### **Introduction/Background**

Advances in medical imaging technology have enabled machine learning applications in healthcare, particularly for detecting metastatic cancer in pathology slides. This project will utilize the **PatchCamelyon (PCam)** dataset that consists of small image patches from whole-slide images (WSIs). PCam offers a binary classification task aimed at detecting tumor tissue within a 32x32 pixel region.

**Dataset**: The PCam dataset contains 262,144 training samples, with 32,768 samples each for validation and testing. Each image is 96x96 pixels, and a label is applied based on the presence of tumor tissue in the central region. The main reason we are using this dataset is the probabilistic sampling as it ensures balance across the dataset. The dataset we used is available on Kaggle, and it’s ideal for ML models because of its non-duplicative nature. Its features include a label for the lymph node section being cancerous (1 or 0) and an image label column with file paths to the image.  
**Dataset Link**: [PCam Dataset.](https://github.com/basveeling/pcam?tab=readme-ov-file#details)

### **Problem Definition**

**Problem**: The problem is automating the identification of metastatic cancer in tissue samples. Detecting cancerous regions in pathology slides is a difficult task that currently requires a lot of time. As a result of the task being conducted by people, there is a risk of errors. These errors in manual interpretation can have grave consequences for patient outcomes.

**Motivation**: An automated system for cancer detection would reduce the time and effort required by pathologists while also improving diagnostic accuracy. The PCam dataset will allow us to use ML techniques in a way that can enhance real-world medical diagnostics. Additionally, it aligns with the trend of applying machine learning in healthcare to improve clinical patient care.

### **Methods**

#### **Data Preprocessing**

1. **Image Resizing**: All images will be resized to 96x96 pixels to ensure uniform input.
2. **Normalization**: Pixel values will be scaled to standardize the dataset across all color channels using normalization techniques common in ML (such as those used in ImageNet).
3. **Data Augmentation**: Random transformations like flips and rotations will be applied to artificially increase the diversity of the training set.

#### **Supervised Machine Learning Algorithms:**

1. **Convolutional Neural Networks (CNNs)**: CNN architectures, such as ResNet and VGG16, are suited for image classification and will perform well in detecting tumors in the PCam dataset (Esteva).
2. **Random Forest**: This traditional ML model will be used as a baseline to compare feature importance and model interpretability, though it’s likely going to be less effective than CNNs.
3. **Support Vector Machines (SVM)**: SVMs, known for their binary classification capabilities, will be employed for comparison. Though their performance will likely lag behind CNNs given the high-dimensional nature of image data (Esteva).

### **Results and Discussion**

**Metrics**: The performance of the models will be evaluated using the following metrics:

1. **Accuracy**: The percentage of correctly classified images.
2. **Precision and Recall**: These metrics are meant to handle imbalanced classes, focusing on how well the model identifies true positives and avoids false negatives.
3. **AUC-ROC**: This curve will measure the model’s ability to differentiate between positive and negative classes (Stokes).

**Project Goals**: The main goal of this project is to develop an efficient ML model that can accurately detect cancerous tissues. Ethical implications will be one of our main considerations, as this will be a determinant of if the model is deployable in clinical settings without introducing bias.

**Expected Results**: We expect that CNNs (Bejnordi), due to their effectiveness in image classification tasks, will outperform the more traditional methods like Random Forest and SVM. Additionally, data augmentation is likely to improve model generalization, leading to higher accuracy and better AUC-ROC scores.

### **References (IEEE format)**

1. H. C. Stokes, “Association of dietary cholesterol or egg consumption with incident cardiovascular disease and mortality,” *JAMA*, vol. 321, no. 11, pp. 1081-1095, Mar. 2019. (Online). Available:<https://jamanetwork.com/journals/jama/fullarticle/2665774>. (Accessed: Oct. 3, 2024).
2. A. Esteva, et al., “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, pp. 115-118, Feb. 2017. [Online]. Available:<https://pubmed.ncbi.nlm.nih.gov/28778026/>. [Accessed: Oct. 3, 2024].
3. B. E. H. Bejnordi, et al., “Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer,” *Nature Medicine*, vol. 24, pp. 1549-1558, Oct. 2018. [Online]. Available:<https://www.nature.com/articles/s41591-018-0316-z>. [Accessed: Oct. 3, 2024].