HIDS7009: Final Project Report

Application of ML models for Brain Cancer classification based on TCGA and Rembrandt MRI data

by Aizhan Uteubayeva

NetID: au198

Table of Content

1. Project Overview

This data science project focuses on classifying different types of brain tumors (Astrocytoma, GBM, Oligodendroglioma) using radiomic features merged with clinical data. The process involves data merging, cleaning, and encoding before training separate models for each disease type. Results include accuracy metrics and detailed classification reports for each model. The project is structured to be reproducible.

1.1 Libraries Used

- Pandas: For data manipulation and analysis.
- NumPy: For numerical operations on arrays.
- SimpleITK: To handle medical images in conjunction with radiomics.
- PyRadiomics: For the extraction of radiomics features from medical images.
- Joblib: To serialize Python objects.
- Imbalanced-Learn: For handling imbalanced datasets.
- XGBoost: For model training and prediction.

2. Data Preprocessing Steps

1. Data Loading:
    - Clinical data is loaded from an Excel file (TCGA_GBM_LGG_clinical_data_for_task2_updated.xlsx) into a pandas DataFrame.
    - Radiomics data is loaded from a CSV file (TCGA_pyradiomics_t1.csv) into another DataFrame.
2. Data Merging:
    - The radiomics and clinical data are merged on the patient ID column.
    - The Row.names column is dropped after the merge.
3. Data Encoding:
    - The Gender column is encoded into integers, with 'MALE' as 0 and 'FEMALE' as 1. Missing values are filled with -1.
    - The Race column is mapped to integers for different races, with 'UNKNOWN' as 4 and missing values filled with 0.
4. Handling Missing Values: Any remaining NaNs in the DataFrame are filled with 0.
5. Descriptive Statistics: The script calculates and displays descriptive statistics for select columns in the DataFrame.
6. Value Counting: The script counts and displays the number of instances for each unique value in the Disease_Type column.

2.1. About the target

The 'Disease_Type' column is the target. We are classifying three categories of brain tumors present in the patient data. This column has three unique values: 'GBM', 'Astrocytoma', and 'Oligodendroglioma'. The disease type column was encoded using one hot encoding.

2.2. Disease Types

1. GBM (Glioblastoma Multiforme):
    - Description: GBM is the most aggressive and common form of primary brain tumor in adults. It is known for its rapid growth and resistance to treatment.
    - Count in Dataset: 102 instances.
    - Implications: The high prevalence of GBM in the dataset is reflective of its common occurrence in clinical settings. Models trained on this dataset may perform better in predicting GBM due to the larger number of examples, but there might be a risk of bias towards this class.
2. Oligodendroglioma:
    - Description: This is a slower-growing tumor that arises from the oligodendrocytes, cells that cover and support nerve cells in the brain. It is less aggressive than GBM.
    - Count in Dataset: 26 instances.
    - Implications: With fewer instances, the model might struggle to learn sufficient patterns for this tumor type. This could lead to lower sensitivity or precision for detecting Oligodendroglioma compared to GBM.
3. Astrocytoma:
    - Description: These tumors originate from astrocytes, another type of glial cell in the brain. They can vary widely in behavior, from slow-growing (low grade) to highly malignant (high grade, such as anaplastic astrocytoma).
    - Count in Dataset: 21 instances.
    - Implications: Similar to Oligodendroglioma, the low number of Astrocytoma cases can hinder the model's ability to generalize well for this tumor type. This might result in a lower recall or increased false negatives for Astrocytoma.

3. Overview of the Models Used

I. SVM (Support Vector Machine):
    - Basic SVM models are used, followed by enhanced versions using SMOTE to address data imbalance. Subsequent models use hyperparameter optimization via GridSearchCV and RandomizedSearchCV.

II. AdaBoost with SVM:
    - AdaBoostClassifier with SVM as the base classifier is employed to improve model performance, especially beneficial for handling imbalanced classes.

III. Isolation Forest:
    - Implemented for anomaly detection in 'Astrocytoma', this model distinguishes outliers, crucial for identifying rare or atypical cases.

IV. Stacking Classifier:
    - Combines SVM and AdaBoost under a Logistic Regression final estimator in a stacking framework, leveraging the strengths of each model to enhance predictions.

V. XGBoost:
    - Utilizes XGBoost for robust classification across each disease type, focusing on managing various data dimensions and complexities.

3.1. Detailed Analysis of the Models

I. SVM Models:
   - Initial SVM Setup: Begins with SVC(kernel='linear', C=1.0), standardizing features using StandardScaler. Models are saved post-training, and performance metrics like accuracy and classification reports are generated.
   - SVM with SMOTE: To tackle class imbalance, SMOTE is applied to balance training data before SVM fitting.

II. AdaBoost with SVM:
   - AdaBoostClassifier using an SVM with RBF kernel as the base estimator is designed to aggregate the predictive power of multiple weak classifiers to strengthen overall classification, particularly beneficial for less prevalent classes.

III. Isolation Forest for Anomaly Detection:
   - Employed exclusively for 'Astrocytoma' to identify anomalies. This model is adept at distinguishing outliers from the majority.

IV. Stacking Classifier:
   - This method integrates SVM and AdaBoost, topped with Logistic Regression, aiming to utilize individual classifiers' benefits to yield more robust outcomes.

V. XGBoost:
   - Configured with 100 estimators and a 0.1 learning rate, XGBClassifier is assessed using accuracy and cross-validation to ascertain its consistency and effectiveness. Outcomes are visualized through confusion matrices and detailed in classification reports.

Note: Hyperparameter Optimization was done using GridSearchCV and RandomizedSearchCV for SVM parameters.

4. Training Results

Accuracy, confusion matrices, and classification reports are the primary metrics used to evaluate model performance. **The most successful outcomes were achieved when the problem was framed as multilabel. Conversely, framing it as multiclass did not attain an accuracy of 80%, unlike the multilabel approach.**

The project employed various machine learning models, including SVM, AdaBoost with SVM, Isolation Forest, Stacking Classifier, and XGBoost, to classify different types of brain tumors ('Astrocytoma', 'GBM', 'Oligodendroglioma'). Each model's performance was assessed using metrics like precision, recall, F1-score, and accuracy to ensure robustness, especially in handling imbalanced datasets.

4.1. Best Performing Models:

- Disease_Type_GBM: The models consistently showed strong performance in predicting 'GBM' with accuracy rates reaching up to 80%. This indicates a robust capability to identify the majority class in the dataset.

- Disease_Type_Oligodendroglioma: Notably high performance was observed with an accuracy of 90% in some models, highlighting effective learning despite this being a minority class compared to GBM.

4.2. Overview of results based on disease type:

1. Astrocytoma:
    - Initial SVM models showed an accuracy of about 73.33%, struggling mainly with the minority class.
    - After applying SMOTE and adjusting with SVM there was an improvement with an overall accuracy of approximately 76.67%. There was a beneficial impact of addressing class imbalance for minority classes.
2. GBM:
    - The consistent accuracy across different SVM configurations for GBM remained around 80%. This consistency underlines the models' capability to handle well-represented classes without significant variance.
    - The precision and recall for identifying true GBM cases remained robust, often reaching or exceeding 75%, indicating that the models are well-tuned for the predominant class in the dataset.
3. Oligodendroglioma:
    - The highest accuracy observed was 90% with initial SVM models, showcasing excellent performance.
    - However, precision for the true class varied significantly, from 100% in some models to 50% in others when using SMOTE, which might indicate a trade-off between recognizing majority and minority instances in the dataset.
4. Model Enhancements:
    - Using GridSearchCV and RandomizedSearchCV with SVM significantly helped optimize parameters for handling imbalanced data. The use of cross-validation further ensured the models' generalizability across different data segments.
    - AdaBoost with SVM and the Stacking Classifier approach were particularly useful in enhancing the model's performance by combining the strengths of multiple classifiers and ensuring better handling of varied data characteristics.

4.3. Initial SVM Model Results

- Disease_Type_Astrocytoma: The model achieved an accuracy of 0.7666666666666667.
- Disease_Type_GBM: This model had an accuracy of 0.8.
- Disease_Type_Oligodendroglioma: The accuracy recorded was 0.7666666666666667.

Post Grid Search Optimization

- Disease_Type_Astrocytoma: The accuracy improved to 0.80.
- Disease_Type_GBM: The model maintained an accuracy of 0.80.
- Disease_Type_Oligodendroglioma: The accuracy remained at 0.7666666666666667.

After Random Search Optimization

- The models for all diseases demonstrated consistent performance, with each achieving an accuracy of 0.80 on the test set.

4.4. SVM with AdaBoost approach:

1. Disease_Type_Astrocytoma:
    - Accuracy: The model achieved an accuracy of 0.80 on the test set.
    - Confusion Matrix: The matrix showed that out of 30 samples:
        - 24 were correctly classified as False (Negative class).
        - 3 were incorrectly classified as True when they were False.
        - 3 True (Positive class) cases were all misclassified as False.
2. Disease_Type_GBM:
    - Accuracy: The model reached an accuracy of 0.80 on the test set.
    - Confusion Matrix for this type:
        - 8 False (Negative class) samples were correctly identified.
        - 2 False samples were incorrectly classified as True.
        - 16 True (Positive class) samples were correctly identified.
        - 4 True samples were misclassified as False.
3. Disease_Type_Oligodendroglioma:
    - Accuracy: The accuracy here was 0.7666666666666667 on the test set.
    - Confusion Matrix for this disease type:
        - 20 False (Negative class) samples were correctly identified.
        - 3 False samples were incorrectly classified as True.
        - 3 True (Positive class) samples were correctly identified.
        - 4 True samples were misclassified as False.

These results indicate a strong performance in identifying the False class across all disease types, particularly for Astrocytoma and GBM. However, the models struggled with the True class for Astrocytoma, indicating a need for further tuning or additional data to improve sensitivity for this minority class. The high accuracies reflect the models' robustness and their potential stability across various datasets.

4.5. IsolationForest approach:

1. Astrocytoma:
    - Accuracy: The model achieved an accuracy of 0.7333333333333333 on the test set.
    - Confusion Matrix: The matrix indicated:
        - 22 False (Negative class) samples were correctly identified.
        - 5 False samples were incorrectly classified as True.
        - 3 True (Positive class) samples were all misclassified as False.

This result shows that while the model is fairly good at identifying the majority class, it fails to correctly identify any instances of the True class, indicative of a need for model adjustments or additional data for this category.

2. GBM:
    - Accuracy: The model reached an accuracy of 0.80 on the test set.
    - Confusion Matrix: For this disease type:
        - 8 False (Negative class) samples were correctly identified.
        - 2 False samples were incorrectly classified as True.
        - 16 True (Positive class) samples were correctly identified.
        - 4 True samples were misclassified as False.
3. Oligodendroglioma:
    - Accuracy: The accuracy here was 0.7666666666666667 on the test set.
    - Confusion Matrix: For this type:
        - 20 False (Negative class) samples were correctly identified.
        - 3 False samples were incorrectly classified as True.
        - 3 True (Positive class) samples were correctly identified.
        - 4 True samples were misclassified as False.

These results highlight that while IsolationForest performs competently in identifying the majority class across all disease types, its performance on the minority class (True) for Astrocytoma is notably weak.

4.6. Stacking Classifier approach:

1. Astrocytoma:
    - Accuracy: The model achieved an accuracy of 0.7333333333333333 on the test set.
    - Confusion Matrix: The matrix showed:
        - 22 False (Negative class) samples were correctly identified.
        - 5 False samples were incorrectly classified as True.
        - 3 True (Positive class) samples were all misclassified as False.
    - This result indicates that the Stacking Classifier, similar to the IsolationForest in this dataset, is effective for the majority class but fails to identify True cases, highlighting a potential area for improvement in sensitivity to the minority class.
2. Disease_Type_GBM:
    - Accuracy: The model reached an accuracy of 0.8666666666666667 on the test set.
    - Confusion Matrix: For this type:
        - 8 False (Negative class) samples were correctly identified.
        - 2 False samples were incorrectly classified as True.
        - 18 True (Positive class) samples were correctly identified.
        - 2 True samples were misclassified as False.
3. Disease_Type_Oligodendroglioma:
    - Accuracy: The accuracy here was 0.80 on the test set.
    - Confusion Matrix: For this disease type:
        - 21 False (Negative class) samples were correctly identified.
        - 2 False samples were incorrectly classified as True.
        - 3 True (Positive class) samples were correctly identified.
        - 4 True samples were misclassified as False.

Overall, the Stacking Classifier showed strong performance, particularly for GBM, where it achieved the highest accuracy among the models. The consistent results across disease types indicate that this approach is robust, but like other models, it could benefit from further tuning to improve detection of minority classes, especially for Astrocytoma.

4.7. XGBoost approach:

1. Disease_Type_Astrocytoma:
    - Accuracy on the test set: 0.9333333333333333
    - Confusion Matrix:
        - 27 samples classified correctly as False (Negative).
        - 0 samples incorrectly classified as True when they were False.
        - 2 samples classified incorrectly as False when they were True.
        - 1 sample classified correctly as True (Positive).
2. Disease_Type_GBM:
    - Accuracy on the test set: 0.7666666666666667
    - Confusion Matrix:
        - 9 samples classified correctly as False (Negative).
        - 1 sample incorrectly classified as True when it was False.
        - 6 samples classified incorrectly as False when they were True.
        - 14 samples classified correctly as True (Positive).
3. Disease_Type_Oligodendroglioma:
    - Accuracy on the test set: 0.7666666666666667
    - Confusion Matrix:
        - 21 samples classified correctly as False (Negative).
        - 2 samples incorrectly classified as True when they were False.
        - 5 samples classified incorrectly as False when they were True.
        - 2 samples classified correctly as True (Positive).

These results highlight that while XGBoost is highly accurate for Disease_Type_Astrocytoma, particularly in identifying the False class, it has challenges in predicting the True class across all disease types.

4.8. Model most important features (generally):

1. diagnostics_Mask-original_VolumeNum: This feature often indicates the number of distinct volumetric segments detected in medical imaging data. In the context of diagnosing diseases, such as tumors from radiomics data, this can indicate the complexity or the number of separate growths, which might correlate strongly with specific disease types or severities.
2. original_glcm_Idmn: Stands for Inverse Difference Moment Normalized from the Gray Level Co-occurrence Matrix (GLCM). This feature measures textural uniformity and the presence of local homogeneity in the image. High values suggest smoother textures, which can be crucial in differentiating between types of tissue or tumor characteristics.
3. original_shape_Sphericity: This is a measure of how spherical a shape is. In medical imaging, particularly in tumor analysis, the more spherical a shape, the more it might differ from irregular

malignant growths, making this a key feature in distinguishing between benign and malignant tumors.

4. original_shape_SurfaceVolumeRatio: This ratio indicates how compact or spread out a shape is relative to its volume. A higher surface-to-volume ratio might indicate a more irregular or branching structure, which is often seen in malignant or aggressive tumor growths.
5. original_gldm_SmallDependenceLowGrayLevelEmphasis: From the Gray Level Dependence Matrix (GLDM), this feature emphasizes areas with low gray levels and small dependencies. It helps in identifying subtle variations in texture that are less pronounced, aiding in the detection of early or less aggressive disease forms.
6. original_glrlm_ShortRunEmphasis: From the Gray Level Run Length Matrix (GLRLM), this measures the prevalence of short runs of similar intensity in the image. This can be important for identifying fine textures and is often used to detect changes in tissue that could indicate disease.
7. original_glcm_Contrast: Also from the GLCM, this measures the intensity contrast between a pixel and its neighbor over the whole image. High contrast can indicate the presence of edges or transitions, which are important in defining the boundaries of a lesion or differentiating between tissues.

5. Testing on Unseen Data:

The test phase of the project involved applying trained models to a new dataset to evaluate their predictive performance on unseen data. This phase is crucial to determine the generalizability and robustness of the models. Here's a detailed overview:

5.1. Data Preparation

- Feature Set for Testing: The new features from the dataset were loaded.
- Unnecessary columns, such as various diagnostics and configuration settings, were removed to match the trained model's expectations.
- Clinical Ground Truth: The clinical data for ground truth validation was loaded from a text file, and the 'Disease_Type' was one-hot encoded to facilitate accuracy measurement.
- The SVM used was the first version, without any optimization or additional algorithms.

5.2. Model Application

1. SVM Models:
    - Scalers and Models: For each disease type ('GBM', 'Astrocytoma', 'Oligodendroglioma'), the respective scaler and SVM model were loaded from saved joblib files.
    - Data Scaling and Prediction: The features were scaled using the loaded scalers, and predictions were made using the SVM models.
    - Accuracy Calculation: Accuracy for each disease was calculated by comparing the model predictions against the true labels encoded from the clinical data.
2. XGBoost Models:
    - Model Loading: XGBoost models for the same set of diseases were loaded from their respective joblib files.

- Prediction: As XGBoost handles feature scaling differently, the raw features were used directly for making predictions.
- Accuracy Calculation: Similar to SVM, accuracy was computed by comparing the predictions with the true labels.

5.2. Results

The accuracies achieved provide insight into how each model generalizes beyond the training dataset:

SVM Model Performance

The SVM models were applied to unseen data for three different disease types: GBM, Astrocytoma, and Oligodendroglioma. Here are the results:

- GBM: The SVM model for GBM achieved an accuracy of 0.484375. This indicates that the model had challenges in accurately predicting this disease type on the new dataset, potentially due to variations or complexities not captured during training.
- Astrocytoma: For Astrocytoma, the SVM model recorded an accuracy of 0.5625. This result shows a moderate level of predictive ability, suggesting some alignment between the training and test data characteristics but also room for improvement.
- Oligodendroglioma: The SVM model performed best for Oligodendroglioma, with an accuracy of 0.734375. This higher accuracy suggests that the SVM model was more effective in capturing and generalizing the features of this disease type compared to the others.

XGBoost Model Performance

The XGBoost models were similarly tested on the same set of diseases, and here are their accuracies:

- GBM: The XGBoost model for GBM had an accuracy of 0.421875. This lower performance, even compared to the SVM model, indicates significant challenges in dealing with this disease type in unseen data.
- Astrocytoma: The accuracy for Astrocytoma was 0.5625, identical to the SVM model's performance. This consistency across model types suggests a limit to how well current features and model configurations can predict this disease.
- Oligodendroglioma: The XGBoost model excelled for Oligodendroglioma with an accuracy of 0.796875, outperforming the SVM model. This superior result highlights the robustness of XGBoost in handling this disease type, likely due to its capacity to manage complex patterns and feature interactions effectively.

Important to note that Xgboost on the multiclass framing way produced poor results on unseen data (Accuracy: 0.3793103448275862)

Conclusion

SVM and XGBoost models showed varying degrees of success across the disease types. While both models struggled with GBM, indicating a need for further model refinement or additional data, they performed moderately well for Astrocytoma. Notably, Oligodendroglioma was the area where both models, particularly XGBoost, demonstrated strong predictive power.

When training the models, accuracy for the GBM disease type had good results but for unseen data probably because of a different class distribution the models didn't perform so well. More data for training would probably help.