WiDS'25 Project Report

AI Powered Anomaly Detection in Medical Imagery

Sayali Mahajan

24B3937

Fundamentals of Image Processing

Image processing is a method of performing certain operations on an image to enhance it or extract useful information. It can be divided into:

- Analog image processing – used for physical images such as photographs and printouts.
- Digital image processing – used for digital images and implemented using computers.

All digital image processing typically follows three stages:

1. Pre-processing – preparing the image for further processing.
2. Enhancement – improving image quality or emphasizing certain features.
3. Display or extraction of information – presenting or using the processed data

An image is represented as a two-dimensional function:
$f(x, y) = i(x, y) \times r(x, y)$
Where $i(x, y)$ = illumination (amount of light incident on the scene) and $r(x, y)$ = reflectance (amount of light reflected by objects).
This representation shows that what we see in an image depends on both lighting conditions and object properties.

There are several types of images:

1. Binary image
   It is the simplest type of image. It takes just two values Black and White or 0 and 1. Binary images consist of a 1-bit image, and only one binary digit represents a pixel.

2. Grayscale image
   Grayscale images are monochrome images, which means they have only one color. Grayscale images do not contain any color information. A standard grayscale image contains 8 bits/pixel data with 256 different grey levels. Each pixel determines available different grey levels.

3. Color images
   Color images are three-band monochrome images where each band contains a different color, and the information is stored in the digital image. The color images have gray-level data in each spectral band. Digital images are represented in red, green, and blue (RGB model). Each color image has 24 bits/pixel, which means 8 bits for each of the three-color bands (RGB).

Purposes and Applications of Image Processing

i) Visualization

Converting (rendering) image pixel/voxel into 2D/3D graphical representation. Most computers support 8-bit (256) grayscale display, sufficient for human vision to resolve 32-64 grayscale. Visualization aims to communicate data or information clearly and effectively to readers.

ii) Image restoration

The purpose of image restoration is to “compensate for” or “undo” defects that degrade an image. Degradation takes many forms, such as motion blur, noise, and camera misfocus. In cases like motion blur, it is possible to estimate the actual blurring function perfectly and “undo” the blur to restore the original image.

iii) Image retrieval

Browsing, searching, and retrieving images from an extensive database of digital images. Most traditional and standard image retrieval methods utilize metadata such as captioning, keywords, or descriptions of the images to retrieve the annotation words.

iv) Pattern recognition

Pattern recognition is classifying input data into objects, classes, or categories using computer algorithms based on key features or regularities. Pattern recognition has applications in computer vision, image segmentation, object detection, radar processing, speech recognition, and text classification.

v) Image Acquisition

In image processing, image acquisition retrieves an image from a source, usually hardware systems like cameras, sensors, etc. It’s the first and most crucial step in the workflow sequence because, without an image, the system makes no actual processing.

vi) Image enhancement

Improves the quality of an image by extracting hidden information from it for further processing.

vii) Image restoration

It is a basic problem in image processing, and it also provides a testbed for more general inverse problems. Image restoration is performed by reversing the process that blurs the image. Such is accomplished by imaging a point source and using the point source image called the Point Spread Function (PSF) to restore the image information lost to the blurring process.

viii) Morphological processing

It explains the shapes and structures of the objects in an image. In the morphological processing of images, pixels are added or removed. The design and shape of the objects are analyzed so that they can be identified. The basic operations in this processing are binary convolution and correlation based on logical operations rather than arithmetic operations.

ix) Image Segmentation

It is the process of dividing an image into multiple segments. Image segmentation is often used to locate objects and boundaries in images. The goal of segmentation is to simplify and change the representation of an image into something more meaningful and easier to analyze.

x) Object recognition

It is a computer-vision technique for identifying objects in images or videos. Object recognition is a crucial output of deep learning and machine learning algorithms. When humans look at a picture or watch a video, we can readily spot people, objects, scenes, and visual details. The objective is to teach a computer to do what comes naturally to humans: understand what an image contains.

Traditional image processing algorithms

i) Morphological Image processing

Morphological image processing removes noise and smooths binary images using non-linear operations based on image structure. It uses a small matrix called a structuring element (made of 0s and 1s) that scans the image to modify pixel regions. The two basic operations are dilation, which adds pixels to object boundaries, and erosion, which removes pixels from object boundaries. The number of pixels added or removed depends on the size and shape of the structuring element, which probes different regions of the image to analyze and modify its structure.

ii) Gaussian Image Processing

Gaussian blur, also known as Gaussian smoothing, reduces image noise and softens details by convolving the image with a Gaussian function, producing a translucent, blurred effect. It is widely used in computer vision for multi-scale image enhancement and as a data augmentation technique in deep learning. In practice, Gaussian blur is efficiently implemented using its separable property, where the image is blurred in two passes using a one-dimensional kernel first in one direction and then in the other making the process faster while achieving the same result.

iii) Fourier Transform in image processing

The Fourier Transform is a key tool in image processing that decomposes an image into its sine and cosine components, enabling applications such as image reconstruction, compression, and filtering. In digital image processing, the Discrete Fourier Transform (DFT) is commonly used. A sinusoidal component is described by its magnitude (contrast), spatial frequency (brightness or detail level), and phase (structural and positional information). In the frequency domain, the image is represented by these components, making it easier to analyze and manipulate specific image features. The principle for 2D discrete Fourier transform is:

$$F(u, v) = \frac{1}{MN} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y)\, e^{-j2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)}$$

iv) Wavelet Image Processing

A wavelet is a short, wave-like oscillation that starts at zero, rises, and returns to zero, making it localized in time unlike a sine wave that extends infinitely. This time localization allows the wavelet transform to capture both time and frequency information, which is especially useful in signal and image processing. Wavelets can be combined with signals using convolution to extract important features, making them effective for analyzing non-stationary or complex signals.

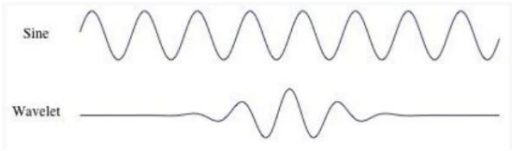
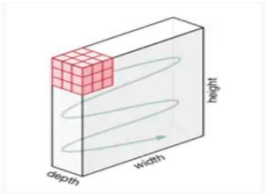


Image processing using Neural Networks

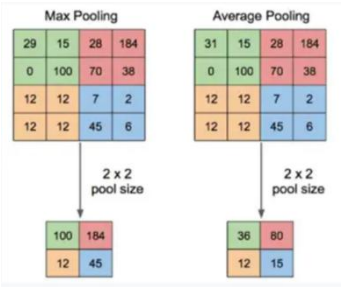
Neural Networks are multi-layered networks consisting of neurons or nodes. These neurons are the core processing units of the neural network. They are designed to act like human brains. They take in data, train themselves to recognize the patterns in the data and then predict the output. A basic neural network has three layers which are Input layer, Hidden layer and Output layer.

The convolutional neural network is based on three primary layers which are:



Movement of the kernel (Source)

1. Convolutional Layer (*CONV*): They are the core building block of CNN, responsible for convolution operations. The element involved in this layer's convolution operation is called the Kernel/Filter (matrix)**.** The kernel makes horizontal and vertical shifts based on the stride rate until the full image is traversed.
2. Pooling layer (POOL): The pooling layer gradually reduces the image's size, keeping only the most essential information. Its purpose is to progressively reduce the spatial dimension of the representation to reduce the number of parameters and computation in the network. There are two types of Pooling: Max Pooling and Average Pooling.



3. Fully Connected Layer (*FC*): The fully connected layer (FC) operates on a flattened input where each input is connected to all neurons. If present, FC layers are usually found towards the end of CNN architectures. CNN is used primarily to extract features from the image with the help of its layers. CNN is widely used in image classification, where each input image is passed through a series of layers to get a probabilistic value between 0 and 1.

Advanced Topics in Computer Vision and Image Processing

Computer vision operates at three main levels: low-level, mid-level, and high-level vision. Low-level vision focuses on basic pixel operations such as resizing, color adjustments, edge detection, oriented gradients, and color segmentation, with applications in photo editing, feature extraction, and zooming. Mid-level vision connects images to the real world or across time through techniques like panorama stitching, multi-view stereo, structured light scanning, LiDAR-based range finding, optical flow, and time-lapse imaging, supporting tasks such as 3D reconstruction, depth mapping, and motion analysis. High-level vision extracts semantic meaning from images, including image classification, object tagging, object detection, and semantic and instance segmentation, and is widely used in robotics, medical imaging, surveillance, and self-driving cars.

Human vision plays a crucial role in inspiring computer vision systems. Eyes evolved from simple light-sensitive spots into complex refractive systems, with human eyes balancing light gathering and visual acuity. The retina contains rods for low-light, peripheral vision and cones for color and fine detail, concentrated in the fovea, while a blind spot exists where the optic

nerve exits the eye. Small eye movements such as microsaccades, ocular drift, and micro tremors prevent sensory adaptation and enhance perception.

Visual information is processed in the brain through compressed signals sent from millions of photoreceptors to about one million ganglia. Two main pathways exist: the ventral ("what") pathway, responsible for object recognition and memory, and the dorsal ("where/how") pathway, responsible for motion, spatial awareness, and action. Different regions of the visual cortex, such as V1 and V2, perform specialized tasks like edge detection and shape or color processing.

Depth perception is achieved by combining multiple cues. Monocular cues include blur, motion parallax, shading, occlusion, and familiar object size, while binocular cues include image disparity between the two eyes and eye convergence. The brain integrates these cues to infer 3D structure from 2D images.

In computer vision, an image is defined as a projection of the 3D world onto a 2D plane, recording the amount of light at each point. This process is explained by the pinhole camera model, which uses a focal point and image plane to describe image formation. In real cameras, sensor grids capture light, and a Bayer filter pattern allows each sensor to record red, green, or blue light, with more green sensors used due to human visual sensitivity. Interpolation is then applied to reconstruct full-color images.

Digitally, a grayscale image is represented as a 2D matrix of pixel values, while a color image is a 3D tensor with dimensions width × height × channels. Pixel addressing uses coordinates (x, y, channel), and images are typically stored in linear memory using row-major order. Coordinate systems usually place the origin at the top-left corner, with the x-axis increasing to the right and the y-axis increasing downward, making consistency in conventions essential.

Color space transformations improve intuitive color manipulation. While RGB is suitable for display, it is not intuitive for editing, so HSV is used instead, separating hue, saturation, and value. This allows easy adjustment of brightness, color intensity, hue shifting, and data augmentation for machine learning.

Finally, image interpolation and resizing are essential when scaling or transforming images, as new pixel locations fall between existing ones. Common methods include nearest neighbor interpolation, which is fast but pixelated, and bilinear interpolation, which produces smoother results by averaging neighboring pixels**.**

CNN, Darknet and YOLO

Darknet is an open-source neural network framework primarily known for powering the YOLO object detection system. It's written in C and CUDA for high performance, supporting both CPU and GPU computation, which makes it fast for training and inference on deep learning tasks like image classification and real-time object detection. The framework is lightweight, easy to install via a simple Makefile (with GPU flags enabled), and excels in convolutional neural networks (CNNs). Darknet serves as the backbone for YOLO models (e.g., YOLOv3, YOLOv7), enabling single-stage detection that's both accurate and speedy.

ImageNet is a large-scale image database essential for advancing computer vision and deep learning research. It organizes over 14 million annotated images hierarchically using WordNet synsets, spanning roughly 22,000 categories from broad concepts like "animal" to specifics like "golden retriever." Key subsets include ILSVRC with 1.2 million training images across 1,000 classes, plus bounding boxes on over a million images for localization tasks. ImageNet enables transfer learning: pre-train on its scale, fine-tune for custom tasks, cutting training time and boosting performance on smaller datasets.

Common CNN architectures revolutionized image recognition by stacking convolutions, pooling, and nonlinearities to extract hierarchical features, with each design tackling depth, efficiency, or accuracy limits. AlexNet (2012) kickstarted deep learning with 8 layers, using large 11x11 and 5x5 filters early to capture broad patterns, ReLU activations to speed training, and dropout to prevent overfitting. Its dual-GPU design halved parameters, proving deep CNNs could crush ImageNet top-5 error from 25% to 15% via data augmentation and overlapping pooling. VGG Intuition VGG (2014) scales depth to 16-19 layers using tiny uniform 3x3 filters, relying on stacking to mimic larger receptive fields while exploding parameter count to 138M. The intuition: deeper uniform blocks build richer representations smoothly, ideal for transfer learning despite high compute. ResNet (2015) breaks 1000-layer barriers with skip connections (identity mappings) that add input to output, mitigating vanishing gradients and degradation in very deep nets. Core idea: residual learning ("learn differences") reuses features, enabling ResNet-152's top ImageNet accuracy by training what to change rather than full mappings. EfficientNet (2019) uses compound scaling to balance depth, width, and resolution via a fixed ratio (e.g., φ for scaling factor), starting from a tiny baseline optimized by neural architecture search. Intuition: uniform scaling respects compute budgets, yielding top accuracy with 10x fewer parameters than prior models by avoiding wasteful dimensions.

YOLO Modes

YOLO modes in Ultralytics YOLOv8/YOLOv26 streamline the ML workflow from training to deployment, with each handling a specific phase on datasets like COCO or custom ones.

Train Mode: Train mode fine-tunes a YOLO model on labeled data using specified hyperparameters, epochs, and image sizes (e.g., model.train(data='coco.yaml', epochs=100)), optimizing for tasks like detection by minimizing loss.

Val Mode: Val mode evaluates a trained model's performance on a validation set post-training, computing metrics like mAP50-95, precision, and recall to tune hyperparameters without overfitting**.**

Predict Mode: It runs inference on new images/videos using a trained model (e.g., model.predict(source='image.jpg', conf=0.25)), outputting detections, labels, and bounding boxes for real-world use.

Export Mode: It converts the trained PyTorch model to formats like ONNX, TensorRT, or CoreML for efficient deployment on edge devices, cloud, or mobile apps.

Track Mode: Track mode extends prediction with multi-object tracking across video frames using algorithms like BoT-SORT, assigning persistent IDs to objects for applications like surveillance

YOLO Tasks

YOLO supports versatile vision tasks beyond basic detection, leveraging unified architectures trained on ImageNet/COCO.

Detect: It draws bounding boxes and class probabilities around objects in one forward pass, ideal for real-time apps like autonomous driving with high speed/accuracy tradeoffs.

Segment: It adds pixel-level masks to detections for instance segmentation, delineating exact object shapes (e.g., polygons around people) for precise tasks like medical imaging.

Classify: Classify predicts the overall class of an input image (no localization), using a CNN backbone for tasks like scene recognition or product tagging**.**

Pose: It estimates keypoints (e.g., joints on humans) on detected objects, enabling skeleton-based analysis for activity recognition or sports analytics.

TILs and TIGER Grand Challenge

The TIGER (Tumor-Infiltrating lymphocytes in breast cancer) challenge is the first international initiative focused on the fully automated assessment of tumor-infiltrating lymphocytes (TILs) in H&E-stained breast cancer histopathology slides. It is organized by the Diagnostic Image Analysis Group (DIAG) of Radboud University Medical Center (Radboudumc), Netherlands, in collaboration with the International Immuno-Oncology Biomarker Working Group. The primary objective of TIGER is to evaluate and advance computer algorithms capable of automatically generating clinically meaningful TIL scores with strong prognostic value.

TILs are immune cells present within and around tumor tissue, and their presence has been shown to correlate with patient prognosis and treatment response, particularly in certain subtypes of breast cancer. By automating TIL assessment, TIGER aims to reduce subjectivity, improve reproducibility, and support clinical decision-making.

Breast Cancer

Breast cancer has become the most commonly diagnosed cancer worldwide, accounting for approximately 12% of all new cancer cases. Among women, it remains the leading cause of cancer-related mortality. However, breast cancer is not a single disease but rather a group of biologically distinct subtypes, each requiring different treatment strategies and having different prognoses.

Breast cancers are commonly classified into four molecular subtypes based on receptor status:

- **Luminal A**: Hormone receptor (HR) positive, HER2 negative
- **Luminal B**: HR positive, HER2 positive
- **HER2-enriched**: HR negative, HER2 positive
- **Triple Negative Breast Cancer (TNBC)**: HR negative, HER2 negative

These subtypes guide treatment decisions, such as the use of hormone therapy, targeted therapy, or chemotherapy. Among these, HER2-positive and TNBC subtypes are associated

with more aggressive disease and poorer outcomes, making them critical targets for research and biomarker development.

HER2-Positive and Triple Negative Breast Cancer

The TIGER challenge specifically focuses on HER2-positive and Triple Negative breast cancers because these subtypes have the worst prognosis and limited treatment options compared to hormone receptor-positive cancers. They are the subject of extensive research in immuno-oncology due to their responsiveness to immune-based therapies and the strong prognostic significance of immune cell infiltration.

In both HER2-positive and TNBC cases, TIL levels have been shown to predict response to chemotherapy and immunotherapy, as well as long-term survival. Therefore, developing reliable and automated methods to quantify TILs in these subtypes can significantly impact patient management and treatment personalization.

Tumour-Infiltrating Lymphocytes (TILs)

Cancer treatment is increasingly influenced by the biological characteristics of the tumour and the patient's immune response. The tumour microenvironment (TME), which includes immune cells, blood vessels, and surrounding tissue, plays a crucial role in tumour progression and treatment response.

Tumour-infiltrating lymphocytes (TILs) are immune cells that migrate into tumour tissue and participate in anti-tumour immune responses. High levels of TILs are associated with improved survival and better response to treatment, particularly in HER2-positive and TNBC patients. TILs are therefore considered a powerful biomarker with both prognostic (predicting outcome) and predictive (predicting treatment response) value.

Accurate measurement of TILs can help clinicians tailor treatment strategies, potentially reducing the need for aggressive therapies such as chemotherapy and improving overall patient outcomes. However, current assessment methods rely on visual scoring by pathologists, which can be subjective and time-consuming. This creates a strong need for automated, standardized, and reproducible assessment methods.

Recommendations from the International TIL Working Group

In 2015, the International TIL Working Group published standardized guidelines for assessing TILs in breast cancer histopathology slides. A seminal paper by Salgado et al. proposed a structured procedure for visually scoring TILs on H&E-stained tissue sections. This approach focuses on evaluating lymphocytes within the stromal region of invasive tumors while excluding areas such as necrosis or in-situ carcinoma.

Subsequent studies, including research by Denkert et al., demonstrated the strong prognostic and predictive value of visually assessed TILs in both surgical specimens and core needle biopsies. These findings reinforced the clinical importance of TILs and highlighted the need for consistent and reliable assessment methods.

The TIGER challenge builds upon these recommendations by seeking to automate the scoring process using artificial intelligence, while maintaining clinical relevance and alignment with established pathology guidelines.

Other Approaches to TIL Quantification

In addition to the recommendations of the TIL Working Group, other quantitative approaches to immune assessment have been developed. One notable example is the **Immunoscore**, proposed by Galon et al., which quantifies immune cell populations (such as CD3+ and CD8+ T-cells) in specific tumor regions, including the tumor center and invasive margin. Although originally developed using immunohistochemistry (IHC) for colorectal cancer, the Immunoscore concept highlights the importance of spatial distribution of immune cells.

Similar spatial and morphological analysis approaches have been explored in breast cancer, lung cancer, and pan-cancer studies, demonstrating that immune infiltration patterns are strongly associated with patient outcomes. These studies emphasize that not only the quantity but also the location and organization of immune cells within the tumor microenvironment are critical for prognostic assessment.

TIGER encourages participants to explore diverse computational strategies, including Immunoscore-like methods, spatial statistics, and morphological pattern analysis, to develop robust and clinically meaningful automated TIL scoring systems.

Goals of the TIGER Challenge

The TIGER challenge has two primary goals:

1. **Development of AI-based TIL Quantification Models**:The first goal is to enable the development of artificial intelligence models capable of automatically quantifying TILs in HER2-positive and TNBC breast cancer histopathology slides. This includes accurate detection of immune cells and segmentation of relevant tissue compartments, such as invasive tumor and tumor-associated stroma.
2. **Validation of Prognostic Value**: The second goal is to validate the clinical utility of AI-generated TIL scores using a large, independent test dataset. This dataset includes both routine clinical cases and samples from a phase III clinical trial, ensuring that the models are evaluated under realistic and clinically meaningful conditions.

By achieving these goals, TIGER aims to establish automated TIL scoring as a reliable biomarker that can be integrated into clinical workflows, supporting precision medicine and improving patient outcomes.

Tasks in the TIGER Challenge

1. **Detection of Lymphocytes and Plasma Cells**: These immune cells are the primary components of tumour-infiltrating lymphocytes. Accurate detection at the cellular level is essential for reliable quantification.
2. **Segmentation of Tumour and Tumour-Associated Stroma**: The algorithm must distinguish between invasive tumour tissue and surrounding stromal regions, as TIL scoring is performed specifically within relevant tissue compartments.

3. **Computation of an Automated TIL Score**: Using the outputs of cell detection and tissue segmentation, the algorithm must compute a single TIL score per slide that reflects immune infiltration in a clinically meaningful way.