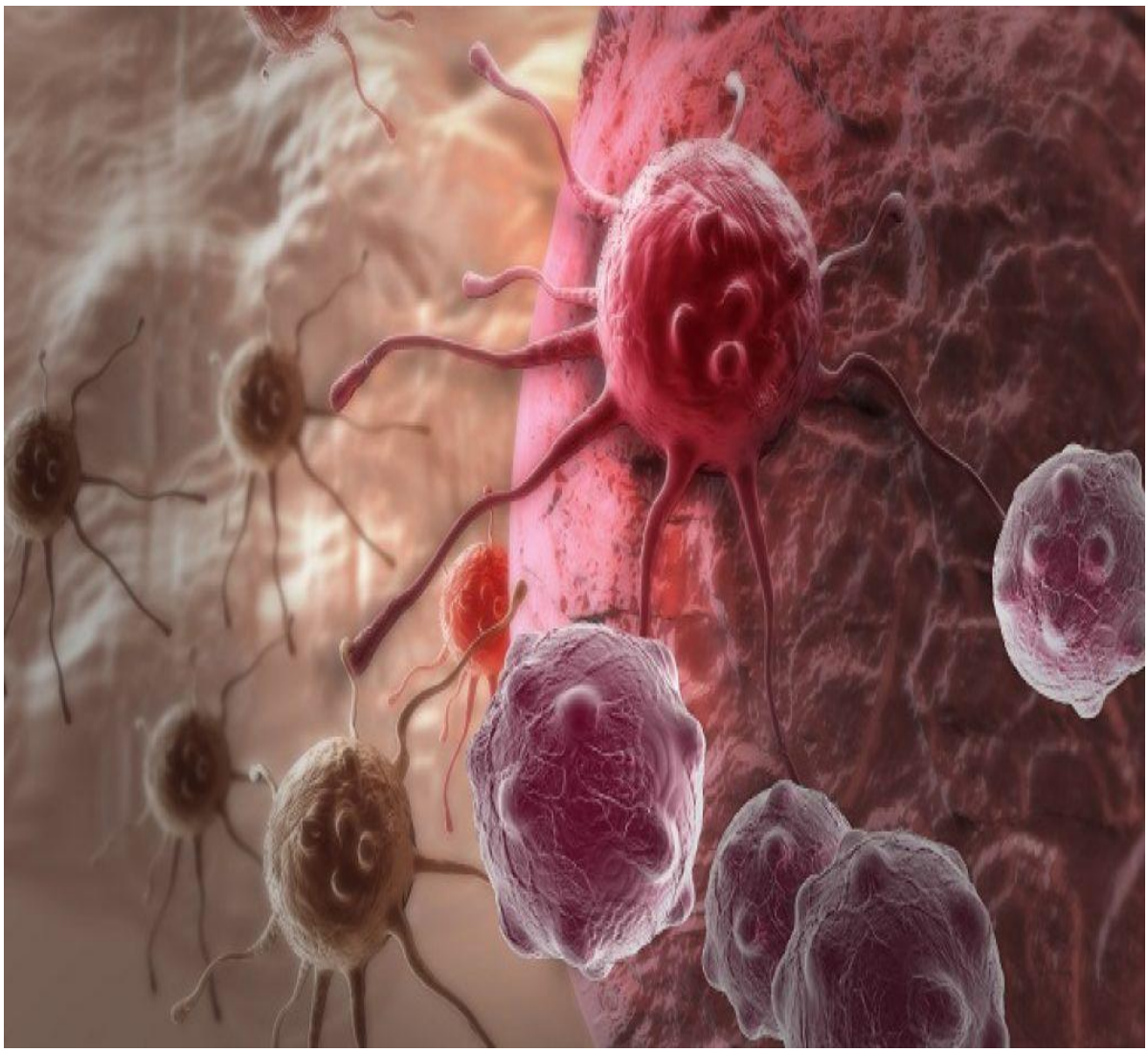


Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks

AUTHOR(S): ANGEL CRUZ-ROA; AJAY BASAVANHALLY; FABIO GONZÁLEZ; HANNAH GILMORE; MICHAEL FELDMAN; SHRIDAR GANESAN; NATALIE SHIH; JOHN TOMASZEWSKI; ANANT MADABHUSHI

Introduction

Intrusive ductal carcinoma (IDC) is the most widely recognized phenotypic subtype of all Breast malignancies (BCa) involving almost 80% of them. This is routinely checked by pathologists through visual examination of tissue slides recolored with hematoxylin and eosin (H&E). Evaluation of sickness (for example tumor reviewing) is normally restricted to areas containing intrusive cancer.2 Hence, the initial phase in the histopathological portrayal of resected breast tissue is to recognize tissue locales comparing to obtrusive tumor and non-intrusive or sound tissues

Locating intrusive breast disease is a tedious and challenging task fundamentally in perspective of time. It includes a pathologist filtering huge patches of considerate regions to at last recognize the malign parts. An aggregate of 51 chart-based highlights from Voronoi graphs, least spreading over tree, and Delaunay triangulation. The atomic centroids were utilized to recognize parts of malignant growth from generous territories on entire slide pictures, yielding a general exactness of 80% related to a help vector machine classifier.

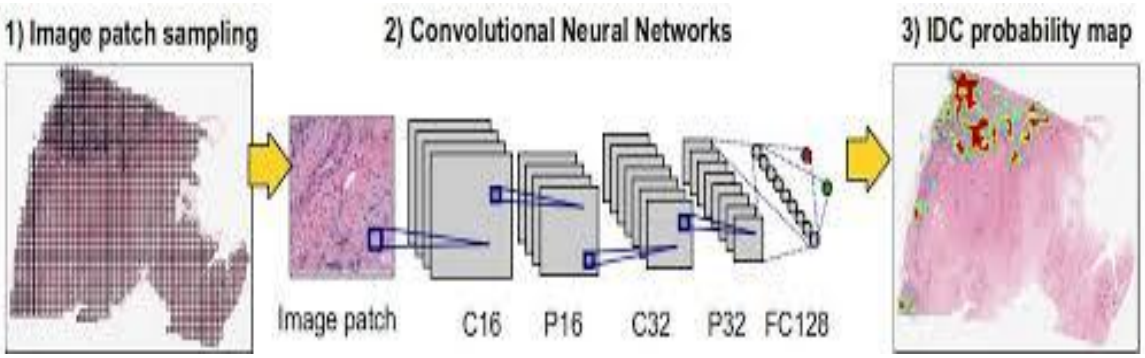
Previous researcher tried using some handcrafted techniques to do some preprocessing on images and later advancing to segmentation and detection. This paper uses the deep learning and CNN in comparison to old techniques.

This Deep learning model contained a visual interpretable technique to feature applicable harmful places on histopathology pictures, for example, a computerized recoloring. Different sorts of profound learning models incorporate convolutional neural systems (CNN), which are a group of multi-layer neural systems especially intended for use on two-dimensional information, for example, pictures and videos.12 CNN are feedforward neural system whose designs are customized for limiting the affectability to translations, rotation and noise in the picture

The dataset selected comprises 162 cases of patients diagnosed with IDC of BCa which 113 slides were selected for training and 49 slides for standalone testing. We compared our approach against a traditional classification approach comprising handcrafted descriptors, global features (color and texture) and histopathology image features (nuclear textural and architectural). Our goal was to compare handcrafted feature approaches against a feature learning approach of deep learning.

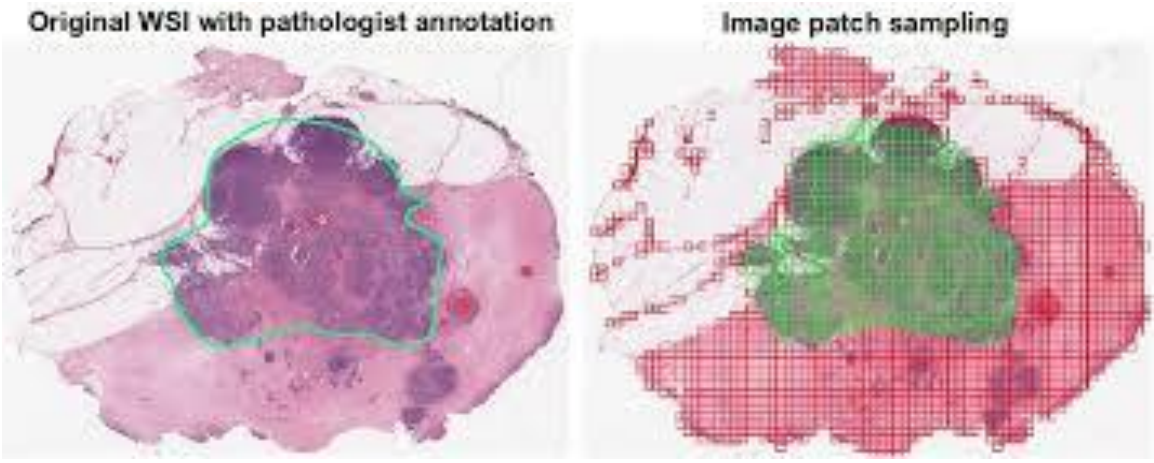
Methodology

The framework described in Figure 1 comprises the following steps: 1) grid sampling of image patches is performed over all tissue-containing regions in the WSI; 2) a Convolutional Neural Network is trained from sampled patches to predict the probability of a patch belonging to IDC tissue; and 3) finally a probability map built over the WSI highlighting the predicted IDC regions.



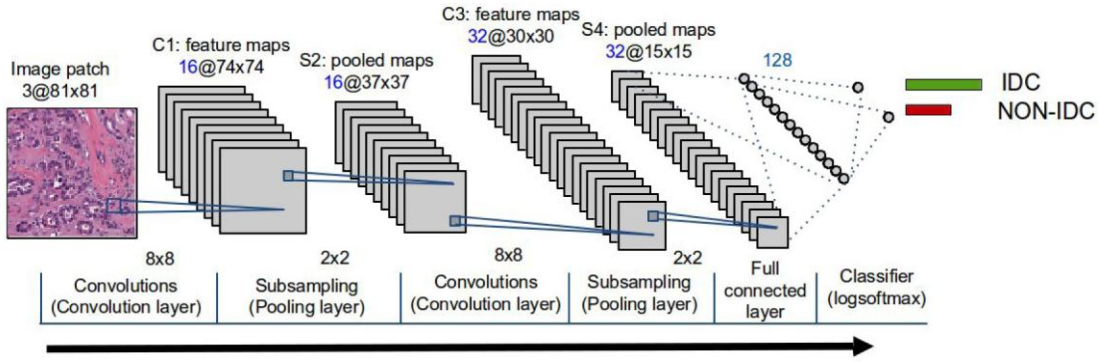
Patch Sampling

Each WSI is split into non-overlapping image patches of 100×100 pixels via grid sampling. Patches with mostly fatty tissue or slide background are discarded. For training purposes, regions containing IDC are manually annotated by a pathologist and used to generate a binary annotation mask. An image patch is considered to be a positive sample if at least 80% of the patch falls within the annotation mask, otherwise it used as a negative sample. Figure 2 illustrates an example of image patch sampling involving the WSI and pathologists' annotations.



Convolutional Neural Network (CNN)

CNN is applied as local feature detectors or filters over the whole image to measure the correlation between individual image patches and signature patterns within the training set. Then, an aggregation or pooling function is applied to reduce the dimensions of the feature space. The image patches collected in Step 1 are then used as inputs to a 3-layer CNN architecture (in Figure below) in which two layers are used for convolution and pooling while the remaining layer is fully-connected.



The detailed steps are:

1. *Patch preprocessing*: Each image patch is converted from RGB to YUV color space and normalized to a mean of zero and variance of one. This minimizes the correlations of raw pixels, allowing to focus on properties that are not dependent on covariances, such as sparseness. The motivation to do that are to attenuate differences between input features and fasten gradient-based learning.

2. *Convolutional layer*: The convolution process is a way to represent larger images (i.e. more than 64 × 64 pixels) since a set of learned features working as feature detectors. To do that we apply a 2D convolution of the input feature maps (i.e. image channels for the first layer) with a squared convolution kernel (i.e. filters or features), is defined as

$$Yj = \tanh(\sum kij * xi),$$

where xi corresponds to the i th input feature map kij is the convolution kernel, and yj corresponds to j th output feature map. The feature map looks like a salient map where learned features was detected. The tanh(·) function is used to rectify non-linearities in the convolution output, a common problem in the modeling of biological processes. A contrast normalization step is applied to each yj independently to help reduce overfitting and generalize overall performance.

3. *Pooling layer*: This stage allows to decrease huge dimension of image representation through a subsampling mechanism which support local space invariances. Then, this layer applies downscaling by (2x2) to allow for optimal learning of invariant features in each window.33

4. *Fully-connected layer*: This layer is typically applied in the top layer of a CNN architecture in order to capture complex relationships between high-level features. In this stage spatial information is ignored to learn correlation between different locations. Thus, the output of pooling layer is the input to a fully connected layer which mixes them into a feature vector. This is like a well-known neural network.

5. *Classification layer*: This final layer is a fully-connected layer with one neuron per each of the two classes (invasive or not) activated by a logistic regression model (i.e. SoftMax classifier). The whole CNN model is trained using Stochastic Gradient Descent to minimize the following loss function:

$L(x) = -\log [e^{xi} \div \sum e^{xj}]$.

where xi corresponds to outputs of full-connected layer multiplied by logistic model parameters. Thus, the outputs of CNN are the log likelihoods of class membership

Step 3.

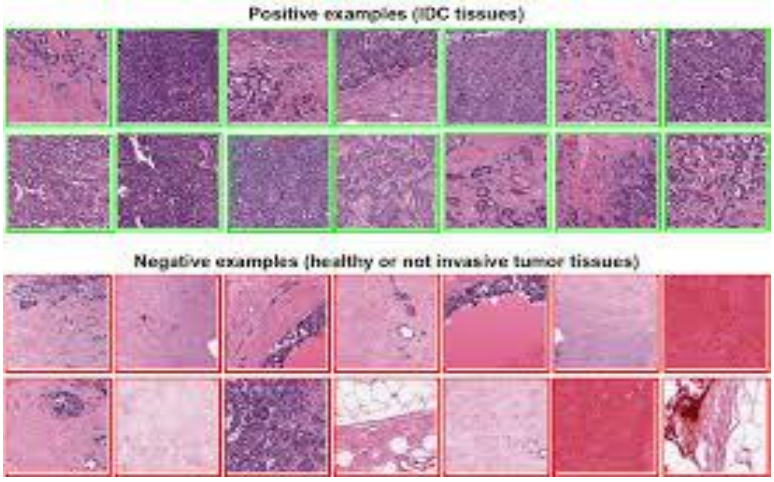
IDC probability map: The exponential function is applied to each output value obtained from positive class neuron of classified patches by CNN to get values between 0 and 1, so that they could be interpreted as probabilities.

Description of dataset

The data cohort comprises digitized BCa histopathology slides from 162 women diagnosed with IDC at the Hospital of the University of Pennsylvania and The Cancer Institute of New Jersey. All slides were digitized via a whole-slide scanner at 40x magnification (0.25 µm/pixel resolution). Operating on entire whole-slide histopathology images is intractable due to their extremely large size (on the order of 1010 pixels). In this work, each WSI was sampled (by a factor of 16:1) to a resolution of 4 µm/pixel.

Image patch-based dataset construction

The original data cohort of 162 slides was randomly split into three different subsets comprising 84 training (D1) and 29 validation (D2) cases for parameter exploration, and 49 test cases for final evaluation (D3). The patch-based datasets comprise 782 and 196 instances for training (D1) and instances for testing (D2). Examples of positive (IDC) and negative (non-IDC) tissue regions from the training and test sets are shown in figure below:



State of the art Handcrafted Features

state-of-the-art handcrafted features used in this work to compare against the deep learn

Handcrafted feature	Visual property captured
Gray Histogram (GH)[34]	Luminance
Fuzzy Color Histogram (FCH)[35]	Color
HSV Color Histogram (HSVCH)[36]	Color
RGB Histogram (RGBH)[34, 37]	Color
JPEG Coefficient Histogram (JPEGCH)[36]	Color and Texture
Local Binary Partition Histogram (LBP)[38]	Texture
MPEG7 Edge Histogram (M7Edge)[39]	Texture
Haralick features[9]	Nuclear textural (NT)
Graph-based features[9]	Nuclear Architectural (NA)

Experiment

In order to evaluate our CNN based system, we compare it against a suite of state-of-the-art handcrafted features (global features or histopathology features). The two different approaches are described below.

CNN based framework:

System adapts a 3-layers CNN architecture employing 32, 64 and 128 neurons, for the first and second convolutional-pooling layers and the fully-connected layer respectively. For all experiments, a fixed convolutional kernel of size 3×3 and pool kernel of size 2×2 were used.

Handcrafted features:

Variety of handcrafted feature sets were selected in order to evaluate different visual properties of histopathology images such as staining, tissue appearance, morphological and architectural arrangement of cells by using color, texture and graph-based descriptors. The first set comprises of global features used in computer vision to represent color and texture information of whole images, having been previously successfully used in content-based image retrieval of natural35–37 and histopathology images.34 The second set comprises features used to describe nuclear arrangement and morphology via using Haralick and graph-based descriptors. Table above details the complete set of handcrafted features, global features and histopathology image features, with the corresponding visual property captured by each one of them.

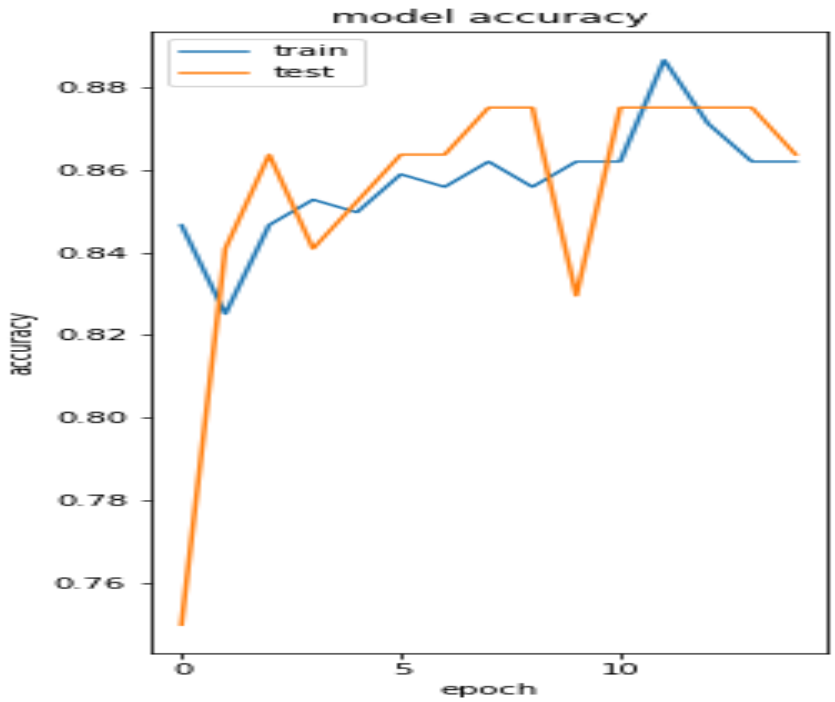
Results:

antitative performance for classification of IDC and healthy tissues. The performance measures showed 'r), Recall (Rc) or Sensitivity (Sen), Specificity (Spc), F-measure (F1) and Balanced Accuracy (BAC).

	Pr	Rc/Sen	Spc	F1	BAC
CNN	0.6540	0.7960	0.8886	**0.7180**	**0.8423**
FCH	0.7086	0.6450	0.9298	0.6753	0.7874
RGBH	0.7564	0.5956	0.9493	0.6664	0.7724
GH	0.7102	0.5240	0.9434	0.6031	0.7337
JPEGCH	0.7570	0.4646	0.9605	0.5758	0.7126
M7Edge	0.7360	0.4372	0.9585	0.5485	0.6979
NT	0.6246	0.2851	0.9547	0.3915	0.6199
LBP	0.7575	0.2291	0.9806	0.3518	0.6048
NA	0.6184	0.2413	0.9606	0.3472	0.6009
HSVCH	0.7662	0.2223	0.9821	0.3446	0.6022

Evaluation

Our completely learn-from-data approach for mechanized IDC tissue locale location was quantitatively assessed on D3, yielding a classification execution of 71.80% (F-measure) and 84.23% (BAC). This execution is point by point in Push 1 of Table over counting exactness, review or affectability and specificity with 65.40%, 79.60% and 88.86% separately. The primary try addresses the address of how comparative the ground truth manual comment is compared to the robotized estimation of IDC. CNN yields the most excellent by and large execution in terms of both F-measure and BAC (71.80%, 84.23%). By comparison, the most excellent handcrafted highlights are Fluffy Color Histogram (67.53%, 78.74%) and RGB Histogram (66.64%, 77.24%). In any case the approach yields moved forward F-measure and BAC comes about (by 4% and 6%, separately) over the finest handcrafted features.



Code Setup

The code is setup in COLAB, framework is keras and I have uploaded one folder of data with name 8863. The folder is number with 0 and 1 in it. 0 for non-cancer image patches and 1 for patches with cancer. If you open the code in colab then you need to follow the steps below:

1) Upload the data folder 8863 in google drive.
2) After mount drive command, in the next command paste this cd /content/gdrive/My Drive/8863.
3) Rest of the commands will be executed in sequential manner.