# ****Histopathologic-Cancer-Detection****

***Breast cancer ranks second in mortality among women*** after lung cancer. According to clinical statistics, 1 out of every 8 women are diagnosed with breast cancer in their lifetime. However, periodic clinical examinations and self-tests help in its early detection, and thus significantly increase the chances of survival. Invasive detection methods cause the tumor to rupture, accelerating the spread of cancer to surrounding areas. Therefore, there is a need for a more reliable, fast, accurate and effective non-invasive cancer detection system (Selvathi, D & Aarthy Poornila, A. (2018). In-depth training methods for breast cancer detection using medical imaging analysis).

Early detection can give patients more treatment options. To detect signs of cancer, breast tissue biopsies are stained to improve the nuclei and cytoplasm for microscopic examination. Pathologists then evaluate the extent of any abnormal structural change to determine if there are tumors.

Architectural distortion is a very thin compaction of the breast tissue and may represent the earliest sign of cancer. Since radiologists are most likely not to notice this, over the years several approaches have been proposed for early detection, but none of them use deep learning methods.

In its "Renaissance Epoch" Deep Learning has become the leading machine learning branch that has been used in many areas of human life for more than 10 years, including health care, a key requirement of deep learning methods is the data that the Kaggle platform presents in this competition to solve histopathological diagnosis of cancer.

**Task**

***What is the challenge?***

**The task of classifying two-dimensional images** . It is necessary to determine the presence of metastases from digital histopathological images of 96x96 pixels. One of the key problems is that metastases can be as small as individual cells in a large area of ​​tissue.

**What is known about the data?**

**Histopathological images are microscopic images of vitreous lymph node slides stained with hematoxylin and eosin (H & E)**. This staining method is one of the most widely used in medical diagnostics, it stains in blue, violet and red. Dark blue - hematoxylin is associated with negatively charged substances, such as nucleic acids, Pink - eosin, with positively charged substances, such as amino acid side chains (most proteins). Typically, the nucleus is stained blue, while the cytoplasm and extracellular parts have different shades of pink.

**What data is provided to solve the problem?**

**220 thousand Training and 57 thousand. Evaluation images** . The data set is a subset of the[PCam](https://github.com/basveeling/pcam) data[set](https://github.com/basveeling/pcam) , and the only difference between the two is that all the duplicate images in the data provided to us were deleted. The PCam dataset is derived from the Camelyon16 Challenge dataset , which contains 400 H & E-stained full slides of sentinel slides of lymph nodes, which were acquired and digitized at 2 different centers using a 40x objective. The PCam dataset, including this one, uses a 10-fold downsampling to increase the field of view, which gives the resulting pixel resolution of 2.43 microns.

The sentinel lymph node is called the lymph node, into which the lymph outflow from the tumor is carried out first, therefore, it is primarily affected. In the literature, you can find two features that characterize SLN. The first is the lymph node nearest to the tumor, and the second is that the node is affected by metastases first.

According to the description of the [data](https://www.kaggle.com/c/histopathologic-cancer-detection) , ***there is a 50/50 balance*** between positive and negative examples ***in the training and test subset****. However, the distribution of education is 60/40 (negatives / positives). A positive label means that there is at least one pixel of tumor tissue in the central area (32 x 32 pixels) of the image.****Tumor tissue in a region distant from the center of the slide does not affect the label*** . This means that a negatively tagged image may contain metastases in the central region. Thus, by cutting off the images in the central region, information will not be lost.

**Description of the provided images** Format: TIF Resolution: 96x96 Channels: 3 Color depth: 8 Data type: Unsigned char Compression type: Jpeg

**How good is the data?**

This data set is a combination of two independent data sets collected at the University Medical Center Radboud (Nijmegen, the Netherlands) and the University Medical Center Utrecht (Utrecht, the Netherlands). images to identify metastases. However, with these samples of small images, some important information could be lost.

According to the description of the data, the data set is devoid of duplicates. However, this was not confirmed by testing.

For the entire data set, when the slide level label was unclear during the H & E dyed glass test, an additional WSI film layer with a sequential tissue cut immunohistochemically stained for cytokeratin was used to confirm the classification.

* <https://academic.oup.com/gigascience/article/7/6/giy065/5026175>

All the vitreous slides included in the CAMELYON dataset were part of the usual clinical diagnosis, which means that the data is of diagnostic quality. However, during the scanning process, scanning may fail or result in out-of-focus images. As a quality control measure, all slides were manually checked after scanning. The test was conducted by an experienced technician (QM and NS for UMCU, MH or R.vd.L. for other centers) to assess the quality of the scan.