ABSTRACT

DEEP LEARNING FOR DEVELOPING CLASSIFICATION PIPELINE TO DETECT METASTATIC BREAST CANCER FROM HISTOLOGICAL WHOLE SLIDE IMAGES (WSI)

ARJUN PUNABHAI VEKARIYA, M.S.

The University of Texas at Arlington, 2017

Supervising Professor: Dr. Junzhou Huang

Cancer, the second most deadliest diseases on the planet is a generalized term for the class of diseases caused by proliferation of abnormal cells in a human body. These abnormal cells are caused due to unwanted growth of new cells and improper recycling process of old or damaged cells. They also have tendency to damage cells in its surrounding areas as well as in the areas far away from them, by spreading (metastasize) through different parts of the body. Breast cancer starts developing in breast cells. Metastatic presence in lymph nodes is one of the most important prognostic variables of breast cancer. Methods available today in medical industry are very time consuming as pathologist has to manually analyze sentinel lymph nodes, which requires him to scan entire whole slide image for detecting metastasis region. Moreover, in some cases it's very difficult to detect these metastasis as sometimes they are only visible under high resolution and remains invisible from visual cortex. Developing computer aided methods for analyzing whole slide images has remained a great interest of computer scientist for decades, but historical approaches to histological image

analysis in digital pathology have focused primarily on low level image analysis tasks (e.g., color normalization, nuclear segmentation, and feature extraction). These classical methods have not been proven useful for practical use in clinical practice as they require several manual parameters to be set manually for accurate results thus proves burdensome for pathologists. Also these techniques can't be generalized for every whole slide image as whole slide images prepared by different clinical laboratories happen to contain variety of staining like Haemotoxylin, Eosin and others.

In this thesis, a deep learning-based classification pipeline for detection of cancer metastases from whole slide images of breast sentinel lymph nodes is proposed and analyzed. The classification pipeline consists of five different stages: 1. Image processing for background subtraction. 2. Tiling. 3. Deep ConvNet for tile based classification. 4. Building Tumor probability heat-maps. 5. Heat-map post-processing for slide based classification. GoogLeNet, a deep 27 layer ConvNet is used to distinguish tumor positive areas from negative ones into digital whole slide images. The main challenge is to discriminate between hard negative areas from positives ones as tiles from hard negative areas mimic tiles from positive areas which results in too many false positives. Ensemble learning method using two Deep ConvNet model is developed to eliminate these false positives. Proposed system achieves an area under the receiver operating curve (AUC) of 0.91 which is quite close to score obtained by human pathologist while reviewing same images. Results in this thesis indicate that proposed platform can automatically scan any WSI for detecting metastatic regions. Moreover, following Deep Learning based approach, it can also be generalized for different types of whole slide images with minimal efforts.