CS 415 Final Project Report
Group 24 - Aryansingh, Tanuj, and Snigdha

Medical Image Diagnostic

Personal View and Course Reflection

We believe that computer vision is a cornerstone of computer science, bridging the gap between data processing and the understanding of visual information. This course was an incredible first step that sparked our collective curiosity and excitement to explore more. It highlighted how crucial computer vision is for enabling machines to interpret and analyze visual data. From autonomous driving and traffic analysis to video and image generation and medical image diagnostics, we are excited to experiment further. It also gave us a chance to deepen our technical skills and see firsthand how computer vision can solve real-world problems, especially in fields like healthcare and autonomous systems.

Project Description

Overview

Medical imaging has revolutionized healthcare by providing crucial insights from disease detection to treatment planning. However, the rapid growth in the use of medical imaging technologies presents several substantial challenges that need to be addressed to fully realize their potential.

Problem Statement

Despite its benefits, the widespread adoption of medical imaging brings several challenges:

- The sheer volume of medical images generated daily poses a challenge for healthcare professionals. Manual interpretation is time-consuming and prone to errors, especially in high-pressure environments such as emergency rooms.
- Many healthcare facilities face resource constraints, such as a shortage of trained radiologists and pathologists. This makes it difficult to provide timely and accurate diagnostics, particularly in underserved regions.

- Diagnostic outcomes often vary based on the expertise of individual practitioners, leading to inconsistencies in patient care.
- There is a growing demand for automated solutions that can complement human expertise by identifying patterns and anomalies with high precision. This includes the ability to handle a diverse range of medical imaging datasets, from X-rays to histopathology slides.

To address these issues, there is an escalating demand for automated solutions that can identify patterns and anomalies with high precision. This research focuses on the critical question: Can sophisticated computer vision models provide robust, reliable, and scalable support for medical diagnostics, enhancing the accuracy, efficiency, and accessibility of healthcare delivery?

Goal

The goal of our project is to systematically leverage and compare advanced computer vision techniques to enhance the diagnostic capabilities of medical imaging systems. Our comprehensive approach to model analysis and evaluation aims to:

- Assist healthcare professionals by developing intelligent diagnostic support tools that:
  - Prioritize critical cases
  - Highlight key regions of interest in medical images, such as potential signs of COVID-19 or Viral Pneumonia. This strategic approach streamlines clinical decision-making processes, enabling faster, more precise, and more consistent diagnostics.
- Conduct a comparative analysis of computer vision techniques, focusing on:
  - Performance metrics across diverse medical imaging modalities
  - Computational efficiency and scalability
  - Generalizability and robustness across various medical imaging modalities

By developing robust computer vision models for medical image analysis, this project supports healthcare professionals in delivering accurate, efficient, and consistent diagnostics. This not only improves patient outcomes but also lays the foundation for integrating AI-driven diagnostics into innovative healthcare solutions.

Datasets

Medical MNIST

The Medical MNIST dataset comprises 60,000 grayscale images with a resolution of 64x64 pixels, categorized into six distinct medical imaging classes: Abdomen CT, Breast MRI, Chest CT, Chest X-Ray, Hand, and Head CT. Each class represents a unique type of medical imaging data, making this dataset a valuable resource for testing the generalization and feature extraction capabilities while allowing us to evaluate the impact of adding complexity (e.g., additional layers and regularization) to the architecture.

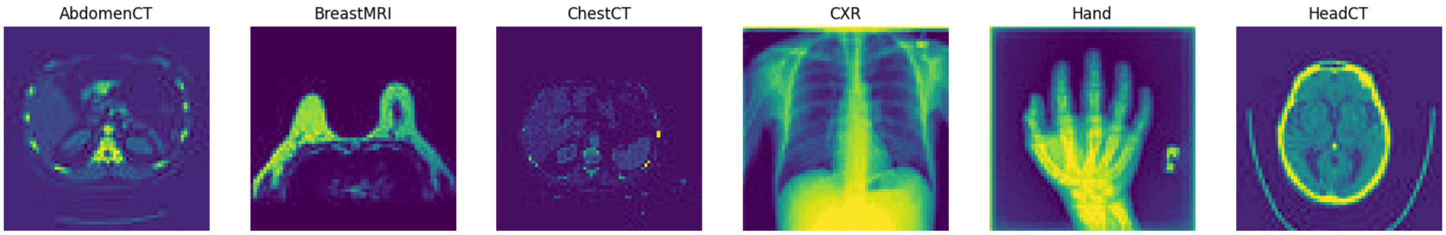


Fig 1: Medical MNIST dataset sample images

We selected this dataset due to its balanced and relatively simple structure, which makes it an ideal starting point for benchmarking our models. This dataset's wide use in the research community also enables direct comparison of our results with established benchmarks, making it a practical choice for performance evaluation.

Covid-19 Radiography

The Covid-19 Radiography dataset consists of 40,000 high-resolution (299 x 299 pixels) medical images, divided into four classes: Normal, COVID, Viral Pneumonia, and Lung Opacity. Each class contains diagnostic images representing varying levels of complexity, with COVID and Viral Pneumonia being particularly challenging due to subtle differences in radiographic patterns.

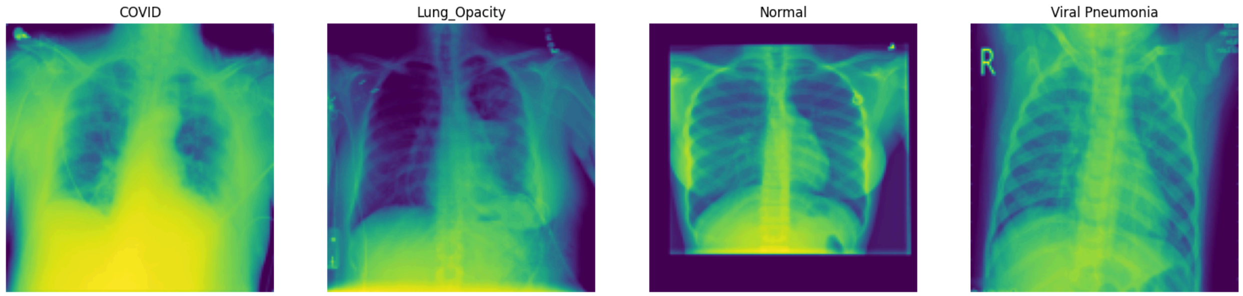


Fig 2: Covid-19 Radiography dataset sample images

The inclusion of this dataset adds depth to our analysis, allowing us to explore how architectural choices, such as the number of layers, regularization techniques, and optimizers, impact the ability of models to generalize across challenging and complex medical imaging tasks. Furthermore, the dataset aligns with the goal of leveraging AI for rapid and accurate diagnostic support in critical healthcare scenarios, particularly during global health crises like the COVID-19 pandemic.

Data Preprocessing

Images were resized to a uniform dimension (128x128 or 64x64) to ensure consistency in input size across the models. However, for Resnet, all dataset images were resized to the ResNet-50 model's default input size of 224x224. Additionally, pixel values were normalized to fall within the range [0,1], which enhanced model convergence and improved numerical stability during training.

Class imbalance in the datasets was addressed by computing class weights based on the frequency of each class. These weights were incorporated during training to ensure that minority classes received adequate attention, preventing biased predictions toward the majority class.

Despite the substantial size of the available datasets, data augmentation techniques such as rotation, flipping, and scaling were explored as an experimental trial. This approach was undertaken to assess their potential impact on model performance and generalization. By artificially introducing variability, these techniques simulated real-world conditions, offering valuable insights into their potential benefits for enhancing robustness, even with already diverse and extensive datasets.

Various train-test splits were explored to optimize model performance across approaches. While an 80-20 split was used as a baseline, a 70-15-15 split (training, validation, test) was also applied for fine-tuning and consistent evaluation. This flexible partitioning helped assess and enhance model performance, ensuring configurations were tailored to each approach.

Approach 1 (Traditional Methods)

Overview

This approach employs traditional computer vision techniques for medical image diagnostics. It focuses on feature extraction and classification using Support Vector Machines (SVMs) combined with methods such as Histogram of Oriented Gradients (HoG) and Scale-Invariant Feature Transform (SIFT). The primary goal is to evaluate the effectiveness of these techniques on two datasets: Medical MNIST and COVID-19 Radiography.

Feature Extraction

HoG Features

- Divided images into 8x8 pixel cells.
- Grouped two cells per block and applied normalization.
- Conducted experiments varying the cell and block sizes to determine their impact on feature representation and model performance.

SIFT Features

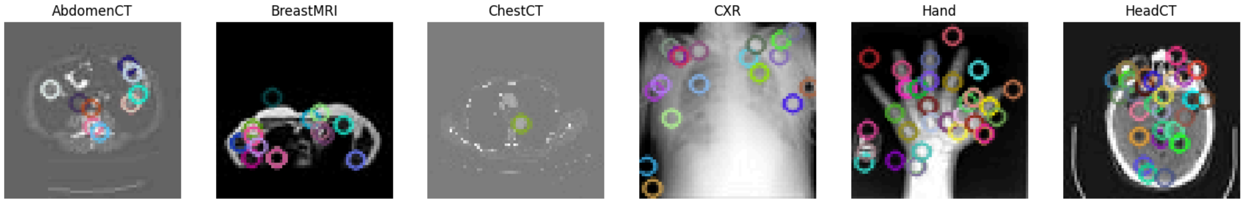


Fig: SIFT features extracted for the Medical MNIST

- Tried and compared both unmasked and masked images for selective focus.

- Experimented to find the optimal number of SIFT features, settling on extracting up to 100 key points per image for both masked and unmasked configurations.
- Experimented with different thresholds for key point detection for computational efficiency and accuracy.

Classification

Support Vector Machines (SVMs):

- Trained using HoG and SIFT features to classify medical images.
- GridSearchCV was employed to tune hyperparameters such as kernel type, regularization strength, and gamma values.
- Conducted multiple experiments:
  - Linear kernel excelled in simpler datasets (Medical MNIST), leveraging its ability to handle linearly separable data efficiently.
  - RBF kernel performed better on complex datasets (COVID-19 Radiography), effectively managing non-linear relationships in high-dimensional feature spaces.
  - Regularization strength tuning: Found optimal values for balancing bias and variance.
- While SVM results varied across configurations, it served as an essential benchmark for evaluating feature extraction methods and informed the use of ensembles.

Bag of Words (BoW) Model:

- Performed k-means clustering on SIFT features to generate vocabularies.
- Experimented with vocabulary sizes (e.g., 50, 100, and 500 words):
  - Smaller vocabularies provided computational efficiency but lower accuracy.
  - Optimal results were achieved with 100-word histograms
- Integrated BoW with a pipeline:
  - Feature extraction → Train k-means → Compute histograms → Train SVM.

- Increased the number of training images to test the scalability of the BoW approach, observing marginal accuracy improvements.

Decision Trees + Ensembles

- Bagging and gradient boosting were applied to decision trees to reduce overfitting and improve classification robustness.
- Conducted trials with varying tree depths.
- Integrated SIFT features within ensembles to improve overall stability and accuracy.

SVM + Ensembles

- Bagging and AdaBoost were applied to SVM classifiers, to analyze their adaptability across datasets.
- Used HoG and SIFT features for SVM with both Bagging and AdaBoost.
- Explored different ensemble sizes, balancing computational time and accuracy

Challenges and Solutions

Parameter Sensitivity: Feature extraction and SVM performance were highly sensitive to hyperparameters. Extensive grid searches were conducted to determine optimal configurations for various datasets, ensuring consistent model performance.

Class Imbalance and Generalization: Imbalanced class distributions skewed model predictions. To mitigate this, data augmentation techniques and ensemble strategies were employed, enhancing the model's generalization capabilities.

Bag of Words (BoW) Effectiveness: Despite numerous experiments, the BoW approach did not significantly improve accuracy. This highlighted the need for more advanced feature extraction methods to better capture the complexity of medical images.

Masking Experiments: Applying masks to focus on regions of interest often failed to improve classification accuracy. Refinement of masking strategies, including dynamic mask generation, was considered to better capture relevant image areas without losing key contextual features.

Although HoG and SIFT worked well for global feature extraction, they struggled with capturing more complex local features, such as lung damage, and failed to build upon basic features to form higher-level representations. These limitations highlighted the need for advanced feature extraction methods like convolutional techniques, explored further in Approach 2.

Approach 2 (Convolution Neural Networks)

Overview

CNN's are deep learning algorithms used for image recognition and classification tasks. We developed two CNN architectures, both rooted in deep learning principles: a simple (base) model and a bigger (advanced) model, tailored to handle the challenges posed by medical imaging datasets. Leveraging the power of deep learning, these models were designed to automatically learn hierarchical features from images, minimizing the need for manual feature extraction.

Implementation

The simple model was built with simplicity in mind, consisting of two convolutional layers with 32 and 64 filters, each followed by max pooling layers to reduce spatial dimensions while preserving critical features. A dropout layer with a rate of 0.6 was included before the dense layer to combat overfitting. This architecture aimed to extract core features from the images efficiently, making it well-suited for smaller datasets or datasets with simpler patterns. Its design exemplifies how deep learning can adapt to scenarios requiring computational efficiency while maintaining robust performance.

The bigger model, on the other hand, utilized a deeper deep learning architecture to tackle more complex datasets. It comprised three convolutional layers with 32, 64, and 128 filters, each followed by max pooling layers for feature downsampling. To enhance generalization and prevent overfitting, the advanced model incorporated L2 regularization in all convolutional and dense layers, coupled with multiple dropout layers at rates of 0.3, 0.5, and 0.5 at different stages. This architecture was designed to learn more intricate patterns and capture latent features, particularly in datasets with higher diversity or complexity. However, the increased depth and regularization demanded larger amounts of data to improve learning, highlighting the interplay between model complexity and data requirements in deep learning.

Both models were trained using the Adam and Nadam optimizers, key tools in deep learning to refine model weights and improve convergence. Adam, with its adaptive learning rate and momentum, provided stable optimization, while Nadam extended it further with Nesterov momentum, offering faster convergence and smoother updates. These optimizers underscored the importance of selecting the right algorithms to maximize learning and training efficiency.

Challenges

Vanishing Gradient Problem: As the depth of the CNN architecture increases, gradients can diminish during backpropagation, making it harder for the model to learn meaningful features.

Feature Extraction Bottleneck: Traditional CNN architectures depend heavily on layers progressively extracting features, which can lead to information loss if earlier layers fail to extract sufficient patterns.

Computational Complexity: Larger CNNs with deeper architectures require significant computational resources and training time, especially for high-resolution datasets like Covid Radiography.

Solutions

Class Balancing Techniques: Applied class weights during training to address the imbalance in datasets, ensuring that minority classes received appropriate focus and attention.

Data Augmentation: Augmented the datasets with techniques like rotation, scaling, and flipping to artificially increase the diversity of the training data, improving generalization.

Optimized Architectures: The simple model was designed for datasets with simpler patterns, while the bigger model targeted complex datasets. This tailored approach helped balance efficiency and performance.

Training with Multiple Optimizers: Leveraged both Adam and Nadam optimizers to compare their effects on model convergence and generalization, selecting the best fit for each dataset.

Despite the promising results with our CNN architectures, particularly the Simple (Base) Model, we encountered challenges with the Bigger (Advanced) Model, which did not consistently outperform the Simple Model. The increased depth and regularization sometimes led to feature loss, especially in datasets with limited diversity like COVID-19 Radiography. These challenges highlighted the need for a more robust approach, leading us to explore transfer learning with ResNet-50 to enhance performance and efficiency.

Approach 3 (Transfer Learning with ResNet-50)

Overview

The model employed as the backbone for transfer learning was ResNet-50, a 50-layer deep residual network, leveraging residual connections to solve the vanishing gradient problem in very deep architectures. The ResNet-50 model used in this implementation was imported from the TensorFlow Keras Applications library, which follows the original paper by He et al. (2016). The model was pre-trained on the ImageNet dataset, making it highly suitable for feature extraction and fine-tuning in our application.

Implementation

The base model is ResNet-50 with an input size of 224x224x3 initialized with ImageNet weights and excluding the top classification layer. Of its 177 functional layers, the last 40 were unfrozen for fine-tuning, while the first 147 remained frozen to retain pre-trained features.

Custom layers included a Global Average Pooling layer, a Dense layer with 128 neurons and ReLU activation, L2 regularization with a penalty coefficient of 0.01, a Dropout layer with a 20% rate, and a final Dense layer with softmax activation for multi-class classification.

Class weights were incorporated into the loss function to address the class imbalance and ensure fair learning across all classes. Optimization employed Categorical Cross-Entropy as the loss function, Adam and Nadam optimizers with an initial learning rate of 0.001, a learning rate scheduler (ReduceLROnPlateau) reducing the rate by 0.5 after 5 epochs without validation loss improvement, and Early Stopping with a patience of 5 epochs to restore the best weights.

Challenges and Solution

Multiple strategies and iterative trials were employed to develop an optimized model to address challenges related to underfitting and overfitting. These experiments involved adjusting various parameters and configurations to achieve a well-balanced and high-performing model.

Underfitting Model (Naive with minimal modifications): This is the baseline for measuring improvements. Only the pre-trained ResNet-50 backbone is used without changing its weights. The convolutional layers remain frozen, and no fine-tuning is applied. A lightweight classifier head is appended to the model to adapt ResNet-50 for the target task. This head includes Global average pooling and a dense layer with a softmax activation for classification tasks (or sigmoid for binary tasks). This underperformed as it didn't fully leverage the model's capacity to learn task-specific features.

Overfitting Model: Building on the naive approach, moderate updates were introduced to improve performance. To address class imbalance, class weights were incorporated into the loss function. Both Keras Sequential and Functional API (Model) were trialed for the custom head layer, with the Functional API being preferred due to its flexibility and better performance for advanced architectures like ResNet-50. Partial fine-tuning was applied by unfreezing the last residual block to allow feature adaptation, using a lower learning rate to stabilize updates. Dropout layers, up to 50%, were added to the custom classifier to mitigate overfitting. Early stopping was utilized to monitor validation loss and prevent overtraining, while input sizes were tested for optimal configuration. These strategies either enhanced the model's performance or, in some cases, led to overfitting, emphasizing the importance of iterative trials to achieve a well-optimized model.

Result and Analysis

Approach 1

GridSearch Results

Experiment	**Dataset**	**Parameters**	**Results** (Best Cross-Validation Score)
SVM (HoG)	COVID-19 Radiography	C=1, kernel=poly	82.75%
	Medical MNIST	C=1, kernel=rbf	99.875%
SVM (SIFT)	COVID-19 Radiography	C=1, kernel=poly	76.875%
	Medical MNIST	C=1, kernel=rbf	99.875%
BoW (SIFT)	COVID-19 Radiography	C=1, kernel=rbf	31.625%
	Medical MNIST	C=1, kernel=rbf	22.989%

Table 1: GridSearch Results

The GridSearchCV method was employed to fine-tune hyperparameters such as kernel type and regularization strength for the SVM models. The best parameters identified for each dataset and feature extraction method are summarized in Table 1. For the SVM with HoG features. The choice of C=1 indicates a balanced regularization, preventing both underfitting and overfitting. The polynomial kernel was effective for the COVID-19 dataset due to its ability to capture complex patterns, while the RBF kernel worked well for the Medical MNIST dataset by handling non-linear relationships efficiently.

SVM

Experiment	**Dataset**	**Parameters**	**Accuracy**	**Recall**	**Precision**
HoG	COVID-19	C=1, kernel=poly	88.25%	0.88	0.88
	Medical MNIST	C=1, kernel=rbf	99.50%	0.99	0.99
SIFT	COVID-19	C=1, kernel=poly	76.00%	0.75	0.80
	Medical MNIST	C=1, kernel=rbf	88.75%	0.89	0.9

Table 2: Simple SVM training results

The SVM models were further evaluated using the best parameters identified through GridSearchCV. The results are detailed in Table 2. The SVM with HoG features achieved an accuracy of 88.25% on the COVID-19 Radiography dataset and 99.5% on the Medical MNIST dataset. This indicates that while the SVM with HoG features performed exceptionally well on the simpler Medical MNIST dataset, it faced challenges with the more complex COVID-19 Radiography dataset. On the other hand, the SVM with SIFT features achieved an accuracy of 76% on the COVID-19 Radiography dataset (with masks) and 88.75% on the Medical MNIST dataset. For SIFT features, the precision slightly exceeded recall on the COVID-19 dataset, indicating a tendency to minimize false positives at the expense of missing some true cases.

The high accuracy on the Medical MNIST dataset suggests that HoG features are effective in capturing the overall shape and structural patterns critical for classification. However, the lower accuracy on the COVID-19 Radiography dataset indicates that **HoG features struggle with finer internal textures**. Similarly, **SIFT features, while effective for localized feature importance, underperformed in summarizing global structural characteristics** compared to HoG.

Bag of Words (BoW)

Experiment	Dataset	Parameters	Accuracy	Recall	Precision
Vocab Size	COVID-19	50 words	46.25%	0.47	0.53
		100 words	**58.75%**	**0.58**	**0.59**
		500 words	22%	0.25	0.06
	Medical MNIST	50 words	20.25%	0.22	0.14
		100 words	21.2%	0.20	0.16
		500 words	15.2%	0.18	0.14
Number of Images	COVID-19	1000 images	21.2%	0.24	0.05
		5000 images	58.75%	0.58	0.59
	Medical MNIST	1000 images	16.5%	0.20	0.15
		5000 images	34.4%	0.36	0.20

Table 3: BoW experiments results

The BoW approach was evaluated with different vocabulary sizes and the number of images. The results are presented in Table 3. For the COVID-19 Radiography dataset, the optimal results were achieved with 100-word histograms, yielding an accuracy of 58.75%. This indicates that a moderate vocabulary size strikes a balance between computational efficiency and accuracy. For the Medical MNIST dataset, the accuracy varied with the vocabulary size, with the highest accuracy of 21.2% achieved with 100 words. The number of images also impacted the results, with 5000 images yielding the highest accuracy of 58.75% for the COVID-19 Radiography dataset and 34.4% for the Medical MNIST dataset. The improvement in recall from 0.05 to 0.58 highlights a significant enhancement in the model's ability to identify true positive cases effectively and decrease false negatives. This shift indicates that increasing the training images to 5000 enabled the model to capture more representative features, improving its sensitivity to relevant cases in the complex COVID-19 dataset while still leaving room for further optimization.

These results suggest that increasing the number of images improves the model's ability to generalize and capture relevant features. The disparity between precision and recall for larger vocabularies further illustrates the approach’s inefficacy in distinguishing meaningful features in simpler datasets. The moderate vocabulary size of 100 words provides a good balance, allowing the model to effectively represent the data without overfitting. But overall, the BoW approach fails to summarize features as well as simple HoG or SIFT.

Decision Trees + Ensemble

Experiment	Dataset	Parameters	Accuracy	Recall	Precision
SIFT and Gradient Boosting	COVID-19	N_ESTIMATORS=50	**55.00%**	**0.55**	**0.57**
	Medical MNIST	N_ESTIMATORS=50	**90.00%**	**0.89**	**0.90**
SIFT + Random Forest	COVID-19	N_ESTIMATORS=50	48.00%	0.49	0.47
	Medical MNIST	N_ESTIMATORS=50	89.50%	0.89	0.89

Table 4: D-Trees + Ensemble experiments results

The performance of decision trees and ensemble methods was evaluated using SIFT features. The results are detailed in Table 4. For instance, the Random Forest with N_ESTIMATORS=50 achieved an accuracy of 48.00% on the COVID-19 Radiography dataset and 89.50% on the Medical MNIST dataset. Gradient boosting with deeper trees captured complex patterns more effectively, with the Gradient Boosting model achieving an accuracy of 55.00% on the COVID-19 Radiography dataset and 90.00% on the Medical MNIST dataset.

These results demonstrate that decision trees benefit from ensemble techniques by leveraging their simplicity and complementing the strengths of SIFT features. Close precision and recall also indicate that models balance between False positive and True cases. Random Forests helps reduce overfitting by averaging multiple models, while gradient boosting improves performance by focusing on difficult-to-classify instances. Experiments also showed that increasing the number of estimators increased performance marginally.

SVM + Ensemble

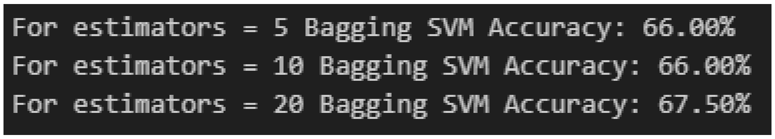
Experiment	Dataset	Parameters	Accuracy	Recall	Precision
SVM + Ensemble + SIFT	COVID-19	Bagging, N_ESTIMATORS=10	51.75%	0.53	0.55
		AdaBoost, N_ESTIMATORS=10	45.25%	0.46	0.47
	Medical MNIST	Bagging, N_ESTIMATORS=10	82.75%	0.83	0.85
		AdaBoost, N_ESTIMATORS=10	86.25%	0.86	0.87
SVM + Ensemble + HoG	COVID-19	Bagging, N_ESTIMATORS=10	75.50%	0.76	0.75
		AdaBoost, N_ESTIMATORS=10	**82.50%**	**0.82**	**0.83**
	Medical MNIST	Bagging, N_ESTIMATORS=10	**99.75%**	**0.99**	**0.99**
		AdaBoost, N_ESTIMATORS=10	86.25%	0.87	0.9

Table 5: SVM + Ensemble experiments results

The SVM models were further enhanced using ensemble methods such as bagging and AdaBoost. The results are presented in Table 5. HoG with bagging achieved strong performance for the Medical

MNIST dataset with an accuracy of 99.75%, and 0.99 precision and recall showcase its stability. HoG with AdaBoost enhanced robustness for the COVID-19 Radiography dataset, achieving an accuracy of 82.50%, precision (0.83) and recall (0.82), highlighting its ability to correct misclassifications effectively compared to bagging for the same. The SVM with SIFT features achieved an accuracy of 51.75% on the COVID-19 Radiography dataset (with masks) and 86.25% on the Medical MNIST dataset.

This suggests that AdaBoost's ability to focus on misclassified instances is useful for handling the complexity and subtle variations in medical radiographs, where precise classification is critical. Ensemble methods still helped improve performance but revealed that SIFT may not capture the global patterns necessary for more complex medical datasets. Precision and recall values here point to the need for more sophisticated feature extraction methods to complement ensemble strategies.



```
For estimators = 5 Bagging SVM Accuracy: 66.00%
For estimators = 10 Bagging SVM Accuracy: 66.00%
For estimators = 20 Bagging SVM Accuracy: 67.50%
```

Fig: Changing the number of estimators in SVM + Ensemble

The figure indicates that increasing the number of estimators can lead to marginal improvements in accuracy, but the gains may plateau beyond a certain point.

The results emphasize that while ensemble methods like Bagging and AdaBoost can boost performance, the choice of feature extraction plays a significant role in determining the effectiveness of these models. Overall, SVM + Ensemble methods are promising, but further refinement in feature extraction and ensemble strategies is needed for complex datasets like COVID-19 Radiography.

Approach 2

Results for simple (base) model

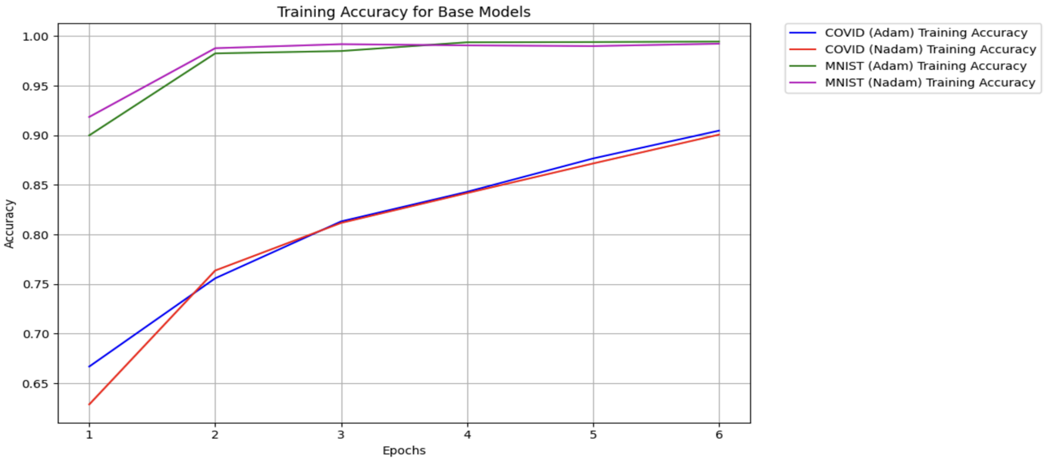


Fig 3: Training accuracy for Simple (Base) CNN Model

Simple (Base) Model	Covid Radiology Dataset	Medical MNIST Dataset
Test Accuracy	Adam: **92.76%** Nadam: 91.51	Adam: 98.73% Nadam: **99.86%**

Table 6: Test accuracy for Simple (Base) CNN Model

Results for Bigger (advanced) model

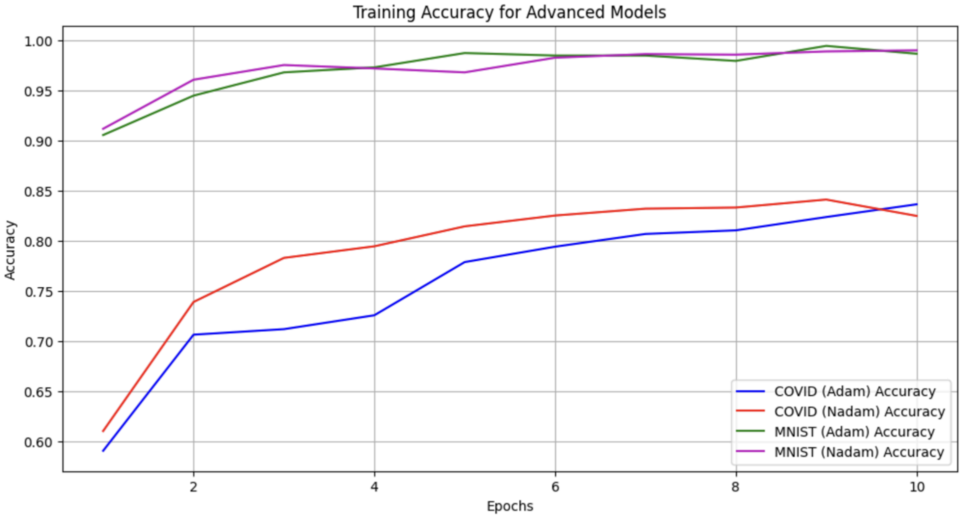


Fig 4: Training accuracy for Bigger (Advanced) CNN Model

Bigger Model	Covid Radiology Dataset	Medical MNIST Dataset
Test Accuracy	Adam: 57.40% Nadam: **78.76%**	Adam: 98.07% Nadam: **98.29%**

Table 6: Test accuracy for Bigger (Advanced) CNN Model

The results from the Simple (Base) and Bigger (Advanced) CNN models revealed interesting insights into the relationship between dataset characteristics and model architectures. On the Covid Radiology Dataset, the Simple Model significantly outperformed the Bigger Model, achieving a test accuracy of 92.76% with the Adam optimizer compared to 57.40% for the Bigger Model. This can be attributed to the Simple Model's lightweight architecture, which effectively extracted dominant features like edges and textures without overfitting. The reduced complexity allowed the Simple Model to converge faster and maintain a balance between feature extraction and generalization, making it more suited for this relatively smaller and less diverse dataset.

For the Medical MNIST Dataset, the results were closer, but the Simple Model still achieved a marginally higher accuracy of 99.86% compared to the 98.29% achieved by the Bigger Model with the Nadam optimizer. Despite its deeper architecture and additional regularization techniques, the Bigger Model's complexity did not lead to a performance advantage on this dataset. The Medical MNIST dataset's simplicity and clearer patterns may have aligned better with the Simple Model's ability to extract core features without unnecessary complexity.

The Bigger Model, designed for datasets with greater variability and noise, likely faced challenges with both datasets due to its reliance on regularization and deeper layers, which may have suppressed the learning of critical patterns. While the additional layers and regularization techniques (e.g., L2 regularization and Dropout) helped prevent overfitting, they may have also introduced feature loss, particularly on datasets with limited diversity or scale, such as the Covid Radiology dataset.

Approach 3

Test Accuracy	Adam	Nadam
COVID	**97.42%**	95.61%
MNIST	98.08%	**99.96%**

Table 7: Test accuracy for ResNet50

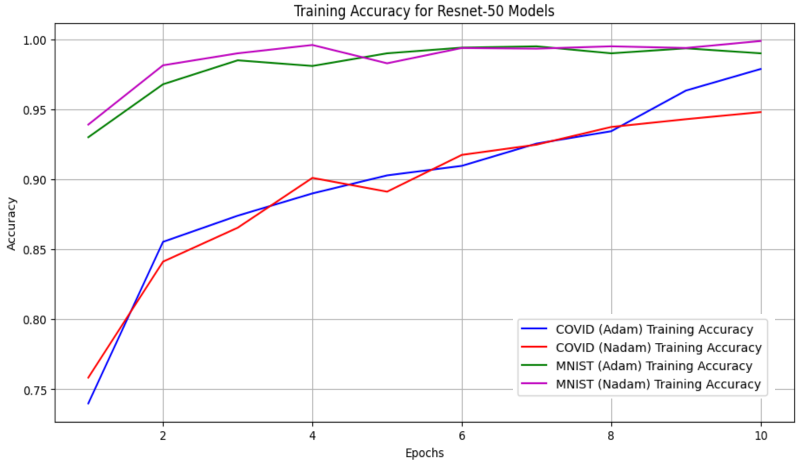


Fig 5: Training accuracy for ResNet-50 models

The results compare the performance of the ResNet-50 models fine-tuned using the Adam and Nadam optimizers. The test accuracy table shows that the Adam optimizer performed slightly better for the COVID dataset, achieving a test accuracy of 97.42%, compared to 95.61% with Nadam. On the other hand, Nadam outperformed Adam on the MNIST dataset, yielding an impressive accuracy of 99.96%, compared to 98.08% with Adam.

The training accuracy graph highlights the learning behavior over 10 epochs for both datasets. While both optimizers demonstrate consistent improvement in training accuracy, Nadam converges slightly faster, particularly for MNIST, reaching near-perfect accuracy earlier. However, the Adam optimizer exhibits steadier performance for the COVID dataset, aligning better with its test accuracy result. This suggests that the Adam optimizer may be more robust for complex datasets like COVID, while Nadam shows exceptional performance for simpler and well-balanced datasets like MNIST.

Computational Efficiency

Training and prediction times varied significantly across the models. Traditional methods like SVM with HoG and SIFT features trained quickly (e.g., 34.7 seconds per fold with GridSearchCV) but struggled with complex datasets. CNNs showed a balance between training time and accuracy. The Simple Model trained on the COVID-19 dataset took ~27 minutes for 6 epochs, while the Advanced

Model required ~48 minutes for the same, making it suitable for more complex data. Transfer learning with ResNet-50 had the longest training time (~110 minutes for 11 epochs) but delivered the highest accuracy.

Prediction times also reflected the models' complexity. Traditional methods excelled in speed (milliseconds per image) but lacked accuracy. The Simple CNN predicted images in ~100 ms, the Advanced Model in ~150 ms, and ResNet-50 in ~200 ms. While smaller models are faster and more cost-effective, larger models like ResNet-50 offer higher accuracy, making them better suited for critical diagnostics. The trade-off between accuracy and efficiency should guide model selection for specific applications.

Conclusion

Our project has been a comprehensive exploration into the application of computer vision techniques for enhancing medical diagnostics. Through this project, we have delved into traditional machine learning methods, advanced convolutional neural networks (CNNs), and state-of-the-art transfer learning models, each offering unique insights and contributions to the field of medical imaging.

In our initial approach, we employed traditional computer vision techniques such as Support Vector Machines (SVMs) combined with Histogram of Oriented Gradients (HoG) and Scale-Invariant Feature Transform (SIFT) for feature extraction. These methods were effective on simpler datasets like Medical MNIST, achieving high accuracy. However, their performance on more complex datasets like COVID-19 Radiography highlighted the limitations of traditional feature extraction methods in capturing intricate patterns. The SVMs, while robust, struggled with the finer details and internal textures of the COVID-19 images, indicating a need for more sophisticated techniques.

The limitations observed with traditional methods, particularly their inability to effectively capture complex and localized features in medical images, prompted us to explore more advanced techniques. The traditional methods' reliance on manual feature extraction and their sensitivity to hyperparameters made them less adaptable to the diverse and intricate nature of medical imaging data. This led us to develop and implement CNNs, which are known for their ability to automatically learn hierarchical features from images, thereby minimizing the need for manual intervention and improving overall performance.

In the second approach, we developed two CNN architectures. A simple (base) model and a bigger (advanced) model. The simple model, with its lightweight architecture, excelled in extracting dominant features from both datasets, achieving impressive accuracy, particularly on the COVID-19 Radiography dataset. Its simplicity allowed for faster convergence and effective generalization. On the other hand, the bigger model, designed for greater complexity, showed mixed results. While it performed well on the Medical MNIST dataset, its deeper architecture and regularization techniques did not translate to a significant performance boost on the COVID-19 dataset. This highlighted the importance of balancing model complexity with dataset characteristics.

Despite the success of CNNs, the advanced model's performance on the COVID-19 dataset revealed the challenges of designing architectures that can generalize well across diverse datasets. The increased depth and regularization, while preventing overfitting, also introduced feature loss, particularly on datasets with limited diversity. These challenges underscored the need for leveraging pre-trained models that have already learned a wide range of features from large-scale datasets. This realization led us to explore transfer learning with ResNet-50, a powerful technique that allows us to fine-tune pre-trained models for specific tasks, thereby enhancing performance and efficiency.

Leveraging the pre-trained ResNet-50 model (He et al. (2016)), we fine-tuned it for our medical imaging tasks. ResNet-50, with its deep residual connections, effectively addressed the vanishing gradient problem and demonstrated exceptional performance, particularly on the COVID-19 dataset. The model's ability to fine-tune and adapt to specific tasks underscored the potential of transfer learning in medical diagnostics. The use of advanced optimizers like Adam and Nadam further enhanced the model's robustness and accuracy, making it a highly effective solution for complex and diverse medical imaging datasets.

Implications for Healthcare Sector

The findings from our project have significant implications for the future of healthcare diagnostics. By integrating advanced computer vision techniques, healthcare professionals can achieve higher diagnostic accuracy, reducing the likelihood of errors and improving patient outcomes. Automated diagnostic tools can handle large volumes of medical images, addressing resource constraints and enabling timely diagnostics, especially in underserved regions. The development of real-time image processing systems can revolutionize critical care scenarios, providing immediate insights and supporting rapid decision-making.

Future Work

Our project lays the groundwork for several future directions. Incorporating a wider variety of imaging modalities and pathological cases will test the scalability and generalization of our models. Exploring cutting-edge architectures like EfficientNet or Vision Transformers could further optimize performance and address the limitations observed in our current models. Developing explainable AI solutions will ensure transparency and trust in clinical applications, providing insights into model decisions and fostering better communication between doctors and patients. The potential integration of our models into futuristic medical diagnostic pods could provide automated, accessible, and comprehensive health assessments, transforming both urban and remote healthcare settings.

Team Contribution

We all worked on the dataset selection, pre-processing, and data augmentation, as well as researching for which computer vision techniques we wanted to implement (like SVM/CNN) and their current implementations by researchers. Each of us worked on one of the approaches / models that we discussed for our project. Tanuj worked on the SVM/KNN model, Aryansingh worked on the CNN Models, and Snigdha worked on the Resnet Model. At the end, we compared the performance and drew results.

References

Covid-19 Radiography dataset:

- M.E.H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M.A. Kadir, Z.B. Mahbub, K.R. Islam, M.S. Khan, A. Iqbal, N. Al-Emadi, M.B.I. Reaz, M. T. Islam, "Can AI help in screening Viral and COVID-19 pneumonia?" IEEE Access, Vol. 8, 2020, pp. 132665 - 132676. Paper link
- Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Kashem, S.B.A., Islam, M.T., Maadeed, S.A., Zughaier, S.M., Khan, M.S. and Chowdhury, M.E., 2020. Exploring the Effect of Image Enhancement Techniques on COVID-19 Detection using Chest X-ray Images. Paper Link

Medical MNIST dataset:

- Polanco, A. (2017). *Medical MNIST Classification* [GitHub repository]. Retrieved from https://github.com/apolanco3225/Medical-MNIST-Classification.

  License: Public Domain
- Splash Image Credit: Photo by Hush Naidoo on Unsplash.
- https://www.kaggle.com/datasets/andrewmvd/medical-mnist

ResNet50:

- He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385*. https://doi.org/10.48550/arXiv.1512.03385
- https://medium.com/@kenneth.ca95/a-guide-to-transfer-learning-with-keras-using-resnet50-a81a4a28084b
- *ResNet50 — Torchvision*. (2017). PyTorch. Retrieved December 12, 2024, from https://pytorch.org/vision/main/models/generated/torchvision.models.resnet50.html