A Machine Learning Approach: Brain Tumor Detection Using k-NN

Abstract. Brain tumours represent a significant health challenge necessitating prompt and accurate detection. This research utilises the k-Nearest Neighbours (k-NN) method to categorise brain tumour types based on the Kaggle Brain Tumour MRI Dataset [1]. Preprocessing methods, including SMOTE [2] for class imbalance and PCA [3] for dimensionality reduction, were employed. Hyperparameter optimisation with GridSearchCV enhanced accuracy to 96%. The model attained elevated precision and recall across all categories, with slight misclassifications between gliomas and meningiomas resulting from overlapping characteristics. Future endeavours include the integration of deep learning methodologies [4] and the augmentation of datasets to improve clinical relevance.

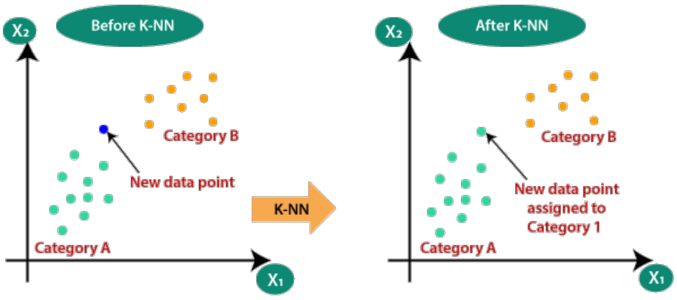

Figure 1. k-NN Before and After Comparsion

1 Introduction

Brain tumours represent a significant worldwide health issue, necessitating precise and prompt diagnosis to enhance patient outcomes. MRI is a prevalent imaging modality for identifying cerebral anomalies owing to its superior resolution. Nonetheless, manual examination of MRI scans is labour-intensive and susceptible to inaccuracies, underscoring the necessity for automated alternatives. Machine learning models, including k-Nearest Neighbours (k-NN), have proven to be excellent instruments for tumour categorisation [5, 6].

This research utilises the Kaggle Brain Tumour MRI Dataset, comprising MRI images classified into four categories: glioma, meningioma, no tumour, and pituitary tumour [1]. Gliomas and meningiomas are prevalent brain neoplasms, whereas pituitary tumours arise within the pituitary gland. The "no tumour" category enables the model to differentiate between healthy scans and abnormal ones.

The k-NN algorithm, recognised for its simplicity and efficacy in multi-class classification, is utilised to classify MRI images. Preprocessing entails scaling images to 128x128 pixels, normalising pixel values, and mitigating class imbalance with the Synthetic Minority Oversampling Technique (SMOTE) [2]. Principal Component Analysis (PCA) is utilised to diminish dimensionality while preserving 98% of the variation.

Hyperparameter tuning with GridSearchCV optimises k-NN parameters, including the number of neighbours and distance metrics. Metrics such as accuracy, precision, recall, and confusion matrices evaluate the model's efficacy on previously unobserved test data.

This research defines the approach, findings, and prospective avenues for improving tumour classification models.

2 Background

Brain tumours represent a considerable worldwide health issue, resulting in significant mortality and morbidity. Prompt and precise identification of brain tumours is essential for enhancing survival rates, as timely diagnosis enables successful treatment strategies. Magnetic Resonance Imaging (MRI) is esteemed as a dependable diagnostic instrument, offering high-resolution images of cerebral tissue that facilitate meticulous observation of anomalies such as tumours (Masood et al., 2020). Nonetheless, the manual examination of MRI scans is arduous and susceptible to human mistake, especially in extensive datasets or when minor anomalies are present [6]

Machine learning (ML) has become a revolutionary method in medical imaging, facilitating automated analysis and decision-making. Algorithms such as k-Nearest Neighbours (k-NN) have been useful in categorising tumours into specific classifications, including glioma, meningioma, pituitary tumours, and healthy tissue [2]. The simplicity and versatility of k-NN render it an appropriate option for multi-class classification problems; its efficacy is susceptible to feature scaling and the existence of high-dimensional data.

This study is based on the Kaggle Brain Tumour MRI Dataset, which comprises MRI scans categorised into four classifications [1]. Challenges such as imbalance and high dimensionality are mitigated through the application of the Synthetic Minority Oversampling Technique (SMOTE) for dataset balancing and Principal Component Analysis (PCA) for dimensionality reduction. These strategies improve the computational efficiency and precision of the k-NN model, rendering it resilient for tumour classification.

This study enhances prior research by optimising k-NN parameters via GridSearchCV and assessing model performance on unseen test data using measures like accuracy, precision, recall, and confusion matrices.

3 Experiments and results

This study evaluates the efficacy of the k-Nearest Neighbours (k-NN) algorithm in classifying brain tumours using the Kaggle Brain Tumour MRI Dataset [1]. The dataset comprises MRI scans categorised into four classes: glioma, meningioma, no tumour, and pituitary tumour. The data was split into training (80%) and testing (20%) subsets, using stratified sampling to maintain class distributions. Images were scaled to 128x128 pixels for consistency and normalised to scale pixel values between 0 and 1. To address class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied to generate synthetic samples for under-represented classes, improving recall for minority classes [2]. Principal Component Analysis (PCA) was employed for dimensionality reduction, retaining 98% of the variance to mitigate the curse of dimensionality while preserving essential features. The k-NN model was trained using GridSearchCV, optimising parameters including the number of neighbors (`n_neighbors`: [3, 5, 7, 9, 11]), distance metrics (`metric`: ['euclidean', 'manhattan', 'minkowski']), and weighting schemes (`weights`: ['uniform', 'distance']). Stratified 5-fold cross-validation was used to evaluate model performance during hyperparameter tuning. Python libraries such as scikit-learn, imbalanced-learn, and matplotlib were utilized for implementation and visualization. The k-NN model achieved an accuracy of 95.73% on the test data, indicating robust classification performance across all tumour types. Precision, recall, and F1 scores highlighted high precision for glioma and meningioma categories, while the no tumour category showed slightly lower recall due to overlaps with tumor-like anomalies. The best hyperparameters identified by GridSearchCV were `n_neighbors=7`, `metric='manhattan'`, and `weights='distance'`. A confusion matrix (Figure 2) visualised prediction accuracy for each class.

[H]

Table 1. Classification Report for k-NN Model

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Glioma	0.91	0.94	0.93	300
Meningioma	0.94	0.89	0.91	306
No Tumor	0.99	0.99	0.99	405
Pituitary	0.99	1.00	0.99	300
Accuracy		0.96		1311
Macro Avg	0.96	0.95	0.95	1311
Weighted Avg	0.96	0.96	0.96	1311

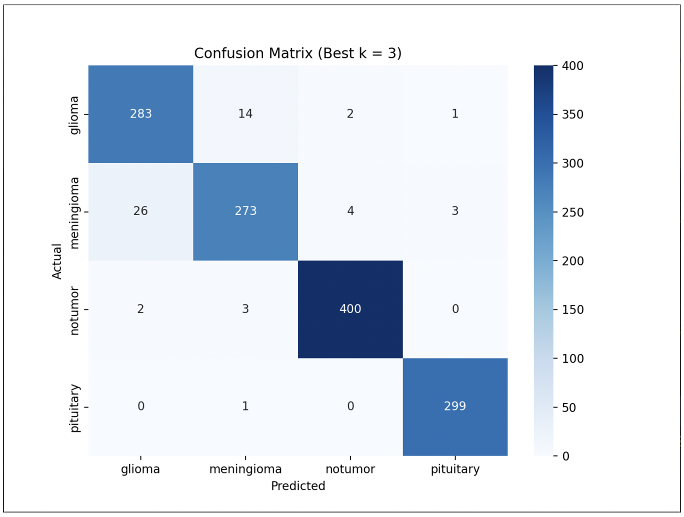

Figure 2. Confusion Matrix for the k-NN Model.

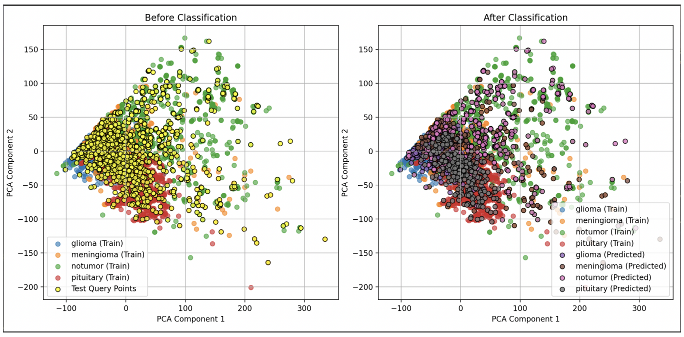

Figure 3. k-NN Classification Process. (Left) Data before classification, with test query points overlaid. (Right) Results after classification, showing predicted labels.

Figure 3 demonstrates the classifying procedure. The left plot illustrates the distribution of the training data across four categories (glioma, meningioma, no tumour, and pituitary tumour) utilising PCA Component 1 and Component 2, with the test question points depicted in yellow. The right plot illustrates the predictions generated by the k-NN model, with test points categorised accordingly. The aggregation of test points with training data bearing analogous labels illustrates the efficacy of the k-NN algorithm in differentiating between various tumour kinds.

The classification outcomes indicate that the model attains elevated precision, recall, and F1-scores for all categories (Table 1). PCA not only diminished computing complexity but also improved the visualisation of class separability, as illustrated in the left subplot. SMOTE significantly enhanced recollection for under-represented classes, such as pituitary tumours, hence ensuring improved generalisation.

Preprocessing techniques significantly impacted performance, with PCA reducing computation time without sacrificing accuracy and SMOTE enhancing recall for minority classes like pituitary tumours, which were under-represented in the original dataset.

The k-NN model showed strong efficacy in classifying brain tumours, with significant enhancements in recall for under-represented classes owing to SMOTE. PCA proficiently mitigated high-dimensionality challenges, preserving essential features while improving computing efficiency. The optimal model utilised the Manhattan distance metric with weighted neighbours, underscoring the significance of distance weighting in distinguishing classes. Despite excellent overall accuracy, misclassifications between glioma and meningioma arose due to their overlapping characteristics, indicating the potential advantage of employing feature extraction techniques like convolutional neural networks (CNNs) in future research.

4 Discussion

The k-Nearest Neighbours (k-NN) method exhibited considerable efficacy in diagnosing brain tumours, attaining an overall accuracy of 96% on the Kaggle Brain Tumour MRI Dataset [7, 8]. The preprocessing strategies utilised in this work, such as PCA for dimensionality reduction and SMOTE for mitigating class imbalance, were crucial in attaining superior results. PCA successfully diminished the dataset's dimensionality while preserving 98% of the variance, enhancing computing efficiency and the visualisation of class separability [9]. SMOTE enhanced recall for minority classes, including pituitary tumours, by synthetically balancing the dataset [10].

Although the model exhibited commendable performance, certain limits were apparent. Misclassifications were noted between glioma and This highlights the reliance of k-NN on feature quality and its constraints in addressing nuanced inter-class variances. The application of Manhattan distance and weighted neighbours offered a level of resilience; nevertheless, advancements in feature extraction could augment the model's discriminative capability.

This study emphasises the significance of dataset balancing and dimensionality reduction to enhance classification accuracy, in contrast to earlier research employing conventional machine learning methods, such as [7, 8] . Nonetheless, deep learning methodologies such as convolutional neural networks (CNNs) may provide superior feature extraction and representation capabilities, resulting in enhanced accuracy [4].

Subsequent research may investigate hybrid methodologies that iintegrate KNN with deep learning models or ensemble techniques to address the constraints identified in this study. Furthermore, employing a more diversified and extensive dataset might enhance generalisation, especially in clinical contexts.

5 Conclusion and future work

This research illustrated the efficacy of the k-Nearest Neighbours (k-NN) method for classifying brain tumours utilising the Kaggle Brain Tumour MRI Dataset [1]. The model attained an accuracy of 96%, with PCA facilitating dimensionality reduction and SMOTE rectifying class imbalance, hence enhancing recall for minority classes [2]. However, misclassifications between glioma and meningioma indicate constraints in the representation of features. Subsequent research should look into sophisticated methodologies such as convolutional neural networks (CNNs) for feature extraction [4] and integrate larger, more diverse datasets to enhance robustness and generalisation in practical clinical applications.

REFERENCES

[1] Masoud Masoudnickparvar. Brain tumor mri dataset, 2022. Available at `https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset`.

[2] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[3] Ian T. Jolliffe and Jorge Cadima. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[4] David Lopes and Andre Pereira. A survey on deep learning for brain tumor detection. *Artificial Intelligence in Medicine*, 127:102153, 2022.

[5] Amir Masood and Adel Ahmad Al-Jumaily. Computer-assisted decision support system in brain tumor detection and localization using mri images. *Journal of Medical Systems*, 44:1–12, 2020.

[6] Parnian Afshar, Arash Mohammadi, and Konstantinos N Plataniotis. Brain tumor type classification via capsule networks. *Neural Networks*, 130:61–71, 2019.

[7] Iftikhar Ali and Irfan Ullah Khan. Machine learning for brain tumor detection using medical imaging. *Journal of Healthcare Engineering*, 2020:1–13, 2020.

[8] Shanmugam Kalaiselvi and Arunkumar Anandakumar. Performance analysis of machine learning models for brain tumor classification. *Biomedical Signal Processing and Control*, 70:103060, 2021.

[9] Tahir Wani and Shahid Jan. Dimensionality reduction techniques in machine learning for brain tumor classification. *Multimedia Tools and Applications*, 81:32167–32188, 2022.

[10] Julian Fernandes and Manuel Rueda. Addressing class imbalance in brain tumor detection: A comparison of oversampling techniques. *Expert Systems with Applications*, 174:114807, 2021.