Analyzing histopathological images by using machine learning techniques

R Madana Mohana
Chaitanya Bharathi Institute of Technology

Applied Nanoscience
https://doi.org/10.1007/s13204-021-02217-4

ORIGINAL ARTICLE

Analyzing histopathological images by using machine learning techniques

Darshana A. Naik[1] · R. Madana Mohana[2] · Gandikota Ramu[3] · Y. Sri Lalitha[4] · M. SureshKumar[5] · K. V. Raghavender[6]

Received: 25 September 2021 / Accepted: 26 October 2021

Abstract
Medical image data have become an important part of every patient's digital health record. With the advancement of microscope technology, pathologists can now handle histopathological tissue slides more quickly with digitized WSI. Manual evaluations of massive histological images are time taking and sometimes error-prone, particularly for pathologists with diverse degrees of skill. Patient can be harmed by a delayed or erroneous analysis. Our research work combines image processing techniques (grayscale, edge-detection) plus supervised machine learning algorithms such as RF, SVM, and KNN for analyzing histopathological images (HI) and finds the optimal algorithm to classify breast cancer. Breast cancer is the major malignant common cancer in women after lung cancer; it is the 2nd biggest cause of death from cancer of women. RF algorithm achieved 98.2 and 98.3% accuracy for Benign, and Malignant cancer compared with other algorithms to classify breast cancer on WSI dataset.

Keywords Machine learning · Histopathological images · Issues · Breast cancer · Classification · RF, SVM, and KNN

Introduction

Medical image data have become an important part of every patient's digital health record. The pathologist examines color specimen on a slide glass under the microscope to a pathology diagnosis. With the advancement of microscope technology, pathologists can now handle histopathological tissue slides more quickly with digitized WSI (Whole Slide Imaging).

Incorporating histopathological imagery into cancer research has improved tailored treatments as well as survival estimations and also provides a fuller insight into cancerous cells phenotypic patterning and genetic mechanisms. Histopathological imaging consists of phenotypic data about tumor morphology, plus, high throughput genomic data has revealed cancer's molecular profiles. As a medical golden primary interface for the prognosis and diagnosis of most tumors, histopathology helps doctors to make precise treatment decisions. Manual evaluations of massive histological images are time taking and sometimes error-prone, particularly for pathologists with diverse degrees of skill. As the volume of WSI grows, efforts were made to evaluate them using deep learning, machine learning, and image processing techniques to aid tasks like diagnosis.

Pathology digitization is an example of the latest advances which generate large amounts of visual data for computer-aided diagnosis. With the help of personal computer technologies, we can observe and analyze diseased cell and tissue samples in high-resolution images. Also, it allows for the application of image processing methods. Such methods, like grading and hosting, would aid pathologists and validate the explanations.

Patient can be harmed by a delayed or erroneous analysis. As a result, having autonomous, accurate, and efficient

✉ R. Madana Mohana
rmmnaidu@gmail.com

[1] Department of CSE, Ramaiah Institute of Technology, Bangalore, India

[2] Department of CSE, Bharat Institute of Engineering and Technology, Ibrahimpatnam, Hyderabad, Telangana, India

[3] Department of CSE, Institute of Aeronautical Engineering, Dundigal, Hyderabad, Telangana, India

[4] Department of IT, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana, India

[5] Department of IT, Sri SaiRam Engineering College, Chennai, Tamil Nadu, India

[6] Department of CSE, G. Narayanamma Institute of Technology & Science, Hyderabad, Telangana, India

Published online: 07 February 2022

machine learning (ML) techniques do medical imaging (MI) evaluations will be advantageous. As the MI data are quite structured and categorized, MI analyses have become a major research area in ML.

For HI analysis, effective ML algorithms are employed to aid pathologists to obtain rapid, reliable, and validated assessment results for the effective diagnosis. Pathologists are using a variety of standard and deep learning methods to gain access to large volumes of tissues to identify the disease class in the images. Furthermore, as ML algorithms were typically semi- or fully automated, these seem efficient, implying technical viability to histopathological analysis in the new big data era (Li et al. 2020). ML algorithms for feature extraction would disclose details and correlations that often go undetected by the naked human eye.

Our research work combines image processing techniques plus supervised machine learning algorithms such as RF, SVM, and KNN for analyzing histopathological images (HI) and finds the optimal algorithm to identify disease early. To do so, we did analysis on breast cancer HI images. Breast cancer is the major malignant common cancer in women after lung cancer; it is the 2nd biggest cause of death from cancer of women. Breast cancer is caused by several variables, including reproductive factors, radiation therapy, hormones, genetics, obesity, and hormones. Annually, 1 million women were diagnosed with breast cancer, as per the World Health Organization report, 50% of them will die, as the cancer is detected at late (Guyon et al. 2002). Malignant tumors spread to nearby cells, that can progress into metastasizing/spreads to different parts of body, but benign masses do not spread to other tissue where as it will expand inside the benign masses (Gokhale 2009; Tang et al. 2009).

Flow of paper is as follows. Literature review was briefly covered in Section "Literature review", and issues in analyzing histopathological images were briefly discussed in Section "Specific issues in analyzing histopathological images". Machine learning techniques were addressed in Section "Machine learning algorithms". Methodology is explained in Section "Methodology", and the findings are discussed in Section "Results". Section "Conclusion" details the research conclusion.

Literature review

Literature review examines at a few research studies which utilized machine learning techniques and neural networks, on Medical images, to identify diseases.

The concerns in breast cancer detection, prognosis risk analysis of metastasis and recrudescence were analyzed by Osareh and Shadgar (2010) authors utilizing three known classifiers: PNN, KNN, and SVM. Based on the gene microarrays dataset, and FNAB dataset, the classifiers were integrated with principal component feature extraction analysis, sequential forward selection-based feature extraction, and signal–noise ratio feature ranking. SVM classifier achieved 98.80% and 96.33% accuracy in the diagnosis of breast cancer, compared to other classifiers.

On Wisconsin Breast Cancer datasets, the Amrane et al. (2018) authors used the Naive Bayes (NB) and the k-nearest neighbour (KNN) classifier for breast cancer classification. After a thorough analysis, it was shown that KNN has a greater efficiency of 97.51%, while NB has 96.19%. However, whereas if the dataset grows larger, KNN's processing time is increased.In the WEKA tool, the authors (Bayrak et al. 2019) used ANN and SVM for classification of the WBC dataset. According to the performance metrics, SVM have achieved, 96.9% accuracy to predict and diagnosis of breast cancer.

The authors in Chi et al. (2007) published a paper on breast cancer survival analyses on 2 breast cancer datasets using ANN and compared with nuclear morphometric features such as nuclear size, nuclear shape, texture and so on. ANNs give more flexibility in predicting survival time compared with conventional approaches since they can readily analyze variable correlations and develop a non-linear predictive model. Their findings reveal that ANNs can accurately estimate the probability of recurrence and distinguish between patients with a favourable and bad prognosis.

The authors of Bourdès et al. (2010), Natarajan et al. (2020a) presented a paper in which they compared ANN to logistic regression. By utilizing AUROC (Area Under Receiver Operating Characteristics: ROC analysis is a helpful method for assessing diagnostic test performance and, more broadly, examining the accuracy of the statistical model), authors compared multilayer perceptron NNs with SLR (Standard Logistic Regression) to find key variables effecting on Disease Recurrence, cancer causes and DFS (Disease-Free Survival) in breast cancer patients.

In Adam and Omar (2006), researchers combined a genetic algorithm with Back propagation NN to build a computerized breast cancer diagnostic that reduces diagnosis time while enhancing accuracy in categorizing breast masses into benign or malignant. On the dataset, two alternative cleaning methods were performed. Set A was taught to just remove records with missing data, whereas Set B were trained to use a conventional statistical clearing approach to find all noisy/missing data. Set A achieved 100% accuracy rate, whereas Set B had an accuracy percentage of 83.36%. As a result, the result concludes that medical data should be retained in its original form because it provides a higher accuracy percentage than modified data.

A study on early diagnosis of breast cancer using the SVM classifier was used in Rejani and ThamaraiSelvi (2009), Ramesh (2020), Swaroopa et al. (2018), Natarajan et al. (2020b). The authors of these papers highlight how mammography (an X-ray image of a breast is called a mammogram. Mammograms are used by doctors to check for early symptoms of breast cancer) can be used to detect tumors. And developed a method for tumor detection which contains mammogram images that have been filtered by a Gaussian filter.

The authors of Fogel et al. (1995), Sreedhar et al. (2020) explored the evolution of neural networks for identifying breast cancer and similar works which use back propagation technique with multilayer perceptron for breast cancer detection. Despite back propagation, authors discovered that evolutionary of computational methods and algorithms outperformed more traditional optimization techniques.

The authors from Zhang et al. (2013) presented a cascade approach with the rejection choice. The dataset with 361 samples was used to evaluate this technique. The accuracy was reported to be around 97%. Color texture characteristics and various classifiers like nearest neighbor classifier, and Decision trees have been suggested for efficient classification of breast cancer. With regard to different classifiers, this technique used ensemble voting, and achieved 87.53% for patient-level recognition (Gupta and Bhavsar 2017).

The authors in Spanhol et al. (2015) showed 85.1% accuracy on the collection of breast cancer histopathology pictures using PFTAS features and SVM for patient analysis. For nucleus classification, a variety of approaches like Gaussian mixture methods, NNs, K-means and fuzzy C-means clustering were used on the dataset of 500 real-world medical images collected in 50 patients. The accuracy of binary classification test (benign versus malignant tumors) was found to be in 96–100% range (Kowal et al. 2013), implying that all these machine learning-based techniques allowed for relatively accurate and objective analysis, and were deemed beneficial in breast cancer diagnosis.

Specific issues in analyzing histopathological images

This section specifies the problems faced while analyzing histopathological images is as follows:

Very large image

Images with larger size usually need to be reduced to the smaller size that allows for significant differentiation, as the size of input image increases, the number of parameters to be calculated, the necessary computing power, and memory requirements will be increased. WSI, on the other hand, has numerous cells and the image might comprise billions and billions of pixels, making it difficult to evaluate as it is. Resizing whole image into smaller size, would result in information loss at the cellular level, results significant reduction in identification accuracy.

Insufficient labeled images

The lack of training data is arguably the most significant challenge in pathological image analysis using machine learning.

Efficient labeling

Reducing pathologists' working time to designate ROIs in WSI is one approach to enhance training data. Active-learning is the supervised-learning approach that automatically selects the most useful unlabeled sample and displays it for pathologists to identify. Because this method is expected to improve discrimination performance with fewer labelled pictures, the total labelling time required to achieve same discrimination efficiency will be reduced.

Various magnification levels correspond to different levels of information

Tissues were typically made up with cells, and various tissues' have varied biological characteristics. High-powered field microscopic picture captures information on cell shape effectively, while structure information like glandular structure made up of numerous cells, is best caught in a lower-powered field microscopic image. Because malignant tissues exhibit both structural and cellular, pictures acquired at various magnifications will include useful information. Pathologists diagnose disorders by obtaining various types of information, ranging from tissue to cellular level, using varying microscopic magnifications. Because handling pictures at their highest resolution is challenging, they are frequently resized to match to different magnification and then utilised as input while analyzing. In terms of diagnosis, the most useful magnification is debatable, however inputting both maximum and minimum magnification photos at the same time can occasionally increase accuracy, depending on the sorts of disorders and tissues.

Artifacts and color variation

Multiple steps are involved in the creation of WSIs: pathologic specimens were cut and placed on slide-glass, tinted with eosin, hematoxylin, and scanned. Unwanted effects that are irrelevant to underlying biological variables might be created at each phase. As instance, tissue slices may be twisted and wrinkled when they put on slide, dust can contaminate slides while scanning; blur due to various thicknesses of tissue sections; and tissue areas might well be tagged with colour markers. Specific methods to identify artifacts-like tissue folds (Kothari et al. 2013), and blur (Wu et al. 2015) had been developed, since these artifacts might have a negative impact on interpretation.

Color variation is another major artifact. Different batches, scanner models, staining circumstances, tissue slice thickness, staining chemical manufacturers are all sources of variance. The efficiency of a machine learning system might be harmed if colour variation is not taken into account. The impact of colour variation on classification results can be lower, if enough data on every stained tissue obtained by each scanner are integrated; however, this appears unrealistic at the present. Conversion to grey scale, colour normalization (Bejnordi et al. 2015; Ciompi et al. 2017; Khan et al. 2014), and colour augmentation (Lafarge et al. 2017) have all been offered as solutions to this problem. The simplest method is to convert to grayscale.

Machine learning algorithms

Unsupervised-learning and supervised-learning are two types of machine learning algorithms commonly employed in digital pathology picture analysis. The purpose of supervised-learning, using train data is to assume a function that maps input images with its relevant label (e.g., breast cancer). Unsupervised-learning, aims to assume a function to unlabeled images, which can reveal hidden patterns. Three supervised-learning classifiers will be discussed in this section.

In two steps, supervised-learning classification methods that maps data to specified class labels:

1. For specific labelled data, a classification model is constructed during the training phase.

2. For data classification, trained classification model built in phase1 will be used.

Several measures like as accuracy, recall, precision, and f-measure are used to assess trained model's performance by employing test data. SVM, KNN, and RF are some of the supervised classification and regression techniques that were employed while analyzing. The Support Vector Machine is the linear model which is used to solve regression and classification problems. By creating hyper-planes, the SVM algorithm divides data into classes. SVM finds the best hyper-plane in high-dimensional space that maximize the margin space between data points and classes.

The nearest-neighbor principle is used in the KNN classification procedure. The classifier trains patterns that are used to classify test patterns based on their similarity to the training patterns. KNN classifier generates the class membership values for each item, either votes the most frequently used label (in the case of classification) or averages the labels (in regression case). The Random Forest constructs decision trees on data samples and later uses majority voting to determine the optimal solution. The overfitting problem is reduced by averaging the results.

Methodology

In Fig. 1, a suitable machine-learning architecture for classifying breast cancer is shown. The components of the proposed technique include data collection, image pre-processing, image segmentation, feature extraction, classification, and performance measure. Each component is described in detail further below.

Data collection

The Wisconsin Breast Cancer (WBC) dataset is obtained and analyzed in this study. There are 699 samples in the dataset that have been categorized as benign or malignant refer Fig. 2 and sample images is shown in Fig. 3. In addition, the dataset contains 11 integer-valued attributes refer Table 1.

Data (Image) pre-processing and segmentation

Image preprocessing is necessary for successful results in subsequent phases. To obtain reliable findings, all of the photos are converted to grayscale images and noisy data is deleted. The image segmentation (IS) algorithm divides the image into multiple objects and extracts useful data for the model's other stages by examining the images. IS were carried out in two ways: by searching for common patterns and anomalies. In addition, the photos are divided into categories depending on pre-determined criteria. As an outcome, image segmentation determines the gradient of an image's intensity to each pixel using a label edge-detection approach. In the case of discontinuities, pictures were segregated based on rapid changes in edge detection intensity values.

Feature extraction and classification

In the feature extraction stage, the object features from input images are extracted. Then it divides them into nine categories:

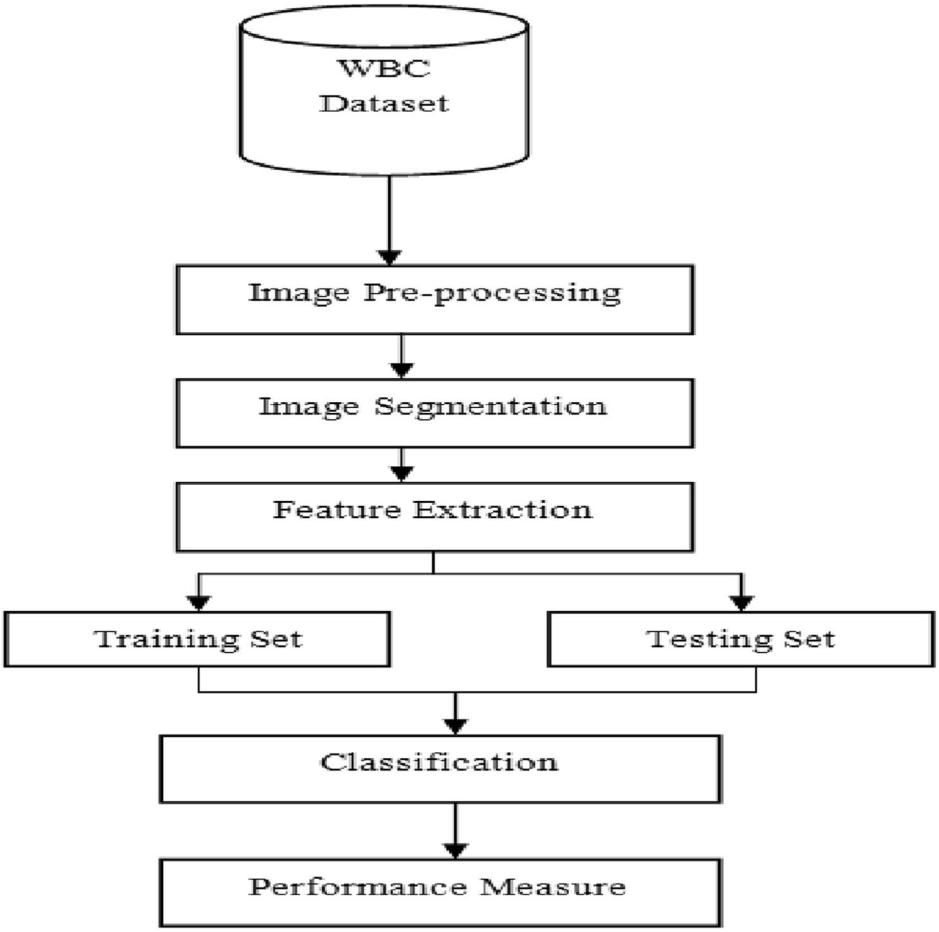
1. Based on clump thickness feature, thickness of the tumour is determined.

2. Uniformity of cell-size feature assess cancer cell size compared to other cells.

3. As cancerous cells differ in shapes, measure the uniformity of cell shapes and identify margin variance by uniformity of cell shape feature.

4. Cancerous cells migrate throughout the organs and attach to healthy cells determined by marginal adhesion feature.

Fig. 1 Breast cancer classification system



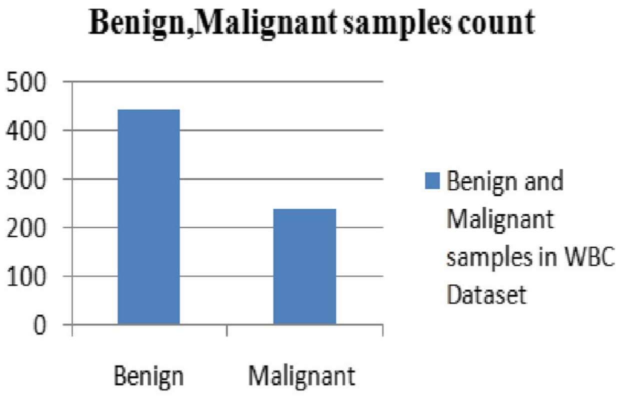


Fig. 2 Total number of samples for Benign and Malignant classes in WBC Dataset

5. Single epithelial cell size feature specifies uniformity measure; larger epithelial cells are a symptom of malignancy.

6. For benign tumour, is not surrounded by cytoplasm, it is specified by Bare Nuclei feature.

7. Describes the nucleus texture, which is homogenous in benign cells. In malignancies, the chromatin is coarser determined by Bland Chromatin feature.

8. The nucleolus is normally undetectable and very tiny in normal cells, obtained by Normal Nucleoli feature.

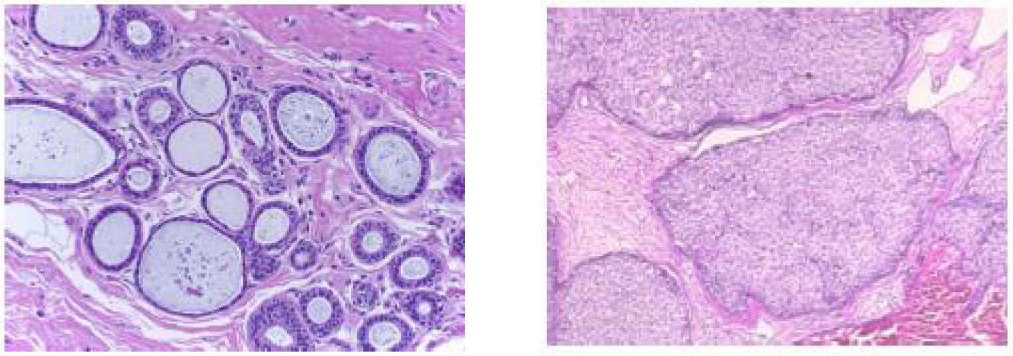
9. Calculate the number of mitoses that have occurred. Higher the value, the higher the risk of malignancy determined by Mitoses feature (Tarca 2007).

The grey scale pixel values (255: white, 0: black), were used in the experiment for further analysis. Finally, to classify breast cancer, machine learning approaches are used.

Results

This section examines the performance of several classification methods such as RF, KNN, SVM on the WBC dataset. The WBC dataset is divided into training data (80%) and testing data (20%).

Table 2 shows class labels along with the number of samples in each class is divided as train and test samples. To train and test model more quickly, all the images are resized uniformed size and maintained a balanced dataset. Table 3



(a) Benign

(b) Malignant

Fig. 3 Sample benign and malignant HI images

Table 1 Integer-valued attributes present in WBC dataset

Number	Attribute name
1	Sample code number
2	Clump thickness (1–10)
3	Cell size uniformity (1–10)
4	Cell shape uniformity (1–10)
5	Marginal adhesion (1–10)
6	Single epithelial cell size (1–10)
7	Bare nuclei (1–10)
8	Bland chromatin (1–10)
9	Normal nuclei (1–10)
10	Mitoses (1–10)
11	Class (2: Benign, 4: Malignant)

Table 2 Class labels and number of samples used to train and test the proposed system to classify breast cancer

Class	No. of train samples	No. of test samples
Benign	141	95
Malignant	141	95
Total	282	190

Table 3 Comparison of classification accuracy between several machine learning classifiers

Class	Accuracy percentage		
	SVM	KNN	RF
Benign	94	97.9	98.2
Malignant	93	97.6	98.3
Total	93.50	97.75	98.25

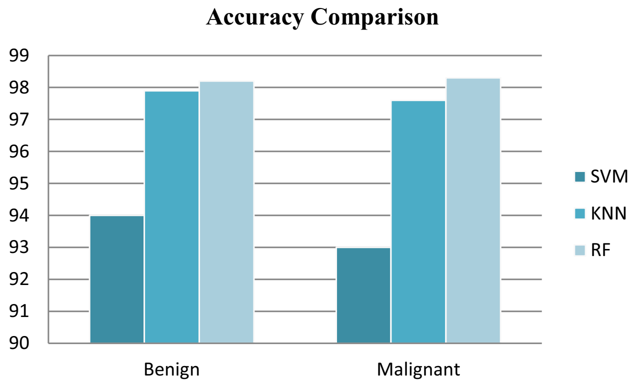


Fig. 4 SVM, KNN, RF algorithms achieved accuracy on breast cancer histopathological images

and Fig. 4 shows the classification accuracy of the examined machine learning algorithms, and it is clear that the RF method outperformed the others. Where KNN also achieved closure accuracy compared to RF but when the image volumes increases, takes a lot of time to process the images. So RF is the best optimal algorithm to classify breast cancer histopathological images.

Conclusion

The aim of this research work is to compare and identify the best supervised machine learning classification method for classifying breast cancer histopathology pictures among RF, KNN and SVM. To train and test the model more quickly, all the images are resized uniformed size and maintained a balanced dataset. RF algorithm achieved 98.2 and 98.3% accuracy for Benign, and malignant cancer compared with other algorithms to classify breast cancer on WSI dataset. In future, we wanted to train and test our proposed model to classify other cancers like lung cancer, brain tumor and so on.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Adam A, Omar K (2006) Computerized breast cancer diagnosis with Genetic Algorithm and Neural Network. In: Proc. of the 3rd International Conference on artificial intelligence and engineering technology (ICAIET). 2006

Amrane, M, et al. "Breast cancer classification using machine learning. In: 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT). IEEE, 2018.

Bayrak EA, Kırcı P, Ensari T (2019) Comparison of machine learning methods for breast cancer diagnosis. In: 2019 Scientific meeting on electrical-electronics & biomedical engineering and computer science (EBBT). IEEE, pp 1-3

Bejnordi BE et al (2015) Stain specific standardization of whole-slide histopathological images. IEEE Trans Med Imaging 35(2):404–415

Bourdès V, Bonnevay S, Lisboa P, Defrance R, Pérol D, Chabaud S, Bachelot T, Gargi T, Négrier S (2010) Comparison of artificial neural network with logistic regression as classification models for variable selection for prediction of breast cancer patient outcomes. Adv Artif Neural Syst 2010:309841. https://doi.org/10.1155/2010/309841

Chi CL, Street WN, Wolberg WH (2007) Application of artificial neural network-based survival analysis on two breast cancer datasets. In: AMIA annual symposium proceedings, vol 2007. American Medical Informatics Association, p 130

Ciompi, F, et al (2017) The importance of stain normalization in colorectal tissue classification with convolutional networks. In: 2017 IEEE 14th International Symposium on biomedical imaging (ISBI 2017). IEEE, 2017

Fogel DB, Wasson EC III, Boughton EM (1995) Evolving neural networks for detecting breast cancer. Cancer Lett 96(1):49–53

Gokhale S (2009) Ultrasound characterization of breast masses. The Indian J Radiol Imaging 19(3):242

Gupta V, Bhavsar G (2017) Breast cancer histopathological image classification: is magnification important? In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 17–24

Guyon I et al (2002) Gene selection for cancer classification using support vector machines. Mach Learn 46(1):389–422

Khan AM et al (2014) A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. IEEE Trans Biomed Eng 61(6):1729–1738

Kothari S, Phan JH, Wang MD (2013) Eliminating tissue-fold artifacts in histopathological whole-slide images for improved image-based prediction of cancer grade. J Pathol Inf 4:1–22

Kowal M et al (2013) Computer-aided diagnosis of breast cancer based on fine needle biopsy microscopic images. Comput Biol Med 43(10):1563–1572

Lafarge MW, Pluim JPW, Eppenhof KAJ, Moeskops P, Veta M (2017) Domain-adversarial neural networks to address the appearance variability of histopathology images. Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer, Cham, pp 83–91

Li C et al (2020) A review for cervical histopathology image analysis using machine vision approaches. Artif Intell Rev 53(7):4821–4862

Natarajan VA, Kumar MS, Patan R, Kallam S, Mohamed MYN (2020a) Segmentation of nuclei in histopathology images using fully convolutional deep neural architecture. In: 2020 International Conference on computing and information technology (ICCIT-1441), pp 1–7, IEEE

Natarajan VA, Babitha M, Kumar MS (2020b) Detection of disease in tomato plant using deep learning techniques. Int J Mod Agric 9(4):525–540

Osareh A, Shadgar B (2010) Machine learning techniques to diagnose breast cancer. In: 2010 5th international symposium on health informatics and bioinformatics. IEEE, pp 114–120

Ramesh G (2020) Detection of Plant diseases by analyzing the texture of leaf using ANN classifier. Int J Adv Sci Technol 29(8s):1656–1664

Rejani Y, ThamaraiSelvi S (2009) Early detection of breast cancer using SVM classifier technique. arXiv preprint arXiv: http://arxiv.org/abs/0912.2314

Spanhol FA et al (2015) A dataset for breast cancer histopathological image classification. IEEE Trans Biomed Eng 63(7):1455–1462

Sreedhar B, ManjunathSwamy BE, Kumar MS (2020) A comparative study of melanoma skin cancer detection in traditional and current image processing techniques. In: 2020 Fourth international conference on I-SMAC (IoT in Social, mobile, analytics and cloud) (I-SMAC). IEEE, pp 654–658

Swaroopa K, Rodda S, Chilukuri S (2018) Differentiated caching for improved QoS in vehicular content-centric networks. Int J Comput Sci Eng 6(10):317–322

Tang J et al (2009) Computer-aided detection and diagnosis of breast cancer with mammography: recent advances. IEEE Trans Inf Technol Biomed 13(2):236–251

Tarca AL et al (2007) Machine learning and its applications to biology. PLoS Comput Biol 3(6):e116

Wu H, et al (2015) Detection of blur artifacts in histopathological whole-slide images of endomyocardial biopsies. In: 2015 37th annual International Conference of the IEEE Engineering in Medicine and biology society (EMBC). IEEE, 2015

Zhang Y et al (2013) Breast cancer diagnosis from biopsy images with highly reliable random subspace classifier ensembles. Mach vis Appl 24(7):1405–1420

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.