Akimuddin Aslam Shaikh

x22123245

MSc. Research Project

National College of Ireland

Title - Analysis of Scalable and Efficient Approaches for Liver disease detection

Question 1. How does the integration of texture-based (GLCM) and deep learning-based (ResNet50) features influence the performance of supervised classification models like Random Forest, SVM, and XGBoost in liver disease detection?

Answer:

The integration of GLCM and ResNet50 features against KMeans and Agglomerative clusters significantly improved the performance of supervised classification models like Random Forest, SVM, and XGBoost in liver disease detection because of the complementary nature of these features. Below is the explaination about the individual features influence:

1. **GLCM Features**: It captures spatial texture information such as contrast, correlation, homogeneity, dissimilarity and energy, which are crucial for analyzing liver tissues. These characteristics can differentiate between healthy and unhealthy areas by highlighting fine-grained patterns in the liver images.
2. **ResNet50 Features**: It provides a high-dimensional images representation through deep learning that focus on global image features and capture complex hierarchical patterns in the image dataset. It helps in capturing minute differences that might not be evident in texture features alone.
3. **Performance**: When GLCM and ResNet50 features were combined against KMeans clusters, it resulted in moderately lower result with 88% for RF, 87% for SVM, 100% for XGBoost. However, the integration of GLCM and ResNet50 against Agglomerative excelled the performance, providing a richer feature space. Random Forest: 97% accuracy, demonstrating its ability to handle combined feature sets effectively. SVM: 99% accuracy, benefiting from the enhanced separability in the feature space. XGBoost: 98% accuracy, leveraging its ability to handle high-dimensional data.
4. **Explanation**:
   * The complementary nature of both the feature sets addresses the limitations of each individual approach and ensures a more comprehensive representation of the dataset.
   * The combined features allow classifiers to learn from both fine-grained patterns, GLCM and ResNet50, resulting in better model generalization and improved accuracy.

This approach enhanced the detection accuracy and also demonstrated the potential for integrating texture-based and deep learning-based features in medical imaging tasks.

Question 2. What are the relative strengths and limitations of KMeans, DBSCAN, and Agglomerative Clustering in pseudo-labeling unlabeled liver image datasets, and how do these cluster labels impact downstream classification performance?

Answer:

1. **KMeans Clustering**

**Strengths:**

* **Scalability**: KMeans is computationally efficient and is able to handle large datasets, making it suitable for ResNet50 and GLCM features.
* **Balanced Clusters**: KMeans produced balanced clusters which is crucial for supervised classifiers.

**Limitations:**

* **Assumption of Spherical Clusters**: KMeans assumes clusters are spherical in shape and are equally sized, which may not align with the actual distribution of liver imaging datasets.
* **Inability to Handle Noise**: KMeans struggles with outliers or noisy data, which can distort cluster assignments.

**Impact on Classification:**

* KMeans with **GLCM features** resulted 100% accuracy across all classifiers (Random Forest, SVM, XGBoost), indicating potential overfitting or an inability to capture meaningful pattern from the data.
* KMeans with **ResNet50 features** achieved moderate results, as its partitioning did not fully align with the complex feature space.

1. **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**

**Strengths:**

* **Noise Handling**: The main strength of DBSCAN is that it can detect clusters of varying density and are robust to outliers, so they can handle noisy data well.
* **Non-Spherical Clusters**: It does not assume spherical clusters, making it more flexible for real-world data distributions.

**Limitations:**

* **High Dimensionality Issues**: DBSCAN performed poorly on our high-dimensional dataset, resulting in poorly defined clusters.
* **Computational Cost**: For large datasets, DBSCAN becomes computationally expensive and inefficient.

**Impact on Classification:**

* DBSCAN failed to generate meaningful cluster labels due to the high dimensionality and sparsity of the liver image dataset. This resulted in poor pseudo-labels that could not be effectively used for classification.

1. **Agglomerative Clustering**

**Strengths:**

* **Hierarchical Structure**: Agglomerative clustering builds a hierarchy, capturing both global and local relationships in the dataset. This worked well for complex features like ResNet50 and combined features.
* **No Assumption of Cluster Shape**: Unlike KMeans, it does not assume clusters to be spherical, making it more adaptable for diverse datasets.

**Limitations:**

* **Scalability**: Agglomerative clustering is computationally expensive, especially for large datasets, limiting its applicability without optimization.
* **Imbalanced Clusters**: It may produce imbalanced clusters, which may affect model performance.

**Impact on Classification:**

* Agglomerative clustering with **ResNet50 features** yielded excellent results (99% for Random Forest and SVM, 100% for XGBoost), showcasing its ability to generate meaningful pseudo-labels for high-dimensional data.
* Agglomerative clustering with **GLCM features** achieved 100% accuracy across all classifiers, suggesting it effectively captured spatial patterns but may have led to overfitting in simpler feature spaces.

**Conclusion:**

* **KMeans** is performs well for balanced clusters but struggles with capturing complex feature relationships, leading to moderate results for ResNet50 features but potential overfitting for GLCM.
* **DBSCAN** failed to extract meaningful cluster labels patterns, due to high dimensionality and parameter sensitivity.
* **Agglomerative Clustering** emerged as the most effective method and provided with meaningful pseudo-labels for complex and combined features.

Question 3. How does the alignment of GLCM features with KMeans and Agglomerative cluster labels lead to potential overfitting, and what measures can be taken to validate model generalizability in real-world scenarios?

Answer:

**Reason for potential overfitting:**

1. **High Accuracy**:
   * Both **GLCM + KMeans** and **GLCM + Agglomerative Clustering** achieved **100% accuracy** across all models (Random Forest, SVM, and XGBoost). While these results suggest excellent alignment between the features and pseudo-labels, however, this accuracy is highly unusual in real-world scenarios and a strong indicator of overfitting.
   * Overfitting occurs when the models learn patterns specific to the training data rather than generalizable patterns. This gets worsen when:
     + **Cluster labels perfectly align** with extracted features.
     + There is **insufficient variability** in the features to challenge the model.
2. **Cluster Label Dependency**:
   * Models become dependent on pseudo-labels for classification, which may not correspond to meaningful groupings. This increases the risk of poor generalization when the dataset distribution changes.

### ****Measures to Validate Model Generalizability****

1. **Cross-Validation**:
   * Perform **k-fold cross-validation** to evaluate model performance across multiple data splits. This ensures that the high accuracy is not an artifact of a specific train-test split.
2. **External Validation Metrics**:
   * Use advanced clustering evaluation metrics like **Dunn Index** or **Davies-Bouldin Score** to assess the quality of clusters and their suitability for downstream tasks.
3. **Dimensionality Reduction**:
   * Apply methods like **Principal Component Analysis (PCA)** or **t-SNE** to reduce feature dimensionality. This prevents overfitting by focusing on the most important features while discarding noise.
4. **Hybrid Clustering**:
   * Explore **hybrid clustering approaches**, combining the strengths of multiple clustering methods to ensure more robust pseudo-labels.
5. **Evaluation on Real-World Scenarios**:
   * Test the model in real-world settings, such as **clinical diagnosis workflows**, to assess its practical applicability.

Question 4. What challenges does the computational cost of ResNet50 feature extraction and clustering pose for large-scale liver disease detection, and how can these issues be mitigated to improve real-time application feasibility?

Answer:

### ****Potential Challenges from Computational Cost of ResNet50 Feature Extraction and Clustering****

1. **High Computational Demand for ResNet50**:
   * **Feature Extraction Complexity**:
     + ResNet50 has over 23 million parameters and so, extracting features from high-resolution liver images is computationally intensive, especially for large-scale datasets.
   * **Hardware Dependency**:
     + For efficient feature extraction, advanced GPUs or TPUs are required to process data within a reasonable time frame. This poses a challenge in real-time applications.
   * **Latency in Real-Time Scenarios**:
     + The time taken to preprocess data and extract features from it, makes it difficult to meet the demands of real-time liver disease detection systems.
2. **Scalability Issues in Clustering**:
   * **Large Dataset Bottlenecks**:
     + Unsupervised clustering techniques like Agglomerative Clustering, perform poorly with large datasets due to their computational complexity (e.g., Agglomerative Clustering has O(n3)O(n^3)O(n3) time complexity).
   * **Memory Constraints**:
     + Clustering large datasets requires significant memory, which can exceed the capacity of standard computational resources, especially when combined with high-dimensional ResNet50 features.
3. **Risk of Model Overfitting**:
   * The high-dimensional ResNet50 features combined with clustering labels can lead to overfitting, especially if the clusters are overly specific and not representative of real-world scenarios.

### ****Mitigation Strategies to Improve Real-Time Feasibility****

1. **Feature Dimensionality Reduction**:
   * Apply **Principal Component Analysis (PCA)** or **t-SNE** to reduce the dimensionality of ResNet50 features, retaining the most significant components while lowering computational cost and memory requirements.
   * Use techniques like **Feature Selection** to focus only on the most impactful features for classification.
2. **Lightweight Feature Extractors**:
   * **MobileNet** or **EfficientNet** models can be used which are considered to be the lightweight as compared to ResNet50. They are optimized for lower computational cost without significant performance loss.
3. **Batch Processing for Clustering**:
   * Dataset can be divided into smaller batches for clustering, and then merge cluster results. This reduces memory requirements and makes clustering scalable to large datasets.
4. **Hardware Optimization**:
   * Use **optimized hardware** like GPUs, TPUs to accelerate computations. Cloud platforms (e.g., Google Colab Pro, AWS) provide scalable solutions for large datasets.
5. **Hybrid Feature Extraction**:
   * Use GLCM features for computationally lightweight scenarios and reserve ResNet50 features for high-priority cases, balancing performance and cost in real-time applications.

### ****Conclusion****

The computational cost of ResNet50 feature extraction and clustering poses significant challenges for large-scale liver disease detection, particularly in real-time scenarios. Mitigation strategies mentioned above can improve scalability and feasibility.

Question 5. You have suggested that your inquiry has ‘..drew inspiration from existing research’, can you expand on this?

Answer:

The study was inspired from existing research that explored both texture-based and deep learning approaches for medical imaging analysis. For instance:

1. **GLCM-Based Approaches**:

* The GLCM approach was derived from **ming Xian (2010) and Priyanka, and Kumar** **(2020)**, which clearly explained GLCM features for the description of the spatial texture for medical image analysis. These studies applied GLCM in identifying pathologies in medical image modalities including liver tumors and kidney disease. The project went a step further in extending this idea by using GLCM features for liver image analysis.

1. **Deep Learning Models**:

* ResNet50 for the feature extraction is adapted from **Byra et al. (2018) and** **Rahman et al. (2022)** that demonstrated the applicability of deep learning models in the learning of high-dimensional features from images of medical nature.

1. **Clustering Techniques**:
   * + K-Means and Agglomerative Clustering adopted by pseudo-labeling of unlabeled datasets are stated by **Hum et al. (2011) & Chi et al. (2010)**. These methods solve the problem of the unavailability of labeled medical data, which is a bottleneck in medical imaging studies.
2. **Feature Integration**:

* The extraction of the GLCM and ResNet50 features was based on the study by **Chou et al. (2021) and Lakshmipriya and Pottakkat (2023**), that proposed separate features and their fusion to improve diagnostic accuracy.

1. **Evaluation Techniques**:
   * Evaluation metrics and clustering validation techniques were incorporated in line with the works by **Poličar et al. (2019)and Bandyopadhyay and Paul (2013),** which make mention of visualization and silhouette scores in the evaluation of clusters.

Question 6. What is the benefit of your research to society?

Answer:

The research project can be beneficial for the following reasons:

1. **Improved Diagnostic Accuracy**:

* The proposed research was based on analyzing the texture features using GLCM and using deep learning features through ResNet50 and the study showed it works by achieving high classification results of liver diseases. The study result can assist healthcare in timely and accurate decisions that would reduce wrongful diagnosis or delay in diagnosis.

1. **Addressing Data Scarcity**:

* The research used large and unidentified liver image database.For scalability, KMeans and Agglomerative Clustering methods were used for the generation of pseudo labels.

1. **Scalability for Broader Access**:

* The other methodologies like ResNet50, the unsupervised clustering that are used here can be used in other medical image analysis. All this creates the opportunity for similar application in identification of other diseases that lead to early detection.

1. **Pathway to Real-Time Applications**:
   * While not yet implemented, the research makes the way for future real-time diagnostic tools, further enhancing the speed and efficiency of medical diagnoses.