ISSN: 2633-4828 Vol. 5 No.4, December, 2023

International Journal of Applied Engineering & Technology

A SURVEY ON APPLYING HANDCRAFTED AND LEARNED FEATURE EXTRACTION FOR BREAST CANCER CLASSIFICATION USING MACHINE LEARNING AND DEEP LEARNING ALGORITHMS

Salini S Nair[1] and Dr. M. Subaji[2*]

[1]Research Scholar, School of Computer Science and Engineering, VIT University, Vellore, India

[2*]Professor and Director, Institute for Industry and International Programmes, VIT University, Vellore, India

[1] salinis.nair2015@vit.ac.in and [2*]msubaji@vit.ac.in

ABSTRACT

In recent years, breast cancer has become a condition that affects more and more people. Because it can spread from the breast to other regions of the body, it is the second most common cause of mortality for women. Therefore, early detection is essential for effective treatment. Expert systems can assist in the precise detection and categorization of benign tumours utilising data mining and machine learning approaches, avoiding the need for needless therapies. The review examines the use of several machine learning techniques, such as regression, Support Vector Machine (SVM), deep learning, random forests, decision trees, and K-Nearest Neighbors (KNN), for the detection of disease. The survey delves into the comparative efficacy of handcrafted versus learned feature extraction methodologies in breast cancer classification, employing a spectrum of machine learning and deep learning algorithms. It examines the nuanced performance variations, elucidating the advantages and limitations of both approaches when applied to the complex task of breast cancer diagnosis. Through comprehensive analysis, it aims to delineate the most promising techniques that could augment the accuracy and robustness of classification models in medical diagnostics. This review explores different breast cancer detection methods and compares their accuracies while summarizing their findings, challenges, and limitations. Overall, automated feature extraction and classification algorithms can assist medical practitioners in diagnosing and detecting breast cancer, improving patient outcomes.

Index Terms – Breast Cancer, Machine Learning, Deep Learning, Computer-Aided Diagnosis (CAD)

INTRODUCTION

Gene abnormalities lead to cancer development. The nucleus of every cell contains the genes, and our bodies' cells replace themselves in a systematic manner through cell proliferation. However, mutations can alter cell growth, the capacity to continue dividing without restriction or order, and the propensity to produce a benign or malignant tumour. Around 2.7 million persons in the United States are estimated to have the condition, according to information in Cancer Statistics, 2020 [19]. Over 13.9 lakh new cases of cancer are recognised each year, and 8.5 lakh cancer-related deaths are reported. For men and women, respectively, the risk of dying from cancer is roughly 4,38,297 and 4,13,381 respectively. Oral cavity, stomach, and lung cancer were found to be the causes of 25% of cancer deaths in men, and uterine cervix, breast, and oral cavity cancer were the causes of 25% of cancer deaths in women.

According to Global Cancer Statistics 2020 [52], there were around 10.0 million cancer deaths and 19.3 million new cancer cases in 2020. A thorough breakdown of cancer cases in 2020 is shown in Fig. 1. As we can see, breast cancer accounts for 11.7% of all malignancies in women and is the most common carcinoma. According to the Globocan 2020 study [2], there were 684,996 fatalities and 2,261,419 new cases of breast cancer.

Achieving a great quality of life is challenging when cancer strikes older people [60]. Ageing is the biggest risk factor for both different types of cancer and cancer in general. Cancer incidence rates increase steadily with age, from less than 25 cases per 100,000 people under the age of 20 to approximately 350 per 100,000 people in the 45-49 age range to more than 1,000 per 100,000 people in the 60+ age range [1]. The incidence rates of all cancer types by various ages are displayed in Fig. 2. The population in the United States (U. S.) aged 85 to 94 years comprised 1.6% of the total population in 2010 and had grown by 30% between 2000 and 2010 [19]. The number

of cancer diagnoses among senior Americans (those in the 85+ age category) has increased as a result of improved cancer screening techniques and the U.S. population's longer life expectancy. 7.7% of all cancers in the US between 2005 and 2009 were cancer.

This survey focuses on current advances in machine learning (ML) and deep learning (DL) approaches for the early identification of breast cancer. It illustrates the advantages and disadvantages of various research techniques, difficulties, and potential paths for future study. CAD systems are created to categorise benign and malignant lesions and automate the identification of breast cancer. With the help of ML and DL algorithms, algorithms have been developed that can more reliably diagnose the condition at an earlier stage, reducing the frequency of readmissions to hospitals and clinics. Thus, using Artificial Intelligence (AI) approaches can facilitate the development of new fidelity protocols in healthcare and lower healthcare expenses resulting from incorrect diagnoses.

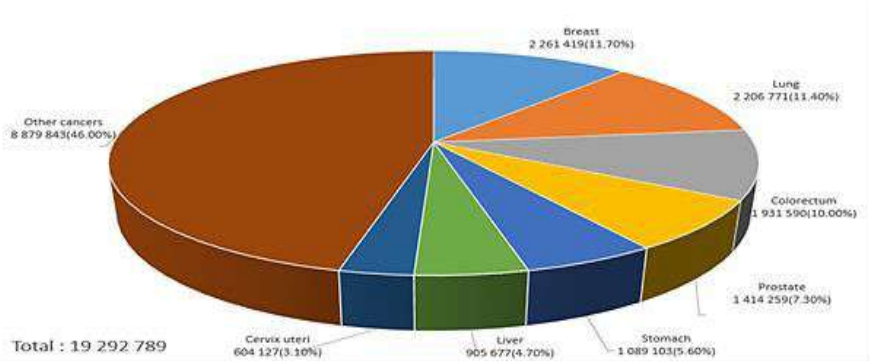

Fig. 1. Global Cancer Statistics 2020

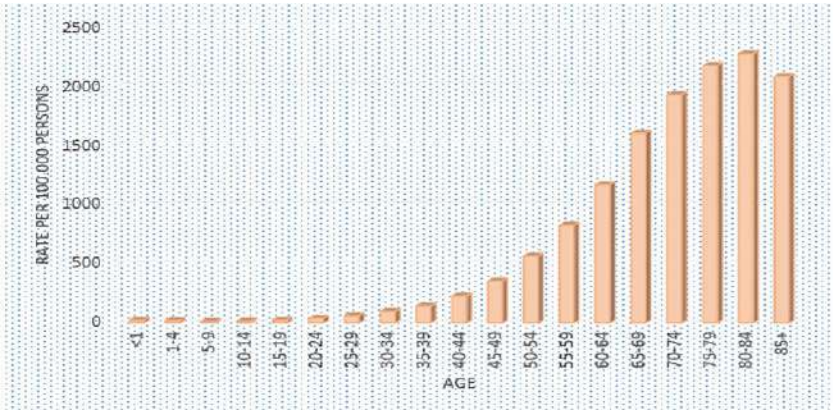

Fig. 2. Incidence rates by age at diagnosis, all cancer types

BREAST CANCER

Breast cancer arises from uncontrolled growth of breast cells. It commonly originates in the lobules or ducts of the breast, responsible for milk production and transportation. In rarer cases, it can develop from the stromal tissues, including fatty and fibrous tissues. Cancer cells can progressively invade nearby healthy breast tissue and reach the underarm lymph nodes, which act as detectors of foreign substances. If cancer cells enter the lymph nodes, they have a potential pathway to spread to other organs. Breast cancer is recognized by the World Health Organization (WHO) as the most prevalent cancer among women and a leading cause of cancer-related deaths in women. In order to save a person's life, cancer must be diagnosed and treated early. 1.7 million women developed cancer in 2012, according to the data [43]. People's inability to recognise diseases at an early stage affects the death rate day by day, as is the case with any healthcare [3]. According to the statistics provided at www.breastcanerindia.net, 144937 women had breast cancer identified, and 70218 of them had passed away from

it [26]. It is challenging to determine the precise diagnosis even when a different test is carried out. Numerous researchers have identified numerous automated breast cancer detection techniques. A few examples, like as CAD (Computer Aided Tool), are also used by doctors. Most Indian women, both in urban and rural areas, are affected by breast cancer, which is a common disease. The type of cancer tissue is determined using a variety of ML and DL methods. Unlike malignant conditions, which are cancerous and can be fatal, benign conditions do not pose a threat to human life. Even though technology has advanced, 20% of women worldwide still pass away each year [35]. The effectiveness of prediction illness models can be improved by enhancing the knowledge data with the help of developing approaches like machine learning and deep learning [34]. The fundamental motivation for the work discussed in this paper is the growing dangers of breast cancer shown by earlier research, which pushes us to explore unresolved and challenging questions. Breast cancer first manifests as a dangerous lump, and as it progresses, it grows uncontrollably and inconsistently. Even though technology has advanced, it is still unknown what causes breast cancer. Pathologists can find it through a few typical risk variables that affect women. Sometimes this disease is thought to have characteristics that come from the patient's genetic makeup. The two main types of breast cancer treatment are often local and symmetric. According to a World Health Organisation survey, 96% of women with cancer may live for an average of 5 years if they receive an early diagnosis.

Progression Stages of Cancer

The breast cancer stage describes the extent of the cancer cells' invasion outside of the primary tumour. It depends on a number of variables, such as the tumor's size and location as well as whether the disease has spread to other parts of the body [16]. Breast cancer is classified into five primary phases, as shown in Table 1 [22].

Table 1: Stages of Breast Cancer

Cancer Stages	Description
Stage 0	The disease is non-invasive. This means abnormal cells are present but have not grown beyond your breast ducts.
Stage I	The cancer cells extend to the nearby breast tissue and are confined only to the lobules and ducts of the breast.
Stage II	The tumor is either smaller than 2 cm or larger than 5 cm and may or may not affect the nearby lymph nodes.
Stage III	Cancer has invaded nearby tissue and lymph nodes at this stage but hasn't spread to distant organs.
Stage IV	Cancer has extended to areas away from your breast, such as your bones, liver, lungs, or brain.

METHODS

In exploring the methodologies relevant to the survey on handcrafted and learned feature extraction for breast cancer classification using machine learning and deep learning algorithms, an array of comprehensive methods comes into play. Diverse datasets comprising breast imaging data, meticulously preprocessed and augmented, serve as the foundation for evaluating the efficacy of these methods. Additionally, interpretability assessments and optimization strategies for hyperparameters within these methodologies are integral, ensuring a comprehensive understanding of model performance and applicability in the crucial domain of medical diagnostics.

Datasets

Table 2 lists the publicly accessible datasets that researchers have recently used.

Table 2: Databanks

Data repository	Dataset name	Description
Breast Histology [7]	Bioimaging Challenge 2015 Breast Histology Dataset	The dataset contains 269 histology images of breast cancer that have been stained with hematoxylin and eosin (H&E). These images have a resolution of 2048×1536 pixels. The dataset includes images from four different groups: normal breast tissue, benign breast conditions, in situ breast cancer, and invasive breast cancer.
BACH's dataset [8]	ICIAR 2018 - Grand Challenge on Breast Cancer Histopathology images dataset	There are 400 total photos in the dataset, which have been divided into 100 normal classes, 100 benign classes, 100 classes of in situ carcinomas, and 100 classes of invasive carcinomas.
CAMELYON16 [9]	CAMELYON16 challenge dataset	There are 400 whole-slide pictures (WSIs) of sentinel lymph nodes in this challenge data set, which was compiled from two different datasets. Both the University Medical Centre Utrecht and Radboud University Medical Centre in Nijmegen, the Netherlands, have collected the data.
CAMELYON17 [9]	CAMELYON17 challenge dataset	Lesion-level training data from CAMELYON16, which was gathered from Radboud UMC and UMC Utrecht, will be used for CAMELYON17. Additionally, there are lesion-level annotations for 10 training slides from each of the 50 annotated medical centres in CAMELYON17.
Cancer Metastases in Lymph Nodes Challenge [9]	Breast cancer metastasis detection dataset for the Cancer Metastases in Lymph Nodes Challenge	The combined dataset consists of images from the CAMELYON16 and CAMELYON17 challenges. The images in this dataset have dimensions ranging from approximately 1×105 to 2×105 pixels at the highest resolution. The main task in this dataset is classification, where each image is labeled as either normal tissue or metastases. There are two sets of training datasets within this combined dataset. The first set comprises a total of 170 images, with 100 images belonging to the normal class and 70 images in the metastases class. The second set consists of 100 images, with 60 images in the normal class and 40 images in the metastases class.
MITOS-ATYPIA-14 [29]	MITOS-ATYPIA-14 dataset	1539×1376 -pixel images of breast cancer on (H&E)-stained slides at 20 and 40 times magnification. A testing set contains 496 images obtained from 5 distinct breast biopsies and a training set has 1200 images obtained from 16 different breast biopsies.

TUPAC16 [56]	TUPAC16 dataset	The dataset comprises of 73 histopathology images of breast cancer at a magnification level of 40× from three Netherlands-based pathology facilities. The dataset consists of 50 training images that are 5657×5657 pixels in size and 23 test images that are 2000×2000 pixels in size. It was assembled from two different pathology centres. The training dataset's images are later randomly cropped to 2000 by 2000 pixels in size.
UCIMLR/University of California Irvine Machine Learning Repository [36]	Breast Cancer Wisconsin (Diagnostic) Data Set (WBCD)	WBCD is a commonly used multivariate classification dataset among researchers working on machine learning research. It contains 569 instances with 32 attributes.
BreakHis [48]	Breast Cancer Histopathological Database	It includes 9,109 microscopic pictures of breast tumour tissue taken at different magnifications (40×, 100×, 200×, and 400×) from 82 individuals. In PNG format (700×460 pixels, 3-channel RGB, 8-bit depth per channel), it currently stores 2,480 benign and 5,429 malignant samples.
SEER [4]	SEER Breast Cancer Data	It contains stage-by-stage survivorship statistics for breast cancer and SEER incidence and population data correlated with age, sex, race, year of diagnosis, and geographic locations

Breast Cancer Diagnosis Techniques

Breast cancer diagnosis is a multimodal process that typically involves physical examinations by the patient and physician, breast screening methods, and other diagnostics. Breast cancer can be diagnosed using a variety of methods [28]. Table 3 lists some of the classic and cutting-edge techniques that researchers have recently employed [37], [6], [31], [14].

Table 3: Traditional and Emerging Techniques

Techniques	Pros	Cons
SBE-Self–Breast examination	1. Raises public consciousness 2. May execute this simple technique at home.	1. Recognise breast cancer early 2. Significant amounts of overdiagnosis and false positives
CBE-Clinical Breast examination	1. Reduced breast cancer mortality. 2. Can detect breast cancer missed by mammography.	1. An increase in false-positive findings. 2. Excessive levels of overdiagnosis and false positives.
Screen Film Mammography (SFM)	1. Identifies cancer in its earliest stages 2. Common modality 3. High sensitivity for fatty-tissue-filled breasts 4. It is affordable	1. Low sensitivity and thick breast 2. Nondigital
Full Field Digital Mammography (FFDM)/Digital Mammography (D.M.)	1. High specificity and sensitivity in detecting cancer at an early stage 2. Effective and standard modality 3. Mobile equipment 4. Temporal reaction (around one minute) 5. Reliable resolution 6. Greater accuracy with digital mammography in thick breasts	1. High specificity and sensitivity in detecting cancer at an early stage 2. Effective and standard modality 3. Mobile equipment 4. Temporal reaction (around one minute) 5. Reliable resolution 6. Greater accuracy with digital mammography in thick breasts

Ultrasound (U.S.)	1. High diagnostic value in breast-dense women 2. Mobile equipment 3. No radiation is present. 4. It is appropriate for pregnant ladies. 5. Low-cost, reliable, and secure	1. High false-positive rates 2. Poor contrast 3. Operator dependent biases 4. Incorrectly indicates a negative result
Magnetic resource imaging (MRI)	1. Almost maximal sensitivity 2. Can identify cancer that has spread intraductally 3. Optimal contrast 4. High quality 5. Nonionizing radiation 6. Applied to patients at high risk	1. Specificity values are lower, and variables require suitable equipment. 2. Biopsies can be challenging. 3. Imaging is limited to the breast's lateral side. 4. Not transportable 5. Expensive gadget. 6. Possibility of a false-positive outcome 7. High body temperature may during long MR
Positron Emission Tomography (PET)/ Computed Tomography (CT) Imaging	1. It is accurate to register and fuse images. 2. Effective contrast 3. Relevant details 4. Extremely sensitive	1. Makes use of radiation with ions 2. Employed radioisotopes 3. Lack of clarity 4. The stationary gadget 5. Expensive gadget
Digital breast tomosynthesis (DBT)/3D mammography	1. Identifies cancer in dense breasted women 2. Decrease the number of false negatives and positives 3. Several 3D images are shown at a single screening	1. Costly 2. Compared to D.M., the average radiation dose is one to two times higher.
Thermography	1. Prompt detection 2. Noninvasive 3. Not radiation-emitting 4. Rapid reaction 5. The ideal imaging technique for thick breasts.	1. Sensitive to changes in room temperature 2. There are many false negatives and false positives 3. Limited Specificity
Electrical impedance tomography	1. Noninvasive 2. Nonradioactive 3. Relatively inexpensive	1. Poor spatial resolution
Histopathology (H.P.) images	1. Use multi-color pictures to diagnose various cancer types rather than only detecting malignancy. 2. Tissues are capable of being thoroughly studied. 3. Due to the creation of several region-of-interest (ROI) images from slide images, there is a reduced risk of missing cancer in its early stages.	1. Because manual analysis is laborious and time-consuming, high proficiency is necessary. 2. Colour fluctuation and various staining techniques might lead to incorrect diagnosis
Microwave imaging	1. Minimally invasive 2. Not radioactive	1. Inadequate resolution at greater depth 2. Fibro glandular tissues with low contrast

Optical imaging	1. Cheap, non-radioactive, and transportable 2. Rapid reaction 3. Strong contrast 4. Safe	1. Low contrast due to high dispersion 2. Insufficient imaging depth 3. Less spatial resolution

COMPUTER AIDED CANCER DIAGNOSIS

Radiologists are increasingly relying on computer-aided diagnosis (CAD) as a tool for analysing mammograms. Several research teams are developing CAD systems to categorise benign and malignant tumours. Ultrasound and magnetic resonance imaging tests are frequently advised to characterise breast lesions found either clinically or during screening. There have recently been more CAD apps created for screening mammography, but more are being tested for ultrasound and MRI breast imaging. The next generation of CAD systems will include CAD characterisation as a key element.

The technology of CAD is used to identify and characterise cancer. Although CAD is not restricted to a single form of disease, many CAD systems have been created and used for breast cancer up to this point. This review will go over how CAD systems are now used to diagnose breast cancer, how they are used as a second reader in clinical settings, and research that have looked at how CAD affects radiologists' performance. For various imaging modalities, many CAD applications are being created. Due to commercially available, FDA (Food and Drug Administration) -approved devices, screen-film mammography has been the primary clinical application of CAD to date. Numerous research has demonstrated how CAD enhances radiologists' abilities. Many academic institutions have invested a significant amount of research time in creating CAD techniques. As a second reader, CAD systems will take on a greater significance in the clinic. Clinical studies have demonstrated that CAD can increase the precision of the identification of breast cancer. Preclinical research has shown that CAD has the ability to more accurately classify malignant and benign tumours. For various breast-imaging modalities, more and more CAD systems are being created. The diagnosis of breast cancer using histopathological pictures has recently piqued the research community's interest. The biggest labelled publicly accessible dataset, "BreaKHis," including 7909 histopathological pictures of both benign and malignant classifications, was made available by the authors [50]. Since then, a large number of academics and professionals have studied the BreaKHis dataset to create automated and trustworthy methods to separate various breast cancer histological images utilising CNNs (Convolutional Neural Networks) as the basic building models. To undertake evaluative comparisons in the next sections, we exclusively present a systematic review of recently published publications that examined the BreaKHis dataset for MD (Magnification Dependent) and MI (Magnification Independent) breast cancer classifications. Spanhol et al. [51] trained a deep CNN model, which is a subset of AlexNet [25], using a set of pixel patches from histopathology pictures. These patches, with sizes of 32×32 and 64×64, were extracted using sliding window and random strategy approaches. Then, utilising sum, product, and maximum procedures, the final probabilities of these patches were used to classify benign and cancerous breast tissues. The highest degree of accuracy was attained between 80.8% and 89.6%. The same authors looked into the utility of employing deep features (DeCAF) derived from a pre-trained CNN (BVLC CaffeNet), then used them as inputs to train a logistic regression classifier in their subsequent study [49]. Their findings demonstrated that DeCAF characteristics might be a practical substitute for a CNN trained from scratch, reaching an accuracy for benign and malignant classification between 81.6% and 84.8%. 'Single task' and'multi-task' are two distinct CNN-based designs that Bayramoglu et al. [11] introduced. While the multi-task was utilised to predict both the malignancy and the magnification factor simultaneously, the single task was made to predict the malignancy alone. According to the patient level score, the reported accuracy when using the single-task CNN ranged from 82.10% to 84.63% and when using the multi-task CNN, it ranged from 80.69% to 83.39%. BiCNN is a revolutionary approach put forth by Wei et al. [59] that is built on CNN. Advanced data augmentation and transfer learning algorithms were used to enhance the classification performance after taking into account prior information based on cancer class and subclass. Between 97.64% and 97.97% accuracy was attained in separating benign from malignant pictures. For classifying breast cancer histopathology pictures into benign or malignant, Pratiher et al. [40] presented a

cascaded strategy based on manifold learning L-ISOMAP and stacked sparse auto-encoder. The reported outcomes ranged from 96.8% to 98.2%. Breast histopathology pictures can be classified as benign or malignant using CNN features, local, and frequency domain features, according to Nahid et al. [33]. Their general accuracy ranged from 94.40% to 97.19%. The same authors presented other deep learning strategies in [32] that were aided by K-Means and Mean-Shift local clustering algorithms. They tested both Softmax and Support Vector Machine (SVM) as output classifiers and used CNN and Long-Short-Term Memory (LSTM) as main building models. The classification of benign and malignant conditions had the highest reported accuracy, which ranged between 90% and 91%. A comparison of the classification performance of handcrafted features and deep features was made by Bardou et al. [10]. While deep features were produced by training a CNN from scratch, handcrafted features were encoded using a bag of words and locality-constrained linear coding before being categorised using SVM. The most accurate results were based on CNN and ranged from 83.31% to 88.23% for multi-class classification and between 96.15% and 98.33% for binary classification. Overall, the research discussed above demonstrated a clear preference for using deep CNNs for the categorization of histological images of breast cancer due to their much-improved performance. However, it is not simple to train a deep CNN because it needs a lot of CPU and memory resources and frequently encounters convergence and overfitting issues. In this context, a recent study by Tajbakhsh et al. [54] shows that a well-tuned CNN performs better than the one trained from scratch or, in the worst scenario, as well as it in medical image processing (computed tomography (CT), ultrasonography, and optical endoscopy). The research for the histopathological imaging modality was expanded by Shallu et al. [44]. They noticed that a fine-tuned VGG-16 [46] with a logistic regression classifier had produced better classification results by achieving an accuracy of 92.60% for MI binary classification when compared to a CNN trained from scratch. In [44], the same authors described a CNN model with three convolutional layers, a max pooling layer, and two fully-connected layers that they fully trained to achieve an accuracy of 85.3% for MI binary classification. To categorise MI breast histopathology pictures into benign or malignant, Xiang et al. [61] tested a fine-tuned Inception-v3 [53] and reported an overall accuracy of 95.7%.

Handcrafted ML Based Computer-Aided Diagnosis

For the purpose of choosing the best ways to apply the CAD system, medical image processing requires prior knowledge of the content and nature of the image. Effective image processing techniques must be used in the major stages of the CAD system in order to obtain a high level of efficiency for automated diagnostics. CAD systems typically include five steps, as indicated in Fig 3. The following is a basic explanation of the key steps in a CAD system: (1) Image pre-processing: - Some imaging modalities, including ultrasound, depend on this step to improve the image and minimise noise while minimising feature distortion. There may not be a pre-processing stage in some CAD programmes. (2) Image segmentation: - A crucial stage in the effective development of CAD systems is image segmentation. The separation of the region of interest (ROI) in accordance with the desired properties is the primary goal of segmentation. In recent years, 3D images have been made possible by a variety of imaging modalities, including computed tomography (CT), 3D ultrasound, and many more. As a result, 3D segmentation techniques are preferred for volumetric picture segmentation that is more precise. (3) Feature extraction and selection: - In this step, several features are taken from the image in accordance with the characteristics of the lesions. These characteristics help distinguish between benign and malignant tumours. The feature set is typically fairly large, and the following stage depends greatly on the choice of the best features. (4) Classification: - The suspicious areas are categorised as benign or malignant using various classification techniques based on the chosen criteria. This section presents the typical classification techniques used in medical imaging. (5) Performance evaluation: - The effectiveness of the CAD system is assessed in this step.

The data extracted from the selected articles about different machine-learning techniques are presented in Table 4, which shows the machine-learning methods for disease prediction. The table summarizes existing literature articles collected in this survey using machine learning methods to manage breast cancer disease. The first column refers to the technique; the second column is the employed machine learning method to identify the benefit; and finally, the last column identifies the drawbacks.

Table 4: Different Machine Learning Techniques to Manage Breast Cancer Diseases

Technique	Benefits	Drawbacks
Logistic Regression (LR) [24]	1. It can offer probability. 2. To categorise new data using continuous and discrete datasets. 3. They are used to categorise the observations based on various forms of data. 4. Identify the classification variables that are most useful in a timely manner.	1. There must be a categorical dependent variable. 2. Multicollinearity should not exist in the independent variable.
Support Vector Machine or SVM [58]	1. Classify the n-dimensional space to establish the decision boundary. 2. Future placement of the new data point in the appropriate class will be simple. 3. It finds the closest point of the lines from both classes. 4. SVM aims to maximize this margin.	1. Not appropriate for massive data sets. 2. SVM might function more effectively if the data set contains more noise. 3. Choosing an appropriate kernel function is complex.
K-Nearest Neighbor(KNN) Algorithm [21]	1. Easily identify the group or class of a particular dataset. 2. Resistant to jittery training data. 3. More efficient if there is a large amount of training data.	1. K's value must always be determined, which might be tricky at times. 2. The high computation cost is caused by the need to calculate the distance between each data point for each training sample.
Decision Tree (DT)[15]	1. Simple to understand as it follows a human's process while making real-life decisions. 2. It may be useful for resolving issues involving decisions. 3. Taking into account all potential solutions to an issue is helpful. 4. In comparison to other methods, there is less room for data cleansing.	1. The decision tree is complicated since it has many nodes. 2. It can have an issue with overfitting. 3. For more class labels, the computational complexity of the decision tree could rise.
Random Forest (RF) [15]	1. Both classification and regression tasks can be carried out by Random Forest. 2. It is able to manage big datasets with lots of dimensions. 3. It eliminates the overfitting problem and improves the model's accuracy.	1. It is inappropriate for regression tasks. 2. It requires much more time to train as compared to decision trees.

Principal Component Analysis (PCA) [15]	1. Mostly used to reduce the data's dimensionality. 2. Using the largest variance vectors, the algorithms condense the amount of features to 3 or 4.	1. It may need to include some information compared to the original list of features. 2. Compared to original features, principal components are harder to read and understand. 3. Before using PCA, standardise your data; otherwise, PCA won't be able to identify the best principal components.

Learned DL Based Computer-Aided Diagnosis

Artificial neural networks have been improved or updated with deep learning techniques [21], [15] which take advantage of plentiful, affordable computing. They are focused on creating larger or more intricate neural networks. Massive collections containing labelled analogue data, such images, text, audio, and video, are a focus of several techniques. Convolutional Neural Network, Recurrent Neural Network, Long Short-Term Memory Network, Stacked Auto-Encoders, Deep Boltzmann Machine, and Deep Belief Network are the common deep learning algorithms [18], [50], [33]. A DL-based CAD system [38] is shown in Fig. 3. The following layers are implemented by CAD system based on CNN: (1) Convolutional operations that automatically extract features from the image use convolutional layers. (2) Dimensionality reduction of features using a pooling layer (3) Fully connected and Softmax layer for classification.

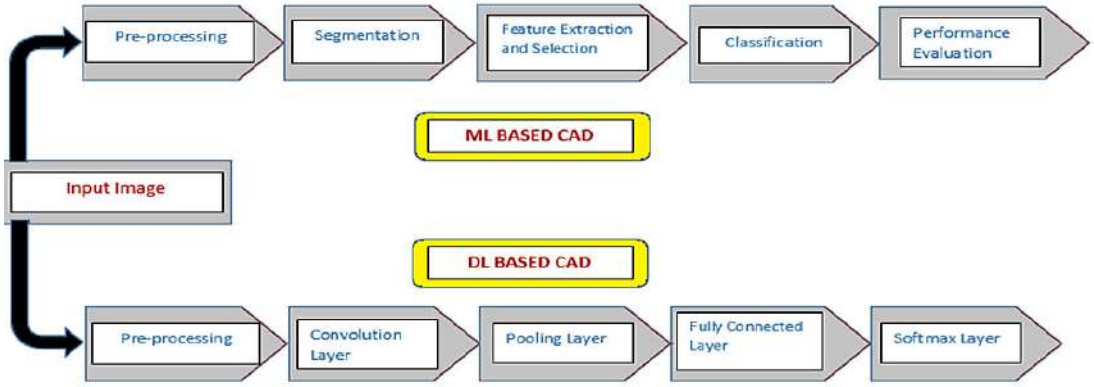

Fig. 3. Machine learning and deep learning framework

DISCUSSIONS

The survey initiates a comprehensive discussion around the application of handcrafted and learned feature extraction techniques in breast cancer classification within the realm of machine learning and deep learning algorithms [27]. It explores the theoretical underpinnings and practical implications of these methodologies, dissecting their impact on model performance and interpretability in the intricate domain of medical diagnostics. Through critical evaluation and comparative analysis, it endeavors to elucidate the optimal approaches that can bolster the accuracy and generalizability of classification models for enhanced breast cancer detection. Worldwide, the prevalence of breast cancer is rising at an alarming rate, and long-term survival depends on early detection and diagnosis. The best method for early diagnosis and therapy has been demonstrated to be CAD image analysis of medical images. When employing deep learning and machine learning to diagnose breast cancer, there are a few major issues that must be resolved [13].

Comparative Analysis Based on Classification Algorithms of Machine Learning

This section includes a comparative analysis of various machine learning classifiers used by authors to diagnose breast cancer. Table 5 provides a summary of some of the well-known studies on the diagnosis of breast cancer.

Table 5: Comparative study on Machine Learning algorithms

Dataset	Method Used	Performance Measure	Pros	Cons	Reference
DDSM database	GA +S3VM	Acc-92.1 % (Gaussian), 87.37% (Triangle), 89.55% Linear)	1. Semi-supervised classification 2. GA reduces feature vector dimensionality	1. Low performance classification	[63]
Wisconsin Breast Cancer Database	GBN, SVM, KNN, LDA, QDA, LR, ET, AB, GB, DT, RF	Acc-98.23%(AB), 95.57%(GB),93.80%(DT),96.45(RF),95.57 %(GBN),51.10 %(SVM),91.11 %(KNN), 95.33%(LDA), 96.51%(QDA), 94.63% (LR), 97.34 %(ET)	1. GB resolves hyperparameter problem 2. Control parameter setting and automated cancer diagnosis	1. Suitable for numerical and categorical values	[17]
MIAS dataset, DDSM dataset	CDTM + KPCA + GOA-KELM	Acc- 97.49% (MIAS) Acc- 92.61% (DDSM)	1. Better classification accuracy 2. Reduced number of features 3. Multi-class classification 4. Minimum computational time	1. Images from CT, ultrasound, and thermography can't be considered. 2. No deep learning approaches	[15]
WBCD dataset	Least Square SVM (LS-SVM)	Acc-97.08	1. Examining medical information in less time 2. Exploration of data	1. Built as a system for offline diagnosis	[39]

UCI machine learning repository	SVM and SVM ensembles (Linear kernel and RBF kernel)	Acc-96.85%, ROC- 0.967, F-measure - 0.966 (GA + linear SVM), F-measure - 0.988(GA + RBF SVM), Acc-98.28% (GA + RBF SVM ensemble), ROC- 0.98(GA + linear SVM ensembles and GA + poly SVM ensembles), Acc-99.52% ROC-0.876, F-measure-0.995 (GA + linear SVM ensembles and RBF SVM ensembles)	1. Better performance for SVM ensembles 2. Small-scale datasets perform better with RBF kernel-based SVM ensembles with the boosting approach and linear kernel-based SVM ensembles based on the bagging method. 3. RBF kernel-based SVM ensembles with boosting-based training perform better for large datasets. 4. After feature selection, a significant reduction in the computational time for SVM classifiers 5. The RBF SVM and poly SVM have the shortest training times, respectively. 6. The linear SVM classifier requires a lot of computing time.	1. Baseline classifier(s) cannot be distinguished from the superior prediction model(s)	[20]
BreaKHis dataset	Sliding window technique + LBP+ SVM + Majority voting technique	Acc- 91.12%, Sn- 85.22%, and Sp- 94.01%.	1. Sliding window for better generalization 2. Fast and effective feature extraction 3. From images of histopathology slide, finely detailed features were retrieved using the fusion of sliding window approach and LBP.	1. No comparison of various machine learning techniques.	[5]

WBCD data set	SVM and KNN	Acc- 98.57% Sp- 95.65% (SVM) Acc- 97.14% Sp- 92.31% (K-Nearest Neighbors)	1. Helpful to both the general public and the medical staff 2. Supervised ML techniques use attribute training 3. Uses 10-fold cross-validation 4. Better results	1. Numerical and categorical values.	[21]
WDBC and BCCD datasets	SVM, LR, KNN, and ensemble classifier	Acc-99.3% (Polynomial SVM) Acc-98.06% (LR) Acc-97.35% (KNN) Acc- 97.61% (Ensemble classifier) for WDBC	1. Shorter training time 2. High accuracy for SVM polynomial kernel 3. LR with recursive feature elimination has higher accuracy. 4. DE (data exploratory) techniques effectively detect higher accuracy	1. Unable to provide malignant features 2. Non-effective results for BCCD dataset 3. Not effective with the Asian patients 4. No deep learning methods	[41]
Images taken using ultrasound technology at the Breast Cancer Unit at the Oncology Specialist Hospital in Baghdad, Iraq	ANN	Prec - 82.04% Sn - 79.39% Sp - 84.75%	1. High positive predicate rates 2. Using the ultrasound images, a variety of texture features were retrieved. 3. ANN classifiers can address a heterogeneous mix of disorders using a single ultrasound image from each patient.	1. The automatic segmentation becomes difficult due to the extreme variance in tumour size, location, and shape. 2. Restrict the feature extraction 3. Limited number of data	[30]

Wisconsin Breast Cancer Dataset from UCI Repository	RF, PCA+R, KNN, ANN, PCA+ANN,NB	Kappa-0.91, acc-0.95, sn-0.92, sp0.98(RF), Kappa-0.95,acc-0.95, sn-0.92,sp-0.97(PCA+RF), Kappa-0.93,acc-0.97,sn-0.92,sp-1(KNN),Kappa-0.93,acc-0.97,sn-0.92, sp-1(ANN) Kappa-0.93,acc-0.97, sn-0.95,sp-0.98 (PCA+ANN), Kappa-0.81,acc-0.91,sn-0.88,sp-0.92(NB)	1. Combine multivariate statistical and machine learning techniques 2. PCA measures variance in data	1. High performance for binary classification	[42]

Comparative Study Using Deep Learning Classification Methods

The advancements made in recent years for using deep learning to diagnose breast cancer are presented in this section. Table 6 provides an overview of some deep learning-based research on the diagnosis of breast cancer.

Table 6: Comparative study on Deep Learning algorithms

Dataset	Method Used	Performance Measure	Pros	Cons	Reference
BACH dataset	EMS-Net	Acc -91.75% (training images) Acc-90.00% (testing images)	1. Patch augmentation and fine-tuning pre-trained DCNN 2. Fast testing	1. Non identified and discriminative image regions 2. Time-consuming in training stage	[62]
ABUS image dataset from JeonbukNational University Hospital (JNUH).	Inception-v3 CNN	Sn-88.6% Sp-87.6% AUC- 0.9468 ±0.0164	1. Learns the image features from raw images 2. CNN for feature extraction and classification	1. No dedicated algorithms for breast lesion detection and classification	[57]

Digital Database for Screening Mammography (DDSM) from the Massachusetts General Hospital (D. Kopans, R. Moore), the University of South Florida (K. Bowyer), and the Sandia National Laboratories (P. Kegelmeyer)	DCNN+GLCM+HOT + XGBoost DCNN+GLCM+HOT + SVM	Acc is 99.6% to Acc- 92.80% (XGBoost) Acc-84% (malignant tumors, XGBoost)	1. DCNN is used as a feature extractor 2. GLCM and HOT are used to extract breast mass information 3. SVM and XGBoost for classification 4. XGBoost deals with a limited number of available training samples and texture features 5. XGBoost handles imbalanced training data	1. Low volume dataset	[47]
Dataset from the India's Visakhapatnam-based M. G. Cancer Hospital & Research Institute	DNNS	Acc-97.21% Pre- 97.9% Rec-97.01%	1. Image performance, effectiveness, and quality are improved with normalization. 2. Based on a deep neural network's support value	1. Not evaluate CNN features	[55]

INBREST and CBIS-DDSM DATABASES	Deep CNN+RGP, Deep CNN+GGP	Acc- 0.919 ± 0.0003 AUC- 0.934 ± 0.0003 (RGP, INBREST DATABASE) Acc-0.922 ± 0.0002 AUC- 0.924 ± 0.0003 (GGP, INBREST DATABASE) Acc-0.762 ± 0.0002 AUC- 0.838 ± 0.0001 (RGP, CBIS-DDSM DATABASE) Acc-0.767 ± 0.0002 AUC- 0.823 ± 0.0002 (GGP, CBIS-DDSM DATABASE)	1. RGP and GGP structures were used for learning features 2. Numerous irrelevant regions are eliminated. 3. DNN to extract hierarchical features 4. Find potentially dangerous areas roughly 5. Automatic annotation to reduce the cost of annotations	1. Limited number of samples. 2. Low performance classification	[45]
BreaKHis dataset	GCN+ ResNet-18+Transfer learning,+Three-fold Data Augmentation	Acc-98.08% - 99.25% (MD binary classification) Acc-89.56%-94.49% (MD eight-class classification) Acc- 98.42% (Binary MI classification) Acc-92.03% (Eight-class MI classification)	1. The binary and eight-class classifications for both Magnification Dependent (MD) and Magnification Independent (MI) may be addressed in an efficient manner.	1. Colour differences in breast images have less of an impact when stain-color normalisation procedures are absent. 2. Low performance of eight-class classification 3. Need to mix handcrafted characteristics with deep CNN intrinsic features	[12]
Dataset from LRH hospital Peshawar, Pakistan	GoogLeNet, VGGNet and ResNet, Transfer Learning	Acc- 93.5% (GoogLeNet) Acc-94.15% (VGGNet) Acc-94.35% (ResNet) Acc- 97.525%	1. Transfer learning utilizes the gained knowledge for improving accuracy 2. Pre-trained CNN is used to extract features from images. 3. A data set's size is increased by data augmentation to increase CNN's effectiveness.	1. Not evaluate the model with hand-crafted features + CNN features to improve the classification accuracy	[23]

LIMITATIONS

Data Pre-processing

To increase the quality of the data used for machine learning, data pre-processing [30] is an essential step. We must pre-process input data before supplying it to our model since the crucial information it extracts directly affects its capacity to learn. Standardisation is a crucial part of the pre-processing phase. When we standardise our data, we change the numbers so that the mean and normal deviation are both 0. Another crucial component of data pre-processing is handling categorical variables. The variables that are discrete and not continuous are known as categorical variables. When features in our dataset are highly interdependent, multi-collinearity occurs, making it impossible to determine the relevance of the features using a weight vector. The interpretability of our model is impacted by multi-collinearity in the pre-processing of the data.

Feature Extraction

Feature extraction [30] automatically reduces the dimensionality [13] of observations into a much smaller set and is modeled it. Predictive modelling algorithms can't directly model some observations since they are significantly too voluminous in their raw form, like image, audio, and textual data. However, tabular data with millions of properties might be included just as quickly. Include projection techniques for tabular data, such as unsupervised clustering and principal component analysis (PCA). Additionally, the image data employs edge or line detection. Observations provide many of the same digital signal processing methods depending on the domain, whether an image, video, or audio. The key to feature extraction is that the methods are automatic, address the issue of too high dimensional data, and are frequently used to analogue observations saved in digital formats. Whenever a data scientist plays with a dataset of many variables, there are chances that some of them are redundant or noisy. That means such variables do not carry any signal useful as a predictor. Noise, as always, affects the overall accuracy of any predictive model.

Massive and Imbalanced Dataset

The magnitude of the data could lead to problems with generalisation, data imbalance, and difficulty achieving the global optimum [31]. In summary, sparse approximation is often seen when there is a lack of training data. This can lead to poor performance, which can stem from two scenarios: an under-constrained model that overfits the sparse training dataset or an over-constrained model that underfits the training dataset. Insufficient test data contributes to an overly optimistic and high variance estimation of model performance. Machine Learning algorithms generally produce inadmissible classifiers when confronted with imbalanced datasets. A large portion of this present reality arrangement issue shows some class awkwardness, which happens when inadequate information relates to both class names.

Learning Strategy

Algorithms for supervised learning [21] examine labelled data in which the proper classifications are determined to learn particular analytical patterns. The unsupervised learning algorithm works independently to find a new model or pattern that might not be apparent to the human eye rather than being guided by labelled data. When working with large datasets devoid of underlying structure, unsupervised learning is very useful for data mining. In datasets that you need assistance understanding or figuring out what to do with, it can help you identify patterns and meanings. Unsupervised learning is a computationally intensive technique; when coupled, it identifies the structures that act as training material for the supervised learning process.

Model Selection

Model selection [13] is the way toward picking one of the models as the final learn model that tends to the issue. The best way to deal with model selection requires adequate information, which might be almost vast and liable to the problem's complicated nature. One of the critical decisions to be made in the model selection procedure relates to our presumption about the shape of the functional relationship between the input variable and the response variable. When we decide to accept the shape of our model, we develop a parametric model, and our concern lessens to assessing many quantifiable elements, known as parameters. The most popular assumptions are that the

data is linear. While we can loosen up the linear assumption when fundamental, we, once in a while, don't have any desire to expect the shape of the function by any means. Non-parametric models help to maintain a strategic distance from the situation where we erroneously expect a function that doesn't match the data. In any case, many perceptions are acquired to make non-parametric techniques successful, which can be expensive or even infeasible.

Notwithstanding the way that non-parametric strategies are regularly not practical, there are different trade-offs to contemplate. One critical trade-off is between interpretability and adaptability. Since non-parametric models follow the data cautiously, they periodically bring about unusually shaped plots, which can be challenging to interpret. Suppose the objective is to comprehend and demonstrate the connection between the input and output variables. In that case, we might be eager to exchange some predictive power for a parametric curve that is increasingly reasonable. Another fundamental trade-off is that of variance versus bias [9].

CONCLUSION

The research survey looks at disease diagnosis using different machine learning methods, including regression, SVM, deep learning, decision trees, random forests, and K-NN. The study recommends comparing techniques to create an efficient algorithm for pattern classification for multiple classes. Machine learning algorithms can detect hidden patterns for classification from a large dataset, making them useful for medical diagnosis. The review aims to answer several research questions, such as existing techniques for disease diagnosis and problems solved using ML and DL algorithms. Machine learning is essential in detecting cancer diagnosis, and the timely automated prediction of specific patterns ensures better resource utilization in biomedical fields. The development of clustering, noise removal, and fuzzy rule-based algorithms for breast cancer disease identification is still required, nevertheless.

REFERENCES

[1] Cancer. Causes and Prevention-Risk Factors-Age and Cancer Risk. [Online], https://www.cancer.gov/about-cancer/causes-prevention/risk/, (2021), Accessed on : March 5, 2021.

[2] GLOBOSCAN. Cancer Fact Sheets-All cancers-Estimated number of new cases in 2020, world, both sexes, all ages. [Online]. https://gco.iarc.fr/today/data/factsheets/cancers/39-All-cancers-fact-sheet.pdf9.Technical Report, (2020), Accessed on: December 17, 2020.

[3] BREAST CANCER INDIA. The Latest Statistics of Breast Cancer in India. [Online] https://www.breastcancerindia.net/statistics/trends.html, (2020), Accessed on: May 22, 2020.

[4] SEER Data. SEER Incidence Data, 1975 – 2020. [Online], https://seer.cancer.gov/data/access.html, (2020), Accessed on: December 12, 2020.

[5] Alqudah A, and Alqudah AM. “Sliding Window Based Support Vector Machine System for Classification of Breast Cancer Using Histopathological Microscopic Images.” IETE Journal of Research 68, no. 1, (2022): 59-67.

[6] Devi R. R, and Anandhamala GS. “Recent Trends in Medical Imaging Modalities And Challenges for Diagnosing Breast Cancer.” Biomedical and Pharmacology Journal 11, no. 3, (2018):1649-1658.

[7] Araújo T, Aresta G, Castro E, Rouco J, Aguiar P, Eloy C, Polónia A, and Campilho A. “Classification of breast cancer histology images using convolutional neural networks.” PloS one 12, no. 6, (2017).

[8] Aresta G, Araújo T, Kwok S, Chennamsetty SS, Safwan M, Alex V, Marami B, Prastawa M, Chan M, Donovan M, and Fernandez G. “Bach: Grand challenge on breast cancer histology images.” Medical image analysis 56, (2019):122-139.

[9] Bandi P, Geessink O, Manson Q, Van Dijk M, Balkenhol M, Hermsen M, Bejnordi BE, Lee B, Paeng K, Zhong A, and Li Q. "From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge." IEEE transactions on medical imaging 38, no.2, (2018):550-60.

[10] Bardou D, Zhang K, and Ahmad SM. "Classification of breast cancer based on histology images using convolutional neural networks." IEEE Access 6. (2018):24680-24693.

[11] Bayramoglu N, Kannala J, and Heikkilä J. "Deep learning for magnification independent breast cancer histopathology image classification." In 2016 23rd International conference on pattern recognition (ICPR), IEEE, (2016):2440-2445.

[12] Boumaraf S, Liu X, Zheng Z, Ma X, and Ferkous C. "A new transfer learning based approach to magnification dependent and independent classification of breast cancer in histopathological images." Biomedical Signal Processing and Control 63, (2021):102192.

[13] Chugh, G., Kumar, S. and Singh, N., "Survey on Machine Learning and Deep Learning Applications in Breast Cancer Diagnosis," Cogn. Comput. 13, (2021):1451–1470.

[14] Debelee TG, Schwenker F, Ibenthal A, and Yohannes D. "Survey of deep learning in breast cancer image analysis." Evolving Systems 11, (2020):143-163.

[15] Dhahri H, Al Maghayreh E, Mahmood A, Elkilani W, and Faisal Nagi M. "Automated breast cancer diagnosis based on machine learning algorithms." Journal of healthcare engineering, (2019).

[16] Feng X, Zhang R, Liu M, Liu Q, Li F, Yan Z, and Zhou F. An accurate regression of developmental stages for breast cancer based on transcriptomic biomarkers. Biomarkers in Medicine 13, no. 01, (2019):5-15.

[17] Ghiasi MM, and Zendehboudi S. "Application of decision tree-based ensemble learning in the classification of breast cancer." Computers in biology and medicine 128, (2021):104089.

[18] Gupta D, Kose U, Khanna A, and Balas VE. "Deep Learning for Medical Applications with Unique Data." 1st ed. Academic Press, (2022).

[19] Gupta N, Pandey AK, Dimri K, Jyani G, Goyal A, and Prinja S. "Health-related quality of life among breast cancer patients in India." Supportive Care in Cancer 30, no. 12, (2022):9983-9990.

[20] Huang MW, Chen CW, Lin WC, Ke SW, and Tsai CF. "SVM and SVM ensembles in breast cancer prediction." PloS one 12, no. 1, 2017:0161501.

[21] M. M. Islam, H. Iqbal, M. R. Haque and M. K. Hasan, "Prediction of breast cancer using support vector machine and K-Nearest neighbors," 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), (2017):226-229.

[22] Kate RJ, and Nadig R. "Stage-specific predictive models for breast cancer survivability." International journal of medical informatics 97, (2017):304-11.

[23] Khan S, Islam N, Jan Z, Din IU, and Rodrigues JJ. "A novel deep learning based framework for the detection and classification of breast cancer using transfer learning." Pattern Recognition Letters 125, (2019):1-6.

[24] Khandezamin Z, Naderan M, and Rashti MJ. "Detection and classification of breast cancer using logistic regression feature selection and GMDH classifier." Journal of Biomedical Informatics 111, (2020).

[25] Krizhevsky A, Sutskever I, and Hinton GE. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25, (2012).

[26] Ladha L, and Deepa T. "Feature selection methods and algorithms." International journal on computer science and engineering 3, no. 5, (2011):1787-1797.

[27] Mehra R. "Breast cancer histology images classification: Training from scratch or transfer learning?." ICT Express 4, no. 4, (2018):247-254.

[28] Mishra J, Kumar B, Targhotra M, and Sahoo PK. "Advanced and futuristic approaches for breast cancer diagnosis.", Future Journal of Pharmaceutical Sciences 6, no. 1, (2020).

[29] MITOS-ATYPIA-14. MITOS-ATYPIA-14 Grand Challenge. [Online], https://mitos-atypia-14.grand-challenge.org/Dataset/, (2021), Accessed on: March 17, 2021.

[30] Mohammed MA, Al-Khateeb B, Rashid AN, Ibrahim DA, AbdGhani MK, and Mostafa SA. "Neural network and multi-fractal dimension features for breast cancer classification from ultrasound images." Computers & Electrical Engineering 70, (2018):871-882.

[31] Murtaza G, Shuib L, Abdul Wahab AW, Mujtaba G, Mujtaba G, Nweke HF, Al-garadi MA, Zulfiqar F, Raza G, and Azmi NA. "Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges." Artificial Intelligence Review 53, (2020):1655-1720.

[32] Nahid AA, Mehrabi MA, and Kong Y. "Histopathological breast cancer image classification by deep neural network techniques guided by local clustering." BioMed research international, (2018).

[33] Nahid AA, and Kong Y. "Histopathological breast-image classification using local and frequency domains by convolutional neural network." Information 9, no. 1, (2018).

[34] Napoleon D, and Pavalakodi S. "A new method for dimensionality reduction using k-means clustering algorithm for high dimensional data set." International Journal of Computer Applications 13, no. 7, (2011):41-6.

[35] Naseriparsa M, and Kashani MM. "Combination of PCA with SMOTE resampling to boost the prediction rate in lung cancer dataset." International Journal of Computer Applications 77, no. 3, (2013):33-38.

[36] Omondiagbe DA, Veeramani S, and Sidhu AS. "Machine learning classification techniques for breast cancer diagnosis." In IOP Conference Series: Materials Science and Engineering 2019 Jun 7, vol. 495, IOP Publishing, (2019):012033.

[37] Patlak M, Nass SJ, Henderson IC, and Joyce C. "Mammography and Beyond: Developing Technologies for the Early Detection of Breast Cancer: A Non-Technical Summary." Institute of Medicine (U.S.) and National Research Council (U.S.) Committee on the Early Detection of Breast Cancer; Washington (D.C.): The National Academies Press (U.S.), (2001).

[38] Peddireddy D, Fu X, Wang H, Joung BG, Aggarwal V, Sutherland JW, and Jun MB. "Deep learning based approach for identifying conventional machining processes from CAD data." Procedia Manufacturing 48. (2020):915-925.

[39] Polat K, and Güneş S. "Breast cancer diagnosis using least square support vector machine." Digital signal processing 17, no. 4, (2007):694-701.

[40] Pratiher S, and Chattoraj S. "Diving deep onto discriminative ensemble of histological hashing & class-specific manifold learning for multi-class breast carcinoma taxonomy." In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, (2019):1025-1029.

[41] Rasool A, Bunterngchit C, Tiejian L, Islam MR, Qu Q, and Jiang Q. "Improved machine learning-based predictive models for breast cancer diagnosis." International journal of environmental research and public health 19, no.6, (2022):3211.

[42] Sahu B, Mohanty S, and Rout S. "A hybrid approach for breast cancer classification and diagnosis." EAI Endorsed Transactions on Scalable Information Systems 6, no. 20, (2019).

[43] Sahu B. "A combo feature selection method (filter+ wrapper) for microarray gene classification." International Journal of Pure and Applied Mathematics 118, no. 16, (2018):389-401.

[44] Shallu, and Mehra R. "Automatic magnification independent classification of breast cancer tissue in histological images using deep convolutional neural network." In Advanced Informatics for Computing Research: Second International Conference, ICAICR 2018, Shimla, India, July 14–15, Springer Singapore, (2018):772-781.

[45] Shu X, Zhang L, Wang Z, Lv Q, and Yi Z. "Deep neural networks with region-based pooling structures for mammographic image classification." IEEE transactions on medical imaging 39, no. 6, (2020):2246-2255.

[46] Simonyan K, and Zisserman A. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556. (2014).

[47] Song R, Li T, and Wang Y. "Mammographic classification based on XGBoost and DCNN with multi features." IEEE Access 8. (2020):75011-75021.

[48] Spanhol FA, Oliveira LS, Petitjean C, and Heutte L. "Breast Cancer Histopathological Database (BreakHis) –LaboratórioVisãoRobótica e Imagem." [Online], (2016). http://www.inf.ufpr.br/vri/databases/BreaKHis_v1.tar.gz. 2016, Accessed on: July 2016.

[49] Spanhol FA, Oliveira LS, Cavalin PR, Petitjean C, and Heutte L. "Deep features for breast cancer histopathological image classification." In 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, (2017):1868-1873.

[50] Spanhol FA, Oliveira LS, Petitjean C, and Heutte L. "A Dataset for Breast Cancer Histopathological Image Classification." IEEE Transactions on Biomedical Engineering 63, no. 7, (2016):1455-1462.

[51] Spanhol FA, Oliveira LS, Petitjean C, and Heutte L. "Breast cancer histopathological image classification using convolutional neural networks." In 2016 international joint conference on neural networks (IJCNN), IEEE, (2016):2560-2567.

[52] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, and Bray F. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries." CA: a cancer journal for clinicians 71, no. 3, (2021):209-249.

[53] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, and Wojna Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE conference on computer vision and pattern recognition, (2016):2818-2826.

[54] Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, and Liang J. "Convolutional neural networks for medical image analysis: Full training or fine tuning?." IEEE transactions on medical imaging 35, no. 5, (2016):1299-1312.

[55] Vaka AR, Soni B, and Reddy S. "Breast cancer detection by leveraging Machine Learning." ICT Express 6, no. 4, (2020):320-324.

[56] Veta M, Heng YJ, Stathonikos N, Bejnordi BE, Beca F, Wollmann T, Rohr K, Shah MA, Wang D, Rousson M, Hedlund M, Tellez D, Ciompi F, Zerhouni E, Lanyi D, Viana M, Kovalev V, Liauchuk V, Phoulady HA, Graham TQ, Rajpoot N, Sjöblom E, Molin J, Paeng K, Hwang S, Park S, Jia Z, Chang EI, Xu Y, Beck AH, Diest PJ, and Pluim JPW. "Predicting breast tumor proliferation from whole-slide images: the TUPAC16 challenge." Medical Image Analysis 54. (2019):111-121.

[57] Wang Y, Choi EJ, Choi Y, Zhang H, Jin GY, and Ko SB. "Breast cancer classification in automated breast ultrasound using multiview convolutional neural network with transfer learning." Ultrasound in medicine & biology 46, no. 5, (2020):1119-1132.

[58] Wang Z, Li M, Wang H, Jiang H, Yao Y, Zhang H, and Xin J. "Breast cancer detection using extreme learning machine based on feature fusion with CNN deep features." IEEE Access 7, (2019):105146-105158.

[59] Wei B, Han Z, He X, and Yin Y. "Deep learning model based breast cancer histopathological image classification." In 2017 IEEE 2nd international conference on cloud computing and big data analysis (ICCCBDA), IEEE, (2017):348-353.

[60] White MC, Holman DM, Boehm JE, Peipins LA, Grossman M, and Henley SJ. *"Age and cancer risk: a potentially modifiable relationship."* American journal of preventive medicine 46, no. 3, (2014):7-15.

[61] Xiang Z, Ting Z, Weiyan F, and Cong L. *"Breast cancer diagnosis from histopathological image based on deep learning."*, In2019 Chinese Control And Decision Conference (CCDC), IEEE, (2019): 4616-4619.

[62] Yang Z, Ran L, Zhang S, Xia Y, and Zhang Y. "EMS-Net: Ensemble of multiscale convolutional neural networks for classification of breast cancer histology images." Neurocomputing 366, (2019):46-53.

[63] Zemmal N, Azizi N, Dey N, and Sellami M. "Adaptive semi supervised support vector machine semi supervised learning with features cooperation for breast cancer classification." Journal of Medical Imaging and Health Informatics 6, no. 1, (2016):53-62.