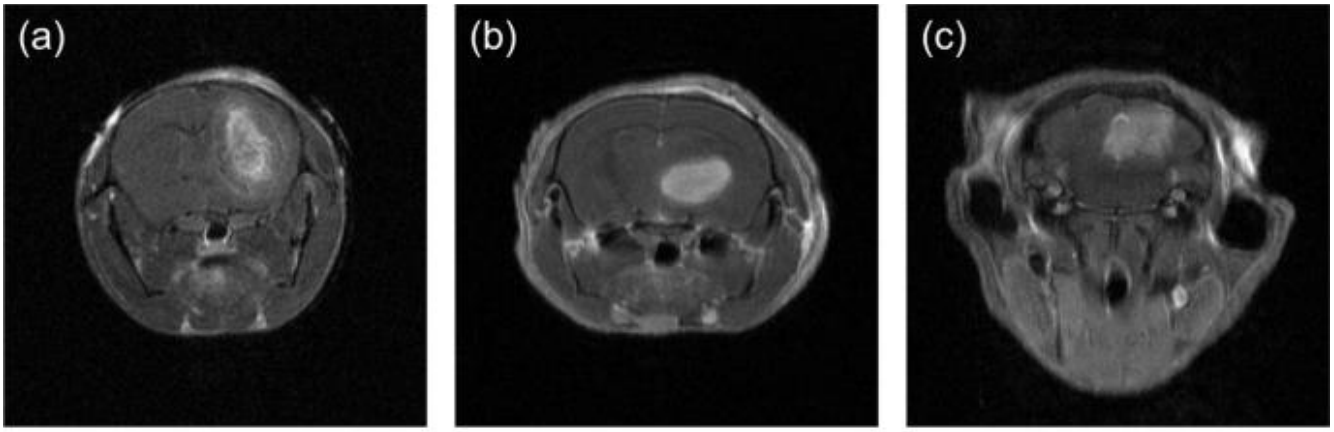
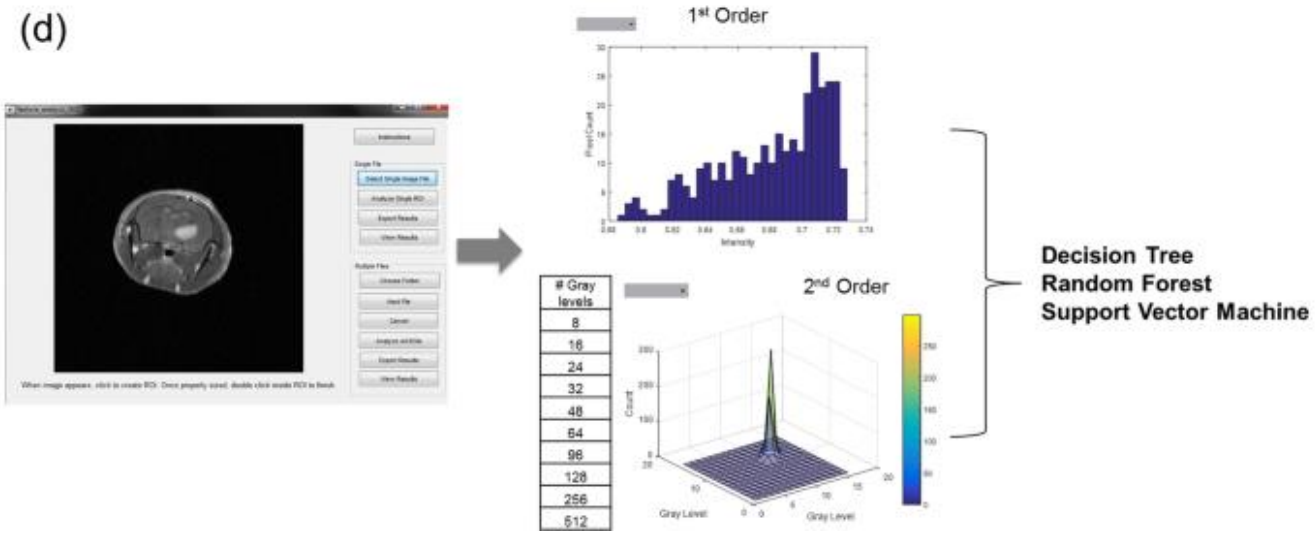
Data Mining in Brain Tumor Detection

By: Tariere Timitimi, Ana Sytina, Tahmidul Haque





Abstract

Brain Tumors have overwhelmed the world's populations due to its impact once detected amongst a human body. Despite a large awareness and assertive effort in addressing brain tumors, without an early enough detection, it becomes a disease of greater complication. Due to the advancements of technology, many fields would utilize Data Mining as its source in order to increase precision and effectiveness of diagnosis, which has improved tremendously. Its application to detection of brain tumors can increase early detection and intervention, which is life-changing to patients with the disease. Data Mining became a way of exploring what we can find and therefore conceive solutions. In the case of Brain tumors, the goal is to identify all information leading up to one patient experiencing brain tumors and conduct methods into hypothesising its timing, detection, and most importantly treatment availability. Various data mining techniques will be enforced such as predictive modelings, data wrangling and model evaluations to promote ways of predicting the formations and therefore predict resolutions. We hope to enhance brain tumor patients' early diagnosis and treatment planning by utilizing these strategies, to hopefully create an accurate model that is able to detect different types of tumors in the brain, and essentially lead to better patient outcomes.

Introduction

Early identification of cancer and other life-altering diseases is crucial for enhancing the quality of life and improving patient outcomes. Like other types of cancer, brain tumors have the ability to spread throughout other parts of the body such as more parts of the brain and the spinal cord. This increases the complexity of treatment and isn’t a favorable outcome. Key Factors of Brain Tumors are its timing and its unpredictability. Just like many others, this form of cancer can start off in the area but then will progress to spread the diseases throughout the whole body at a fast pace. That is why, timing is crucial for early detection of possible cancer cells and thereby finding possible and fast treatment plans. Even Public health decision makers require a significant amount of analytical information to effectively manage health services and achieve the goal of providing treatment to all citizens. (Santos, Malheiros, Cavaleiros).

When it comes to brain tumor diagnosis, early detection can open possibilities to better treatment options and reduce the risk of the tumor progressing and impacting the patient's life. Using MRI brain scan images to detect brain tumors in patients through the application of data mining techniques we can develop a model(s) capable of accurately detecting brain tumors, which can lead to the reduction of human error and enable a faster method of diagnosis using medical imaging/ is to develop a predictive model using data mining techniques to identify indicators of brain tumors from MRI images. By analyzing dataset of brain tumor images and applying deep learning algorithms, we hope to improve the accuracy and efficiency of tumor detection which facilitates timely diagnosis and treatment.

Some of the various different forms of Data Models we can enforce includes Decision Trees, SVM's, Neural Networks. These are all useful models we can use to help detect possible signs of brain tumors for patients. Drawing from the results of these models, we plan to assess their performance using various metrics, including precision, recall, and F1-score, to summarize our findings and, hopefully, identify solutions to the problem.

Literature Review

Brain tumor detection is critical for improving patient outcomes, and recent advancements in deep learning and data mining offer promising tools for enhancing diagnostic accuracy, efficiency, and accessibility.

Deep learning models, especially Convolutional Neural Networks (CNNs), have shown significant potential in automating the detection process by learning complex features directly from medical images. Abdusalomov et al. refined the YOLOv7 model for accurate detection of meningioma, glioma, and pituitary gland tumors, achieving an overall accuracy of 99.5%. This refined model incorporates image enhancement techniques to improve the visual representation of MRI scans and uses data augmentation to enhance the training dataset. The inclusion of a Convolutional Block Attention Module (CBAM) improves feature extraction, while the Spatial Pyramid Pooling Fast+ (SPPF+) layer and Bi-directional Feature Pyramid Network (BiFPN) enhance sensitivity and multi-scale feature fusion. The enhanced YOLOv7 model demonstrates its potential as a decision-making tool for experts in diagnosing brain tumors and its usefulness in

monitoring and detecting brain tumors using MRI. Additionally, Saeedi et al. used 3264 MRI brain images to develop a new 2D CNN and a convolutional auto-encoder network, achieving optimal accuracy of approximately 95% to 96%. Furthermore, the principles of quantum rotation matrices have been applied to traditional algorithms for feature selection. Bilal et al.'s Q-BGWO-SQSVM model, which uses quantum-infused techniques, achieved top-tier results in accuracy, sensitivity, specificity, precision, F1 Score, and Matthews Correlation Coefficient (MCC) across diverse medical image datasets.

Image pre-processing techniques play a crucial role in improving the accuracy of brain tumor detection in MRI images. These techniques aim to remove noise and extraneous elements from the images, with common steps including noise reduction, grayscale conversion, and smoothing and sharpening procedures. Some techniques also apply global thresholding, adaptive thresholding, Sobel filters, and high-pass filters to improve image quality, as well as normalization to convert brain images into intensity brain images using a min–max normalization rule. To improve the readability of low-resolution MRI images, a three-stage image preparation strategy can be used.

Various data mining and machine learning algorithms also contribute to brain tumor detection. Saeedi et al. showed that:

- Support Vector Machines (SVMs) classify brain tumors with high accuracy when combined with other techniques.
- Extreme Learning Machines (ELM) apply Regularized Extreme Learning Machine (RELM) classification to identify and classify tumor types.
- K-Nearest Neighbors (KNN) is used for brain tumor classification and has high precision and recall rates.
- Random Forest (RF) classifiers demonstrate effectiveness in classifying different types of tumors.
- Multi-layer Perceptron (MLP) classifiers are also used, though with lower accuracy rates compared to other methods.

Despite these advancements, there are still challenges to address. One such challenge is detecting small-size brain cancers, though integrating SPPF+ and BiFPN components helps focus

on localized tumors. Data diversity can be increased through data augmentation strategies when available data is low. Future studies should incorporate a diverse collection of clinically relevant brain lesions and additional imaging modalities to enhance models' segmentation accuracy. Moreover, challenges remain in the accurate diagnosis of tumors due to the radiologist's experience and potential fatigue, which computational intelligence-oriented techniques can assist with. Future research should focus on developing deep neural networks that are robust, simple, and have less execution time for a more rapid and accurate diagnosis.

Methodology

This study will work on creating a predictive model by integrating data mining techniques and deep learning algorithms to extract and analyze features indicative of brain tumor presence in MRI images. The following sections detail the procedures for data collection and preprocessing, exploratory data analysis, model development and training, model evaluation and optimization, and the final documentation of results.

Data Collection and Preprocessing

Using the Crystal Clean: Brain Tumors MRI Dataset acquired from Kaggle, we will form the basis of our analysis. During this phase, the dataset is carefully examined to ensure it is comprehensive and structured for subsequent analysis. Data preprocessing involves handling missing values, normalizing pixel intensities, and performing data augmentation where necessary. Standardization and normalization techniques are employed to ensure that the input features are on comparable scales.

Exploratory Data Analysis

During exploratory data analysis, we use a variety of data mining tools to identify underlying patterns, trends, and correlations within the dataset. This involves generating summary statistics and visual representations. This phase provides insights into the distribution of tumor types, the variability of imaging features, and potential outliers. These insights help the feature selection process and assist in understanding the complex nature of brain tumor imaging data.

Model Development and Training

The core of this research lies in the development of several predictive models using a combination of traditional machine learning methods and advanced deep learning techniques. The models under consideration include Decision Trees, Support Vector Machines (SVMs), and Neural Networks. In order to apply the research done by Abdusalomov et al. and Saeedi et al., we can apply a 2D CNN to extract and classify features from the MRI images and a YOLOv7 model for localization and detection. The combination of these two models enhances the feature extraction and classification processes, which results in a more accurate identification of brain tumors. During model development, the dataset is partitioned into training, validation, and test sets. The training process involves tuning of hyperparameters to optimize performance. Cross-validation is used to ensure the robustness of the model, while dropout and regularization techniques are applied to prevent overfitting.

Model Evaluation and Optimization

Upon training the predictive models, the performance is evaluated using key metrics such as accuracy, precision, recall, and F1-score where the diagnostic ability is assessed by comparing the model's predictions against the ground truth labels from the dataset. By completing this step, we are able to measure the effectiveness of the models and identify the most efficient algorithm for early brain tumor detection. We can also use these evaluations to tune model parameters to improve sensitivity and specificity.

Final Phase

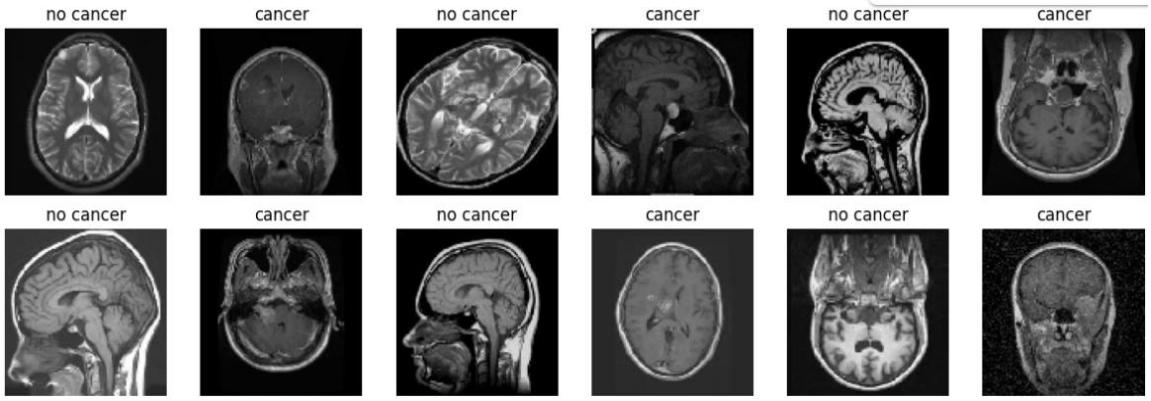
The final phase of this research involves a report on the methodology, experimental setup, results, and conclusions where a discussion of the findings, the limitations of the study, and recommendations for future research are presented.

Analysis and Results

The CNN architecture was structured as a sequential model that extracts hierarchical features from the input MRI images and makes a binary classification. The first convolutional block consisted of a Conv2D layer, which applied 32 convolutional filters to the input image to detect basic features such as edges/textures and a MaxPooling2D layer, to down-sample the feature maps, reduce spatial resolution and reduce computational load. Followed by the second convolutional block consisting of another Conv2D layer with 64 filters to capture more complex features and another MaxPooling 2D layer. Finally a third convolutional layer was applied with 128 filters on the Conv2D later and reduce spatial dimensions to a manageable size with a MaxPooling2D layer. Finally we had a flatten, dense, dropout and final dense layer. The flatten layer prepared out data for the dense layer, which then captured the most significant patterns for classification. The dropout layer was applied in order to reduce overfitting by preventing the network from relying too much on any one feature. The final dense layer used a sigmoid activation function to output a probability, and finally classify the images as cancerous or non-cancerous. The model had a total of 1,731,905 trainable parameters, a layered structure, and a layer to control overfitting, which demonstrates a good architecture for binary classification of MRI images.

Firstly, you visually want to acknowledge how you can tell clearly whether a MRI image displays a Normal or a Tumor Brain. As you can see below, the images are columned to distinguish a multi-viewed perspective between a normal brain and a Tumor brain. This is extremely important for doctors and even analytics to enforce these techniques in order for medical personnel to distinguish if a patient has any form of brain cancer or not visually. This is whats important about displaying MRI images for research. You are referring technology in order to display imaging to which help define your analysis on how you caa determine if some form of tumor is embedded within a patients brain.

Figure 1. Distinguishing Normal and Brain Tumor Images



With a total of 21,672 images, 18,606 tumor images and 3,066 normal images, the CNN model showed promising performance in detecting cancer vs non-cancer brain MRI images despite the significant class imbalance in the dataset. The CNN was trained over 10 epochs, and its performance improved consistently over time. Early epochs started with training accuracy around 85.8% and validation accuracy around 82.8%. By epoch 5, validation accuracy reached about 90.8%, and continued to improve, peaking around 92.6% by the final epoch. The decrease in training loss and the increase in training accuracy indicate that the CNN is effectively learning from the data. Overall, the trend was strongly positive.

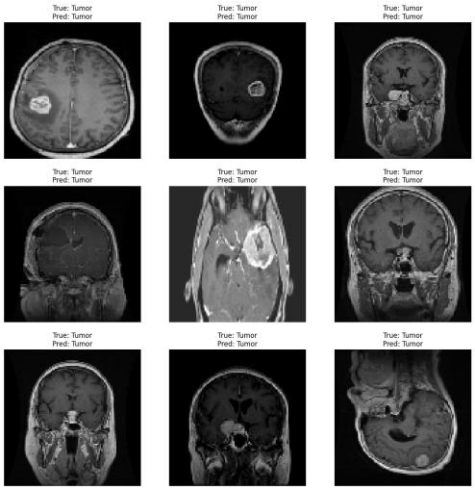
Figure 2: Evaluation of the CNN model

```
Epoch 1/10
407/407 ━━━━━━━━━━━━━━━━━━━━ 355s 863ms/step - accuracy: 0.8588 - loss: 0.3994 - val_accuracy: 0.8238 - val_loss: 0.4217
Epoch 2/10
407/407 ━━━━━━━━━━━━━━━━━━━━ 357s 877ms/step - accuracy: 0.8713 - loss: 0.3125 - val_accuracy: 0.8055 - val_loss: 0.4586
Epoch 3/10
407/407 ━━━━━━━━━━━━━━━━━━━━ 342s 840ms/step - accuracy: 0.8878 - loss: 0.2694 - val_accuracy: 0.8987 - val_loss: 0.2265
Epoch 4/10
407/407 ━━━━━━━━━━━━━━━━━━━━ 342s 840ms/step - accuracy: 0.8965 - loss: 0.2368 - val_accuracy: 0.8265 - val_loss: 0.4093
Epoch 5/10
407/407 ━━━━━━━━━━━━━━━━━━━━ 409s 908ms/step - accuracy: 0.9076 - loss: 0.2130 - val_accuracy: 0.8881 - val_loss: 0.3327
Epoch 6/10
407/407 ━━━━━━━━━━━━━━━━━━━━ 342s 841ms/step - accuracy: 0.9161 - loss: 0.1929 - val_accuracy: 0.8955 - val_loss: 0.2843
Epoch 7/10
407/407 ━━━━━━━━━━━━━━━━━━━━ 338s 830ms/step - accuracy: 0.9213 - loss: 0.1845 - val_accuracy: 0.9170 - val_loss: 0.2103
Epoch 8/10
407/407 ━━━━━━━━━━━━━━━━━━━━ 336s 824ms/step - accuracy: 0.9207 - loss: 0.1794 - val_accuracy: 0.8851 - val_loss: 0.2989
Epoch 9/10
407/407 ━━━━━━━━━━━━━━━━━━━━ 409s 1s/step - accuracy: 0.9255 - loss: 0.1712 - val_accuracy: 0.9123 - val_loss: 0.2200
Epoch 10/10
407/407 ━━━━━━━━━━━━━━━━━━━━ 343s 843ms/step - accuracy: 0.9378 - loss: 0.1495 - val_accuracy: 0.9156 - val_loss: 0.2081
136/136 ━━━━━━━━━━━━━━━━━━━━ 30s 222ms/step - accuracy: 0.9224 - loss: 0.1943
Test Accuracy: 0.9217993021011353
```

In consideration to the dataset imbalance, it was important to consider metrics beyond accuracy (e.g., precision, recall, F1-score). With a final test accuracy of 93.75% and test loss of 0.1436, the CNN not only matches the linear SVM in terms of accuracy but also benefits from

not needing any intermediate steps or feature engineering, which is particularly advantageous for image data.

Figure 3 and 4: CNN Sample Prediction Images and Confusion Matrices



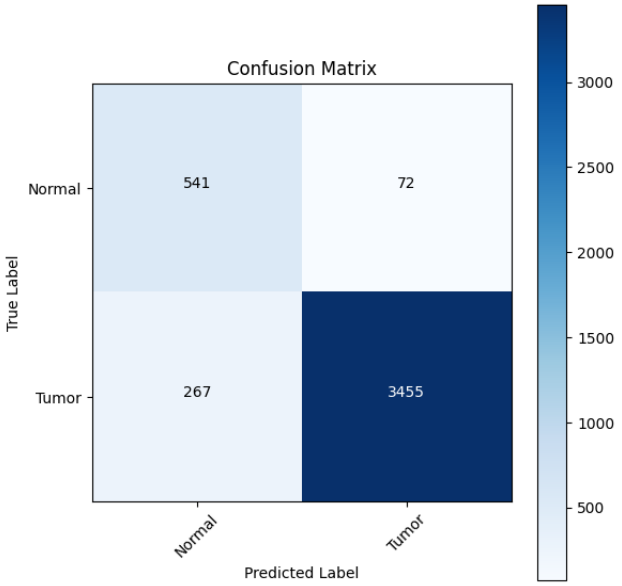


Figure 5: CNN Precision, Recall, F1-score and ROC AUC Score

	Precision	**Recall**	**F1- Score**	**Support**
Normal	**0.67**	**0.88**	**0.76**	**613**
Tumor	**0.98**	**0.93**	**0.95**	**3722**
Accuracy			**0.92**	**4355**
Macro. Avg	**0.82**	**0.91**	**0.86**	**4355**
Weighted Avg	**0.94**	**0.92**	**0.93**	**4355**

ROC AUC Score: 0.9674699967478761

When predicting a normal image, the model is correct 67% of the time as shown in its precision value, and it successfully identifies 88% of the normal cases as shown in the recall value, indicating a high possibility of false negatives. When classifying the tumor images, the

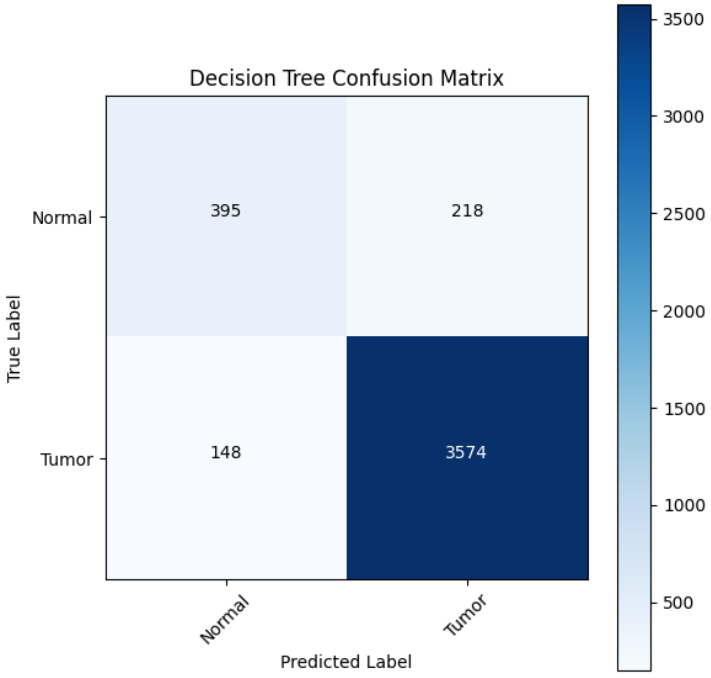
model is correct in identifying them 98% of the time and detects 93% of the tumor cases successfully. According to precision value, when the model predicts a tumor, it is correct 98% of the time and according to the recall value, the model detects 93% of tumor cases successfully. Having an F1-score of 0.76 for Normal images and 0.95 for Tumor images, the model performs much better on the Tumor class. The ROC AUC Score of 0.967 indicated that there is an excellent overall ability for the model to discriminate between the two classes. Overall, the model being able to detect tumors in the brain scans most likely has to do with the imbalance of data where the dataset consists of about 86% of tumor images, making it biased towards the tumor class.

Following the CNN, we decided to use a decision tree and an SVM to weigh out which option would make the best classifier in our case. Decision trees being a good option to visualize and understand how decisions are being made, and SVMs being useful in high-dimensional spaces extending to handle non-linear decision boundaries, both options were important to gauge the performance of the CNN and assess whether or not the additional complexity of a CNN leads to improvement.

Figure 6: Decision Tree, Precision, Recall, F1-score and ROC AUC Matrix and Calculation

Decision Tree Validation and Test Accuracies

- **Decision Tree ROC AUC Score:** 0.7747292889187433 6
- **Decision Tree Validation Accuracy:** 0.9121107266435986
- **Decision Tree Test Accuracy:** 0.9155709342560554
- **Decision Tree Precision**: 0.9425105485232067
- **Decision Tree Recall**: 0.9602364320257926



	Precision	**Recall**		
Normal	**0.73**	**0.64**		
Tumor	**0.94**	**0.96**	**0.95**	**3722**
Accuracy			**0.92**	**4355**
Macro. Avg	**0.83**	**0.80**	**0.82**	**4355**
Weighted Avg	**0.91**	**0.92**	**0.91**	**4355**

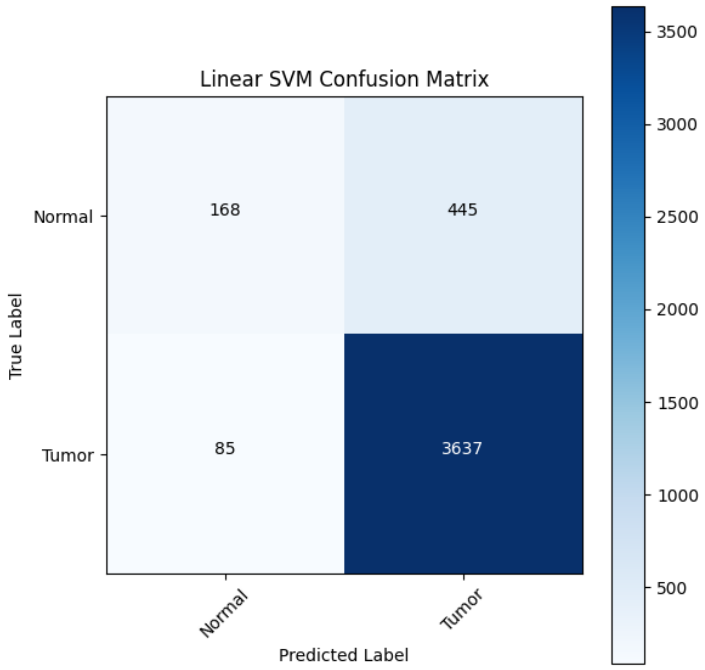
- **Decision Tree F1**: 0.9512909236092627

The Decision Tree classifier achieves 92% accuracy, demonstrating its ability to classify the majority of the samples accurately. The classifier's limitations are revealed when considering the classifier's overall discriminative power, as provided by its ROC AUC score of 0.7747, the lowest of the classifiers considered. This suggests that while the Decision Tree can be powerful in tumor class image classification, it is less powerful in its ability to consistently separate tumor and normal classes. Therefore is less suitable for more subtle medical image classification issues.

Figure 8 and 9: SVM Precision, Recall, F1-score and ROC/AUC Score Matrix and Calculations

SVM Validation and Test Accuracies

- **Linear SVM Validation Accuracy:** 0.8747404844290657
- **Linear SVM Test Accuracy:** 0.8777393310265282
- **Linear SVM Precision**: 0.8909848113669769
- **Linear SVM Recall:** 0.977162815690489
- **Linear SVM F1 Score:** 0.9320861096873398
- **Linear SVM ROC AUC Score:** 0.8443639643651389



	Precision	Recall	F1- Score	Support
Normal	**0.66**	**0.27**	**0.39**	**613**
Tumor	**0.89**	**0.98**	**0.93**	**3722**
Accuracy			**0.88**	**4355**
Macro. Avg	**0.78**	**0.63**	**0.66**	**4355**
Weighted Avg	**0.86**	**0.88**	**0.86**	**4355**

The SVM classifier recorded 88% overall accuracy, yet it does not sufficiently discriminate between tumor class images and normal class images, thus lacking detection performance reliability for both classes. Its ROC AUC measure, which is a measure of its discriminative power, is 0.844, indicating higher overall predictive power compared to the decision tree classifier. Though the SVM presents a worthy enhancement over the decision tree by way of classification quality, neither models came close to the high results obtained by CNN.

Discussion

The development of these CNN models can help predict or detect tumors and they can provide the possibility of reducing mortality. Furthermore, the integration of genetic data with imaging could help in the identification of treatment plans that are most suitable to the patient, thus improving the outcomes of the patient.There is a great potential for this critical interface of medicine, technology, and data science to improve diagnostic precision and therapeutic strategies by enforcing these models and networks

When comparing the three classifiers, a convolutional neural network (CNN), a decision tree, and a linear support vector machine (SVM)—for brain tumor detection from MRI images. The CNN achieved an overall accuracy of 92% with a remarkable ROC AUC of 0.967, demonstrating excellent discriminative ability between normal and tumor classes. Although its performance on the normal class (precision: 0.67, recall: 0.88, f1-score: 0.76) was lower than on the tumor class (precision: 0.98, recall: 0.93, f1-score: 0.95), it demonstrated good performance overall. The Decision Tree classifier also reached an accuracy of 92%, performing strongly in identifying Tumor cases (precision: 0.94, recall: 0.96, f1-score: 0.95) but with lower ROC AUC of 0.775. Therefore, it is effective under a fixed threshold to discriminate across varying decision thresholds. The Linear SVM classifier, with an accuracy of 88%, showed high performance for the tumor class (precision: 0.89, recall: 0.98, f1-score: 0.93) but significantly underperformed on the normal class (f1-score: 0.39), showing its susceptibility to an unbalanced datasets.

Conclusion

Early diagnosis of brain tumors remains an essential role in improving outcomes and quality of life for survivors. New findings in deep learning and data mining have established promising paths to employing the diagnostic process for automation with fewer human errors and faster and more precise treatments. By using powerful predictive models, e.g., Decision Trees, Support Vector Machines (SVMs), Neural Networks, and Convolutional Neural Networks (CNNs), the detection of brain tumors from MRI images by medical practitioners becomes more precise. Image pre-processing, data augmentation, and feature selection help to improve such models to handle complex cases, e.g., localized and small-sized tumors.

The CNN model demonstrated superior performance, particularly in terms of its high ROC AUC, making it a possible candidate where accurate tumor detection is needed. The validation metrics, while slightly variable, generally mirror this trend, suggesting good generalization.With further refinement working out the CNNs limited ability to manage the dataset, it can be very reliable.

In summary, the application of data mining techniques in medical imaging is a breakthrough time for healthcare. With the development of models that are not only accurate but also efficient and cost-effective, we are able to unveil doors to early diagnosis and personalized treatment plans, ultimately leading to better patient outcomes for patients with brain tumors. Support from ongoing collaboration between medical experts and data experts will be at the forefront of unlocking the full potential of such technologies and assisting no patient have to endure brain cancer.

References

Abdusalomov, A. B., Mukhiddinov, M., & Whangbo, T. K. (2023). Brain tumor detection based on deep learning approaches and magnetic resonance imaging. Cancers, 15, 4172. https://doi.org/10.3390/cancers15164172

World Health Organization. (2021). Artificial intelligence is changing the health sector. *WHO Consultation Towards the Development of Guidance on Ethics and Governance of Artificial Intelligence for Health.* Retrieved from https://www.jstor.org/stable/resrep35680.7

Saeedi, S., Rezayi, S., Keshavarz, H., & Niakan Kalhori, S. R. (2023). MRI-based brain tumor detection using convolutional deep learning methods and chosen machine learning techniques. BMC Medical Informatics and Decision Making, 23, 16. https://doi.org/10.1186/s12911-023-02114-6

Manjunathan, N., & Gomathi, N. (2025). A comparative analysis of the health monitoring process using deep learning methods for brain tumor. *Measurement: Sensors, 37*(101807). https://doi.org/10.1016/j.measen.2025.101807

Crystal Clean: Brain Tumors MRI Dataset. (n.d.). Www.kaggle.com. https://www.kaggle.com/datasets/mohammadhossein77/brain-tumors-dataset/data

Hemanth, G., Janardhan, M., & Sujihelen, L. (n.d.). Design and implementing brain tumor detection using machine learning approach. IEEE Xplore. Retrieved from https://ieeexplore.ieee.org/

Santos, A., Malheiros, B., Cavalheiros, C., & Olivera, D. (n.d.). A data mining system for providing analytical information on brain tumors to public health decision makers. ScienceDirect. Retrieved from https://www.sciencedirect.com/

Analytics Vidhya. (2021, May). Convolutional neural networks (CNN). https://www.analyticsvidhya.com/blog/2021/05/convolutional-neural-networks-cnn/

Spiceworks. (n.d.). What is support vector machine? https://www.spiceworks.com/tech/big-data/articles/what-is-support-vector-machine/

Analytics Vidhya. (2021, August). Decision tree algorithm. https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/

ScienceDirect. (n.d.). Confusion matrix. https://www.sciencedirect.com/topics/engineering/confusion-matrix

Restack. (n.d.). *Sequence-to-sequence models: Answer what is sequential model in CNN? cat AI*. https://www.restack.io/p/sequence-to-sequence-models-answer-what-is-sequential-model-in-cnn-cat-ai