augment the other classes until they reach the same number of instances, thus balancing our training set. Admittedly, balancing all classes in this manner leads to the previously leading class having no augmented instances. This can in turn impair the augmentation's initial purpose by creating some bias. However, we decided not to augment further in avoidance of introducing too many synthetic instances that would smoothen our parameters and increase the gap between the train and the other two sets.

During training, the augmentation pipeline randomly selects one transformation from a predefined set for each image. The intensity and specific details of the chosen transformation are then determined by sampling from preset parameter ranges, creating a unique augmented version of the image each time.

Specific transformations included:

Geometric Rotations: Images were subjected to minor in-plane rotations, with the angle of rotation randomly sampled from a uniform distribution ranging between -15 and +15 degrees. This simulates slight variations in patient positioning or scan orientation.

Scale Variations (Zoom): To account for differences in effective field-of-view or subject distance, a random isotropic zoom factor was applied, varying between 0.8x (zoom out) and 1.2x (zoom in) of the original image dimensions. The resulting images were either cropped or padded with reflected boundary pixels to maintain the original image size.

Image Blurring: A gentle Gaussian blur was introduced with a randomly selected kernel size (from 3x3, 5x5, or 7x7 pixels) and a sigma value randomly chosen between 0.1 and 1.0. This simulates minor focus imperfections or subtle noise smoothing effects.

Brightness Adjustments: The overall image brightness was randomly modulated, introducing variations up to a +20% change from the original pixel intensities. This addresses potential differences in scanner calibration or image exposure levels. Figure 5 displays the different augmented image possibilities.

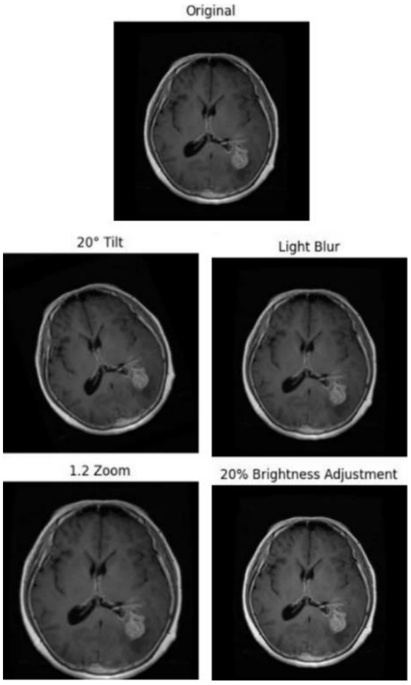


Figure 5: Augmented images

6. Model Architectures

In this paper, we introduce four custom models trained and evaluated on our dataset. The first is a simple fully connected neural network (Multilayer Perceptron), which serves as our baseline. In addition, we present two custom-built Convolutional Neural Networks (CNNs) that differ in complexity and architectural design. The final model is a Support Vector Machine (SVM) pipeline, which we further optimize using Bayesian hyperparameter optimization. These models are compared against one another.

6.1 Fully Connected Neural Network

The Multilayer Perceptron (MLP) we constructed consists of three layers: input, hidden, and output. It begins with a Flatten layer that reshapes the 2D input image into a 1D vector, preparing the data for the dense layers. The hidden *Dense* layer contains 128 units and uses the *ReLU* activation function, which introduces non-linearity while maintaining computational efficiency, allowing the model to learn more complex patterns. The output Dense layer uses a Softmax activation function to produce a probability distribution over the target classes. Although simple and fast to train, we expect this model to underperform compared to the other architectures, due to its inability to capture complex patterns such as spatial relationships between pixels.

6.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a class of deep learning models specifically designed to process data with a grid-like structure, such as images. Unlike fully connected networks, CNNs can capture spatial hierarchies by learning local patterns within their receptive fields, making them particularly effective for image modeling.

Our second model contains two convolutional layers followed by pooling layers to summarize the extracted features and reduce computational load. For the CNN layers we use ReLu as an activation function and include padding to preserve edge information during filtering. Additionally, we employ Batch Normalization which helps maintain consistent activation distributions, enhances gradient flow, and provides a regularizing effect. Moreover, we add two

Dropout layers to prevent overfitting and enhance generalizability. After employing the feature extraction stages, we flatten the image data and add two dense layers with *ReLu* and *Softmax* activation functions, to produce class probabilities.

The second, more complex CNN model includes an additional convolutional layer compared to the simpler architecture, allowing it to extract more hierarchical features. Furthermore, *L2 regularization* is introduced in deeper convolutional layers to penalize large weights and reduce overfitting. This model uses a larger number of filters in its convolutional layers, namely 32, 64, and 64 filters, whereas the simpler CNN used 16 and 32 filters. This increase in depth and filter size allows the model to capture more abstract and high-level patterns from the input images, at the cost of increased computational complexity. The remaining architecture mirrors the simpler model, with ReLU activations, MaxPooling, Dropout layers for regularization, and a fully connected Dense layer before the final Softmax output. In both our models, we use a 64-batch size and an Adam optimizer with a learning rate LR =1e-4, due to its adaptive learning capabilities. Lastly, we use a categorical cross-entropy loss function which is commonly used for multi-class classification tasks with mutually exclusive categories, such as image-based tumor classification. Given the demonstrated superiority of CNNs in image modelling, we contend these models ought to exceedingly outperform our formerly introduced baseline model and perform competitively against the SVM model and its variations presented below.

6.3 Support Vector Machine

The SVM pipeline consists of a HOG Transformer which extracts the texture and shape of the features using Histogram Oriented Gradients. The HOG parameters are selected based on heuristic considerations. Furthermore, we apply a Standard Scaler to ensure that all features contribute equally during classification which is fundamental for SVMs due to their sensitivity to the scale of input features. Then, we apply Principal Component Analysis for dimensionality reduction, while preserving 95% of the variance. The final classification layer is a Support Vector Machine using a Gaussian Radial Basis Function (RBF) kernel, selected to address the non-linear nature of the problem and initially configured with random-based parameters. This kernel enables the model to capture complex similarity relationships by implicitly mapping the input data into a higher dimensional space, without incurring the computational cost of explicitly performing the transformation.

For the selection of the number of iterations we performed several experiments, adding 2000, 1000, 800 and 400 iterations. The 2000 iterations failed due to limited processing resources, whereas the 1000 and 800 did not yield meaningful improvements, and rather consumed resources redundantly.

To optimize this model further, we evaluated different hyperparameter search strategies. We discarded Grid Search due to its exhaustive nature, which is computationally prohibitive in high-dimensional spaces. We then tested Random Search, which is more resource-efficient but still exceeded our resource constraints. Ultimately, we adopted Bayesian Optimization, which efficiently explores the hyperparameter space by modeling the objective function with a probabilistic surrogate (Géron 2022). Bayesian optimization gradually focuses on the most promising regions of the search space, allowing us to minimize computational overhead while still tuning the model effectively. Multiple variations of this optimization strategy were tested to further economize resource usage. The SVM model presented in the results is the outcome of the most efficient of such optimization.

6.4 Model Validation Criteria

To evaluate and compare the performance of our models, we employ a set of metrics commonly used in the literature of brain tumor classification models (Rainio, Teuho, & Klén, 2024; Li, Li, & Su, 2021; Contreras-Luján et al., 2022).

a) Accuracy: the ratio of the number of correct predictions in the sample to the total number of samples. The calculation formula is as follows:

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

b) *F1-Score*: this index considers both the precision and recall rate of the classification model and can be regarded as a weighted average of the precision and recall rate of the model. The calculation formula is as follows:

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

c) *Precision*: Precision: the ratio of the samples with positive prediction to all samples with positive prediction. The calculation formula is as follows:

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

d) *Recall*: it refers to the ratio of positive samples correctly predicted to all positive samples. The calculation formula is as follows:

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

Primarily, we focus on the accuracy metric across all dataset splits to assess overall model performance and detect potential overfitting. Additionally, we monitor the validation loss as an indicator of overfitting or issues such as vanishing gradients. To provide a more detailed evaluation, we also report precision, recall, and the F1 score, which together offer a more balanced perspective on the model's classification performance—particularly in the presence of class imbalance.

7. Results

Model	*Acc.*	*V. Acc.*	*Prec.*	*Recall*	*F1*
MLP	88.1%	81.4%	79%	78%	78%
CNN1	97.1%	86.7%	87%	85%	86%
CNN2	91.6%	89.0%	88%	89%	89%
SVM	99.60%	93.6%	93%	93%	93%

Table4: Model evaluation metrics.

7.1 Model Summary & Comparison

Table 4 displays the performance of the different models, showcasing the best performing version of each model out of many executions.

The multilayer perceptron we used as baseline accomplished satisfactory results, reaching a validation accuracy of 81.4% and an F1-score of 78% after 25 epochs. Despite its simplicity, it required a relatively long training time and underperformed across all key metrics.

The first CNN showed superior performance than the baseline by obtaining a validation accuracy of 86.7% and an F1-score of 81% after 11 epochs. However, the difference between train accuracy and validation accuracy is very high, showing signs of overfitting.

CNN2 partially solves this bottleneck, by reducing the difference between train and validation accuracy to just 1.7% after 8 epochs. On the test set, CNN2 achieved an accuracy and F1-score of 90%. While performance plateaued beyond 8 epochs, the model offered a more balanced trade-off between depth and generalization than its predecessor.

The Support Vector Machine (SVM), enhanced through feature extraction using Histogram of Oriented Gradients (HOG) and Principal Component Analysis (PCA), as well as Bayesian hyperparameter optimization, outperformed all other models. It achieved a validation accuracy of 93.6% and an F1-score of 93.0%, with balanced precision and recall. Even without optimization, the SVM performed competitively. Table 5 summarizes the training duration for each model, where the SVM stands out once again, completing its optimized training in only 3 minutes, compared to 8 minutes without optimization. All models were trained on a machine with 53GB RAM and a NVIDIA L4 GPU with 22.5 GB VRAM. During model development and testing, multiple machines were used to save resources and optimize running time.

Model	***Duration (min)***	**Epochs**
MLP	7	25
CNN1	6	10
CNN2	6	7
SVM	3	-

Table 5: Duration of models.

Among the four classes, the *meningioma* class seems to be the worst classified by all the models. Even the SVM shows an F1-score of 86% for meningioma class images. This underperformance may stem from class imbalance, as meningioma had the fewest images (1,251), compared to the largest class, glioma (2,023 images).

It is also notable that meningiomas are the least malignant and often more visually distinguishable by human experts (Ostrom et al., 2022), potentially leading models to focus disproportionately on subtle features rather than broader patterns.

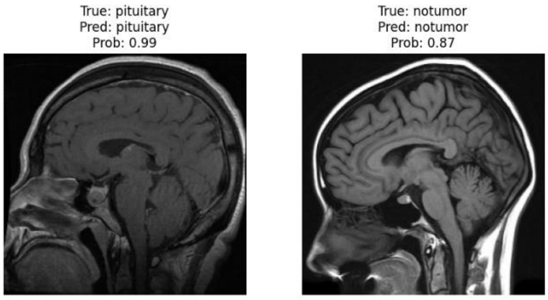


Figure 6: Execution of CNN2 model

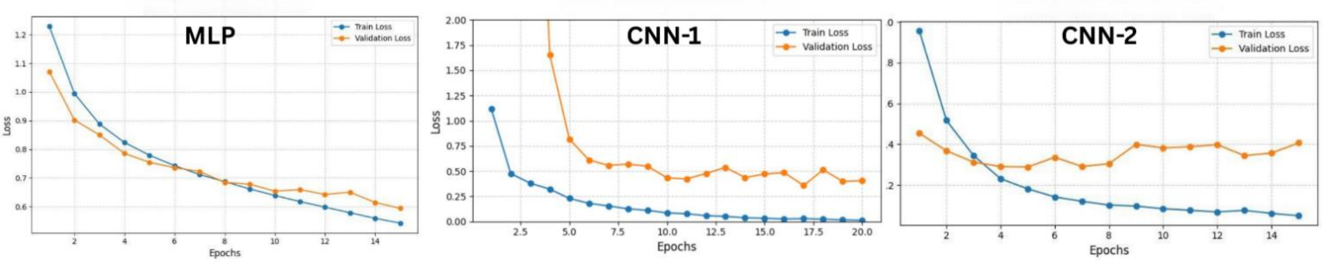


Figure 7: Training loss results

8. Discussion

The results of our study reflect a clear hierarchy in model performance and efficiency, with the optimized Support Vector Machine (SVM) outperforming all other models across key metrics. The SVM achieved the highest validation accuracy and F1-score while also training in just 3 minutes, making it the most computationally efficient option. Its strong performance can be attributed to the use of well-established feature extraction techniques coupled with the deployment of the Bayesian Optimization algorithm.

Convolutional Neural Networks (CNNs), typically considered the ideal standard in image classification tasks, ranked second. CNN1 outperformed the MLP baseline but showed strong signs of overfitting, as reflected by the significant gap between training and validation accuracy. CNN2, which featured a deeper architecture and improved regularization, mitigated this overfitting and achieved solid test performance (90% accuracy and F1-score). However, its training time and resource demands were significantly higher, and performance plateaued early, likely due to dataset limitations.

We hypothesize that this performance hierarchy would shift with access to a larger and more diverse dataset. CNNs tend to scale better with data volume, and a more complex CNN architecture would likely outperform the SVM in both accuracy and generalization.

Nonetheless, our focus in this study was on exploring how models built on different methodologies and with different levels of complexity perform. when trained on a moderately sized, preprocessed dataset. Within this scope, the SVM proved to be the most effective and efficient model.

Class-wise performance also revealed consistent misclassification of meningioma cases. Meningioma, while the least malignant and often visually distinct to human observers (Ostrom et al., 2022), achieved the lowest F1-scores across all models. This can likely be attributed to class imbalance, as meningioma was the least represented in the dataset. The underrepresentation may have prevented the models from learning strong distinguishing features, leading to lower recall and precision for this class. Further data augmentation or targeted sampling could help address this issue in future work.

This study is subject to several limitations. To begin with, the dataset size although relatively sufficient for training and evaluation, was not large enough to assess the upper limits of performance achievable of deeper and more complex architectures and thus to explore the full competences of CNN-based methods. Moreover, our team lacks the domain expertise to delve deeper into the real-world diagnostic workflows. Lastly, application of machine learning in the brain tumor classification context is a well-researched subject. As a result, attempting to contemplate the full aspects of this topic given the assignment's limitations would be disproportionate.

Ethical considerations are also fundamental in the creation of machine learning models in a medical context. Although all datasets we use did not offer any details as to the patients the MRI scans originated from, any future deployment must ensure strict patient privacy and data security throughout the machine learning pipeline. Techniques such as federated learning may help mitigate risks by decentralizing model training. Additionally, developers and doctors must remain vigilant regarding dataset diversity. Models trained on datasets that underrepresent certain ethnic or racial groups may develop hidden biases, which can impair diagnostic performance.

Overall, this study demonstrates that model selection in medical image classification must account for both diagnostic performance and practical feasibility. While CNNs offer strong modeling capabilities, their success is closely tied to data availability and computational power. In contrast, the optimized SVM provided a strong balance of accuracy, recall, and training efficiency, making it an appealing alternative.

9. Conclusion

In this paper, we developed four models using different methodologies for brain tumor multi-class classification, with the aim of achieving both high accuracy and computational efficiency.

We began by introducing four publicly available datasets containing MRI scans of the most common brain tumor types. At first, we loaded the datasets and did the necessary conversions for them to be in readable form. Subsequently we filtered and normalized the data ensuring their integrity and uniformity. Then we balanced the dataset through data augmentation before merging the datasets and splitting them to train and evaluate the models.

As anticipated, the Multilayer Perceptron baseline model, with its relatively simple architecture, achieved moderate performance. The two Convolutional Neural Network (CNN) models performed significantly better, with the more complex architecture outperforming the simpler variant.

Additionally, we trained a Support Vector Machine (SVM) model, which initially served as a mid-range solution in terms of efficiency. However, after applying Bayesian optimization, the SVM achieved a validation accuracy of 93.6%.

In terms of average F1 scores, the CNN models reached 87.5%, while the optimized SVM attained 93.3%. These results demonstrate strong capabilities in accurately classifying the various brain tumor types.

Finally, we address the limitations of our models and discuss the ethical complications of developing machine learning models in a medical context.

10. Future Work

In the past years, brain tumor classification through machine learning has boomed as a field of prominent research advancements. A comprehensive review by Ghorbian et al. (2024) outlines the taxonomy, challenges, and evaluation criteria used in this field, highlighting the widespread efforts toward accurate tumor classification using ML. Specifically, there have been numerous publications with advanced models achieving strong accuracy metrics (Ahmad et al., 2023; Luo et al., 2023; Rahmathunneesa & Ahammed, 2023), while others have addressed the issue of data limitation through robust preprocessing techniques or models requiring smaller datasets (Alam et al., 2023; Patro & Acharya, 2023; Zhang et al., 2022). However, we contend that future research should invest in the consistency of metrics, evaluation methods, and dataset uniformity. This will allow both researchers and practitioners to understand and compare different methods on a common base, and will ensure the validity of the models developed. Although such libraries already exist (e.g., the BraTS dataset), there has not yet been a library with the capacity to truly unlock machine learning models' potential.

Additionally, researchers need to delve deeper into the concerns of physicians regarding the use of AI models during their diagnosis. Adopting explainable AI methods will allow physicians to better comprehend how the models work, thus facilitating the adoption of these techniques (Verma, Rana, & Agrawal, 2023).

11. Data availability

The datasets are openly available on the following links:

a) Brain Tumor data: https://data.mendeley.com/datasets/w4sw3s9f59/1

b) PMRAM:https://www.kaggle.com/datasets/orvile/pmram-bangladeshi-brain-cancer-mri-dataset

c) China_Dataset: https://figshare.com/articles/dataset/brain_tumor_dataset/1512427

d) AD_VS_CN: https://www.kaggle.com/datasets/annemsony/ad-vs-cn-dataset

12. References

1. Ahmad, I., Sharif, M., Raza, M., & Muhammad, N. (2023). Multiple brain tumor classification with Dense CNN architecture using brain MRI images. *Computerized Medical Imaging and Graphics, 104*, 102208. https://doi.org/10.1016/j.compmedimag.2022.102208

2. Alam, F., Manogaran, G., & Ahmed, K. (2023). A multi-category brain tumor classification method based on improved ResNet50. *Computers in Biology and Medicine, 160*, 106907. https://doi.org/10.1016/j.compbiomed.2023.106907

3. Alam, M., Akhtar, M. T., & Hussain, M. (2024). Brain tumor classification using deep learning: A hybrid model with improved accuracy. *Scientific Reports, 14*, Article 56708. https://doi.org/10.1038/s41598-024-56708-x

4. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., ... & Menze, B. H. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv preprint arXiv:1811.02629*. https://arxiv.org/abs/1811.02629

5. Chang, A., Yu, T., & Liang, J. (2023). Radiological error in brain tumor diagnosis: Causes and implications. *Frontiers in Neurology, 14*, 102411. https://doi.org/10.3389/fneur.2023.102411

6. Contreras-Luján, E. E., García-Guerrero, E. E., López-Bonilla, O. R., Tlelo-Cuautle, E., López-Mancilla, D., & Inzunza-González, E. (2022). Evaluation of machine learning algorithms for early diagnosis of deep venous thrombosis. *Mathematical and Computational Applications, 27*(2), 24. https://doi.org/10.3390/mca27020024

7. Forghani, R. (2020). Artificial intelligence and radiology: Where do we stand? *Journal of the American College of Radiology, 17*(5), 555–562. https://doi.org/10.1016/j.jacr.2019.12.021

8. Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (3rd ed.). O'Reilly Media.

9. Ghaffar, A., Khan, A., & Mehmood, A. (2024). EfficientNet-B1-based enhanced architecture for brain tumor classification. *Neurocomputing, 528*, 57–66. https://doi.org/10.1016/j.neucom.2022.12.001

10. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

11. Jabbar, M. A., Ali, M. R., & Khan, S. (2023). Delay in brain tumor diagnosis: Causes and clinical impact. *BMC Neurology, 23*, 185. https://doi.org/10.1186/s12883-023-03298-3

12. Kale, D., Sharma, S., & Patel, K. (2024). Cancer screening using ML: A survey. *Journal of Biomedical Informatics, 139*, 104431. https://doi.org/10.1016/j.jbi.2023.104431

13. Li, L., Li, S., & Su, J. (2021). A multi-category brain tumor classification method based on improved ResNet50. *Computers, Materials & Continua, 69*(2), 2356–2364. https://doi.org/10.32604/cmc.2021.019409

14. Luo, J., Luo, Y., Liu, Y., Liu, B., & Xia, Y. (2023). A hybrid CNN–SVM model optimized with PSO for accurate and non-invasive brain tumor classification. *Computer Methods and Programs in Biomedicine, 233*, 107480. https://doi.org/10.1016/j.cmpb.2023.107480

15. National Institute of Standards and Technology. (2015). *Secure hash standard (SHS).* https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.180-4.pdf

16. Ostrom, Q. T., Price, M., Neff, C., Cioffi, G., Waite, K. A., Kruchko, C., & Barnholtz-Sloan, J. S. (2022). CBTRUS Statistical Report: Primary brain and other central nervous system tumors diagnosed in the United States in 2015–2019. Neuro-Oncology, 24(Suppl 5), v1–v95. https://pubmed.ncbi.nlm.nih.gov/36196752/

17. Patro, S. S., & Acharya, B. (2023). A CNN architecture with reduced complexity for early brain tumor classification. *Health and Technology, 13*, 495–507. https://doi.org/10.1007/s12553-022-00729-9

18. Price, D. J., Sharma, R., & Tang, C. (2024). A four-class classification system for MRI-based brain tumor diagnosis. *Journal of Medical Imaging and Health Informatics, 14*(2), 78–86. https://doi.org/10.1166/jmihi.2024.4410

19. Rahmani, A. M., Shafiei, M., Safari, M. S., Yousefzadeh, M., & Mohammadi, M. (2021). Machine learning in medicine: A comprehensive review, applications, and challenges. *Journal of Biomedical Informatics, 113*, 103655. https://doi.org/10.1016/j.jbi.2020.103655

20. Rahmathunneesa, K. P., & Ahammed, R. (2023). Biomedical image classification made easier thanks to transfer and semi-supervised learning. *Biomedical Signal Processing and Control, 84*, 104946. https://doi.org/10.1016/j.bspc.2022.104946

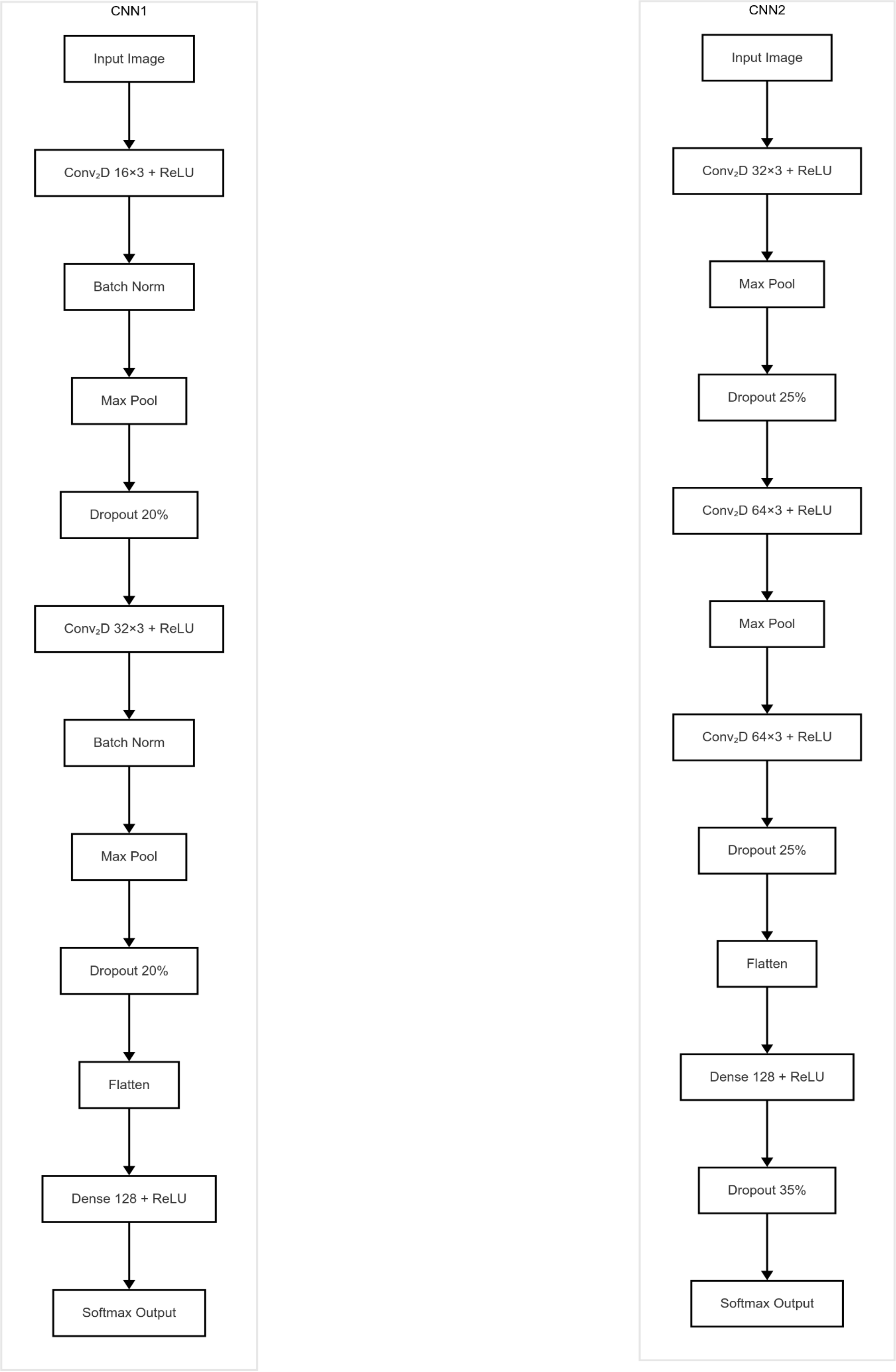
21. Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports, 14*, 56706. https://doi.org/10.1038/s41598-024-56706-x

22. Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning*, *97*, 6105–6114. PMLR. https://proceedings.mlr.press/v97/tan19a.html

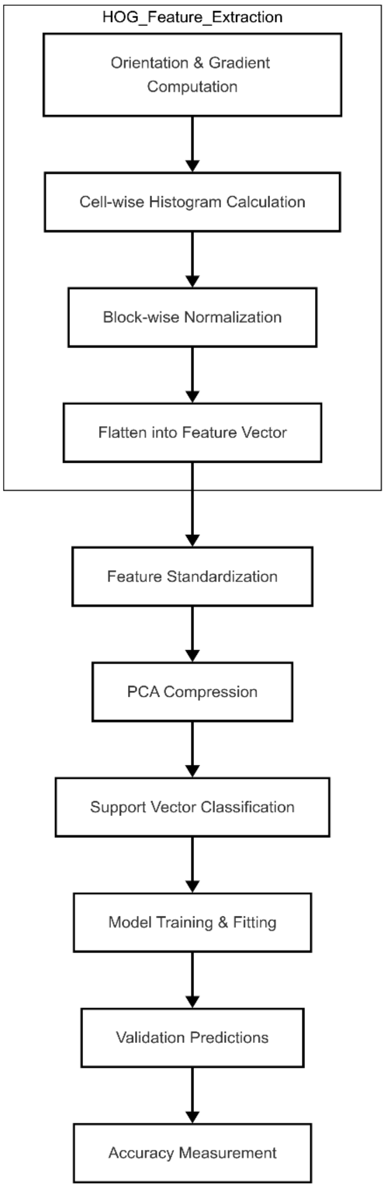
23. Verma, S., Rana, N., & Agrawal, S. (2023). Advanced AI/ML applications for brain tumor detection: A survey of explainability, architectures, and clinical integration. *Biomedical Signal Processing and Control, 84*, 104899. https://doi.org/10.1016/j.bspc.2023.104899

24. Zhang, L., Huang, Y., & Tan, C. (2022). Lightweight deep learning model for early brain tumor detection using MRI. *Healthcare Analytics, 2*, 100020. https://doi.org/10.1016/j.health.2021.100020

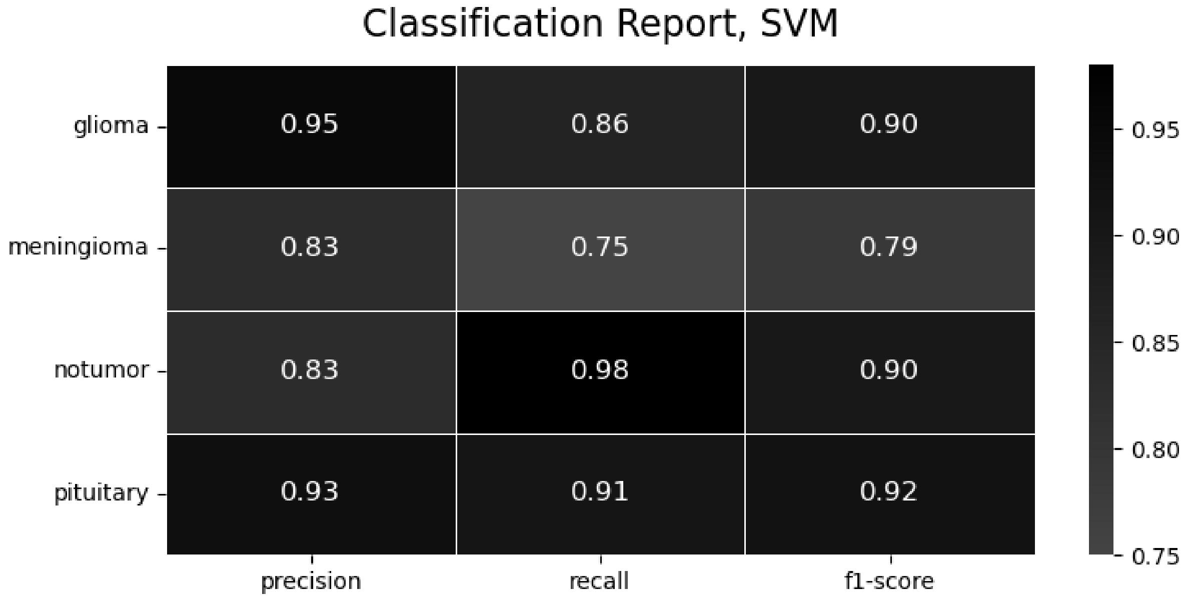
Appendix



Architecture of the two Convolutional Neural Networks models



Architecture of the SVM pipeline



Per class metrics of the SVM model