International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307

OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEIT2410274

# Breast Cancer Classification Using Machine Learning

**Ankit[1], Harsh Bansal[2], Dhruva Arora[3], Kanak Soni[4], Rishita Chugh[5], Swarna Jaya Vardhan[6]**

[1]Assistant Professor, Department of CSE, Lovely Professional University, Phagwara, Punjab, India

[2,3,4,5,6] B.Tech Scholar, Department of CSE, Lovely Professional University, Phagwara, Punjab, India

ARTICLEINFO

**Article History:**

Accepted: 10 April 2024

Published: 19 April 2024

**Publication Issue**

Volume 10, Issue 2

March-April-2024

**Page Number**

575-588

ABSTRACT

In the pursuit of precise forecasts in machine learning-based breast cancer categorization, a plethora of algorithms and optimizers have been explored. Convolutional Neural Networks (CNNs) have emerged as a prominent choice, excelling in discerning hierarchical representations in image data. This attribute renders them apt for tasks such as detecting malignant lesions in mammograms. Furthermore, the adaptability of CNN architectures enables customization tailored to specific datasets and objectives, enhancing early detection and treatment strategies. Despite the efficacy of screening mammography, the persistence of false positives and negatives poses challenges. Computer-Aided Design (CAD) software has shown promise, albeit early systems exhibited limited improvements. Recent strides in deep learning offer optimism for heightened accuracy, with studies demonstrating comparable performance to radiologists. Nonetheless, the detection of sub-clinical cancer remains arduous, primarily due to small tumor sizes. The amalgamation of fully annotated datasets with larger ones lacking Region of Interest (ROI) annotations is pivotal for training robust deep learning models. This review delves into recent high-throughput analyses of breast cancers, elucidating their implications for refining classification methodologies through deep learning. Furthermore, this research facilitates the prediction of whether cancer is benign or malignant, fostering advancements in diagnostic accuracy and patient care.

**Keywords :-** Breast Cancer Classification, Convolutional Neural Networks, Naïve Bayesian Classifier, k-Nearest Neighbors

## I. INTRODUCTION

A tumor, also known as a neoplasm, is an irregular mass of cells within the body, resulting from either excessive cell division or the failure of cells to undergo programmed death. Tumors are categorized as either benign or malignant. Detecting and analyzing breast cancer at an early stage significantly enhance the

chances of survival and reduce mortality rates. According to the American Cancer Society's data, it was projected that in 2020, there would be an estimated 327,610 cases diagnosed, including 276,480 cases of invasive breast cancer in women, 2,620 in men, and 48,530 cases of ductal carcinoma in situ among women. The estimated number of deaths for 2020 is around 42,690, comprising 42,690 women and 520 men. [1].

Breast ultrasound serves two main purposes: diagnostic and therapeutic. Diagnostic ultrasound, which is noninvasive, is primarily used for imaging purposes. On the other hand, therapeutic ultrasound doesn't generate images but is utilized for treatment purposes [2]. Breast cancer is the most common disease among women aged 20 to 59 years and the second most common cancer in the United States [3].

Benign tumors remain localized without spreading to other areas of the body, exhibiting slow growth and well-defined boundaries. While typically not problematic, they can grow large enough to compress nearby structures, leading to discomfort or medical issues. For instance, a sizable benign lung tumor might compress the windpipe, causing breathing difficulties and necessitating urgent surgical intervention. Once removed, benign tumors are unlikely to reoccur. Examples include fibroids in the uterus and lipomas on the skin. Some benign tumors have the potential to transform into malignant ones, requiring close monitoring and possibly surgical removal. Colon polyps, for instance, are commonly removed due to the risk of malignancy.

Malignant tumors consist of cells that proliferate uncontrollably and have the ability to spread locally or to distant locations. They are cancerous, invading surrounding tissues and potentially metastasizing to other parts of the body through the bloodstream or lymphatic system. Metastasis can occur in various organs, with common sites being the liver, lungs, brain, and bones. Due to their aggressive nature, malignant tumors often require prompt treatment to prevent further spread. Early detection typically involves surgical intervention, possibly followed by chemotherapy or radiotherapy. In cases where the cancer has already metastasized, systemic treatments like chemotherapy or immunotherapy are often administered [4]. The second main cause of women's death is breast cancer (after lung cancer)[5]. In the United States, it is projected that 246,660 new cases of invasive breast cancer will be diagnosed among women in 2016, with an estimated 40,450 female deaths attributed to the disease [6]. Breast cancer accounts for approximately 12% of newly diagnosed cancer cases overall and constitutes about 25% of all cancers diagnosed in women. [7]. Information and Communication Technologies (ICT) can play potential roles in cancer care. In fact, Big data has advanced not only the size of data but also creating value from it; Big data, that becomes a synonymous of data mining, business analytic and business intelligence, has made a big change in BI from reporting and decision to prediction results [8]. The rapid rise of data mining methodologies, particularly within the realm of medical science, is attributed to their exceptional performance in predicting outcomes, cutting medicine costs, enhancing patient health, elevating healthcare value and quality, and facilitating real-time decision-making crucial for saving lives.

## II. LITERATURE REVIEW

To attain precise forecasts in the field of machine learning-based breast cancer categorization, numerous algorithms and optimizer have been utilized. Convolutional Neural Networks (CNN's) are a popular algorithm that is used often. CNN's are excellent at recognizing hierarchical representations in image data, which makes them suitable for applications like detecting malignant lesions in mammograms. CNN's also provide versatility in terms of architecture,

enabling researchers to customize models for particular datasets and goals.

Random Forest is another frequently used algorithm. During training, the Random Forest ensemble learning approach builds a large number of decision trees and outputs the class that is the mean of the classes of the individual trees. It is well-suited for tasks involving the categorization of breast cancer due to its reputation for being resistant to over fitting and having the capacity to manage high-dimensional data efficiently.

Stochastic Gradient Descent (SGD) is a basic and widely used optimizer in the field of machine learning-based breast cancer classification because of its efficiency and ease of use. SGD makes iterative changes to the model parameters in order to minimize the loss, depending on the gradient of the loss function with respect to the parameters. On the other hand, the Adam optimizer has attracted a lot of interest in more complex problems, such breast cancer classification, where datasets may be high-dimensional and have subtle patterns. Compared to SGD, Adam—short for Adaptive Moment Estimation—offers several advantages. Based on previous gradients and their squared gradients, it adaptively modifies the learning rates for every parameter.Adam can more quickly and effectively navigate the optimization terrain thanks to this adjustable learning rate mechanism, which also improves performance, especially in deep learning models that are frequently used for breast cancer classification tasks. Furthermore, Adam's momentum term helps to escape local minima and smooth out the optimization process by acting as a memory of previous gradients, which speeds up convergence. Adam is a top option for optimizing intricate neural network topologies in the categorization of breast cancer because of these qualities, which will ultimately lead to greater generalization and more precise predictions. KNN is a straightforward but efficient technique that uses the majority class of its k nearest neighbour to classify data items. When it comes to classifying breast cancer, KNN can examine features that have been taken out of mammography pictures and determine a class label by comparing these features to those of nearby data points.

In contrast, SVM is a potent supervised learning algorithm that divides several classes of data points into distinct areas by building a hyperplane in a high-dimensional space. Because SVM can handle high-dimensional data and is flexible in selecting multiple kernel functions to capture intricate correlations between characteristics, it has been widely used in breast cancer classification applications.

To summaries, the incorporation of machine learning, namely deep learning, has the potential to improve the precision of breast cancer categorization. Through the application of innovative techniques like whole-image analysis and the resolution of data annotation issues, scientists hope to create more efficient CAD systems that will enhance the results of breast cancer screening. The use of Python and related modules makes it easier to apply and use these cutting-edge machine learning methods in medical imaging research.

## III. METHODOLOGY

In this study, we provide an approach that improves the accuracy of breast cancer categorization in screening mammography by utilizing machine learning techniques, specifically deep learning. Our first discussion will focus on the shortcomings of conventional screening mammography, which has a high proportion of false positives and false negatives. Although Computer-Aided Design and diagnosis (CAD) software has been around since the 1990s, performance was not greatly enhanced by the early systems. But fresh developments in deep learning have spurred enthusiasm once again in creating more useful tools to help radiologists.

Our approach is centred on the difficult problem of identifying breast cancer that is not yet clinical, in which the tumour only take up a small percentage of the mammography image. Scalability and generalization to bigger, unannotated datasets are limited by the reliance of traditional techniques on manually annotated regions of interest (ROIs). In order to tackle this, we put forth a novel strategy that integrates whole image analysis and patch-based classification, allowing for end-to-end training and minimizing the need for ROI annotations.

Fundamental to our approach is the transformation of patch-based classifiers into entire picture classifiers, which enables smooth integration and optimization inside a single framework. This methodology not only optimizes the training procedure but also enables transfer learning between datasets with different annotation levels. Patch classifiers can be efficiently trained by using publicly accessible databases with ROI annotations, such the Digital Database for Screening Mammography (DDSM). These classifiers are then modified for whole image analysis, which allows for the reliable categorization of mammograms even in the absence of precise ROI markings.

Python is a flexible and extensively used programming language in the machine learning community, and we use it to implement our methods. We pre-process and analyse mammography data using tools like NumPy and Pandas, identifying pertinent characteristics for classification. Essential methods for selecting and evaluating models are offered by Scikit-learn, including accuracy rating and train-test splitting. In addition, we use the flexibility and user-friendliness of Keras, a high-level neural network library, to create and train deep learning models.

All in all, our technique offers a thorough strategy for raising the accuracy of breast cancer classification in screening mammography. Our goal is to improve the accuracy of cancer screening procedures by combining deep learning methods with effective data processing and model validation. This will help CAD systems work better.

## IV. OBJECTIVE

1) Be mindful of potential cancer risks and prioritize proactive health measures.
2) Seek out and disclose the truth, fostering transparency and honesty in all endeavors.
3.)Take precautions to avoid emergencies by staying vigilant and prepared.
4) Provide informative resources and communication to empower informed decision-making and understanding.

One subset of artificial intelligence (AI) is machine learning (ML)that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so.

Machine learning algorithms use historical data as input to predict new output values.The extensive spread of faux news can have a significant negative impact on individuals and society. First, fake news can shatter the authenticity equilibrium of the news ecosystem for instance.Understanding the truth of new and message with news detection can create positive impact on the society.

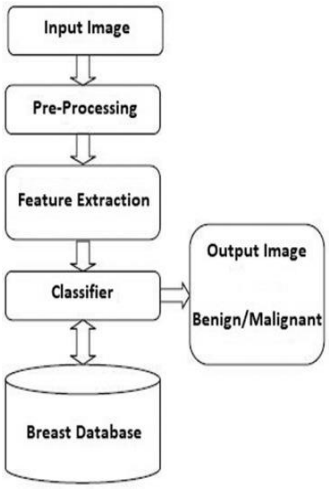


**Fig 1 : Workflow**

## V. LANGUAGE AND LIBRARIES USED

For our project, we have opted for Python, a widely recognized and extensively used language in machine learning. Python is characterized as an interpreted, object-oriented, high-level programming language with dynamic semantics. Its appeal lies in its built-in data structures, dynamic typing, and dynamic binding, making it ideal for Rapid Application Development and as a scripting or integration tool for connecting various components. Python's simplicity and readability in syntax contribute to reduced program maintenance costs. It promotes program modularity and code reuse through support for modules and packages.

Python's popularity among programmers often stems from its ability to enhance productivity. With no compilation step, the edit-test-debug cycle is notably swift. Debugging Python programs is straightforward as errors prompt exceptions rather than segmentation faults. The interpreter provides a stack trace when an exception goes uncaught. Additionally, Python offers a source-level debugger enabling variable inspection, expression evaluation, break-point setting, and line-by-line code traversal. Notably, the debugger itself is written in Python, highlighting Python's introspective capabilities. Conversely, adding print statements to the source code is often an effective and quick debugging method due to Python's fast edit-test-debug cycle.

A) Breast cancer classification (BCC)

1) Malignant (Types of breast cancer grows faster than normally and has irregular borders)
2) Benign (Types of breast cancer which grows smoothly, and has regular border)

Depending on the type of cancer, BCC attempts to identify the most appropriate course of action, which may involve more or less aggressive treatment.

Nine characteristics are needed for a breast cancer categorization to produce a good prognostic: 1. calculate the lump thickness, or layered architectures; 2. Assess the uniformity and sample size (Uniformity of Cell Size); 3. Determine the marginal variances and estimate the equality of cell shapes because cancer cells often have varying shapes (Uniformity of Cell Shape); 4. Normal cells are bonded to one another by marginal adhesion, while cancer cells proliferate throughout the organ; 5. Uniformity measurement: larger epithelial cells indicate cancer (Single Epithelial Cell Size); 6. The cytoplasm does not envelop the nuclei in benign tumors (Bare Nuclei); 7. Characteristics the texture of the nucleus; in benign cells, it has a consistent form. 8. The nucleolus is often tiny and inconspicuous in normal cells; the chromatin in tumors tends to be coarser (Bland Chromatin). There are several nucleoli in cancer cells, and they become considerably more noticeable (Normal Nucleoli); 9. An estimate of the total number of mitoses that have occurred. The greater the value, the higher the likelihood of cancer (mitoses)[6].

B) Methods of machine learning

Machine learning is branch of artificial intelligence, ML methods can employ statistics, probabilities, absolute conditionality, Boolean logic, and unconventional optimization strategies to classify patterns or to build prediction models [7]. Machine learning can be divided into two categories: supervised learning (classification) and unsupervised learning. Depending on the used data and their availability [8]. In this section, we will see two supervised learning classifiers.

1) Naïve Bayesian Classifier (NBC)

A Bayesian method is a basic result in probabilities and statistics, it can be defined as a framework to model decisions. In NBC, variables are conditionally independent; NBC can be used on data that directly influence each other to determine a model. From known training compounds, active (D) and inactive (H), Given representation B, the conditional probability distribution P(B/D) and P(B/H) are

estimated, respectively. Bayesian classifiers are additionally well adapted for ranking of compound databases all with consideration to probability of activity [9].

Bayesian classifiers use Bayes theorem, which is:

$$p(h|d) = p(d|h)p(h)/p(d) \quad (1)$$

In Eq. 1, P(h) is the priori probability that event h will occur. P(d) is the prior probability of the training data. The conditional probability of d when p (d | h) is given. P(h | d) is the conditional probability of h when given d training data. P (h | d) is the probability of generating instance d given class h. In the equation above Bayesian decision theorem is used to determine whether a given xi belongs to Si where Si represents a class [10]:

$$P(x|Si)P(Si) > P(x|Sj)P(Sj) \quad (2)$$

In the Eq. 2, j ≠ I which means that Si and Sj are two different classes and X belongs to Si..

2) KNN, or k-Nearest Neighbors

The KNN algorithm is used to predict the class or property of data. Given N training vector, suppose we have A and Z as training vectors in this bi-dimensional features space, we want to classify c which is feature vector. Classifying c depends on its k neighbors, and the majority vote, k is a positive integer, k is generally smaller then 5, if k=1 the class of c is the closest element from the two sets to c [11]. We use the Euclidean distances to evaluate the distance of a sample with other points,

Euclidean distance is given in equation 3.

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (3)$$

3) Random forest:

Random Forest (RF), proposed by Leo Breiman [12], is fast, highly accurate, noise resistant classification method. Bagging and random feature selection are combined together. Every tree in the forest is influenced by the values of random vectors sampled separately and has identical distribution as any other tree in the forest [12]. RF consists of outsized number of decision trees where decision tree select their separating features from bootstrap training set Si where i represent ith internal node. Trees in RF are grown by means of Classification and Regression Tree (CART) method with no pruning. As number of trees in the forest turns into outsized number, generalization error will also increase until it converges to some boundary level. More details about RF can be found in[12, 13, 14].

4) Decision Tree:

A decision tree is classifier understood as an instance space recursive partitioning. It is made up of nodes that form a directed tree with a node called "root" that has no incoming edges. Alternatively put, it is a tree that is rooted. There is only one incoming edge for the remaining nodes.Internal or test node is node which has outgoing edges. Decision nodes, or also known as leaves are all other nodes. In decision trees, all internal nodes divide the pattern space into more subspaces depending on a specific discrete function of the input attribute variables. All leaves are given to one class denoting the most suitable target value. Classification of patterns is performed by directing them from the tree root down to a leaf, based on the tests' outcome along the path [15].

Cross Validation

By dividing data into two sets—a testing set for model evaluation and a learning set for model training—cross-validation is a statistical approach that is commonly used to verify and assess learning algorithms or models.

In cross-validation, the training and testing sets are split into partitions at random (i.e., 60% of the data belong in the training sets and 40% in the testing sets). Subsequent crossover rounds ensure that every instance is evaluated against the training and testing sets. The most basic type of cross-validation is K-fold

cross-validation, where a validation set is one of the K partitions. With the k-fold as the foundation, there are more intricate variations of cross-validation. [16].

C) Machine Learning Libraries

1) NumPy: The core library for scientific computing in Python is called NumPy. It's a Python library that offers a multidimensional array object, different derived objects (like matrices and masked arrays), and a variety of routines for quick array operations, like sorting, choosing, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation, and much more. [17]

Pandas: pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive.It seeks to serve as the essential high-level building block for using Python to undertake useful, real-world data analysis. Its overarching objective is to become the most potent and adaptable open source data analysis and manipulation tool accessible in any language.It is already well on its way toward this goal.[18]

2) Sklearn: Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python.Through a Python consistency interface, it offers a range of effective tools for statistical modelling and machine learning, including as regression, clustering, classification, and dimensionality reduction. This library is based on NumPy, SciPy, and Matplotlib and is mostly developed in Python. [19]

3) Train Test Split: Split arrays or matrices into random train and test subsets. Quick utility that wraps input validation, next(ShuffleSplit().split(X, y)), and application to input data into a single call for splitting (and optionally subsampling) data into a one-line[20].

4) Accuracy Score: The accuracy is calculated using the accuracy_score function, which returns the count (normalize=False) or the fraction (default) of accurate predictions.The subset accuracy is returned by the function in multilabel classification. The subset accuracy is 1.0 if all of the predicted labels for a sample exactly match the true labels; otherwise, it is 0.0 [21].

$$\texttt{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)$$

5) Keras: Keras is a deep learning API written in Python and capable of running on top of either JAX, TensorFlow, or PyTorch.[22]

## VI. TRAINING AND VALIDATION RESULT

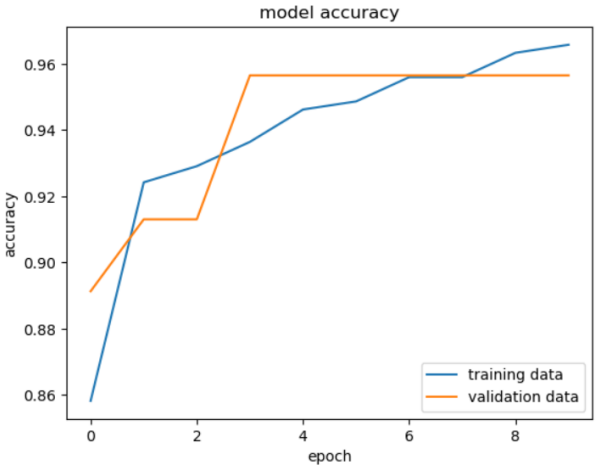


**Fig 2: Accuracy result**

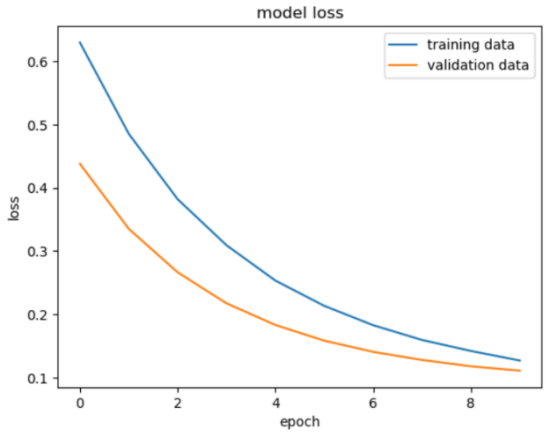


**Fig 3: Model loss result**

## VII. RELATED WORKS

Many studies in the subject of Ml and BCC have been conducted; however, some of these studies used mammography images, which have the drawback of missing approximately 15% of cases of breast cancer.[23], some techniques are more specific and

used genome or phenotypes to do classification [24, 25]. Several methods, including the Softmax Discriminant Classifier (SDC) and Linear Discriminant Analysis (LDA), are used to classify breast cancer [26], and Fuzzy C Means Clustering [27]. The knearest neighbors algorithm is one of the most used algorithms in machine learning [28, 29]. Before classifying a new element, we must compare it to other elements using a similarity measure [23]. In cancer classification, KNN can be used to measure the performance of false positive rates [30,31]. Naïve Bayesian classifiers are generally used to predict biological, chemical and physiological properties. In cancer classification, NBC are sometimes combined to other classifiers such as decision tree to determine prognostics or classification models. Different classification techniques were developed for breast cancer diagnosis, the accuracy of many of them was evaluated using the dataset taken from Wisconsin breast cancer database [32]. For example, in [33] the optimized learning vector method's performance was 96.7%, big LVQ method reached, SVM for cancer diagnosis's accuracy is 97.13% is the highest one in the literature .

## VIII. TRADITIONAL BREAST CANCER MORPHOLOGIC CLASSIFIERS

In routine practice for early-stage breast cancer, the traditional prognostic factors currently utilized include TNM staging information and histologic grade. Additional variables found in most breast cancer synoptic reports, such as tumor histologic type, lymphovascular channel invasion, tumor focality, and features of associated in situ disease, along with margin status, completeness of excision, patients' age, family history, and menopausal status, are considered. However, it's important to note that these factors may not always impact decisions related to systemic therapy.

A) Lymph Nodes

Histologically confirmed loco-regional lymph node (LN) status consistently emerges as the predominant prognostic factor in early-stage/operable breast cancer. Over a 10-year span, 15% to 30% of patients lacking nodal involvement may face recurrence, while approximately 70% of those with axillary nodal engagement are susceptible. Prognosis is further influenced by the absolute count of positive nodes, with a higher count linked to diminished patient survival, as discerned through Histologically examination. Furthermore, involvement of nodes in the upper axillary levels, particularly the apex, and the internal mammary nodes, is associated with a less favorable prognosis.

Further refinement of lymph node (LN) staging can be achieved by taking into account the size of metastatic deposits and the ratio of positive nodes to the total number of harvested nodes. In symptomatic breast cancer (BC), approximately one-third (ranging from 30% to 40%) of operable BC patients present with positive nodes, among them 7% to 15% have more than three positive nodes. The prevalence of node positivity diminishes in patients identified through established breast mammography screening programs, dropping to levels below 20%.

B) Tumor Size

The size of a tumor emerges as a significant predictor of its behavior in breast cancer (BC). Tumor size correlates directly with the likelihood of nodal metastases. For tumors smaller than 1.0 cm, nodal positivity occurs in 10% to 20% of cases, increasing to 40% at 2.0 cm, and reaching 50% for tumors exceeding 2.0 cm. Tumors under 1.0 cm exhibit a commendable 10-year disease-free survival rate of around 90%, which diminishes to 75% for tumors measuring 1 to 2 cm, and further drops to 60% for those ranging from 2 to 5 cm. Tumors exceeding 5 cm may warrant consideration for systemic therapy, with or without local control. Accurate assessment of tumor size is vital for appropriate patient stratification, particularly with

the rising prevalence of pT1 cancers due to screening mammography. When evaluating prognosis, tumor size should be determined exclusively from pathologic specimens, as clinical measurement is notoriously unreliable. Clinical assessment of tumor size, supplemented by ultrasonic measurement, may be undertaken for preoperative therapeutic planning.

C)Tumor Differentiation

Invasive breast carcinomas are presently categorized morphologically according to their growth patterns and degree of differentiation, reflecting their similarity to normal breast epithelial cells. This classification involves evaluating histologic type and histologic grade. Unlike stage variables, the assessment of tumor differentiation brings qualitative distinctions, functioning as an intrinsic biological prognostic factor rather than a time-dependent one. Furthermore, it furnishes crucial prognostic and predictive information, particularly for tumors within similar staging categories.

Histologic tumor grade is a classification method that hinges on the level of differentiation observed in tumor tissue and is universally applicable. In the realm of breast cancer (BC), it involves a semi-quantitative assessment of morphologic characteristics, offering a straightforward and cost-effective approach. Adequate tissue fixation and examination of high-quality hematoxylin-eosin (H&E)-stained tumor tissue sections by a trained pathologist, following a standard protocol, are prerequisites for this evaluation. The Nottingham Grading System (NGS) is employed for histologic grading, focusing on three pivotal biology-dependent morphologic features: (i) the degree of tubule or gland formation, (ii) nuclear pleomorphism, and (iii) mitotic count. This grading system provides a morphologic assessment of tumor biological characteristics and has proven effective in furnishing crucial information about the clinical behavior of BC. The clinical and biological relevance of NGS is underscored by genome-wide microarray-based expression profiling studies, suggesting that the features encapsulated by histologic grade significantly influence tumor behavior. The independent prognostic value of NGS has been consistently validated across multiple independent studies.

In early-stage breast cancer (BC), the prognostic value of the Nottingham Grading System (NGS) matches that of lymph node (LN) status and exceeds that of tumor size. However, its significance becomes more prominent in specific BC subgroups where determining the need for adjuvant chemotherapy is crucial. This is notably evident in patients with LN-negative ER-positive/HER2-negative or those with low-volume LN metastatic disease (pN1), where decisions regarding chemotherapy cannot rely solely on the risk associated with a more advanced tumor stage. Beyond its impact on patient outcomes, NGS is associated with other clinicopathologic prognostic variables like LN stage, tumor size, vascular invasion (VI), and the expression of biomarkers with prognostic and predictive value, such as hormone receptors.

D)Other Morphologic Variables

While the assessment of lymphovascular invasion in breast cancer (BC) is still a subject of debate, with some authors not finding a significant correlation with clinical outcomes, various independent studies highlight its role in predicting both recurrence and long-term survival. Additionally, it has been identified as a predictor of axillary lymph node metastasis and early recurrence in patients without lymph node involvement.

It emerges as a valuable tool in pinpointing a subgroup of axillary node-negative patients with an unfavorable prognosis, potentially benefiting from adjuvant chemotherapy. Our investigation within the Nottingham series revealed that the presence of VI, as assessed in routine H&E sections in the node-negative patient cohort, holds prognostic significance for both recurrence development and survival. Notably, this significance closely aligns (without statistical difference) with that observed in patients with 1 or 2

positive LNs (unpublished observation). The inclusion of VI in the St. Gallen criteria for selecting adjuvant systemic therapy in operable breast cancer (BC) underscores its importance. Furthermore, VI serves as a crucial predictor of local recurrence in patients treated with breast conservation and guides decisions on the use of radiotherapy.

However, it's essential to acknowledge challenges in assessing VI in routine H&E sections, particularly in identifying VI and distinguishing true vessels from artefactual soft tissue spaces. These challenges contribute to the wide variation in the reported frequency of VI in the literature, ranging from 20% to 54%. While immunohistochemical (IHC) detection of VI may offer a more objective alternative, its application in routine practice remains a subject of debate.

Several other morphologic features of breast carcinoma, although proposed as prognostic factors, carry relatively less significance but can be examined using traditional histopathologic methods. These features encompass angiogenesis, tumor necrosis, tumor-associated inflammation, and the presence and extent of ductal carcinoma in situ associated with invasive carcinomas. Traditional histopathologic assessment of tumors also encompasses margin status, completeness of excision, and tumor focality, all crucial in guiding local control strategies for BC and determining the necessity for additional surgery or local radiotherapy. These considerations gain prominence with the widespread use of wide local excision and the increased adoption of oncoplastic surgery, especially in the context of screen-detected early-stage tumors.

D) Traditional Molecular Classifiers

Traditional molecular factors that guide predictions and prognoses in early-stage breast cancer (BC) include the status of hormone receptors (HR) and HER2. These factors are crucial components of the diagnostic workup for all BC patients, with routine determination using standardized techniques and established guidelines. The current emphasis in treatment decisions lies in assessing endocrine responsiveness. Adjuvant endocrine therapy constitutes a significant portion, contributing to almost two-thirds of the overall benefit in patients with HR-positive BC. For low-risk cases characterized by endocrine-responsive HR-positive disease, primary therapy involves endocrine treatment. In contrast, high-risk cases with uncertain endocrine response necessitate a combination of endocrine therapy and chemotherapy, while HR-negative, endocrine non-responsive disease is treated with chemotherapy alone. The utilization of anti-HER2 therapy depends on risk stratification and the HER2 status of the tumor.

Since the mid-1970s, determining the ER status has been a crucial aspect of the clinical management of breast cancer (BC), serving as both an indicator of endocrine responsiveness and a prognostic factor for early recurrence. The established gold standard for assessing ER status involves immunohistochemistry (IHC) conducted on formalin-fixed, paraffin-embedded cancer tissue. While this diagnostic test is routinely employed in clinical settings, and major therapeutic decisions rely on its results, its reliability is not absolute. Existing IHC assays have reported only modest positive predictive values (30% to 60%) for responses to single-agent hormonal therapies. However, the negative predictive value of ER expression is substantial; in other words, ER negativity, found in 20% to 30% of BC cases, effectively identifies patients unlikely to benefit from endocrine therapy. Therefore, it is crucial to identify variables that can accurately pinpoint patients who can safely forego adjuvant therapy or those who might benefit from hormone therapy alone or in combination with chemotherapy and/or targeted therapy.

The progesterone receptor (PR), being an estrogen-regulated gene, is thought to signify a functioning ER pathway. However, approximately 40% of ER-positive

tumors lack PR expression. The absence of PR in ER-positive tumors may serve as a surrogate marker for aberrant growth factor signaling, potentially contributing to tamoxifen resistance. Generally, ER+/PR tumors are considered less responsive than ER+/PR+ tumors. PR status proves valuable in predicting the response to hormone treatment, both in patients with metastatic disease and in the adjuvant setting. Numerous studies have provided evidence supporting the prognostic and predictive importance of assessing PR in BC.

Amplification of the HER2 gene is detected in 13% to 20% of breast cancer (BC) cases, with more than half (approximately 55%) of these cases being hormone receptor (HR)-negative. Numerous studies indicate that HER2 gene amplification or protein overexpression is indicative of a poor prognosis and predicts the response to systemic chemotherapy. Following the development of a humanized monoclonal antibody against HER2 and clinical trials demonstrating the advantages of anti-HER2 agents in HER2-positive BC patients, the importance of determining HER2 status in routine clinical practice has evolved. It is now a prerequisite for the clinical use of anti-HER2 agents in patients with HER2-positive advanced disease and in the adjuvant setting for HER2-positive early-stage disease. Routine assessment of HR and HER2 aims to provide information on the response to endocrine therapy and anti-HER2-targeted therapy, respectively. However, the biomarker expression of HR and HER2 often overlaps, and their prognostic and predictive value can be enhanced by considering them in combination.

Most immunohistochemistry (IHC) studies have employed a combination of estrogen receptor (ER), progesterone receptor (PR), and HER2 as IHC surrogates to categorize the molecular classes initially identified by gene expression profiling (GEP). For example, ER/PR positivity is used as a surrogate for the luminal class, HER2 expression for HER2-positive tumors, and the triple-negative phenotype (ER–, PR–, HER2–) to define the basal-like molecular class. Some authors have further classified HR-positive tumors that are also HER2-positive as the luminal B subclass. Thus, the assessment of ER, PR, and HER2 statuses serves as a readily accessible biological molecular classifier of BC with well-defined prognostic and predictive value. Their combined use offers a practical surrogate for GEP-defined molecular classes. Additionally, incorporating other established IHC markers, such as proliferation-associated markers, may contribute additional prognostic and predictive value to existing classification systems. The IHC expression of Ki67 is widely employed as an objective molecular measure of proliferation, addressing challenges related to tumor fixation and identification of mitotic figures.

## IX. CONCLUSION

This review underscores the efficacy of machine learning, especially deep learning methods like Convolutional Neural Networks (CNNs), in accurately categorizing breast cancer, including benign and malignant types. It addresses challenges in traditional screening mammography and proposes a novel approach combining whole image analysis and patch-based classification to mitigate reliance on manually annotated regions of interest (ROIs). Utilizing Python and libraries such as NumPy, Pandas, and Scikit-learn, the study emphasizes model development and evaluation. Overall, the review highlights the potential for machine learning to enhance diagnostic accuracy, contribute to CAD system improvements, and ultimately improve patient care in screening for benign and malignant breast cancer types.

## X. REFERENCES

[1]. American Cancer Society, Cancer Facts & Figures 2020, no. 4, American Cancer Society, Atlanta, 2020.

[2]. A.Sennoga, “Ultrasound imaging,” in Bioengineering Innovative Solutions for Cancer, pp. 123–161, Academic Press, 2020.

[3]. R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," CA: a cancer journal for clinicians, vol. 69, no. 1, pp. 7-34, 2019.

[4]. Aisha Patel, MBBS, MRCP . Howard (Jack) West, MD.London North West University Healthcare NHS Trust, London, United Kingdom.

[5]. U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999–2008 Incidence and Mortality Web-based Report. Atlanta (GA): Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute; 2012.

[6]. Siegel RL, Miller KD, Jemal A. Cancer Statistics , 2016. 2016;00(00):1-24. doi:10.3322/caac.21332.

[7]. “Globocan 2012 - Home.” [Online]. Available: http://globocan.iarc.fr/Default.aspx. [Accessed: 28-Dec-2015].

[8]. Asri H, Mousannif H, Al Moatassime H, Noel T. Big data in healthcare: Challenges and opportunities. 2015 Int Conf Cloud Technol Appl. 2015:1-7. doi:10.1109/CloudTech.2015.7337020.

[9]. L. Adi Tarca, V.J.C., X. Chen, R. Romero, S. Drăghici, "Machine Learning and Its Applications to Biology", PLoS Comput Biol,, Vol. 3, pp. 116- 122, 2007.

[10]. JF McCarthy, M.K., PE Hoffman, "Applications of machine learning and high-dimensional visualization in cancer detection, diagnosis, and management", Ann N Y Acad Sci, Vol.62, pp. 10201259, 2004.

[11]. AC. Tan, D. Gilbert, "Ensemble machine learning on gene expression data for cancer classification", Appl. Bioinform, Vol. 2, pp. 75-83, 2003.

[12]. S. Kanta Sarkar, A.N., "Identifying patients at risk of breast cancer through decision trees", International Journal of Advanced Research in Computer Science. Vol. 08, pp. 88-96, 2017.

[13]. JA. Cruz, W.D, "Applications of Machine Learning in Cancer Prediction and Prognosis". Cancer Inform, Vol. 2, pp. 56-77, 2006.

[14]. M. Sugiyama, "Introduction to Statistical Machine Learning "1ed, ed. T. Green: Morgan Kaufmann, 2006.

[15]. L. Breiman, "Random Forests," Machine Learning, vol. 45, p. 5–32, 2001.

[16]. E. Alickovic and A. Subasi, "Medical Decision Support System for Diagnosis of Heart Arrhythmia using DWT and Random Forests Classifier," Journal of Medical Systems, vol. 40, no. 108, 2016.

[17]. E. Alickovic and A. Subasi, "Breast cancer diagnosis using GA feature selection and Rotation Forest," Neural Computing and Applications, pp. 1-11, 2015.

[18]. L. Rokach and O. Maimon, Data Mining and Knowledge Discovery Handbook, 2nd ed., M. Oded and R. Lior, Eds., New York: Springer, 2010.

[19]. A. Lavecchia, "machine-learning approaches in the context of ligand-based virtual screening for addressing complex compound classification problems and predicting n

[20]. https://books.google.co.in/books?hl=en&lr=&id=c-xzkDMDev0C&oi=fnd&pg=PR2&dq=numpy&ots=Z8PKCWqeuh&sig=Gb7ti9uUKPFNrhYh3mCKJWHv51c&redir_esc=y#v=onepage&q=numpy&f=false

[21]. https://pandas.pydata.org/pandas-docs/version/0.7.3/pandas.pdf

[22]. https://www.tutorialspoint.com/scikit_learn/scikit_learn_introduction.htm

[23]. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

[24]. https://scikit-learn.org/stable/modules/model_evaluation.html#accuracy-score

[25]. https://keras.io/about/

[26]. P. Baldi, S.R.B., Bioinformatics: The machine learning approach. 2 ed, ed. S.r.B. Pierre Baldi, 2001.

[27]. N. Bhatia, "Survey of Nearest Neighbor Techniques", International Journal of Computer Science and Information Security, Vol. 8, No. 2, 2010.

[28]. A. Francillon, P.R., "Smart Card Research and Advanced Applications": 12th International Conference, CARDIS 2013, Berlin, Germany, 2013.Revised Selected Papers. 1 ed. Lecture Notes in Computer Science 8419 Security and Cryptology2014: Springer International Publishing, November 27-29.

[29]. A. Alarabeyyat, A.M., "Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm", in 9th International Conference on. IEEE, v.i.e.E. (DeSE), pp. 35-39, 2016.

[30]. MF. Akay. "Support vector machines combined with feature selection for breast cancer diagnosis". Expert Syst Appl Vol. 36, Issue. 2, Part. 2, pp. 3240-3247, March 2009.

[31]. S.K. Prabhakar, H. Rajaguru, "Performance Analysis of Breast Cancer Classification with Softmax Discriminant Classifier and Linear Discriminant Analysis", In: Maglaveras N., Chouvarda I., de Carvalho P. (eds) Precision Medicine Powered by pHealth and Connected Health. IFMBE Proceedings, vol 66. Springer, Singapore, 2018.

[32]. J. S. Snchez, R.A.M., J. M. Sotoca. "An analysis of how training data complexity affects the nearest neighbor classifiers", Pattern Analysis and Applications, Vol. 10, Issue 3, pp 189–201, August 2007.

[33]. M. Raniszewski, "Sequential reduction algorithm for nearest neighbor rule", Computer Vision and Graphics, 2010.

[34]. P.BhuvaneswariaA, B. Therese, "Detection of Cancer in Lung with K-NN Classification Using Genetic Algorithm", Procedia Materials Science, Vol. 10, pp. 433-440, 2015.

[35]. Z. Zhou, Y.J., Y. Yang, S.F. Chen, "Lung Cancer Cell Identification Based on Artificial Neural Network Ensembles Artificial Intelligence", Medicine Elsevier, Vol. 24, pp. 25-36, 2002.

[36]. A. Pradesh, A.o.F.S.w.C.B.C.D.," Indian J. Comput. Sci. Eng., vol. 2, no. 5, pp. 756–763, 2011.

[37]. Mook S, Schmidt MK, Rutgers EJ, et al. Calibration and discriminatory accuracy of prognosis calculation for breast cancer with the online Adjuvant! program: a hospital-based retrospective cohort study. Lancet Oncol. 2009;11:1070–1076

[38]. Goldhirsch A, Ingle JN, Gelber RD, et al. Thresholds for therapies: highlights of the St. Gallen International Expert Consensus on the primary therapy of early breast cancer 2009. Ann Oncol 2009;8:1319–1329.

[39]. Hayes DF, Bast RC, Desch CE, et al. Tumor marker utility grading system: a framework to evaluate clinical utility of tumor markers. J Natl Cancer Inst. 1996;20:1456–1466.

[40]. Rakha EA, El-Sayed ME, Powe DG, et al. Invasive lobular carcinoma of the breast: response to hormonal therapy and outcomes. Eur J Cancer. 2008;1:73–83.

[41]. Westenend PJ, Meurs CJ, Damhuis RA, Tumour size and vascular invasion predict distant metastasis in stage I breast cancer: grade distinguishes early and late metastasis. J Clin Pathol. 2005;2:196–201.

[42]. Ichikura T, Tomimatsu S, Okusa Y, et al. Comparison of the prognostic significance between the number of metastatic lymph nodes and nodal stage based on their location in patients with gastric cancer. J Clin Oncol. 1993;10:1894–1900.

[43]. Allred DC, Carlson RW, Berry DA, et al. NCCN Task Force Report: Estrogen Receptor and Progesterone Receptor Testing in Breast Cancer by Immunohistochemistry. J Natl Compr Canc Netw. 2009:S1–S21; quiz S22-S23

[44]. Pathology reporting of breast disease. A Joint Document Incorporating the Third Edition of the NHS Breast Screening Programme's Guidelines for Pathology Reporting in Breast Cancer Screening and the Second Edition of The Royal College of Pathologists' Minimum Dataset for Breast Cancer Histopathology. January 2005. NHSBSP Pub. No 58.

[45]. Harris L, Fritsche H, Mennel R, et al. American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. J Clin Oncol. 2007;33:5287–5312.

[46]. Ma XJ, Salunga R, Tuggle JT, et al. Gene expression profiles of human breast cancer progression. Proc Natl Acad Sci U S A. 2003;10:5974–5975979.

[47]. Warwick J, Tabar L, Vitak B, et al. Time-dependent effects on survival in breast carcinoma: results of 20 years of follow-up from the Swedish Two-County Study. Cancer. 2004;7:1331–1336.

[48]. Balslev I, Axelsson CK, Zedeler K, et al. The Nottingham Prognostic Index applied to 9,149 patients from the studies of the Danish Breast Cancer Cooperative Group (DBCG). Breast Cancer Res Treat. 1994;3:281–290

[49]. Smith JA III, Gamez-Araujo JJ, Gallager HS, et al. Carcinoma of the breast: analysis of total lymph node involvement versus level of metastasis. Cancer. 1977;2: 527–532.

[50]. Reed J, Rosman M, Verbanac KM, et al. Prognostic implications of isolated tumor cells and micrometastases in sentinel nodes of patients with invasive breast cancer: 10-year analysis of patients enrolled in the prospective East Carolina University/Anne Arundel Medical Center Sentinel Node Multicenter Study. J Am Coll Surg. 2009;3:333–340.

[51]. Putti TC, El-Rehim DM, Rakha EA, et al. Estrogen receptornegative breast carcinomas: a review of morphology and immunophenotypical analysis. Mod Pathol. 2005;1:26–35.

[52]. Wirapati P, Sotiriou C, Kunkel S, et al. Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. Breast Cancer Res. 2008;4:R65.

[53]. Mohammed RA, Martin SG, Mahmmod AM, et al. Objective assessment of lymphatic and blood vascular invasion in lymph node-negative breast carcinoma: findings from a large case series with long-term follow-up. J Pathol. 2011;3:358–365.