**Breast cancer detection from histopathology images through deep learning models**

**Research Questions:**

* What are the challenges in modifying CNN architectures so that they identify complex multi-dimensional patterns such as the boundaries between cancer and healthy tissue in breast histopathology images?
* What are the best settings of the hyperparameters, such as type of kernel, number of filters in that kernel, and learning rate, so that CNN models can detect the presence of breast cancer in histopathology images with the highest degree of accuracy?
* How does increasing the depth of neural networks for breast cancer detection in histopathology images impact up/down sampling?

**Objectives of the Project:**

* Examine the difficulties of adjusting CNNs to find intricate patterns in pictures of breast histology.
* For the best possible detection accuracy, adjust the CNN hyperparameters (learning rate, filters, and kernel type).
* Study the impact of CNN depth on upsampling/down sampling in breast cancer detection.
* Enhance CNN architectures to recognize tumour borders in pictures obtained from histology.
* Determine the best kernel types and number of filters to use while extracting features from breast cancer photos.

**Project Overview and Background:**

In this project, I aim to create a Convolutional Neural Network (CNN) diagnostic tool for detecting breast cancer from histopathology images. Histopathology images display tissue samples at a very high resolution so that cancerous and normal cells can be readily identified. However, the manual diagnosis of histopathology images requires too much time and is prone to human error (Litjens et al., 2017). In recent years, deep learning, especially CNNs, has given good performance and prediction power in automating cancer detection and improving precision (Spanhol et al., 2016). In this project, I hope to modify and optimize the architecture of CNNs so that the network can uncover the complex patterns related to the patterns between cancer and healthy tissue. By modifying the hyperparameters and anchor boxes, such as kernel types, filter size and learning rates, I intend to create an incredibly accurate and robust model to help pathologists analyze the patterns of breast cancer diagnosed from histopathology images.

References:

* Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B. and Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. *Medical image analysis*, 42, pp.60-88.
* Spanhol, F.A., Oliveira, L.S., Petitjean, C. and Heutte, L., 2016. Breast cancer histopathological image classification using Convolutional Neural Networks. *International Journal of Computer Assisted Radiology and Surgery*, 11(11), pp.1985-1993.

**Project Timeline and Tasks:**

A screenshot of a computer screen

Description automatically generated

**Project Timeline**

**Project Task list:**

* **Literature Review and Dataset Acquisition:** Examine CNNs in histology and breast cancer diagnosis in detail, and obtain the dataset required for the research.
* **Data Preprocessing and Augmentation:** To enhance the dataset, preprocess histopathological photos (e.g., stain normalization, downsizing), then use augmentation techniques.
* **CNN Model Design and Development:** Create and develop CNN architectures, adjusting the filters and kernel size hyperparameters.
* **Model Training and Optimization:** Utilizing the dataset, optimize learning rates, and test various architectures, train the CNN models.
* **Model Evaluation and Testing:** Analyze accuracy, precision, and recall as well as other performance metrics to assess the trained models using validation and test sets.
* **Final Report and Presentation:** Compile results, write the final project report, and prepare a presentation summarizing the project outcomes.

**Data Management Plan:**

**Dataset Overview:** The dataset commonly used for breast cancer detection in histopathology images is the BreakHis (Breast Cancer Histopathological Image Classification) dataset. It contains 7,909 microscopic images of breast tissue samples, divided into benign and malignant categories. These images were captured at different magnification levels (40x, 100x, 200x, and 400x), making them suitable for deep learning tasks, especially Convolutional Neural Networks (CNNs), due to their complexity and variability.

Eight classifications of tumors are included in the dataset: four classes (adenosis, fibroadenoma, phyllodes tumor, and tubular adenoma) represent benign instances, while four classes (ductal carcinoma, lobular carcinoma, mucinous carcinoma, and papillary carcinoma) indicate malignant cases. The high-resolution format of each image provides comprehensive details on the cellular structure of the tissue, an essential component for precise classification.

Even for seasoned pathologists, it can be challenging for machine learning algorithms to discriminate even minute variations in tissue appearance, as demonstrated by the BreakHis dataset. Since models have to learn information at many sizes, the variety in magnification levels also helps the models to generalize more effectively. In general, this dataset offers a strong foundation for CNN-based model development, training, and testing to enhance the identification and categorization of breast cancer from histopathological pictures, hence expanding the capabilities of automated diagnostic systems.

**Origin and format of dataset:** 7,909 high-resolution PNG photos of breast tissue histopathology were produced by the Pathological Anatomy and Cytopathology Laboratory in Brazil and are part of the BreakHis collection. With four different magnification levels (40x, 100x, 200x, and 400x), the dataset which has a total size of about 2.25 GB offers a variety of scales for deep learning applications. The dataset is publicly available on the Kaggle as <https://www.kaggle.com/datasets/ambarish/breakhis>.

**Project Documentation, Version Control, and Ethical Compliance:**

Code and documentation for the project will be hosted on GitHub, with file names clearly labelled with the version and content (e.g., Code\_v2.1.ipynb, Code\_v1.0.ipynb). Version control will be handled via Git, which has distinct branches for different features and fixes as well as clear commit statements. The project's installation, setup, and operation instructions, as well as information on inputs and outputs and contact details for support, are all included in a README file located in the repository. Code and data will be routinely transferred to GitHub for security and storage purposes, and weekly local backups will be kept on an external hard drive. All changes will be tracked in a changelog. The research will follow ethical guidelines, making sure no personal data is included under institutional ethics and Kaggle's data usage policy. It will do this by using the GDPR-compliant breast Cancer dataset from Kaggle.