

Breast Cancer Dataset, Classification and Detection Using

Deep Learning

Abstract:

Incorporating scientific research into clinical practice via clinical informatics, which includes

genomics, proteomics, bioinformatics, and biostatistics, improves patients' treatment. Computational

pathology is a growing subspecialty with the potential to integrate whole slide images, multi-omics

data, and health informatics. Pathology and laboratory medicine are critical to diagnosing cancer.

This work will review existing computational and digital pathology methods for breast cancer

diagnosis with a special focus on deep learning. The paper starts by reviewing public datasets related

to breast cancer diagnosis. Additionally, existing deep learning methods for breast cancer diagnosis

are reviewed. The publicly available code repositories are introduced as well. The paper is closed by

highlighting challenges and future works for deep learning-based diagnosis.

Introduction :

Computational pathology (CP) has the potential to improve clinical workflow efficiency and diagnostic quality thanks to information integration and advanced digital communication networks . CP is accompanied by several challenges, such as efficient data fusion, limited processing capabilities, and compliance with ethical practices .

Over 2 million women were examined for breast cancer in 2018, among whom approximately

0.6 million died worldwide. Most intrusive breast cancer diseases are chemical

receptor-positive . Chemical therapies targeting the trauma center flagging pathway

often help patients with chemical receptor-positive tumors . After delicately segmenting

a patient's example onto magnifying instrument slides for staining, a pathologist draws a

visual conclusion based on hematoxylin and eosin (H&E) staining, and subatomic markerexplicit

stains are used for confirmation and subtyping. Trauma centers are identified using

atomic ImmunoHistoChemistry (IHC). However, IHC staining is both time-consuming and

expensive . Moreover, test quality can vary significantly due to differences in tissue, the skill level of the expert taking the tissue sample, and specialist ability levels . Finally, pathologists' decisions are prone to error. These factors contribute to misdiagnosis. About 20% of current IHC-based trauma center and PR test results are incorrect , putting patients at risk of receiving subpar treatment. Recent research has shown that emergency room tests can be resolved using morphological stains. However, these studies rely on single-focus tissue microarray datasets (TMAs).

This review examines the application of deep learning (DL) in understanding breast cancer images. We start by pointing out the significance of imaging in nervous system science and its clinical advantages. The review is continued by discussing DL advancements in breast cancer diagnosis.

Digital Pathology and Deep Learning :

Pathology is represented by a variety of terms, including “computerized pathology”, “AI”, and “computational pathology”. With the advancement of fluorescent slide scanners, entire glass slides can be virtualized and digitized . The data from the slides can be saved in cloud storage, allowing pathologists to analyze the data with ease and the benefit of assistance from AI-based diagnosis tools . To this end, researchers have already developed a variety of AI methods for medical diagnosis .

Breast cancer is the most widely recognized malignant growth in women, accounting for nearly half of cancer cases diagnosed in women . HR-positive and lymph node-negative infections also account for nearly half of all cases . Following widespread

clinical approval, multigene tests such as the OncotypeDX 21-gene test, PAM50, and Mamma

Print are used to examine patients and guide ACTx in HR-positive and lymph node-negative breast cancer . The clinical benefit of the 21-gene test is debatable in patients

with HR-positive, lymph node-negative, and early-stage breast cancer [23,24]. Furthermore,

the fragility of RNA extracted from formalin-fixed paraffin-embedded (FFPE) tissue may jeopardize

its precision and prevent proper interpretation of recurrence score (RS) results .

As a result, a simpler and more effective strategy for determining the risk of repetition based

on super-durable tissue is required. Considering that the RS from the 21-gene test is not entirely determined by the expression qualities of the gene set (MKI67, STK15, BIRC5, CCNB1,

and MYBL2) and that the mitotic count is linked to the RS7, a careful and extensive evaluation of mitosis and other cell-cell collaborations includes the RS7. Recently, the Lunit Extension has been demonstrated to predict mitosis accurately in every cell in breast malignant growth , as well as recognized cancer cells and other cells in a microenvironment.

Breast carcinoma is the most common malignant growth in women worldwide, and it encompasses a wide range of diseases with varying histological, prognostic, and clinical outcomes . Metastatic infections, such as liver and cellular breakdowns in the lungs,

affect a majority of patients with malignant bosom growth

Automated Breast Cancer Diagnosis :

In this section, the search strategy for gathering existing papers related to breast cancer

diagnosis is explained. To conduct our search, an AND/OR combination of multiple

keywords was used: (breast cancer diagnosis OR malignant growth OR tumor) AND (deep learning OR machine learning). A total of 514 papers were gathered. Inclusion/exclusion

of the gathered papers was performed based on authors' voting. Papers with at least

three votes were considered for inclusion in this survey. The number of selected papers

categorized by their publishers were 10, 15, 28, and 19, corresponding to Elsevier, Springer,

IEEE, and other publishers. These statistics correspond to the first blue row of Figure 1.

We repeated our search among the references of the selected papers. Among the selected

papers, 9, 9, 16, and 13 belonged to Elsevier, Springer, IEEE, and other publishers, which

have been added to the statistics in the first blue row of Figure 1 to yield the values in the

second blue row of the same.

DL Application in Breast Cancer Diagnosis :

AI has recently demonstrated promising results in terms of precision and accuracy

for the automated diagnosis of diseases such as breast cancer . Among AI methods,

DL stands out for processing high-dimensional data such as medical images .

An extensive search has been conducted to gather articles related to breast cancer diagnosis.

The majority of these articles were gathered from the Nature database, bosom malignant growth. Significant effort has been put into covering recently published articles, especially the ones with publicly available source codes. The remainder of this section is devoted to the overview of the investigated papers.

Wang et al. (the winning team in the CAMELYON16 challenge) created various models using 256 256-pixel patches from positive and negative areas of whole slide images of bosom sentinel lymph hubs . Pathologists reported that having a profound learning framework as an assistant decreases the human error rate by 85% . Other studies reported that estrogen receptor status (trauma centers) is a fundamental atomic marker used to diagnose and select treatment options .

During clinical administration, pathologists examine biopsied tissue for the designated receptor with immunohistochemistry (IHC) to detect cell surface antigens .

Due to the importance of tissue analysis, attempts have been made to automate it using DL. For example, two deep neural networks (DNNs) were attached end-to-end for local and global feature extraction from microscopy images [. The first network acts as an autoencoder for efficient dimensionality reduction, and the second network takes the job of

classification. .

Determining the factor with a high impact on cancer patients' survival is vital for slowing down the cancer progression and increasing the life expectancy of the patients.

To this end, Cho et al. [49] investigated the correlation between HE-stained tissue slides and

adjuvant chemotherapy benefits for cancer patients. A CNN was trained on 1343 patients to identify histological parameters based on HE-stained whole slide images.

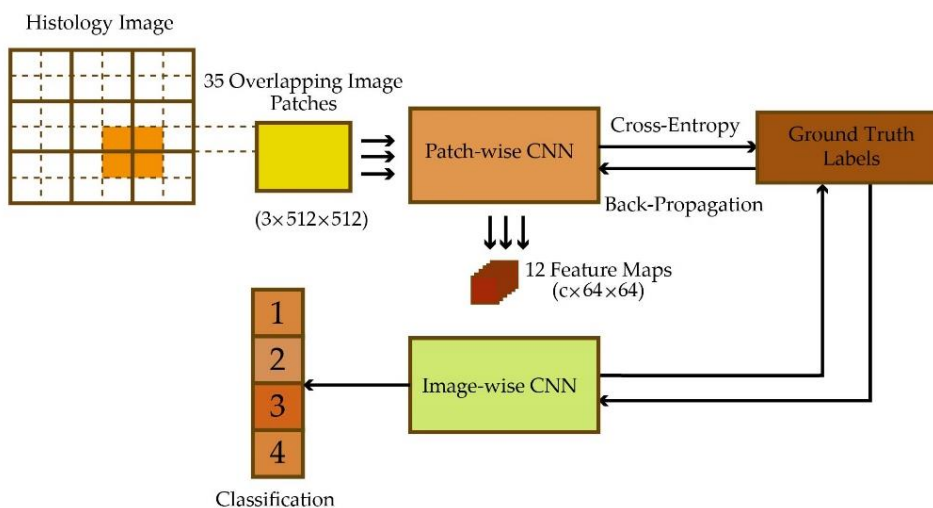
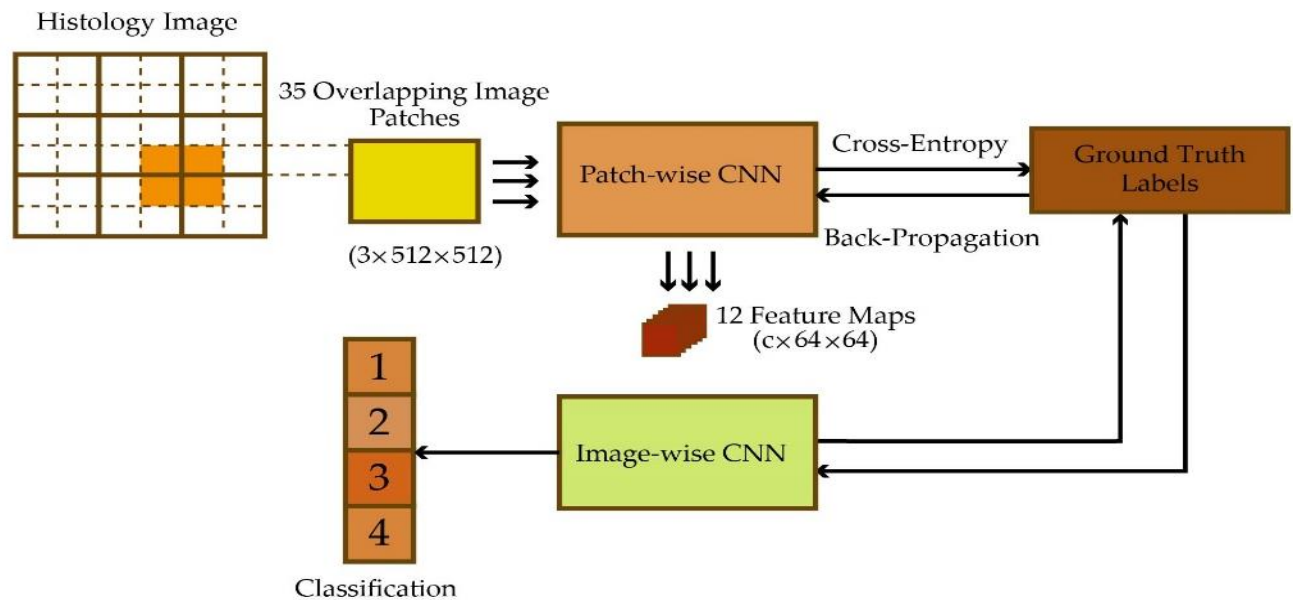


Figure 1. Chemotherapy in hormone receptor-positive breast cancer patients

Another examination approach is a mammogram, which is an X-ray picture of the breast. This approach is even useful for regular examinations of women with no signs of breast cancer. This is particularly important for early diagnosis and taking preventive actions to reduce the potential threat of breast cancer. To this end, Shen et al. [50] utilized DL to diagnose breast cancer based on mammograms. To reduce the cost of preparing a sufficient amount of training data, two sets of training data with different annotations were considered. A limited set of samples with lesion-level annotation was used in the first phase of training. In the second phase, only samples with image-level annotation were used. The cost of image-level annotation is much less than lesion-level annotation, which is appealing.

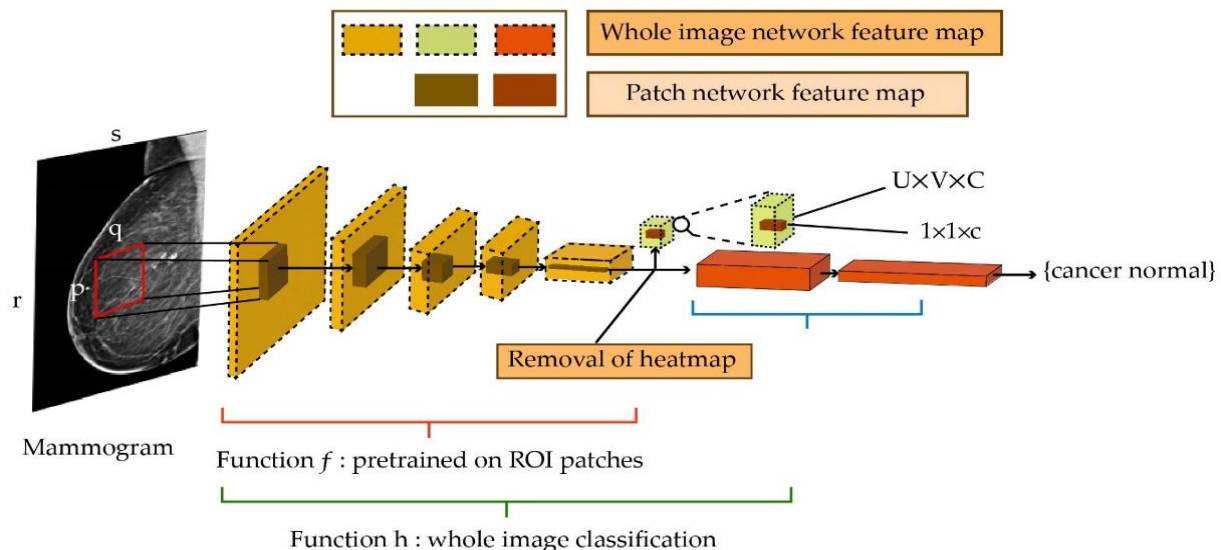


Figure 2 . Whole image classification and prediction of cancer or normal.

Given that mammography is a reliable approach for breast cancer diagnosis,

Petrinet al. have utilized two mammography images (bilateral craniocaudal and mediolateraloblique views) to enhance the diagnosis performance. Their method is based on EfficientNet and has two major components, which are the patch classifier and the whole-image classifier. The patch classifier inspects small sub-images, and the whole classifier uses the patch classifier to scan the whole mammogram. The high-level schematic of this method is depicted in Figure . As can be seen, the two mammograms are processed in parallel.

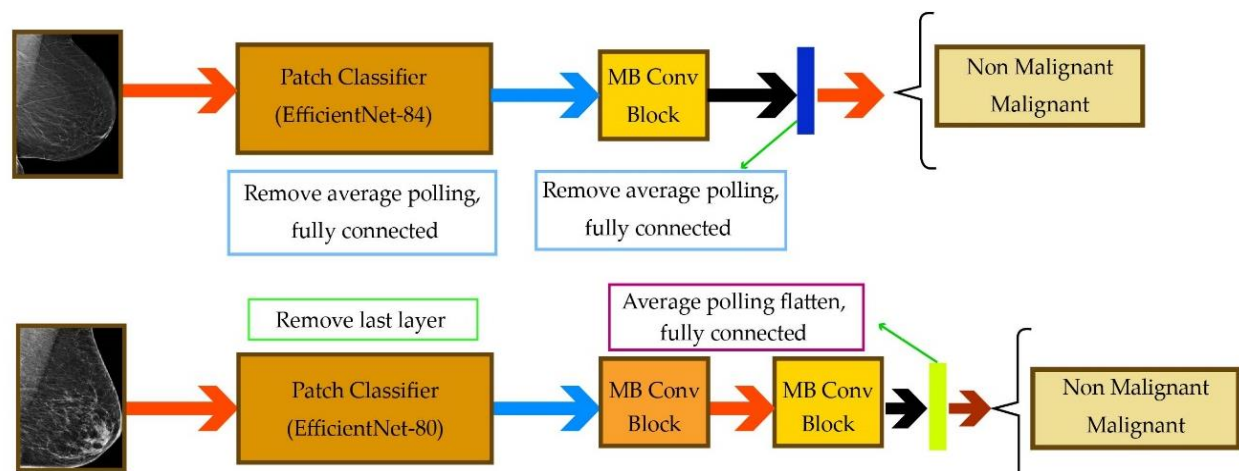


Figure 3. Diagrams of the single-view classifier for the “CV test” (top) and “OD test” (bottom).

In addition to mammography, the detection of small tumors helps with the early diagnosis of breast cancer. To this end, the STAN [52] has been proposed, which utilizes multiple convolution operations with different kernel sizes to capture breast tumors of various sizes (including small ones). The architecture of STAN is illustrated in Figure 6, in which convolutions with different sizes have been marked with different colors.

Researchers have observed that nuclear protein Ki-67 and tumor-infiltrating lymphocytes (TILs) are important factors for breast cancer diagnosis. Due to the lack of publicly available datasets for Ki-67 stained cell detection, Negahbani et al. gathered such a dataset for public use. Additionally, a DNN named PathoNet was proposed which is a light backbone for cancer diagnosis. To facilitate experimenting with different DL models, a generic pipeline for cancerous cell detection was proposed that is compatible with a variety of DL models.

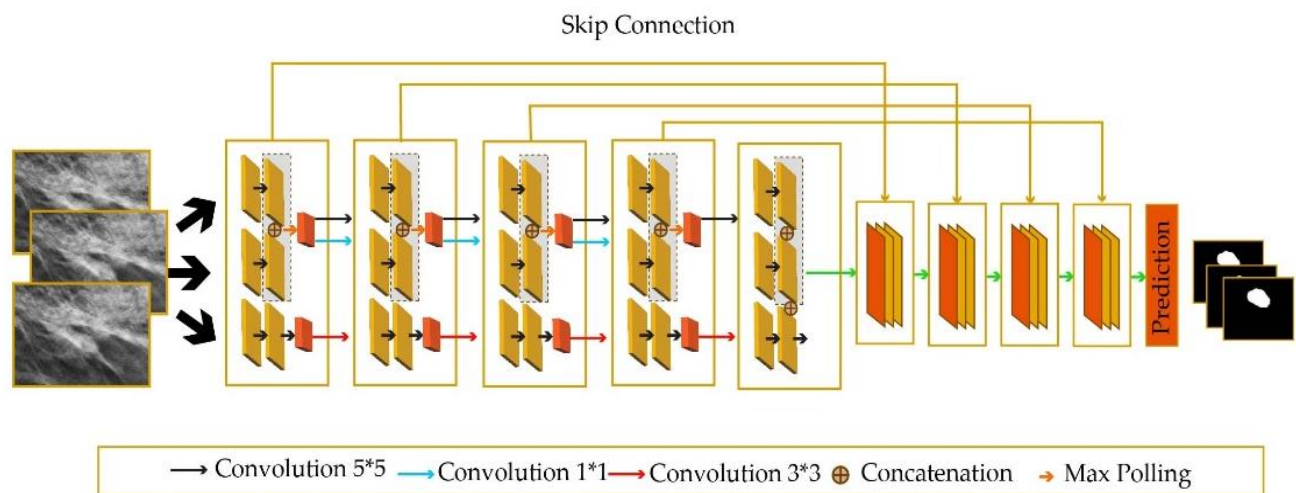


Figure 4. Small Tumor-Aware Network (STAN) to improve the performance of segmenting tumors

Although achieving state-of-the-art diagnosis performance is important, the ability to interpret the decision-making of DL models should not be overlooked. Being able to reason about the decision-making process is useful to gain better insight into the strengths and weaknesses of DL models. To this end, Patil et al. [54] took a multi-instance learning approach in a weakly supervised manner for the classification of breast cancer histology images. As shown in Figure 7, each input image is partitioned into multiple smaller patches. Feeding these patches to the feature extractor module, attention scores are computed, which are used to compute bag-level features. The classification is performed based on the bag-level

Features.

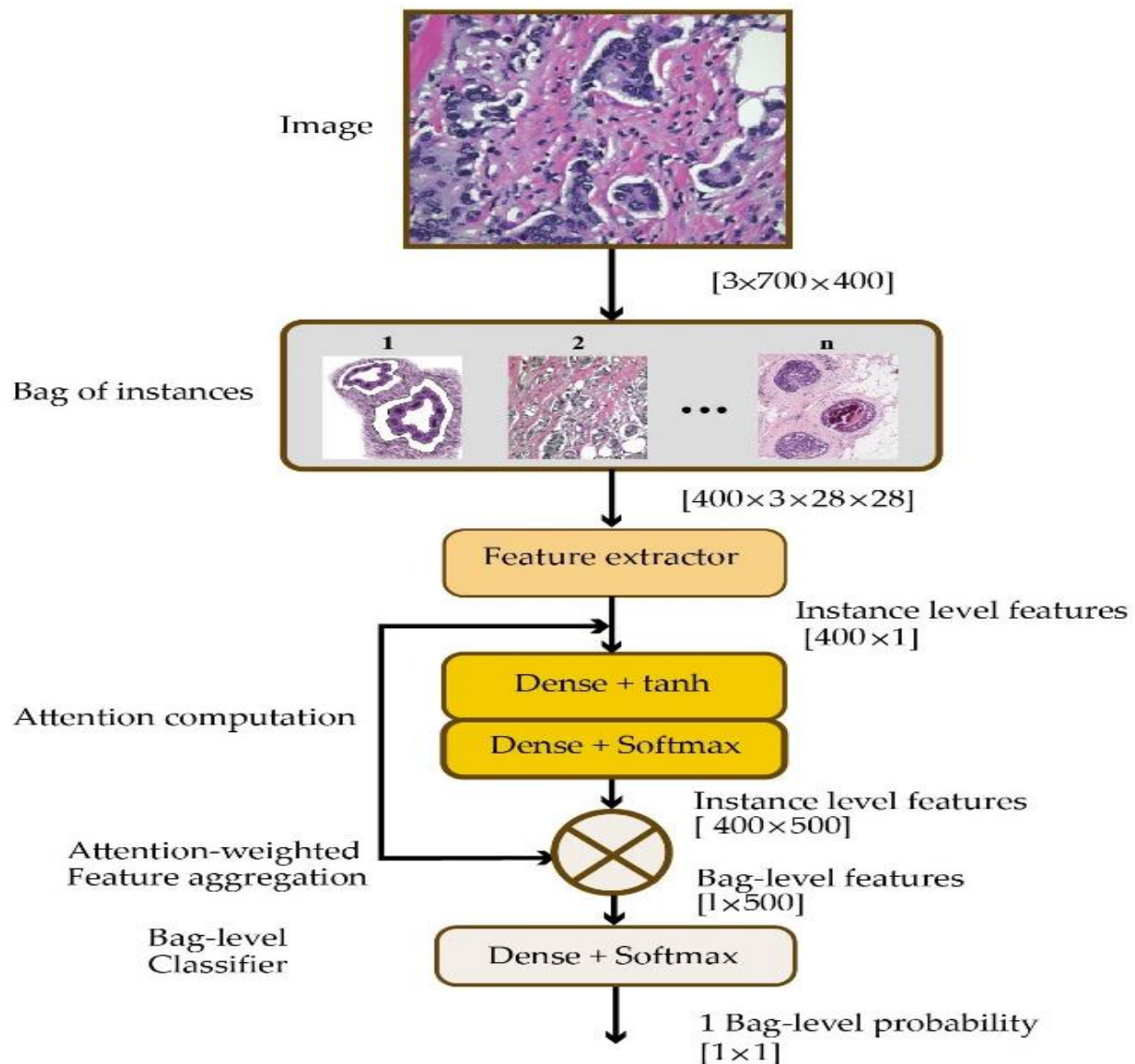


Figure 5. Multi-instance learning architecture for classification of breast cancer histopathology images

Discussion :

Early diagnosis and treatment of breast cancer heavily contribute to increasing life expectancy [91]. In developed countries, age-normalized breast cancer mortality fell by 40%

between the 1980s and 2020 [92]. Breast cancer mortality has been reduced by 2 to 4 percent

per year in nations that have taken effective treatment strategies [93,94]. Assuming that the breast cancer mortality rate is decreased by 2.5 percent per year, it is anticipated that 2.5 million more patients will stay alive from 2020 to 2040 [95,96].

As a worldwide issue, breast cancer took more than 600,000 lives in 2018. Screening mammography is very effective at reducing breast disease mortality by 20–40%, and it is recommended by health organizations worldwide for early detection of malignant growth locations [97,98]. Information obtained from our provincial disease reconnaissance framework

revealed the status of breast cancer growth endurance and mortality rate in northwestern Iran [99]. Generally, Iran has better breast cancer explicit endurance and a lower mortality rate compared to the country's general breast cancer growth explicit endurance.

However, breast cancer endurance is still lower than in developed nations [100,101].

Breast cancer was reported as the third most common malignant growth in the studies carried out in Iran [102]. The US and Western Europe have reported the highest breast cancer rate, while East Asia has reported the lowest [8,103]. Iran is one of the countries with

a rising cancer rate and mortality rate. On the other hand, in agricultural nations, these rates are lower [30]. The aging population, variation to the Western way of life, no fullterm pregnancy, late age at first pregnancy, lack of breast healthcare services, hormonal pregnancy control, and being overweight have contributed to these patterns [104–106].

Over the last decade, early diagnosis and efficient treatment have increased the age-normalized

life expectancy of patients in developed countries.

Conclusions :

In this review, we looked at the most recent research on breast cancer diagnosis using

DL in image modalities. Various well-known DL methods such as CNN, RNN, GoogLeNet, ResNet, and ANN have been used in the literature for breast cancer diagnosis. In addition to reviewing existing DL-based diagnosis methods, the publicly available datasets and source code repositories were introduced as well. Inspection of the existing approaches reveals the

significant progress toward automated diagnosis using DL. However, the reliability of these automated systems is yet to be improved before full deployment in real-world applications.

Over the years, the field of DL has made significant progress to the point that model representation power is rarely the limiting factor. However, without having a sufficient number of training samples, these powerful models will be of no use. Dealing with limited training data is an ongoing research field and can be tackled using different approaches.

The most obvious way of addressing data shortage is gathering high-quality datasets that are publicly available. However, data collation is not always possible. Image composition is an alternative promising approach that can be used to create new samples by merging two images [138]. In this technique, several background and foreground images are combined in different ways to generate new training samples. Transfer learning is another strategy to

deal with data scarcity. It is highly desirable to make transfer learning domain-aware [139]. Oftentimes, existing pre-trained models have been trained on general-purpose datasets such as ImageNet, which bears little resemblance to medical images. To address this issue, it is better to pre-train models on datasets that share common features with our target dataset.

While DL models are general-purpose learners, relying solely on image data is a short-sighted strategy. Investigating the possibility of performance improvement via fusing multiple sources of data [140] is worth investigating. A different but related approach is utilizing an ensemble of DL models for more robust decision making. The challenge is reducing the complexity of ensemble DL models in order to achieve better performance with manageable computational complexity. Knowledge distillation approaches [141,142] may be useful in making ensemble methods computationally efficient without losing much performance.