Project Report

Introduction

Histopathological image classification plays a pivotal role in the early detection and accurate diagnosis of various types of cancer. Among the wide array of cancers, distinguishing between Low-Grade Glioma (LGG) and Glioblastoma Multiforme (GBM) is of particular clinical importance due to their vastly different prognoses and treatment strategies. Accurate classification can significantly impact patient outcomes by guiding personalized treatment plans and improving diagnostic accuracy.

In this project, we focus on classifying histopathological images into two categories: LGG and GBM. To achieve this, we utilized KimiaNet, a pre-trained deep learning model specifically designed for medical image analysis, as a feature extractor. The high-resolution histopathological images were sourced from the TCGA tumor dataset, a reliable and comprehensive repository of cancer-related images.

The classification process involved feeding the histopathological patches into KimiaNet, which extracted feature vectors that encapsulate the critical patterns and structures in the images. These feature vectors were then used to train and test machine learning classifiers, enabling us to distinguish between LGG and GBM with precision. By leveraging the power of pre-trained models and robust feature extraction, this approach offers an efficient and scalable solution to a critical challenge in cancer diagnostics.

KimiaNet: A Pre-trained Model for Histopathological Image Analysis

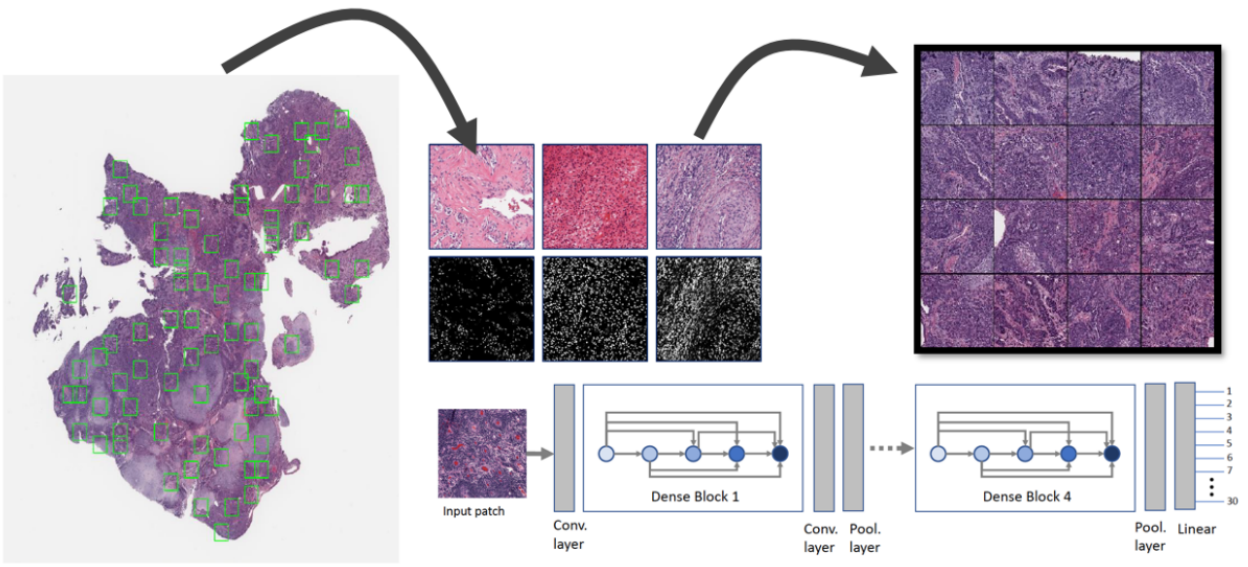
KimiaNet is a powerful deep convolutional neural network (CNN) tailored specifically for analyzing histopathological images. It was developed to address the unique challenges of medical imaging, particularly cancer detection and classification. Unlike general-purpose image models, KimiaNet leverages domain-specific pre-training, making it highly effective at extracting meaningful features from complex medical images. This capability makes it a valuable tool for tasks requiring precise image classification, such as distinguishing between cancer types.

Architecture of KimiaNet

KimiaNet is based on the **DenseNet-121 architecture**, a state-of-the-art design known for its efficient feature reuse and propagation. Key architectural components of KimiaNet include:

- **Four Dense Blocks**: DenseNet connects each layer to all preceding layers within a block, facilitating the flow of information and improving gradient propagation during training. This design ensures that features learned at earlier stages are directly available to subsequent layers, enhancing the model's ability to detect subtle patterns in medical images.
- **Transition Layers**: Positioned between dense blocks, these layers reduce the spatial dimensions of feature maps using pooling operations while retaining the essential information needed for accurate classification.

DenseNet's architecture is well-suited for medical imaging tasks, where intricate and hierarchical patterns in histopathological slides must be captured with high reliability. By building on DenseNet-121, KimiaNet combines robustness with computational efficiency, enabling it to process high-resolution image patches effectively.



Training and Testing Data

KimiaNet was pre-trained on a massive dataset of **240,000 image patches**, each measuring **1000×1000 pixels** at **20× magnification**. This extensive training ensures the model is adept at identifying critical pathological features, such as cell nuclei shapes and tissue patterns, which are vital for cancer diagnosis.

- **Training Dataset**: The patches were derived from **8,611 Whole Slide Images (WSIs)** available in the **TCGA (The Cancer Genome Atlas)** repository. To optimize the training process, the patches were carefully selected using a technique called **High-Cellularity Mosaic**, which focuses on regions densely populated with cell nuclei, as these areas hold the most diagnostic value.
- **Testing Datasets**: KimiaNet's performance was validated on multiple datasets, including:
  - **TCGA Pan-Cancer** images
  - **Endometrial Cancer**: 3,300 patches from 4 distinct classes
  - **Colorectal Cancer**: 5,000 patches from 8 distinct classes

This extensive training and testing regimen ensures that KimiaNet is highly generalized and capable of handling diverse cancer types and histopathological variations.

Dataset Preparation

For this study, we utilized **Whole Slide Images (WSIs)** from the TCGA Brain Tumor Dataset, a comprehensive resource widely used for cancer research. Our dataset includes **50 high-resolution WSIs**, evenly distributed between two cancer types: **25 Low-Grade Gliomas (LGG)** and **25 Glioblastoma Multiforme (GBM)**. These images are invaluable for understanding the pathological characteristics of brain tumors and developing classification methods to aid in cancer diagnosis.

Whole Slide Images (WSIs) are digitized versions of histopathological slides, capturing the entire tissue sample at various levels of magnification. Unlike conventional medical images, WSIs represent vast areas of tissue in a single file, allowing pathologists to examine intricate cellular and structural details. However, this level of detail comes at a cost—WSIs are typically very large, often exceeding gigapixel resolutions. Their immense size is necessary to preserve the fine details of tissue morphology, which are critical for accurate diagnosis and analysis.

WSIs are typically structured in multiple levels of resolution:

- **Level 0**: The highest magnification, offering intricate details of cellular structures.
- **Level 1**: A slightly lower magnification, balancing detail, and computational efficiency.
- **Level 2** (and beyond, if available): Progressively lower magnifications, useful for an overview of broader tissue regions.

This multi-level structure allows researchers to select an appropriate resolution level based on the task and computational constraints.

To ensure robust and balanced evaluation, the dataset was split into:

- **Training Set**: 36 WSIs (18 LGGs and 18 GBMs)
- **Testing Set**: 14 WSIs (7 LGGs and 7 GBMs)

This division ensures that the model learns from a diverse set of images while maintaining a separate and balanced testing set for evaluation.

Given the computational challenges associated with processing gigapixel images, we conducted our analysis at **Level 1 resolution**. This resolution offers a practical balance between retaining sufficient pathological detail and reducing computational resource requirements. By leveraging Level 1, we were able to extract meaningful features while ensuring efficient processing, making it an optimal choice for this study.

Methodology

The analysis of Whole Slide Images (WSIs) in their entirety presents significant challenges due to their immense size and complexity. Each WSI can contain gigapixels of data, making direct analysis computationally infeasible. To address these challenges, we implemented a systematic data preparation process that involved dividing the WSIs into smaller, manageable patches. This method allowed us to maintain high-resolution detail while ensuring the data could be efficiently processed by our model.

Patch Extraction

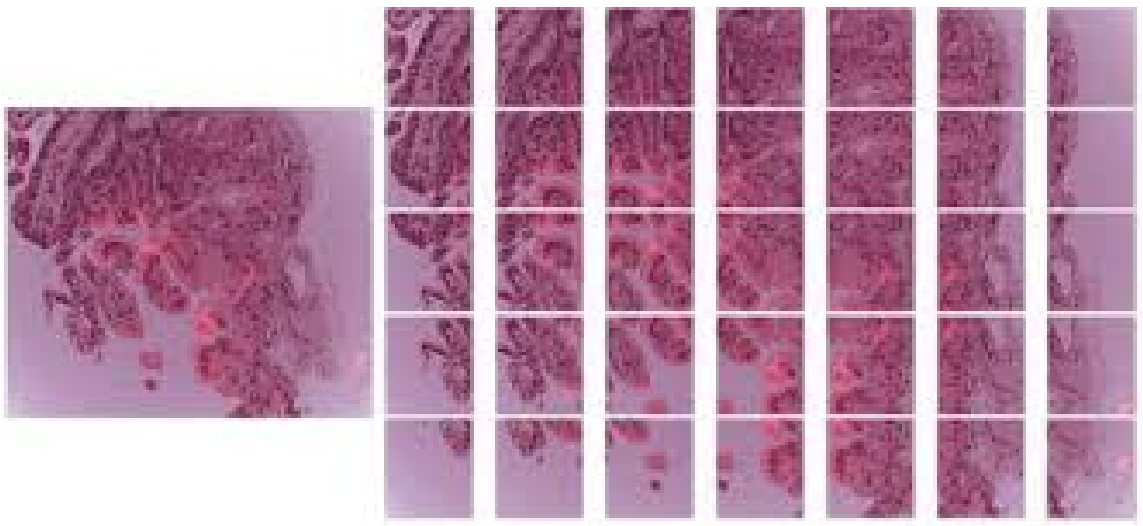
The first crucial step in preparing the WSIs for analysis was **patch extraction**. Since WSIs are typically very large, dividing them into smaller, uniform patches allows for more focused and localized analysis of tissue regions. This process also enables the use of pre-trained models, which typically require fixed input dimensions. By extracting patches from the original WSIs, we ensured that the model could efficiently process each image, one patch at a time.

Dividing the WSI into patches is essential for several reasons:

- **Computational Efficiency**: WSIs are often several gigabytes in size, which makes it impractical to feed the entire image into a model. By dividing the image into smaller patches, we reduce the data's computational load and make it manageable for analysis.
- **Localized Analysis**: Patches allow for more detailed, localized analysis of tissue regions, which is critical for identifying subtle patterns and abnormalities that are indicative of cancer.
- **Consistency for Pre-trained Models**: Pre-trained models, like KimiaNet, require consistent input sizes. By extracting patches of a uniform size, we ensure compatibility with the model's input requirements.

We divided each WSI into **25 non-overlapping patches**. The patches were extracted in a grid-like fashion, ensuring that each patch represented a distinct section of the WSI without overlap. This division was done uniformly across the entire WSI to maintain consistency in data size and ensure all tissue regions were covered.

- Each patch was **assigned the same label** as the parent WSI (either **Low-Grade Glioma (LGG)** or **Glioblastoma Multiforme (GBM)**). This ensures that every patch carries the same class information, enabling accurate classification when fed into the model.



This patch extraction ensured that we were able to leverage all available data within the WSIs while avoiding the computational bottleneck posed by their large sizes. It also improved the robustness of the training and testing process by providing the model with multiple data points per WSI, ultimately enhancing the model’s ability to generalize across various tissue regions.

White Patch Removal

In histopathological images, many patches often contain large white regions that lack any cellular or tissue information. White patches are typically devoid of any diagnostic content and can negatively impact the training process. Including such patches may:

- **Increase Noise**: White patches add no meaningful information to the model and instead act as noise, making it harder for the model to focus on important tissue structures.
- **Bias the Model**: Including too many white patches could result in the model learning irrelevant features, leading to poor generalization to real, informative tissue regions.

Thus, removing white patches is crucial for retaining only those with significant tissue content, ensuring the model focuses on informative data during training.

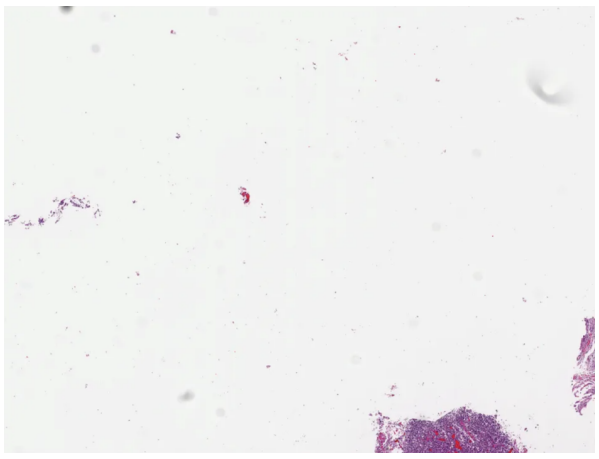
Thresholding for White Regions

To systematically remove these white patches, we employed a thresholding technique based on pixel intensity values. The key steps in this process were:

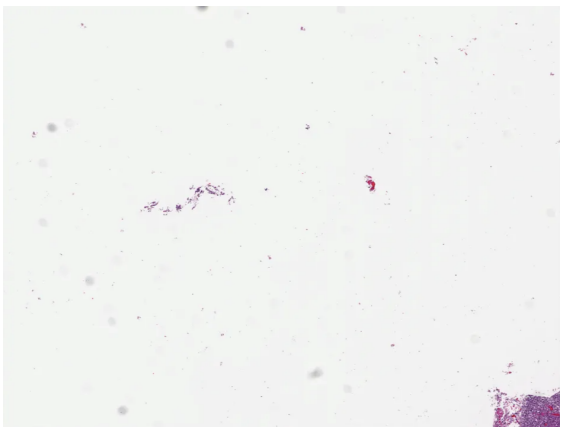
1. **Pixel Intensity Calculation**: The average pixel intensity of each patch was calculated by considering the RGB values of each pixel. Since white regions typically have RGB values close to (255, 255, 255), patches that were predominantly white had very high-intensity values across all channels.

2. **Thresholding**: A threshold was established to identify and exclude patches that were mostly white. Specifically, a patch was considered "white" if more than **98% of its pixels** had intensity values greater than **200** in all three RGB channels (red, green, blue). This threshold effectively captures largely white patches, with minimal tissue information.
3. **Patch Exclusion**: Any patches that met this criterion (i.e., those with more than 98% of pixels classified as white) were removed from the dataset. This process ensured that only patches containing meaningful tissue information, such as cellular structures, remained for further analysis.

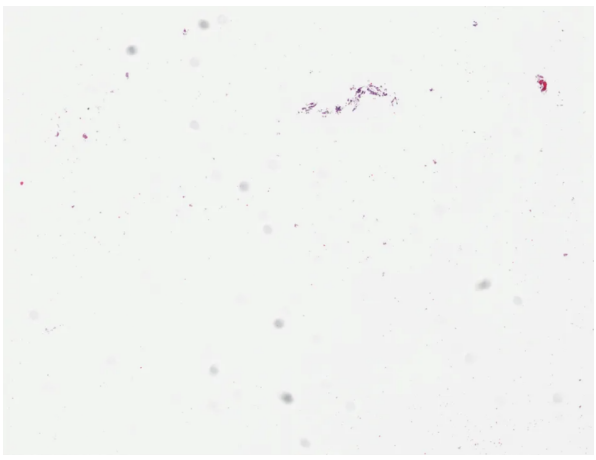
These are some example patches and their corresponding white pixel intensity percentages calculated by the given process.



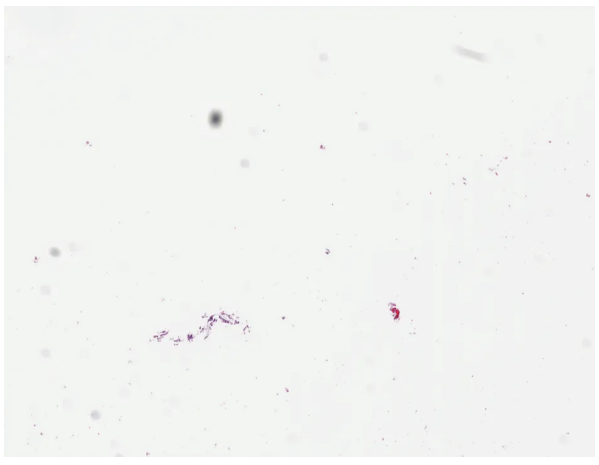
Intensity: **98%**



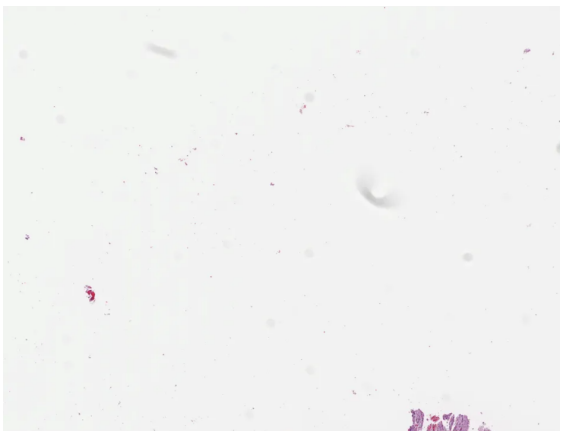
Intensity: **99%**



Intensity: **100%**



Intensity: **100%**



Intensity: **100%**

Intensity: **100%**

After applying this process, we were left with a more refined dataset. Initially, we started with **50 WSIs**, but after removing the white patches, only the patches with substantial tissue content were retained. This resulted in a dataset of patches that are more representative of the tumor tissues and ready for feature extraction using the KimiaNet model.

Feature Extraction using KimiaNet

After the white patch removal process, the next step involved feeding the refined patches into the **KimiaNet** model. KimiaNet, a pre-trained deep learning model based on the DenseNet-121 architecture, was used to extract meaningful features from each patch. The model, having been pre-trained on a large dataset of histopathological images, is highly effective at identifying intricate patterns and structures present in tissue regions, such as cell shapes and tissue organization.

The feature extraction process involved:

- Processing each patch through KimiaNet's convolutional layers and dense blocks to capture tissue patterns and structures
- Generating a **feature vector** for each patch, representing critical characteristics like texture, shape, and cell arrangement - essential for distinguishing between LGG and GBM
- Applying this process across all patches in both training and testing datasets

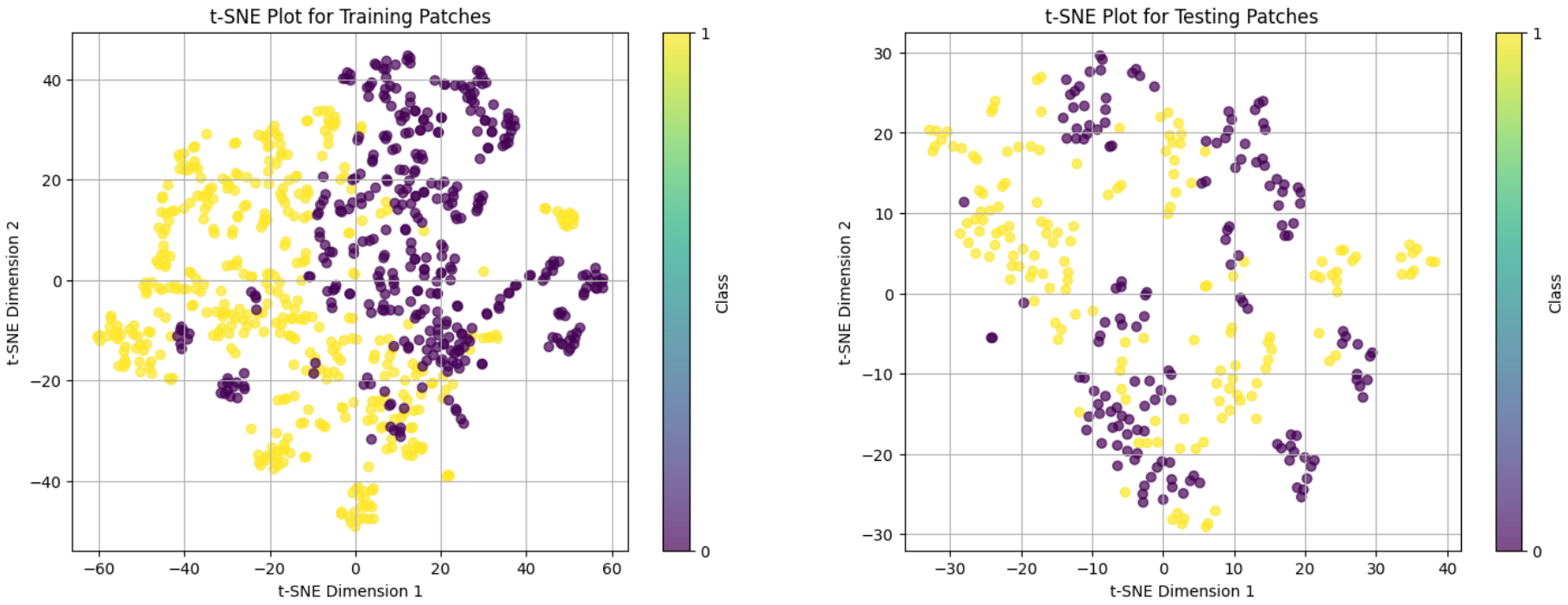
These feature vectors served as inputs for subsequent cancer classification, providing the classifier with rich, relevant information from the histopathological images.

Feature Visualization with t-SNE

To evaluate the effectiveness of the feature vectors generated by the KimiaNet model, we employed **t-distributed Stochastic Neighbor Embedding (t-SNE)**. t-SNE is a dimensionality reduction technique that transforms high-dimensional data (like the 1024-dimensional feature vectors) into a 2D or 3D space for visualization. This enables us to assess how well the extracted features separate the two cancer types, **Low-Grade Glioma (LGG)** and **Glioblastoma Multiforme (GBM)**.

t-SNE visualizations serve as a tool for understanding the structure and distribution of the data in the feature space. Specifically, they help to:

- **Visualize Clusters**: Highlight the grouping of similar feature vectors and their distinction based on cancer type.
- **Evaluate Separability**: Show whether the features extracted by KimiaNet are effective at distinguishing between LGG and GBM.
- **Check Generalization**: Assess whether the separation observed in training data is also evident in testing data, which reflects the model's ability to generalize to unseen data.



These visualizations provided an intuitive way to validate the quality of the feature vectors and guided the subsequent classification process, ensuring the robustness of the model.

Classification and Hyperparameter Tuning

After extracting the feature vectors using the KimiaNet model, we employed multiple machine-learning classifiers to categorize the patches into two categories: **Low-Grade Glioma (LGG)** and **Glioblastoma Multiforme (GBM)**. The classifiers used in this step included:

- **XGBoost (Extreme Gradient Boosting)**: A powerful boosting algorithm that builds decision trees sequentially, with each tree correcting the errors of its predecessors. Known for its speed and accuracy, XGBoost is well-suited for structured data classification tasks.
- **Gradient Boosting**: Similar to XGBoost, this method builds decision trees in a gradient descent manner to optimize a loss function and enhance model performance.
- **Random Forest**: An ensemble learning method that constructs multiple decision trees using random subsets of features and data. The final prediction is made based on the aggregated outputs of all trees, offering robustness and reducing the risk of overfitting.

Hyperparameter Optimization

To maximize the performance of these classifiers, we implemented a **Random Search** for hyperparameter tuning:

- A **parameter grid** was defined for each classifier, including key hyperparameters such as the number of estimators, learning rate, and maximum depth of trees.
- Random Search explored various combinations of these parameters to identify the best configuration.
- The optimal hyperparameter set was selected based on **validation accuracy**, ensuring the classifiers achieved their best performance on unseen data.

Image-Level and Patch-Level Classification

To evaluate the effectiveness of our approach, we analyzed the performance of the classifiers at both the **patch level** and the **image level**, using feature vectors extracted by **KimiaNet**. These results reflect the performance at **Level 1 WSIs** with hyperparameter tuning via **Random Search CV** for all classifiers.

Patch-Level Classification

At the patch level, each extracted patch from the Whole Slide Image (WSI) was treated as an individual data point. The classifiers predicted whether each patch belonged to the **Low-Grade Glioma (LGG)** or **Glioblastoma Multiforme (GBM)** class.

Patch-level accuracy was computed by comparing the predicted labels for all patches with their corresponding true labels. This metric measures how effectively the classifiers identify cancerous features at the microscopic level, where each patch serves as a representation of a small tissue region.

- **Training Accuracy:** 100% for all models, indicating that the models fit the training data perfectly.
- **Test Accuracies:**
  - XGBoost Classifier: 69.7%
  - Gradient Boosting Classifier: 77.5%
  - Random Forest Classifier: 77.85%

Image-Level Classification

For image-level classification, the goal was to assign a single class label to the entire WSI. Since each image is composed of multiple patches, we aggregated the patch-level predictions to determine the final label for the whole image.

We used a **majority voting strategy** for this purpose:

- Each patch within a WSI was classified by the trained model.
- The class with the most votes across all patches was assigned as the label for the entire image.

This approach ensured that the image-level classification reflected the dominant tissue characteristics across the WSI, providing a more comprehensive view of the cancer type.

- **Training Accuracy:** 100% for all models, indicating a strong fit on the training data.
- **Test Accuracies:**
  - XGBoost Classifier: 71.43%
  - Gradient Boosting Classifier: 85.71%
  - Random Forest Classifier: 85.71%

Evaluating Classifier Performance

To validate the effectiveness of our approach, we computed classification metrics at both levels:

1. **Patch-Level Metrics**: These metrics focused on the accuracy of predictions for individual patches, offering insights into how well the classifiers captured localized features.
2. **Image-Level Metrics**: These metrics evaluated the accuracy of majority-vote-based predictions for whole images, emphasizing the classifiers' ability to generalize and provide robust results at the WSI level.

By analyzing both patch-level and image-level results, we demonstrated that the features extracted by KimiaNet were not only effective for localized tissue classification but also provided sufficient information for accurate WSI-level predictions. This highlights the versatility and reliability of using KimiaNet for histopathological image analysis.

Analysis of WSI Levels

Resolution plays a critical role in histopathological analysis, significantly impacting feature quality and computational feasibility. Whole Slide Images (WSIs) are available in multiple resolution levels, with **Level 0** being the highest resolution, offering the finest tissue details. However, analyzing Level 0 WSIs requires substantial computational resources due to their immense file sizes. In this study, we focused on **Level 1** and **Level 2** WSIs to strike a balance between computational feasibility and image quality.

- **Level 1** provides moderately high resolution, allowing detailed tissue analysis.
- **Level 2**, with further reduced resolution, offers smaller file sizes but less detailed imagery.

Key Observations

1. **Performance of Level 1 WSIs**: Across all classifiers—**XGBoost**, **Gradient Boosting**, and **Random Forest**—**Level 1 WSIs consistently outperformed Level 2 WSIs** in terms of test accuracy and other classification metrics.
2. **Performance of Level 2 WSIs**: Classification performance declined significantly with Level 2 WSIs, particularly for hyperparameter-tuned models, highlighting the importance of image resolution in maintaining classifier efficacy.

Building on these observations, we drew several key inferences about the relationship between resolution and classifier performance.

Inferences from the Results

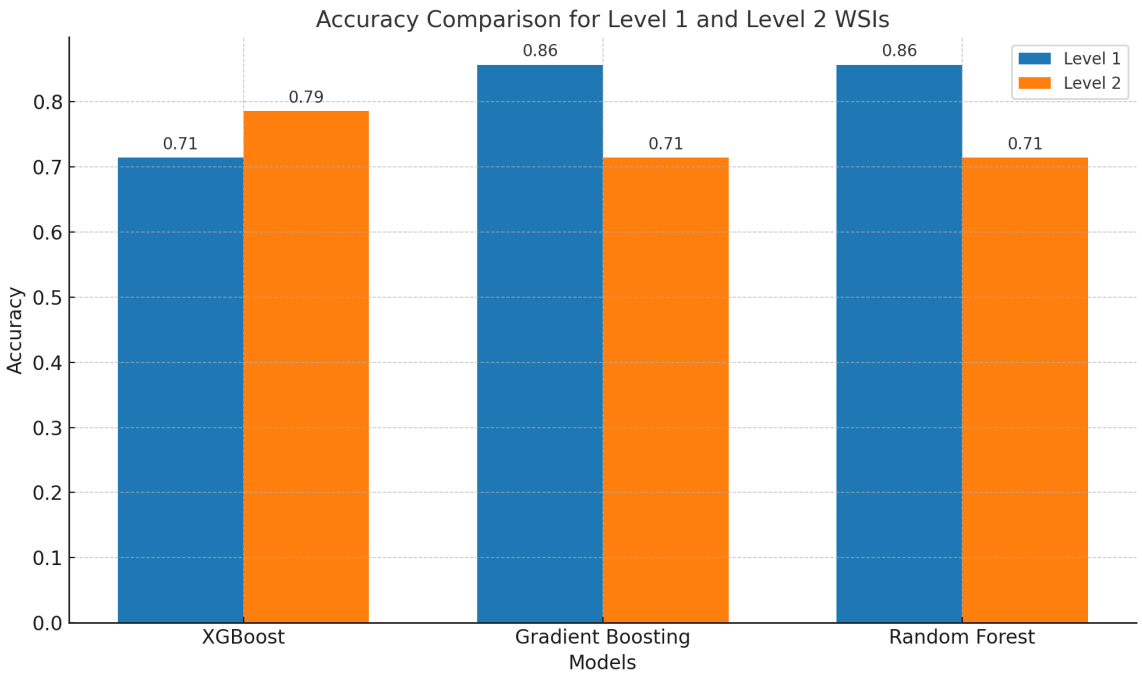
1. **Impact of Resolution on Performance**
    - **Superior Results with Level 1 WSIs**:
        - The higher resolution of Level 1 WSIs allowed KimiaNet to extract detailed and meaningful features. This enabled the classifiers to effectively distinguish between **Low-Grade Glioma (LGG)** and **Glioblastoma Multiforme (GBM)**.
        - On the other hand, the lower resolution of Level 2 WSIs led to a loss of critical details, making it harder for the classifiers to identify subtle differences. This resulted in lower accuracy and less consistent outcomes.
2. **Consistency Across Classifiers**
    - Models like **Random Forest** and **Gradient Boosting** performed well with Level 1 WSIs but showed noticeable drops in performance when using Level 2 WSIs. This highlights how the quality of extracted features, driven by resolution, directly impacts classification accuracy.
3. **Effect of Hyperparameter Tuning**
    - For Level 1 WSIs, hyperparameter tuning significantly improved the performance of Gradient Boosting and Random Forest classifiers, leading to higher accuracy and better precision-recall balance.
    - For Level 2 WSIs, hyperparameter tuning provided little to no improvement and, in some cases, even reduced performance. This was due to the lower quality of features extracted from the less detailed images.

Performance Comparison

The following table summarizes the performance of the tuned classifiers for Level 1 and Level 2 WSIs:

Classifier	WSI Level	Accuracy	Precision (LGG)	Recall (LGG)	Precision (GBM)	Recall (GBM)
XGBoost	Level 1	0.7143	0.67	0.86	0.80	0.57
Gradient Boosting	Level 1	0.8571	0.86	0.86	0.86	0.86
Random Forest	Level 1	0.8571	0.86	0.86	0.86	0.86
XGBoost	Level 2	0.7857	0.75	0.86	0.83	0.71
Gradient Boosting	Level 2	0.7143	0.67	0.86	0.80	0.57
Random Forest	Level 2	0.7143	0.67	0.86	0.80	0.57

Here are the bar plots showing a side-by-side comparison of metrics for Level 1 and Level 2 WSIs across different models:



Conclusion

This study shows that the resolution of Whole Slide Images (WSIs) plays a crucial role in classification performance. **Level 1 WSIs are the best choice**, balancing manageable image sizes and strong model performance. Their higher resolution provides enough detail for effective feature extraction and accurate classification, making them practical for situations with limited computational resources.

In comparison, **Level 2 WSIs** have lower resolution, reducing the quality of extracted features. This loss of detail affects the model's ability to classify accurately, highlighting how important resolution is for capturing the subtle details needed for reliable cancer diagnosis.

Generalization Capability of KimiaNet

KimiaNet, a pre-trained foundation model for histopathological image analysis, demonstrates robust generalization to diverse cancer types beyond its original training dataset. Despite being specifically trained on certain histopathological image datasets, KimiaNet's extracted features enable accurate classification for other cancer types, such as **Low-Grade Glioma (LGG)** and **Glioblastoma Multiforme (GBM)**, with high accuracy.

Key Observations

1. **Effectiveness of Pre-trained Features**
   - KimiaNet excels in extracting meaningful feature representations from histopathological images, regardless of whether the cancer type matches its training data.
   - The extracted features capture critical structural and biological patterns, facilitating accurate classification across different cancer datasets.
2. **Cross-Cancer Applicability**
   - KimiaNet generalizes well to new datasets, showcasing its adaptability in identifying cancer-specific differences in tissue samples.
   - For LGG and GBM, the feature vectors retained essential discriminatory information, enabling downstream classifiers to achieve consistent and reliable performance.
3. **Efficiency and Scalability**
   - Using pre-trained features eliminates the need for retraining or fine-tuning, saving computational resources and time.
   - This efficiency highlights KimiaNet's value as a scalable solution for diverse histopathological classification tasks.

Inferences from Results

1. **Foundation Model Strength**
   - As a foundation model, KimiaNet is inherently equipped to handle variations in cancer datasets due to its comprehensive pre-training on histopathological images.
   - The model's ability to transfer learned representations makes it a reliable tool for tasks that involve new cancer types or unseen datasets.

2. **Validation Through Performance**
    - The successful classification of LGG and GBM using KimiaNet's features confirms its generalizability and robustness.
    - Even without domain-specific fine-tuning, the features extracted by KimiaNet enabled classifiers to achieve strong performance across multiple metrics, reinforcing the hypothesis.
3. **Implications for Research and Diagnostics**
    - KimiaNet's adaptability to new cancer types indicates its potential to accelerate advancements in pathology research.
    - Reducing the need for custom feature extraction pipelines facilitates faster deployment of AI-driven diagnostic tools in diverse clinical and research settings.

Conclusion

KimiaNet's ability to generalize effectively across different cancer histopathology datasets highlights the strength of pre-trained models in biomedical image analysis. Its success in classifying LGG and GBM underscores the utility of foundation models in addressing varied diagnostic challenges, demonstrating how transfer learning can enhance accuracy while minimizing computational costs. This capability paves the way for broader applications of such models in cancer diagnostics and other medical imaging domains.