EÖTVÖS LORÁND UNIVERSITY

FACULTY OF INFORMATICS

THESIS TOPIC DECLARATION

Student:

Name: **Marko Bauer** Code: **R7GNA0**

Type: **full-time student** Course: **Computer Science BSc**

Supervisor:

Name:

Affiliation with address: ELTE Faculty of Informatics
(department name goes here)
1117 Budapest, Pázmány Péter sétány 1/C.

Status and qualification:

Title of the Thesis work: Histopathologic Cancer Detection: Identifying metastatic tissue in histopathologic slides using Deep Neural Networks

Topic of the Thesis work:

Overview: Goal of this Thesis Work is creating an application capable of classifying histopathologic slides of lymph node sections and determining whether the patient has metastatic tissue (metastatic cancer) or not. Application would allow user to load a histopathologic slide, and as a result receive a category to which it belongs (whether it is malignant or benign). Classification would be accomplished using a class of deep neural networks called convolutional neural networks. Plan is to build multiple models trained on three different datasets, using transfer learning (VGG-19 Neural Network), as well as constructing models which are trained from scratch.

Application Development Semester Project: Thesis work is an extension of semester project from subject Application Development. That project is an application for landscape classification, trained on one small dataset (Intel Image Classification dataset), with single neural network and simple graphical user interface.

Datasets: Networks will be trained on three different datasets:

1. Breast Cancer Histopathological Dataset (BreakHis) – composed of 9,109 microscopic images of breast tumor tissue collected from 82 patients using different magnifying factors, containing 2,480 benign and 5,429 malignant samples.
2. PatchCamelyon (PCam) Dataset – consists of 327,680 color images extracted from histopathologic scans of lymph node sections, where each image is annotated with a binary label indicating presence of metastatic tissue.
3. NCT-CRC-HE-100K Dataset - contains 100,000 non-overlapping image patches from hematoxylin & eosin (H&E) stained histological images of human colorectal cancer (CRC) and normal tissue.

Plan: Thesis will be divided into four major parts:

1. Assembling datasets, data preprocessing (loading data, removing noise, normalization, whitening) and data augmentation
2. Building the networks (both the ones which have to be trained from scratch and the ones which use transfer learning), and training them on data
3. Improving neural network prediction accuracy (reducing overfitting with hyperparameter tuning, regularization and batch normalization)
4. Creating graphical user interface which allows user to load histopathologic slide and select network which has be applied on it, and to get as output category to which that slide belongs, along with image which superimposes heatmap on the original one

Budapest, 2019.11.25.