IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ჯ Peer-reviewed & Refereed journal ჯ Vol. 12, Issue 8, August 2025

DOI: 10.17148/IARJSET.2025.12807

CNN-Aided Hybrid Clustering for Enhanced Detection of Lung and Breast Cancer

Shivangi Dubey[1], Prof. Vineeta Singh[2], Rajat Kumar Pachauri[3]

Research Scholar, Department of Statistics, Institute of Social Sciences, Dr. Bhimrao Ambedkar University, Agra, Uttar Pradesh, India[1]

Professor, Department of Statistics, Institute of Social Science, Dr. Bhimrao Ambedkar University, Agra, Uttar Pradesh, India[2]

Research Scholar, Department of Statistics, Institute of Social Sciences, Dr. Bhimrao Ambedkar University, Agra, Uttar Pradesh, India[3]

Abstract: Accurate diagnosis of lung and breast cancer is crucial for effective patient treatment and management. This study presents a novel framework that integrates hybrid clustering and Convolutional Neural Network (CNN) based classification for improved diagnosis of lung and breast cancer. The integration of hybrid clustering allows for the identification of intricate patterns within the lung and breast cancer datasets, while CNN ensures effective feature extraction and classification. The results verified the effectiveness of the proposed approach in accurately clustering and classifying lung and breast cancer instances. Classification results reveal a high level of accuracy for both lung and breast cancer datasets, with lung cancer achieving an accuracy score of 0.9847 and breast cancer reaching an accuracy score of 0.9986. Precision, recall, and F1 scores further validate the robustness of the approach. The proposed approach demonstrates promising potential for accurate cancer diagnosis and prognosis.

Keywords: Breast Cancer, Lung Cancer, Clustering, Classification, Data Mining

1. INTRODUCTION

On a global scale, cancer is presently responsible for one out of every six fatalities, making it a serious and rapidly expanding public health concern [1]. There might be an increase in both the number of new cases, around 18.1 million and fatalities, around 9.6 million, in the next decades. As a result of improvements in disease control and longer life expectancy, non-communicable chronic diseases, including cancer, have emerged as a major public health concern in the current epidemiological shift. It is estimated that cancer would account for almost 30% of all fatalities caused by non-communicable diseases [2].

Although cancer could impact everyone, it is essential to examine the effects based on the significant healthcare disparities that exist globally. Where the population is mostly vulnerable, access to healthcare would almost surely increase the cancer death rate. Access to medical treatment is still a major issue, and diagnoses are often established at the late stages of the disease. Figure 1 shows the types of cancer-based on worldwide occurrence.

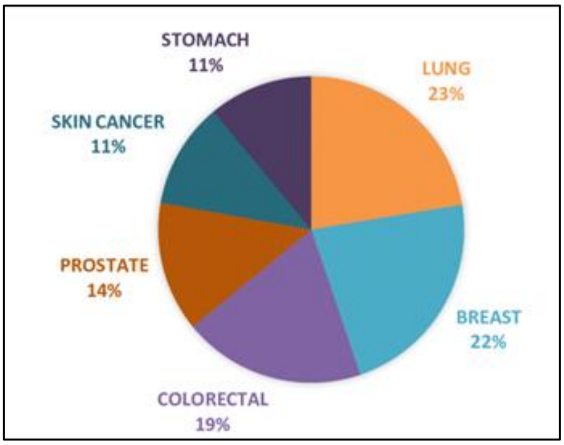

Figure 1. Leading cancer cases globally [3]

An increase in advanced-stage illness and death might result from delayed diagnosis and treatment [4]. Hence, the identification of cancer at an early stage needs to be the primary focus of very specific attention [5].

When it comes to detecting abnormalities in various organs of the body, such as skin cancer [6-8], blood cancer [9]-[10], brain tumour [11]-[12], breast cancer [13-15], retina [16] and lung cancer [17]-[18] and so on, early cancer diagnosis plays a vital role. Worldwide, tumours are the main cause of mortality, and organ anomalies almost always play a role in their fast growth [19]. Nearly 18.1 × 106 new instances of cancer were detected in 2018, leading to 9.6 × 10^6 cancer-related deaths, as reported by Global Cancer Observatory: Cancer Today (GLOBOCAN) [20]. The GLOBOCAN reported that the most common cause of death is lung cancer, accounting for around 18%, followed by breast cancer, which accounts for 6.6% of cases. Additionally, the study emphasizes the fact that more than ½ of all cancer fatalities occur in Asia, while just 23% of cancer death cases have been reported in Europe. Multiple methods are often used to investigate abnormalities in human organs, including magnetic resonance imaging (MRI) [21], mammography [22], computed tomography (CT) [23] and so on [24–28].

Hence, in this study, the authors established an integrated clustering and classification system for precise cancer grouping for breast and lung cancers and detailed the correlation between cancer patterns and different risk variables. Clustering is a method for dividing datasets into smaller sets defined by shared characteristics. Data items that are the same within a cluster but different from one another outside of it are called clusters [29]-[30]. A new approach to determining if cancer is present in a given patient is developed in this study by combining data mining techniques with hybrid clustering algorithms. Frequent occurrences of data and item sets in the database are called frequent patterns. Significant frequent patterns are those patterns that are most strongly associated with certain cancer types and may be used to forecast both the disease and its kind. The CNN is utilized for extracting these significant features. These significant patterns are further used to cluster the data set appropriately.

Since lung and breast cancer are the most prevalent causes of mortality from cancer worldwide, it is essential to come up with a categorization system for each of these kinds of cancer. So, the next part would discuss the different approaches that were utilized for recognizing lung and breast cancer.

1.1 Lung cancer detection

In order to provide patients with an increased chance of survival, it is essential to diagnose lung cancer at an early stage. Possible indications of cancer include the presence of a nodule in the lungs. A nodule can be either benign or malignant [31]. A nodule is an item that has a spherical shape. There is a quick development of the malignant nodule, and the rapid growth of malignant nodules may also affect other organs at this time.

For this reason, it is essential to have malignant nodules treated from the beginning of the process. When it comes to identifying lung cancer, the CT scan is the diagnostic method that is mostly used. Detecting abnormal spots inside the lung requires further studies to be performed after a CT scan [32]-[33]. Several studies about the identification and categorization of lung nodules have been conducted, some of which are given below.

A deep learning (DL) based model that might detect lung cancer on chest radiographs was suggested and tested by Shimazaki et al. [34] using the segmentation technique. The study used mean false positive indications (mFPI) to assess the performance of the DL-based model while training and validating it using a five-fold cross-validation method. The DL-based model was found to be sensitive to change in as small a value as 0.73 mFPI, thereby affecting the identification of lung tumours using chest radiography. Agarwal et al. [35] used CNN along with the AlexNet Network Model to classify lung tumours, which is one of the transfer learning models. The suggested CNN has a higher accuracy than traditional neural network systems do, thus being the best option available.

Naqi et al. [36] suggest that there are four stages of detecting and classifying nodules. These stages include the extraction of the lung area, identification of potential nodules, development of a feature descriptor based on hybrid geometric and textural features, and lastly, use of deep learning for feature reduction and classification. The approach proposed has a sensitivity of 95.6% and a notable reduction in false positives to 2.8 per scan. This study emphasizes the necessity for automated lung nodule detection and classification.

Asuntha and Srinivasan [37] introduced an advanced deep-learning approach for the detection of lung nodules by using a combination of several feature extraction methods to extract attributes. After extracting features, the best one is identified using the flexible Particle swarm optimization (FPSO) algorithm. The last technique used to group these traits is dark leather. Recent FPSO-CNN algorithms simplify CNN computations.

The technique presented by Tahoces et al. [38] is an extension of their previous works for the precise measurement of the aortic lumen's 3D geometry from an initial contour inside, using a simple incremental approach. It is shown that the proposed method consistently outperforms traditional techniques on entire datasets and 3D sections of 16 CT instances, where it gives an average accuracy of 0.951. In addition to common cases, the suggested method has great precision allowing it to be applied even for rare scenarios.

An improved method for automatically detecting pulmonary nodules using CT scans was suggested by Xie et al. [39]. The system makes use of a 2D-CNN. Each network's output is combined to get the final categorization. A sensitivity of 86.42% in detecting nodule candidates was achieved after extensive trials on the Lung Nodule Analysis (LUNA 16) dataset. The suggested approach proved that it was possible to achieve precise lung nodule identification.

Shen et al. [40] introduced a method for dividing lung nodules into two categories: those that require high suspicion and those that do not. By applying maximum pooling times and cutting areas from convolutional feature maps, this approach utilizes the Multi-Crop Convolution Neural Network (MC-CNN) to extract significant information from nodules. The proposed approach yielded a commendable outcome of 87.14% classification accuracy and 0.93% CUP score.

According to Jiang et al. [41], nodule detection in the lungs may be achieved using multi-patches extracted from the Frangi filter's lung image. Integrating the 2 image classification techniques into a four-layer neural network model, this approach can gather data from radiologists to identify nodules. The proposed method achieved 80.06% sensitivity at 4.7 false positives (FP) and 94% at 15.1 FP for each scan.

The approach to pulmonary nodule detection presented by Setio et al. [42] relies on a multi-view Convolution network (MCN) for training models. The accurate identification of all suspicious nodules was achieved by fusing three algorithms for candidate nodule detection. At 1 FP per scan, the suggested method achieves a sensitivity level of 85.4%, and at 4 FP per scan, it reaches 90.1%.

Dou et al. demonstrated that Automated nodule identification using 3-deep CNNs from volumetric CT images had a lower false-positive rate [43]. By reducing the false-positive track and achieving the maximum CPM score, this method has been thoroughly verified in the LUNA16 challenge.

1.2 Breast cancer detection

One of the most prevalent causes of death from cancer in women is breast cancer. After the age of 50, the majority of patients diagnosed with this cancer pass away. This disease is responsible for almost 2×10^6 new cases annually and, in 2018, made up 11.6% of all cancer cases. Among women, it accounted for 24.2% of cases, making it the deadliest disease affecting women worldwide [44][5]. Irregular cells in the breast may be either benign or malignant. Cancer cells, also known as malignant cells, pose a greater threat when they metastasize or multiply in other parts of the body.

In contrast to the tiny number of cancerous cells, the benign cells are big and have a well-established type. Early detection of malignant tumours is challenging due to the small size and abundance of adipose and dense tissue. Advanced automated or computerized technologies also require breast tumour end-match detection.

Classification accuracy is crucial for breast cancer diagnosis. The characteristics that categorize objects as either benign or malignant are trained and tested using a variety of DL and Machine Learning (ML) techniques [45][46]. Here are a few of the numerous studies that have been done on breast cancer identification.

In 2020, Zhou et al. [47] presented the Inception-ResNet V2 and Inception V3 configurations and tested the performance of the model by evaluating the precision, adaptability, and specificity. Using the independent test technique, the CNN was able to estimate the therapeutic final diagnosis of axillary node metastasis with 85% responsiveness 0.89 AUC. In order to deal with errors in diagnosis by improving picture quality and processing time, Acharya et al. (2020) [48] used K-means, DL, enhanced loss feature (ELF) and autoencoder in the classification. With the use of the cluster and auto-encoders, K-means, which made use of a latent image feature, was able to provide superior cluster outcomes. Breast cancer diagnosis accuracy increased to 97% using the DL algorithm, while the processing time increased from 30 to 40 seconds.

Sun et al. (2017) [49] presented a breast cancer categorization approach. An SSL system that relies on graphs and a deep CNN was established. There were 3,158 ROIs in all, with an average size of 1,874 mammographic pairs. The remaining ROIs were considered unmarked, whereas 100 were considered identifiable data. For both labelled and unlabeled data, CNN achieved an accuracy of 0.82, and the AUC was 0.881.

Etemadi et al. (2016) [50] proposed a hybrid selection model as a means of locating biased genes. By implementing the decision tree algorithm, the issue of having an excessive number of groups is resolved, and the subtype prediction of breast cancer with the same or fewer genes consistently yields accurate results.

Abdel-Zaher et al. (2016) [51] presented a CNN-based approach using a backward propagation route and an unmonitored pathway network of deep-faith beliefs to identify breast cancer. The trials were conducted using the Wisconsin Breast Cancer Dataset (WBCD), which boasts an accuracy rate of 99.68%.

Finally, in order to fill in the gaps that were established by earlier studies and to provide a solution to the issue of breast and lung cancer detection and early diagnosis, this study makes use of the clustering data mining approach in order to determine the health of patients who have lung and breast cancer.

2. METHODS

In this section, first, a brief outline of the lung and breast cancer datasets used in this paper is given, and then the various methods that are utilized in this study, along with the proposed architectures for precise cancer grouping and classification, are presented. Firstly, the lung and breast cancer data are collected from the dataset and preprocessed to remove the noise and to ensure uniformity in size, resolution, and quality. After preprocessing, CNN is utilized to extract features from these preprocessed datasets. By combining Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and K-Means clustering, a hybrid clustering approach is used to split the features derived from CNN into numerous clusters. After that, using the features retrieved by the CNN, the cluster labels for each training data are generated using the DBSCAN and K-Means clustering algorithms. Finally, the CNN is utilized for the classification of lung and breast cancer.

2.1 Datasets

Two datasets are being utilized in this study; one is for lung cancer, and another is for breast cancer, which are given as follows:

- **Lung Cancer Dataset**

This dataset is aimed at exploring Lung Cancer occurrences in individuals, featuring 16 columns. Each column represents a specific attribute related to the individual's health status, lifestyle, and symptoms that could potentially influence the diagnosis of Lung Cancer. The application of clustering algorithms like DBSCAN and K-Means clustering could help in identifying patterns or groups based on the severity and combination of these attributes, potentially uncovering hidden relationships between the symptoms and the occurrence of Lung Cancer [52].

- **Breast Cancer Wisconsin (Diagnostic) Dataset**

It is a well-known dataset commonly used in machine learning and cancer research. It typically includes features derived from breast cancer biopsies and is often used for classification tasks. This dataset contains detailed measurements of breast masses, with 32 columns providing information that describes the characteristics of the cell nuclei. This dataset contains 32 columns that describe the characteristics of the cell nuclei. It contains features such as radius, texture, smoothness, concavity, etc. The detailed features provided in this dataset are particularly suited for CNN-based feature extraction as CNN can leverage the spatial relationships in data for classification purposes, distinguishing between benign and malignant diagnoses effectively [53].

2.2 Data mining approach for feature extraction and Classification from breast and lung cancer datasets

Data mining techniques have been useful in extracting significant features. The current study utilized data mining approaches, including CNNs, for feature extraction from lung and breast cancer datasets. These methods are good at identifying certain patterns and features that relate to these types of data. By using these, it aids in accurate diagnosis, prognosis, and treatment planning for lung and breast cancer.

In this study, the CNN architecture shown in Figure 2 is utilized, which contains a convolutional layer, a max pooling layer and a fully connected layer. The two-part feature extractor is performed through sequences of convolutions and max-pooling. An initial section has 3-3 max-pooling layers, a Relu activation function and Convolution layers with 32−32 units each. Relu is a well-known activation function that is often used in neural networks, particularly that which is known as CNNs. This nonlinearity is introduced into the model via the Relu layer.

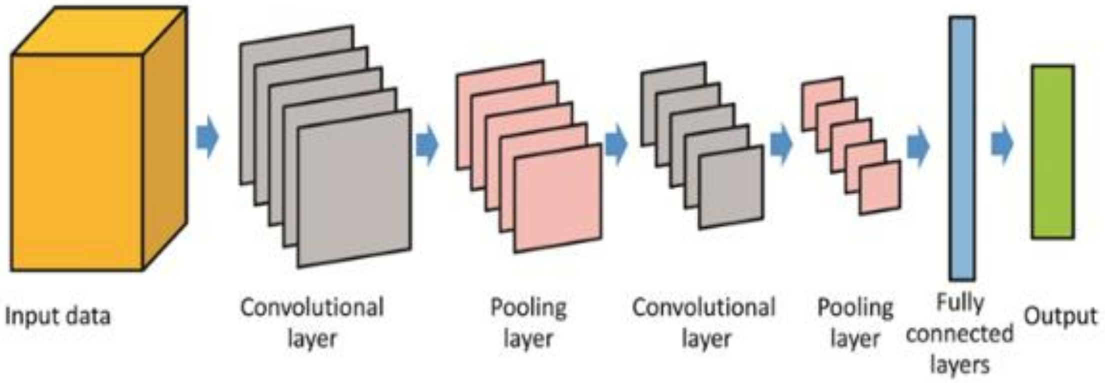

Figure 2. CNN architecture [54]

The output of the CNN is converted into a feature space, in which every value can be represented as a vector of features that have been retrieved from previous analysis. When this is complete, the features that were extracted from the CNN's feature extractor are provided as input for the subsequent clustering process.

2.3 The Structure of Clustering Process for Lung and Breast Cancer Dataset

This section presents the general idea of the proposed structure and clustering process. The study employed a hybrid clustering technique by combining DBSCAN and K-Means clustering for lung and breast cancer patients' stratification.

2.3.1 DBSCAN

DBSCAN is one of the most prevalent clustering methods developed in 1996. The density of data points within a certain region is analyzed to identify clusters by the algorithm. The core principle of density-based clustering is that a certain minimum number of cluster instances must be included within the range of a defined radius for each cluster instance. Core, Border, and Noise are the three categories into which DBSCAN divides data points. In order to find a cluster, DBSCAN takes a random instance (p) from the dataset (D) and finds all instances of D that are within the specified radius (r) and minimal number of instances (m) [55].

2.3.2 K-Means clustering

K-Means clustering is a partitioning algorithm used to group data points into K clusters based on similarity. It operates by iteratively assigning each data point to the nearest cluster centroid and then updating the centroids based on the mean of the data points assigned to each cluster. In the context of integrating hybrid clustering and classification for lung and breast cancer diagnosis, K-Means clustering can be applied as a preprocessing step to identify distinct groups or patterns within patient data. These identified clusters can then serve as features or inputs for subsequent classification algorithms, such as support vector machines (SVM) or neural networks, enabling more targeted and accurate cancer diagnosis based on learned patterns from the clustered data [56].

2.4 Hybrid Clustering Architecture

The hybrid clustering technique is used in the context of clustering a dataset containing features extracted from lung and breast cancer cases. This approach combines two separate techniques, namely DBSCAN and K-Means clustering, to successfully form clusters based on the extracted features. DBSCAN relies on the density of feature points to identify clusters within the feature space without requiring a predefined number of clusters. K-Means clustering, on the other hand, assigns membership degrees to each data point across multiple clusters, allowing for varying levels of belongingness to different clusters based on feature similarities. The hybrid clustering method divides the feature space into several clusters by combining the results of DBSCAN and K-Means clustering. Consequently, each data point in the dataset is assigned a cluster label based on the combined findings of these clustering approaches. This labelling technique provides a systematic way to analyze and understand the correlations and differences among lung and breast cancer cases by grouping them according to their feature similarities within the dataset.

2.5 Statistical analysis

There are different ways to evaluate clustering results, which are mainly divided into internal and external standards. Internal indicators analyze clustering outcomes without any prior information, uncovering the possible distribution and inherent structure of dataset samples. This research lacks a definitive truth for lung and breast cancer stratification. Therefore, the investigation utilized the Davies–Bouldin Index (DBI) and Silhouette Coefficient (SC) to measure the effectiveness of all algorithms under comparison. DBI and SC determine the closeness within a cluster and the distinction between various clusters, respectively [56].

- **Davies–Bouldin Index (DBI)**

The DBI metric determines the mean distance within each category and divides it by the distance that exists among the centres of two clusters, aiming to maximize this value [57]. The DBI index is calculated to measure the efficiency of a cluster, and it is inversely proportionate to the cluster's efficiency. The DBI index can be calculated as follows:

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{j\neq i}\left(\frac{\bar{C}_i+\bar{C}_j}{\left\|w_i-w_j\right\|_2}\right) \tag{1}$$

The variables w_i and w_j measure the centroids of the cluster of ith and jth class, whereas C_i and C_j indicate the inner mean distance of the ith and jth class [56].

- **Silhouette Coefficient (SC)**

Peter J. Rousseeuw, in 1986, presented the silhouette coefficient as an assessment measure [58]. It is used to assess the accuracy of clustering outcomes in this study. The combination of separation and cohesion may be determined using the following formula:

$$S = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{b(i)-a(i)}{\max\{a(i),b(i)\}}\right) \tag{2}$$

The parameter (i) denotes the mean Euclidean distance between sample i and the other samples within the same cluster. A lower value of (i) indicates that sample i should be assigned to a certain cluster. The variable b(i) denotes the dissimilarity among the cluster containing sample i and the other clusters [56].

2.6 Proposed Framework for Integrated Hybrid Clustering and Classification

The proposed approach utilized CNN for feature extraction. Initially, the CNN is applied to extract meaningful features from the combined lung and breast cancer dataset. These extracted features serve as the basis for subsequent clustering using hybrid DBSCAN- K-Means clustering algorithms. The clustering process categorizes instances based on the derived features, enabling the identification of potential patterns or groups within the breast and lung cancer datasets, as shown in Figure 3 below.

Subsequently, the dataset is filtered to distinguish between instances that may indicate malignant or benign conditions in breast cancer and potential indications of lung cancer. This refined dataset is then subjected to further classification using the trained CNN model. CNN, having been initially employed for feature extraction, now plays a pivotal role in categorizing instances based on the distinctive features identified during the clustering phase. This integrated hybrid approach leverages the strengths of both clustering and CNN-based classification to enhance the understanding of complex relationships within the lung and breast cancer datasets, offering potential insights for diagnostic and prognostic purposes.

DOI: 10.17148/IARJSET.2025.12807

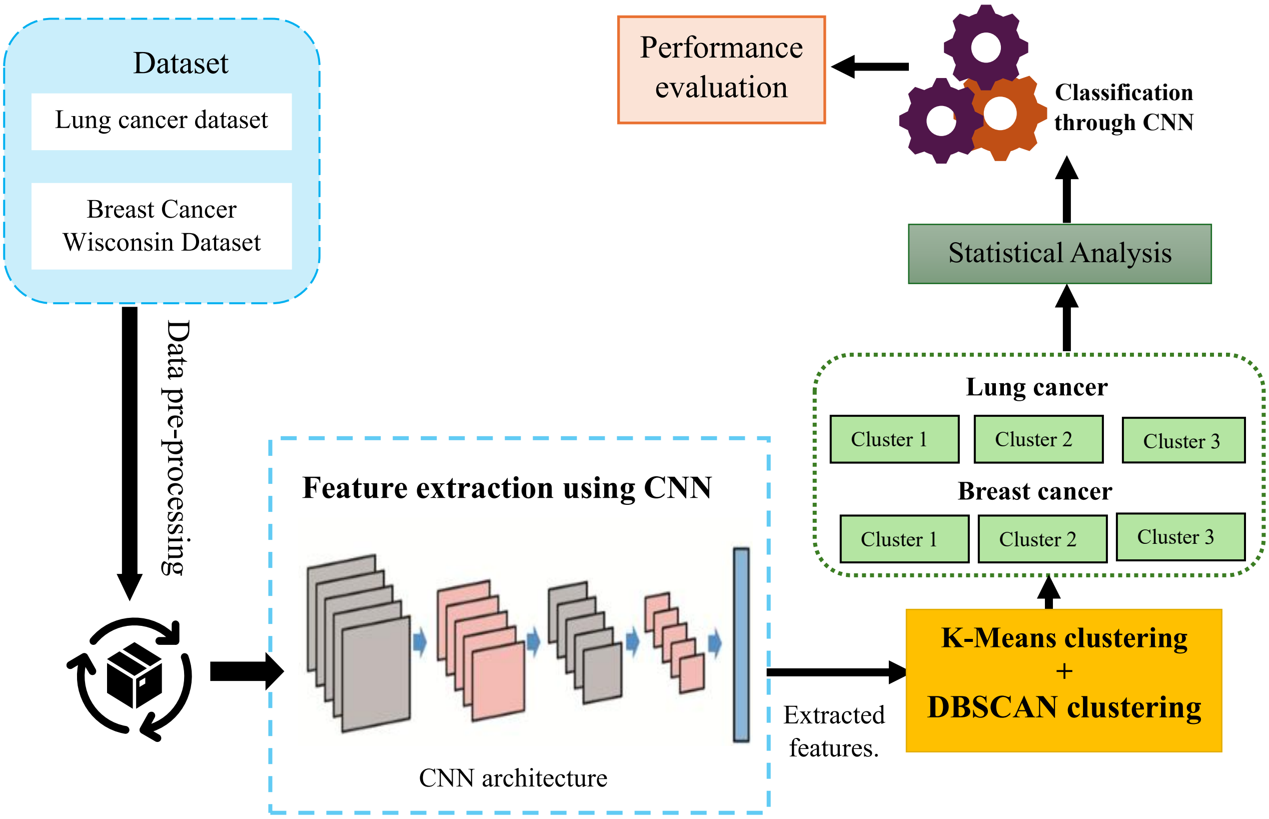

Figure 1. Proposed framework

3. RESULT AND DISCUSSION

This section presents the key findings of the research and provide a thorough analysis and interpretation of those results within the broader context of the field.

3.1 Evaluation Metrics

The performance evaluation of a model typically involves assessing its effectiveness across several key metrics. Accuracy measures the overall correctness of the model's predictions. Precision gauges the model's ability to correctly identify positive cases among all cases it labels as positive. F1 Score combines precision and recall, offering a balanced measure of a model's performance. Recall, also known as sensitivity, quantifies the model's ability to identify all positive instances correctly. Evaluating a model across these metrics provides a comprehensive understanding of its predictive power and reliability.

1. Accuracy: The percentage of cases that are accurately classified out of all the instances is called accuracy.

$$\boldsymbol{Accuracy} = \frac{\mathrm{TP+TN+FP+FN}}{(\mathrm{TP+TN})} \tag{3}$$

2. Precision: The preciseness of correct predictions, or the percentage of relevant examples among the obtained instances, is measured by precision.

$$\boldsymbol{Precision} = \frac{\mathrm{TP}}{(\mathrm{TP+FP})} \tag{4}$$

3. F1 Score: A score that harmonizes recall and precision is known as the F1 Score. With its unified metric for false positive and false negative rates, it strikes a good interference between recall and precision.

$$\boldsymbol{F1-Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

4. Recall (also known as True Positive Rate or Sensitivity): The model's recall is a measure of its capacity to catch all positive occurrences. It is calculated by dividing the total number of relevant instances by the fraction of instances that have been recovered.

$$\boldsymbol{Recall} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} \tag{6}$$

Where,

TP for true positives, TN for true negatives, FP for false positives, and FN for false negatives.

Table 1 illustrates the hyperparameter table for a specific machine learning model or algorithm integrating hybrid clustering and classification for lung and breast cancer diagnosis would require detailed information about the model's architecture, the clustering and classification algorithms used, and their respective hyperparameters.

Table 1. Hyperparameter Table

Hyperparameter	Description	Possible Values
Number of Clusters	Number of clusters to be formed in clustering	Integer > 1
Cluster Algorithm	Algorithm used for clustering	K-Means, DBSCAN, etc.
Clustering Distance Metric	Distance metric for clustering	Euclidean, etc.
Classification Algorithm	Algorithm used for classification	Neural Network, etc.
Neural Network Architecture	Number of layers, units per layer, activation functions, etc.	Varies based on architecture

This table provides a framework for capturing the hyperparameters relevant to a hybrid clustering and classification model. The actual values for these hyperparameters would depend on factors such as the dataset, computational resources, and desired performance.

The breast cancer dataset had two classes: benign and malignant. Hybrid clustering is used for grouping the data based on three clusters which are malignant, benign, and "Maybe" (cancer may be present or not). The prediction class on malignant and benign was gathered from the representation of the malignant and benign clusters, which resulted from a hybrid clustering approach. The result shows that the hybrid clustering approach could identify the beginin and malignant breast cancer clusters. The maybe cluster is the smallest cluster which indicates only a few of the cases remained unidentified. Figure 4 shows the result of the proposed hybrid clustering approach on the breast cancer dataset.

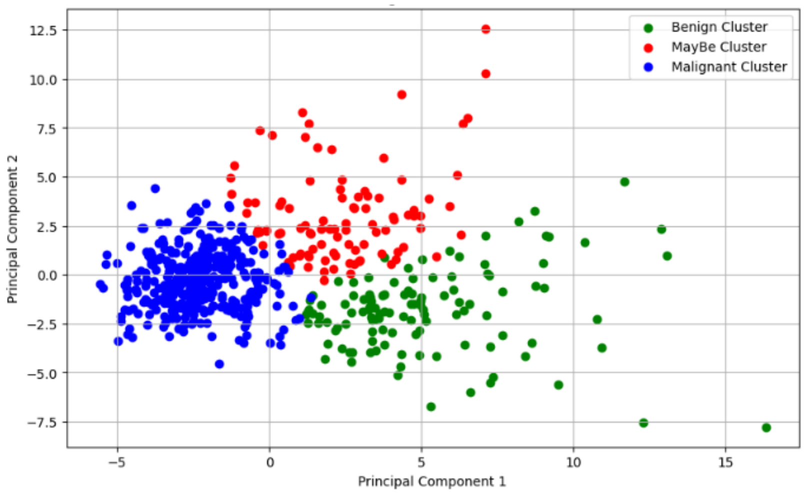

Figure 4. Hybrid clustering of breast cancer dataset

Further, the Lung cancer Dataset had two outcomes: either cancer is present, or cancer is not present. The proposed hybrid clustering is used for grouping the data based on three clusters which are yes (lung cancer is present), No (lung cancer is not present), and "Maybe" (uncertainty of cancer). The prediction class on Yes and No was gathered from the cluster representation, which resulted from the first phase. The result suggests that the hybrid clustering approach effectively separated the data points and formed distinct clusters, indicating a clear separation between cancer and non-cancer cases. The cluster region that corresponds to the Maybe cluster is characterized by its limited extent, indicating an absence of instances where the distinction between cancer and non-cancer is confusing. Figure 5 shows the result of the proposed hybrid clustering approach on the lung cancer dataset.

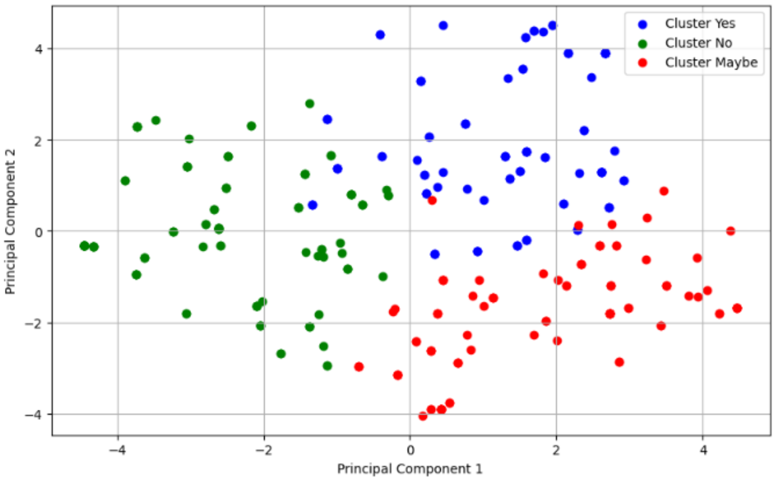

Figure 5. Hybrid clustering of lung cancer dataset

3.2 Results

The proposed approach integrates DBSCAN, and K-means clustering with CNN and demonstrates impressive classification results for lung and breast cancer datasets for uncertain clusters. The classification results for lung and breast cancer datasets are shown in Table 2, given below.

Table 2. Evaluation metric of the proposed approach

Dataset	Performance Metric	Value
Lung Cancer	Accuracy	0.9847
	Precision	0.9628
	Recall	0.9412
	F1 Score	0.9541
Breast Cancer	Accuracy	0.9986
	Precision	0.9815
	Recall	0.9637
	F1 Score	0.9741

For lung cancer, the model achieved an accuracy of 0.9847, with a high precision of 0.9628, a recall of 0.9412, and an F1 score of 0.9541. Similarly, in the breast cancer dataset, the model exhibited a remarkable accuracy of 0.9986, coupled with a precision of 0.9815, a recall of 0.9637, and an F1 score of 0.9741, as shown in Figure 6.

These results highlight the performance of the proposed approach in accurately categorizing uncertain instances into their respective classes. These classification results collectively indicate the robustness of the proposed approach in accurately categorizing uncertain instances in both lung and breast cancer datasets. The high values across accuracy, precision, recall, and F1 score underscore the potential utility of the hybrid methodology in supporting breast and lung cancer diagnosis.

DOI: 10.17148/IARJSET.2025.12807

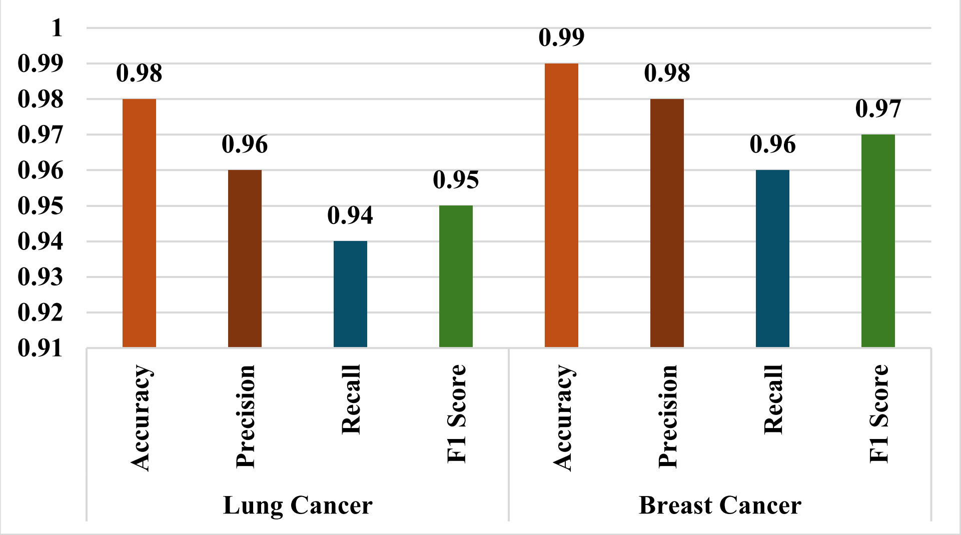

Figure 6. Performance of the proposed approach on both datasets.

3.3 Comparative Analysis

In this section, Table 3 provides comprehensive details regarding authors, methodologies employed, and accuracy pertaining to Clustering and Classification of Lung Cancer. Similarly, Table 4 presents a comparative analysis focusing on Breast Cancer. When assessing the efficacy of a model, accuracy stands out as a crucial metric. A higher classification and detection rate implies greater success of the model. Typically, the accuracy rate serves as a primary measure of a model's performance. The proposed model notably attained an impressive accuracy of 98.47% for lung cancer and an outstanding 99.86% for breast cancer.

Table 3. Comparative Analysis for Lung Cancer.

Authors	Technique	Outcomes
Tahoces et al. (2019) [38]	Energy-Based Optimization Technique	95.1%
Xie et al. (2019) [39]	2D-CNN	86.42%
Shen et al. (2017) [40]	MC-CNN	87.14%
Jiang et al. (2017) [41]	Neural Network	94%
Setio et al. (2016) [42]	MCN	90.1%.
Proposed Method	DBSCAN and K-means clustering	98.47%

Table 4. Comparative Analysis for Breast Cancer.

Authors	Techniques	Outcomes
Zhou et al. (2020) [47]	Inception-ResNet V2	85%
Acharya et al. (2020) [48]	K-means	97%
Sun et al. (2017) [49]	deep CNN	82%

Abdel-Zaher et al. (2016) [51]	CNN	99.68%.
Proposed Method	DBSCAN and K-means clustering	99.86%

4. CONCLUSION AND FUTURE SCOPE

The performance evaluation results for the proposed hybrid approach, integrating DBSCAN and K-means clustering along with CNN, on the lung and breast cancer datasets are highly promising. The combined methodology effectively captures intricate patterns within the lung and breast cancer dataset. The clustering phase, utilizing DBSCAN and K-means clustering, demonstrates excellent performance in grouping instances based on the identified patterns derived from the lung and breast cancer dataset. The cooperative use of clustering and CNN-based feature extraction enhances the precision and significance of the identified clusters. Furthermore, the subsequent CNN-based classification yields more accurate and reliable results in distinguishing between benign and malignant instances of breast cancer, as well as potential indications of lung cancer for uncertain datasets. The classification findings demonstrate a significant level of accuracy for both the lung and breast cancer datasets. Specifically, the accuracy score for lung cancer is 0.9847, while the accuracy score for breast cancer is 0.9986. The precision, recall, and F1 scores provide additional confirmation of the strength and reliability of the technique. The proposed hybrid approach shows promising results in capturing intricate patterns in lung and breast cancer datasets. Future enhancements involve integrating advanced ML techniques and exploring real-time applications for improved diagnosis and patient outcomes.

REFERENCES

[1]. Ferlay, Jacques, M. Ervik, F. Lam, M. Colombet, L. Mery, M. Piñeros, A. Znaor, I. Soerjomataram, and F. Bray. "Global cancer observatory: cancer today." Lyon, France: International Agency for Research on Cancer 3, no. 20 (2018): 2019.

[2]. World Health Organization. "WHO report on cancer: setting priorities, investing wisely and providing care for all." (2020).

[3]. Hamdani, Syed Suhail, Syed Ishfaq Yaseen, Naveed Nabi, Mehreen Syed, Syed Naveed Hamdani, Tahsin Hassan, Ovais Rashid, Akanksha Sharma, and Ruksana Khursheed. "Incidence and Etiology of Various Cancers in Kashmir Valley-A Comprehensive Review of Literature." International Journal for Research in Applied Sciences and Biotechnology 7, no. 6 (2020): 152-158.

[4]. Yabroff, K. Robin, Xiao-Cheng Wu, Serban Negoita, Jennifer Stevens, Linda Coyle, Jingxuan Zhao, Brent J. Mumphrey, Ahmedin Jemal, and Kevin C. Ward. "Association of the COVID-19 pandemic with patterns of statewide cancer services." JNCI: Journal of the National Cancer Institute 114, no. 6 (2022): 907-909.

[5]. Barrios, Carlos H. "Global challenges in breast cancer detection and treatment." The Breast 62 (2022): S3-S6. Saba, Tanzila, Muhammad Attique Khan, Amjad Rehman, and Souad Larabi Marie-Sainte. "Region extraction and classification of skin cancer: A heterogeneous framework of deep CNN features fusion and reduction." Journal of Medical Systems 43, no. 9 (2019): 289.

[6]. Saba, Tanzila, Muhammad Attique Khan, Amjad Rehman, and Souad Larabi Marie-Sainte. "Region extraction and classification of skin cancer: A heterogeneous framework of deep CNN features fusion and reduction." Journal of Medical Systems 43, no. 9 (2019): 289.

[7]. Khan, Muhammad Qasim, Ayyaz Hussain, Saeed Ur Rehman, Umair Khan, Muazzam Maqsood, Kashif Mehmood, and Muazzam A. Khan. "Classification of melanoma and nevus in digital images for diagnosis of skin cancer." IEEE Access 7 (2019): 90132-90144.

[8]. Javed, Rabia, Tanzila Saba, Mohd Shafry, and Mohd Rahim. "An intelligent saliency segmentation technique and classification of low contrast skin lesion dermoscopic images based on histogram decision." In 2019 12th International Conference on Developments in eSystems Engineering (DeSE), pp. 164-169. IEEE, 2019.

[9]. Abbas, Naveed, Tanzila Saba, Amjad Rehman, Zahid Mehmood, Nadeem Javaid, Muhammad Tahir, Naseer Ullah Khan, Khawaja Tehseen Ahmed, and Roaider Shah. "Plasmodium species aware based quantification of malaria parasitemia in light microscopy thin blood smear." Microscopy Research and Technique 82, no. 7 (2019): 1198-1214.

[10]. Rehman, Amjad, Naveed Abbas, Tanzila Saba, Toqeer Mahmood, and Hoshang Kolivand. "Rouleaux red blood cells splitting in microscopic thin blood smear images via local maxima, circles drawing, and mapping with original RBCs." Microscopy research and technique 81, no. 7 (2018): 737-744.

DOI: 10.17148/IARJSET.2025.12807

[11]. Amin, Javaria, Muhammad Sharif, Mudassar Raza, Tanzila Saba, and Muhammad Almas Anjum. "Brain tumour detection using statistical and machine learning method." Computer methods and programs in biomedicine 177 (2019): 69-79.

[12]. Iqbal, Sajid, M. Usman Ghani Khan, Tanzila Saba, and Amjad Rehman. "Computer-assisted brain tumour type discrimination using magnetic resonance imaging features." Biomedical Engineering Letters 8, no. 1 (2018): 5-28

[13]. Saba, Tanzila, Sana Ullah Khan, Naveed Islam, Naveed Abbas, Amjad Rehman, Nadeem Javaid, and Adeel Anjum. "Cloud-based decision support system for the detection and classification of malignant cells in breast cancer using breast cytology images." Microscopy research and technique 82, no. 6 (2019): 775-785.

[14]. Mughal, Bushra, Muhammad Sharif, Nazeer Muhammad, and Tanzila Saba. "A novel classification scheme to decline the mortality rate among women due to breast tumour." Microscopy research and technique 81, no. 2 (2018): 171-180.

[15]. Mughal, Bushra, Nazeer Muhammad, Muhammad Sharif, Tanzila Saba, and Amjad Rehman. "Extraction of breast border and removal of pectoral muscle in the wavelet domain." Biomedical Research 28, no. 11 (2017): 5041-5043

[16]. Jamal, Arshad, Mohammed Hazim Alkawaz, Amjad Rehman, and Tanzila Saba. "Retinal imaging analysis based on vessel detection." Microscopy research and technique 80, no. 7 (2017): 799-811

[17]. Saba, Tanzila, Ahmed Sameh, Fatima Khan, Shafqat Ali Shad, and Muhammad Sharif. "Lung nodule detection based on an ensemble of handcrafted and deep features." Journal of Medical Systems 43 (2019): 1-12.

[18]. Khan, Sajid A., Muhammad Nazir, Muhammad A. Khan, Tanzila Saba, Kashif Javed, Amjad Rehman, Tallha Akram, and Muhammad Awais. "Lungs nodule detection framework from computed tomography images using support vector machine." Microscopy Research and Technique 82, no. 8 (2019): 1256-1266

[19]. Fahad, H. M., M. Usman Ghani Khan, Tanzila Saba, Amjad Rehman, and Sajid Iqbal. "Microscopic abnormality classification of cardiac murmurs using ANFIS and HMM." Microscopy research and technique 81, no. 5 (2018): 449-457.

[20]. Bray, Freddie, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey A. Torre, and Ahmedin Jemal. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries." CA: A Cancer Journal for Clinicians 68, no. 6 (2018): 394-424.

[21]. Kurihara, Yasuyuki, Shin Matsuoka, Tsuneo Yamashiro, Atsuko Fujikawa, Shoichiro Matsushita, Kunihiro Yagihashi, and Yasuo Nakajima. "MRI of pulmonary nodules." American journal of roentgenology 202, no. 3 (2014): W210-W216.

[22]. Abdelrahman, Leila, Manal Al Ghamdi, Fernando Collado-Mesa, and Mohamed Abdel-Mottaleb. "Convolutional neural networks for breast cancer detection in mammography: A survey." Computers in biology and medicine 131 (2021): 104248.

[23]. Iftikhar, Saman, Kiran Fatima, Amjad Rehman, Abdulaziz S. Almazyad, and Tanzila Saba. "An evolution-based hybrid approach for heart diseases classification and associated risk factors identification." Biomedical Research 28, no. 8 (2017): 3451-3455.

[24]. Liaqat, Amna, Muhammad A. Khan, Muhammad Sharif, Mamta Mittal, Tanzila Saba, K. Suresh Manic, and Feras NH Al Attar. "Gastric tract infections detection and classification from wireless capsule endoscopy using computer vision techniques: A review." Current medical imaging 16, no. 10 (2020): 1229-1242.

[25]. Al-Ameen, Zohair, Ghazali Sulong, Amjad Rehman, Abdullah Al-Dhelaan, Tanzila Saba, and Mznah Al-Rodhaan. "An innovative technique for contrast enhancement of computed tomography images using normalized gamma-corrected contrast-limited adaptive histogram equalization." EURASIP Journal on Advances in Signal Processing 2015, no. 1 (2015): 1-12.

[26]. Rahim, Mohd Shafry Mohd, Alireza Norouzi, Amjad Rehman, and Tanzila Saba. "3D bones segmentation based on CT images visualization." Biomedical Research 28, no. 8 (2017): 3641-3644.

[27]. Marie-Sainte, Souad Larabi, Tanzila Saba, Deem Alsaleh, and Mashael Bin Alamir Alotaibi. "An improved strategy for predicting diagnosis, survivability, and recurrence of breast cancer." Journal of Computational and Theoretical Nanoscience 16, no. 9 (2019): 3705-3711.

[28]. Saba, Tanzila, Khalid Haseeb, Imran Ahmed, and Amjad Rehman. "Secure and energy-efficient framework using Internet of Medical Things for e-healthcare." Journal of Infection and Public Health 13, no. 10 (2020): 1567-1575.

[29]. Joshi, Jahanvi, Rinal Doshi, and Jigar Patel. "Diagnosis of breast cancer using clustering data mining approach." International Journal of Computer Applications 101, no. 10 (2014): 13-17.

[30]. Ramachandran, P., N. Girija, and T. Bhuvaneswari. "Early detection and prevention of cancer using data mining techniques." International Journal of Computer Applications 97, no. 13 (2014).

DOI: 10.17148/IARJSET.2025.12807

[31]. Husham, Ahmed, Mohammed Hazim Alkawaz, Tanzila Saba, Amjad Rehman, and Jarallah Saleh Alghamdi. "Automated nuclei segmentation of malignant using level sets." Microscopy research and technique 79, no. 10 (2016): 993-997.

[32]. Saba, Tanzila. "Automated lung nodule detection and classification based on multiple classifiers voting." Microscopy research and technique 82, no. 9 (2019): 1601-1609.

[33]. Mittal, Ansh, Deepika Kumar, Mamta Mittal, Tanzila Saba, Ibrahim Abunadi, Amjad Rehman, and Sudipta Roy. "Detecting pneumonia using convolutions and dynamic capsule routing for chest X-ray images." Sensors 20, no. 4 (2020): 1068.

[34]. Shimazaki, Akitoshi, Daiju Ueda, Antoine Choppin, Akira Yamamoto, Takashi Honjo, Yuki Shimahara, and Yukio Miki. "Deep learning-based algorithm for lung cancer detection on chest radiographs using the segmentation method," Scientific Reports 12, no. 1 (2022): 727.

[35]. Agarwal, Aman, Kritik Patni, and D. Rajeswari. "Lung cancer detection and classification based on alexnet CNN." In 2021 6th International Conference on Communication and Electronics Systems (ICCES), pp. 1390-1397. IEEE, 2021.

[36]. Naqi, Syed Muhammad, Muhammad Sharif, and Arfan Jaffar. "Lung nodule detection and classification based on geometric fit in parametric form and deep learning." Neural Computing and Applications 32 (2020): 4629-4647.

[37]. Asuntha, A., and Andy Srinivasan. "Deep learning for lung Cancer detection and classification." Multimedia Tools and Applications 79 (2020): 7731-7762.

[38]. Tahoces, Pablo G., Luis Alvarez, Esther González, Carmelo Cuenca, Agustín Trujillo, Daniel Santana-Cedrés, Julio Esclarín et al. "Automatic estimation of the aortic lumen geometry by ellipse tracking." International Journal of computer assisted radiology and Surgery 14 (2019): 345-355.

[39]. Xie, Hongtao, Dongbao Yang, Nannan Sun, Zhineng Chen, and Yongdong Zhang. "Automated pulmonary nodule detection in CT images using deep convolutional neural networks." Pattern Recognition 85 (2019): 109-119.

[40]. Shen, Wei, Mu Zhou, Feng Yang, Dongdong Yu, Di Dong, Caiyun Yang, Yali Zang, and Jie Tian. "Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification." Pattern Recognition 61 (2017): 663-673.

[41]. Jiang, Hongyang, He Ma, Wei Qian, Mengdi Gao, and Yan Li. "An automatic detection system of lung nodule based on multigroup patch-based deep learning network." IEEE Journal of Biomedical and Health Informatics 22, no. 4 (2017): 1227-1237.

[42]. Setio, Arnaud Arindra Adiyoso, Francesco Ciompi, Geert Litjens, Paul Gerke, Colin Jacobs, Sarah J. Van Riel, Mathilde Marie Winkler Wille, Matiullah Naqibullah, Clara I. Sánchez, and Bram Van Ginneken. "Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks." IEEE Transactions on Medical Imaging 35, no. 5 (2016): 1160-1169.

[43]. Dou, Qi, Hao Chen, Lequan Yu, Jing Qin, and Pheng-Ann Heng. "Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection." IEEE Transactions on Biomedical Engineering 64, no. 7 (2016): 1558-1567.

[44]. Wild, Chris, Elisabete Weiderpass, and Bernard W. Stewart, eds. World cancer report: cancer research for cancer prevention. International Agency for Research on Cancer, 2020.

[45]. Sadad, Tariq, Asim Munir, Tanzila Saba, and Ayyaz Hussain. "Fuzzy C-means and region growing based classification of tumour from mammograms using hybrid texture feature." Journal of Computational Science 29 (2018): 34-45.

[46]. Vijayarajeswari, R., P. Parthasarathy, S. Vivekanandan, and A. Alavudeen Basha. "Classification of mammogram for early detection of breast cancer using SVM classifier and Hough transform." Measurement 146 (2019): 800-805.

[47]. Zhou, Li-Qiang, Xing-Long Wu, Shu-Yan Huang, Ge-Ge Wu, Hua-Rong Ye, Qi Wei, Ling-Yun Bao et al. "Lymph node metastasis prediction from primary breast cancer US images using deep learning." Radiology 294, no. 1 (2020): 19-28.

[48]. Acharya, Smarika, Abeer Alsadoon, P. W. C. Prasad, Salma Abdullah, and Anand Deva. "Deep convolutional network for breast cancer classification: enhanced loss function (ELF)." The Journal of Supercomputing 76, no. 11 (2020): 8548-8565.

[49]. Sun, Wenqing, Tzu-Liang Bill Tseng, Jianying Zhang, and Wei Qian. "Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data." Computerized Medical Imaging and Graphics 57 (2017): 4-9.

[50]. Etemadi, Roohollah, Abedalrhman Alkhateeb, Iman Rezaeian, and Luis Rueda. "Identification of discriminative genes for predicting breast cancer subtypes." In 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1184-1188. IEEE, 2016.

[51]. Abdel-Zaher, Ahmed M., and Ayman M. Eldeib. "Breast cancer classification using deep belief networks." Expert Systems with Applications 46 (2016): 139-144.

[52]. https://www.kaggle.com/datasets/nancyalaswad90/lung-cancer/data

[53]. https://www.kaggle.com/code/anandhuh/breast-cancer-prediction-accuracy-98-24/input

[54]. Monkam, Patrice, Shouliang Qi, He Ma, Weiming Gao, Yudong Yao, and Wei Qian. "Detection and classification of pulmonary nodules using convolutional neural networks: a survey." Ieee Access 7 (2019): 78075-78091.

[55]. Devi, R. Delshi Howsalya, and P. Deepika. "Performance comparison of various clustering techniques for diagnosis of breast cancer." In 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1-5. IEEE, 2015.

[56]. Zhao, Zijuan, Juanjuan Zhao, Kai Song, Akbar Hussain, Qianqian Du, Yunyun Dong, Jihua Liu, and Xiaotang Yang. "Joint DBN and Fuzzy C-Means unsupervised deep clustering for lung cancer patient stratification." Engineering Applications of Artificial Intelligence 91 (2020): 103571.

[57]. Sutramiani, Ni Putu, I. Made Teguh Arthana, Shana Aurelia, Muhammad Fauzi, and I. Wayan Agus Surya Darma. "The Performance Comparison of DBSCAN and K-Means Clustering for MSMEs Grouping based on Asset Value and Turnover."

[58]. Dey, Debangana, Thamar Solorio, Manuel Montes y Gómez, and Hugo Jair Escalante. "Instance selection in text classification using the silhouette coefficient measure." In Advances in Artificial Intelligence: 10th Mexican International Conference on Artificial Intelligence, MICAI 2011, Puebla, Mexico, November 26-December 4, 2011, Proceedings, Part I 10, pp. 357-369. Springer Berlin Heidelberg, 2011.