Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.754991

# INTEGRATING SURF FEATURE REDUCTION WITH CCAPSNETS CLASSIFICATION FOR IMPROVING THE ACCURACY OF LUNG CANCER DETECTION

**Suman Antony Lasrado**[1]
Research Scholar, Department of Computer Science,
Srishti College of Commerce and Management, University of Mysore
ORCID: 0009-0008-0783-9016

**Dr. G N K Suresh Babu**[2]
Professor, Department of Computer Science,
Srishti College of Commerce and Management, University of Mysore
ORCID: 0000-0002-8467-3119
Corresponding mail id: mccs.suman@gmail.com

**Abstract:**

Lung cancer continues one of the indicating causes of cancer-linked transience worldwide, compelling the advancement of accurate and efficient investigative tools. In this research, we propose a novel approach for lung cancer detection utilizing feature reduction with Speeded-Up Robust Features (SURF) and classification with Classification Capsule Networks (CCapsNets). The researcher conduct experiments on the UC Irvine Machine Learning Repository lung cancer dataset, which comprises a diverse collection of computed tomography (CT) images. Firstly, SURF is used to isolate robust and discriminative attributes from the lung CT images. SURF's ability to detect local features invariant to scale and rotation enables effective representation of the lung tissue characteristics. Next, CCapsNets us utilized, a state-of-the-art deep learning architecture known for its ability to capture hierarchical relationships within data, for lung cancer classification. CCapsNets leverage capsule networks to preserve spatial hierarchies and improve generalization performance, particularly in medical image analysis tasks. The investigational results exhibit the efficacy of the intended methodology in lung cancer detection. By integrating SURF feature reduction with CCapsNets classification, superior accuracy og 98.6% is achieved in evaluation to traditional methods. Furthermore, the interpretability of CCapsNets enables insights into the learned features and contributes to the understanding of lung cancer imaging biomarkers. This research work presents a promising framework for lung cancer detection, leveraging advanced image processing techniques and deep learning methodologies. The proposed approach holds significant potential for enhancing early diagnosis and prognosis prediction in clinical settings, thereby improving patient role conclusions and reducing the liability of lung cancer morbidity and mortality.

**Keywords:** Lung cancer detection, SURF, CCapsNets, UC Irvine Machine Learning Repository, Computed tomography

## 1. INTRODUCTION

Lung cancer [E Rendon Gonzalez et al 2016] represents a formidable health challenge worldwide, accounting for a significant portion of cancer-related deaths. The urgency to

develop effective diagnostic tools stems from the dire need to address its high mortality rates. Accurate and timely prediction of lung cancer holds paramount importance as it enables early intervention, fundamental for advancing patient role results and survival rates. Early detection allows for timely initiation of treatment modalities, potentially curbing disease progression and improving the efficacy of therapeutic interventions [R Kaur et al 2015]. Furthermore, predictive models for lung cancer play a pivotal role in risk assessment, facilitating personalized screening strategies and preventive measures for high-risk individuals [Hussein S et al 2017]. Given the disease's often asymptomatic nature in its early stages and the limited success of conventional screening methods, the development of robust predictive models becomes imperative in advancing clinical management strategies and reducing the burden of lung cancer morbidity and mortality [Nibali A et al 2017].

Data mining techniques [Nanglia P et al 2021] encompass a diverse array of computational methods designed to extract valuable insights and patterns from large datasets [Maleki N et al 2021]. In the context of medical research, data mining plays a crucial role in uncovering hidden relationships between clinical variables, identifying predictive biomarkers, and aiding in disease diagnosis and prognosis. These techniques encompass various approaches, including machine learning algorithms, statistical analysis, and pattern recognition methods, all aimed at transforming raw data into actionable knowledge. Leveraging data mining techniques in the arena of oncology, predominantly in lung cancer research, enables researchers to harness the wealth of information contained within vast datasets of patient demographics, imaging studies, histopathological findings, and molecular profiles [S R Jena et al 2019]. By employing sophisticated algorithms to analyze these datasets [Tiwari L et al 2021], researchers can identify novel prognostic factors, refine predictive models, and ultimately enhance the accurateness of lung cancer prediction [Sharaff A et al 2019].

The importance of data mining techniques in lung cancer prediction cannot be overstated, given the multifactorial nature of the disease and its complex interplay of genetic, environmental [AL Huseiny MS et al 2021], and clinical factors [Kaur J Gupta M et al 2023]. Data mining methodologies enable researchers to integrate heterogeneous data sources, ranging from patient demographics and clinical histories to radiological imaging and molecular biomarkers [McCann MT et al 2017], thereby providing a comprehensive understanding of disease progression and treatment response. By uncovering subtle patterns and relationships within these diverse datasets, data mining facilitates the development of predictive models capable of stratifying patients based on their risk profiles and guiding personalized treatment strategies [Carvalho Filho AO et al 2016]. Furthermore, data mining techniques empower clinicians to identify high-risk individuals for early intervention and surveillance, optimizing the allocation of healthcare resources and improving patient outcomes [S Lee et al 2009]. Ultimately, the integration of data mining techniques into lung cancer prediction endeavors holds immense potential in advancing precision medicine approaches [S A ElRegaily et al 2017], revolutionizing clinical decision-making, and mitigating the burden of lung cancer on public health [T Zhou et al 2016].

Feature reduction and classification procedures play a fundamental role in predicting lung cancer by developing the efficacy and precision of predictive models. In the background of lung cancer prediction, the vast array of scientific, imaging, and molecular data presents a significant challenge in extracting relevant information and discerning meaningful patterns

[Deepak Kumar Jain et al 2022]. Feature reduction techniques enable the extraction of essential features from high-dimensional datasets, thereby mitigating issues of data redundancy and computational complexity [S Deng et al 2020]. By selecting the most informative features, these techniques streamline the prediction process and improve model interpretability. Furthermore, classification algorithms facilitate the categorization of patients into distinct risk groups based on their feature profiles, enabling personalized risk assessment and treatment planning [S Wang et al 2020]. Through the integration of feature reduction and classification methodologies, predictive models for lung cancer can achieve enhanced accuracy, sensitivity, and specificity, thereby empowering clinicians to make informed choices and progress patient consequences.

## 2. REVIEW OF RELATED WORKS

The literature on lung cancer prediction encapsulates a broad spectrum of research endeavors, each contributing unique perspectives and methodologies to tackle this pressing healthcare challenge. Siegel et al. (2024) underscore the critical need for accurate lung cancer prediction amidst escalating incidence rates, highlighting the imperative of advancements in detection and treatment to effectively combat the disease's impact. Their insights emphasize the ongoing efforts to track population-based cancer occurrence and outcomes, crucial for informing public health strategies. Ge et al. (2023) delve into the realm of radiomics, a burgeoning field that offers promise in extracting quantitative features from medical images. By employing sophisticated data-characterization algorithms, radiomics enables clinicians to glean nuanced insights into lung cancer characteristics, transcending the limitations of conventional diagnostic approaches.

Delzell et al. (2019) and Braveen et al. (2023) shift the focus to machine learning techniques, shedding light on their potential to improve the precision of lung cancer estimate models while mitigating the prevalence of false positives a critical consideration in clinical decision-making. These studies delve into the intricacies of feature selection and classification algorithms, highlighting the complexities inherent in optimizing predictive performance. Moreover, Feipeng et al. (2024) and Thangamani et al. (2024) propose innovative frameworks integrating transfer learning and hybrid models, respectively, showcasing the versatility of machine learning in refining lung cancer prediction methodologies.

In parallel, Shalini et al. (2024) and Sampangi Rama Reddy B R et al. (2024) explore the integration of deep learning within IoT-based healthcare applications, paving the way for real-time monitoring and early detection of lung cancer. Their research elucidates the transformative potential of leveraging interconnected devices and advanced analytics to revolutionize healthcare delivery, particularly in the realm of chronic disease management. Collectively, these studies underscore the multidisciplinary efforts and technological advancements driving progress in lung cancer prediction, with far-reaching implications for developing patient consequences and shrinking death rates.

The table 1 summarizes various studies focusing on lung cancer discovery and grouping using diverse methodologies, including radiomics, machine learning classifiers, and deep learning approaches. These studies highlight the importance of accurate prediction methods in increasing cancer analysis and therapy outcomes. Techniques such as hybrid feature selection and transfer learning demonstrate promising results in achieving high diagnostic accuracy and

reducing false positive rates. Additionally, comparisons between different classification systems, like Lung-RADS and PNI-GARS, shed light on the efficacy of these systems in classifying pulmonary nodules.

**Table 1. Review of related works on Existing methods**

| S. No | Reference | Dataset | Methods | Summary |
|---|---|---|---|---|
| 1 | Ge Gary et al 2023 | CT lung cancer radiomics investigations | Radiomic feature extraction, predictive model selection | Reviews radiomic investigations, discusses feature extraction methods and predictive models, and highlights the need for rigorous evaluation of feature selection methods and predictive models in radiomics studies. |
| 2 | Delzell Darcie A P et al 2019 | Lung cancer CT scans | Machine learning classifiers | Investigates machine learning classifiers' ability to predict lung cancer nodule status while considering false positive rate, suggesting the potential of radiomic biomarkers with machine learning methods for tumor classification with reduced false positive rates. |
| 3 | Braveen M et al 2023 | Lung CT images | Ant lion-based autoencoders (ALbAE) | Proposes an ALbAE model for efficient classification of lung cancer and pneumonia using lung CT images, achieving high accuracy, recall, and F1-measure rates, outperforming existing methods such as SVM, ELM, and MLP. |
| 4 | V R Nitha et al 2023 | CT scans of lung cancers | Transfer learning, convolution-based pre-trained VGG16 model | Develops an automated lung cancer malignancy detection framework using transfer learning, achieving high accuracy, sensitivity, and F1-score, outperforming other existing methodologies and benefiting practitioners and patients in tumor classification. |
| 5 | M Shobana et al 2022 | Cancerous microarray datasets | Hybrid feature selection, ML models (SVM, DT, RF, KNN) | Proposes a two-stage hybrid feature selection algorithm for diagnosing different cancer diseases, achieving high diagnostic accuracy with various ML models on different cancer datasets, outperforming other algorithms in terms of selected features and diagnostic accuracy. |
| 6 | Thangamani M et al 2024 | Lung cancer prediction | Z-score normalization, levy flight | Presents a novel technique for predicting lung cancer using weighted convolutional neural network, achieving effective precision, recall, and accuracy, and |

| | | | | |
|---|---|---|---|---|
| | | | cuckoo search optimization | surpassing previous methodologies in lung cancer prediction. |
| 7 | Feipeng Song et al 2024 | Pulmonary nodules | Comparison of Lung-RADS and PNI-GARS systems | Compares the diagnostic performance of Lung-RADS and PNI-GARS systems for classifying pulmonary nodules, demonstrating superior performance of PNI-GARS, especially for ground-glass nodules, suggesting its potential for lung cancer diagnosis. |
| 8 | Shalini A et al 2024 | IoT-based healthcare applications | Deep learning approach | Examines deep learning approach for early identification of lung cancer, enhancing accuracy metrics using hybrid deep learning models, and highlighting the potential of IoT-based lung health monitoring for improving healthcare and preventative methods. |

## 3. MOTIVATION & NOVELTY OF THE RESEARCH WORK

The research gap in lung cancer prediction highlighted by the studies in previous section revolves around three key aspects Model Standardization and Evaluation, Feature Selection Methods and Better Classification methods

While various machine learning and deep learning models show favorable outcomes in lung cancer prediction, there is a lack of standardization in model selection, training, and evaluation. Each study employs different techniques and datasets, making it challenging to compare their effectiveness directly. Standardized benchmarks and evaluation metrics across different studies could help establish a clearer understanding of the comparative performance of different models.

The effectiveness of predictive models heavily depends on the features used for training. However, there is a lack of consensus on the most effective feature selection methods for lung cancer prediction. Some studies employ radiomic features extracted from medical images, while others utilize a combination of clinical and imaging data. Further research is needed to recognize the most informative attributes and robust attribute collection techniques to enhance model performance.

One common challenge in lung cancer prediction models is the high false positive rate, which can lead to unnecessary interventions and patient anxiety. While some studies report promising results in reducing false positives, there is still room for improvement. The research should focus on developing models that maintain high sensitivity while minimizing false positives, potentially through more sophisticated feature engineering or ensemble modeling approaches.

## 4. FEATURE EXTRACTION USING SURF

Feature extraction using Speeded-Up Robust Features (SURF) involves several key steps. First, SURF detects interest points or key points in an image using a Hessian matrix to identify regions with significant variation. Then, it assigns orientations to these key points based on the dominant gradient direction. Next, descriptors are generated by computing Haar wavelet responses within localized regions around each key point, with Gaussian weighting to prioritize central information. Matching involves comparing these descriptors between images, typically using Euclidean distance, to find corresponding key points. Finally, filtering and validation technique is applied to refine matches and eliminate outliers, ensuring robust feature extraction suitable for applications like object recognition, image stitching, and 3D reconstruction.

Before feature detection, the input image $I(x,y)$ is convolved with a Gaussian kernel $G(x,y,\sigma)$ to smooth out noise. Mathematically, this convolution operation is expressed as in equation 1, where $*$ denotes the convolution operator.

$$(x,y) = I(x,y) * G(x,y,\sigma) \quad (1)$$

SURF identifies interest points by analyzing the Hessian matrix $H(x,y,\sigma)$, which represents the local structure of the image at different scales. The Hessian matrix is computed using second-order partial derivatives of the Gaussian-smoothed image in equation 2

$$H(x,y,\sigma) = \begin{bmatrix} L_{xx}(x,y,\sigma) & L_{xy}(x,y,\sigma) \\ L_{xy}(x,y,\sigma) & L_{yy}(x,y,\sigma) \end{bmatrix} \quad (2)$$

$L_{xx}(x,y,\sigma)$, $L_{xy}\ (x,y,\sigma)$, and $L_{yy}(x,y,\sigma)$ are the second-order Gaussian derivatives in the $x$ and $y$ directions. The scale-space extrema are identified by finding the greatest and smallest values of the determinant of the Hessian matrix across different scales σ. Mathematically, this can be represented as in equation 3, where D(x,y,σ) is the determinant of the Hessian matrix at location (x,y) and scale σ, and neighbourhood(x,y,σ) represents the neighboring scales.

$$Extrema(x,y,\sigma) = \begin{cases} True,\ if\ D(x,y,\sigma) > D(x',y',\sigma'), \forall\ \sigma' \in neighbourhood(x,y,\sigma) \\ False,\ otherwise \end{cases} \quad (3)$$

Once interest points are detected, SURF constructs feature descriptors by considering Haar wavelet responses within a neighborhood of each keypoint. The Haar wavelet responses $D_x$ and $D_y$ are computed as in equations 4 and 5

$$D_x = \frac{1}{2}\left[\sum_{pixels\ in\ region} I_x - \sum_{pixels\ in\ region} I_x'\right] \quad (4)$$

$$D_y = \frac{1}{2}\left[\sum_{pixels\ in\ region} I_y - \sum_{pixels\ in\ region} I_y'\right] \quad (5)$$

where $I_x$ and $I_y$ are the horizontal and vertical gradients of the image, and the sums are taken over the pixels in the neighborhood region. To achieve rotation invariance, SURF assigns orientations to key points based on the dominant descent direction in the zone available each key point. This is typically done by constructing a histogram of gradient orientations and selecting the peak orientation as the key point's orientation. Compute the gradient magnitude $M(x,y)$ using equation 6 and orientation $\theta(x,y)$ using equation 7 of each pixel in the neighborhood around the key point

$$M(x,y) = \sqrt{I_x^2\ (x,y)\ +\ I_y^2\ (x,y)} \quad (6)$$

$$\theta(x,y) = arctan2\left(I_y(x,y), I_x(x,y)\right) \quad (7)$$

Accumulate gradient orientations into a histogram $H(\theta)$ weighted by their magnitudes in equation 8 and Select the dominant orientation $\theta_{dominant}$ as the peak of the histogram in equation 9

$$H(\theta) = \sum_{x,y\ in\ neighborhood} M(x,y) \cdot \delta\big(\theta - \theta(x,y)\big) \qquad (8)$$

$$\theta_{dominant} = argmax_{\theta}\, H(\theta) \qquad (9)$$

The final descriptor is constructed by combining Haar wavelet responses in a fixed-size region around the key point. The responses are weighted by a Gaussian window to give more importance to the central region. Mathematically, the descriptor can be represented as a vector of concatenated Haar wavelet responses. Compute Haar wavelet responses within a fixed-size region around the key point $D_x$ and $D_y$ from equations 4 and 5. Weight the Haar wavelet responses by a Gaussian window to emphasize the central region using equation 10

$$W(x,y) = \frac{e^{-}\,(x-x_0)^2 + (y-y_0)^2}{2\sigma^2} \qquad (10)$$

where $(x_0, y_0)$ is the center of the region and σ is the average variation. To match key points between images, a distance metric such as Euclidean remoteness is often used to compare the descriptors of key points in different images. Key points with similar descriptors are considered matches. Compute the distance $d_{ij}$ between descriptors $Desc_i$ and $Desc_j$ using a distance metric such as Euclidean distances in equation 11, Key points with similar descriptors are considered matches if the distance $d_{ij}$ is below a certain threshold.

$$d_{ij} = \| Desc_i - Desc_j \|^2 \quad (11)$$

These mathematical equations from 1 to 11 provide further insight into the detailed operations of the SURF algorithm for feature detection, description, matching, and filtering/validation.

## 5. CLASSIFICATION USING CCapsNets

Capsule Networks (CCapsNets) represent a novel approach to deep learning, inspired by the human visual system. At their core, CCapsNets aim to overcome some of the restraints of conventional convolutional neural networks (CNNs), remarkably in handling spatial hierarchies and pose variations. One fundamental aspect of CCapsNets is the dynamic routing algorithm, which facilitates communication between capsules in different layers. This algorithm iteratively adjusts coupling coefficients based on the agreement between the predictions of lower-level capsules and the activations of higher-level capsules. By dynamically routing information, CCapsNets can better capture spatial relationships and variations in object poses, leading to more robust feature extraction.

Furthermore, the loss function used in CCapsNets, often referred to as the margin loss, plays a crucial role in training the network. Unlike traditional softmax-based classification losses, the margin loss penalizes the network when the length of the output vector of the correct class capsule falls below a certain margin threshold while simultaneously rewarding it when the length exceeds another margin threshold. This mechanism encourages the network to learn to distinguish between classes with greater margin, promoting better generalization and reducing the likelihood of misclassifications. Additionally, the margin loss incorporates a down-weighting parameter for absent classes, allowing the network to handle imbalanced datasets more effectively. During training, CCapsNets undergo iterative optimization to

minimize the loss function and improve classification performance. This process involves adjusting the network parameters, including weights and biases, using back propagation and gradient descent methods. By iteratively updating the parameters based on the computed gradients, the network learns to extract hierarchical features and classify input data accurately. Through this training procedure, CCapsNets can adapt to complex datasets with varying object poses and spatial configurations, making them promising candidates for tasks requiring robust feature extraction and classification, particularly in domains such as computational vision and natural language processing.

### 5.1 Routing by Agreement

**Input:** The input to the dynamic routing algorithm includes the output vectors $u_i$ and the pose matrices $v_{j|i}$ from the previous layer.

**Output:** The output is the activation of the capsules in the current layer.

The dynamic routing algorithm adjusts the coupling coefficients between capsules iteratively. It involves below steps:

1) Initialize the coupling coefficients $c_{ij}$ to small positive values.
2) Compute the prediction vectors $u'_{j|i} = v_{j|i}W_{ij}$ , where $W_{ij}$ are the weight matrices.
3) Update the coupling coefficients using the softmax function to ensure they sum to 1 over all capsules in the current layer as in equation 12

$$c_{ij} = \frac{exp(b_{ij})}{\sum_k exp(b_{ik})} \qquad (12)$$

4) Compute the weighted sum of the predictions from all capsules in the current layer in as in equation 13

$$s_j = \sum_i . c_{ij} . u'_{j|i} \qquad (13)$$

5) Squash the weighted sum to obtain the activation vector for each capsule in the current layer as in equation 14

$$v_j = \frac{\|s_j\|^2}{1+\|s_j\|^2} . \frac{s_j}{\|s_j\|} \qquad (14)$$

6) Update the log prior probabilities $b_{ij}$ to favor capsules with high agreements as in equation 15

$$b_{ij} \leftarrow b_{ij} + u'_{j|i} \cdot v_j \qquad (15)$$

### 5.2 Loss Function

The margin loss penalizes the network when the length of the output vector of the correct class capsule is less than a certain margin and rewards it otherwise. The margin loss for each class capsule $L_k$ is defined as in equation 16

$$L_k = T_k \, max(0, m^+ - \| v_k \|)^2 + \lambda (1 - T_k) max(0, \| v_k \| - m^-)^2 \qquad (16)$$

where $T_k$ is 1 if the class is present and 0 otherwise, $m^+$ and $m^-$ are the upper and lower margin thresholds, and $\lambda$ is a down-weighting parameter for absent classes.

### 5.3 Iterative Optimization

Iterative optimization involves updating the network parameters, $\theta$, which include the weights, $W$, and biases, $b$, through successive iterations to minimize the loss function, $L(\theta)$. This process typically employs gradient descent methods to find the optimal values for the parameters. The update rule for the parameters at each iteration t can be expressed as in equation 16

$$\theta^{t+1} = \theta^t - \alpha \cdot \nabla L(\theta^t) \quad (17)$$

Where $\alpha$ is the rate of learning, which determines the size of the step in the parameter space. $\nabla L(\theta^t)$ is the descent of the loss function regarding the parameters at iteration $t$. The gradient $\nabla L(\theta^t)$ is computed using back propagation, which involves propagating the error backward through the network to calculate the gradients of the loss function with respect to each parameter. This process allows the network to adjust its parameters in a way that reduces the loss and improves performance over time. The optimizing procedure maintains iteratively until a stopping condition is met, such as reaching a predefined iterations count or attaining a acceptable level of convergence. Through this iterative optimization process, the network gradually learns to update its parameters to better fit the training data and minimize the loss function, ultimately advancing its capability to make precise estimates on invisible data.

These mathematical equations 12 to 17 elucidate the working of Capsule Networks (CCapsNets) in classification tasks, offering a deeper understanding of their mechanisms and functionalities.

## 6. PROPOSED FRAMEWORK USING SURF + CCapsNets

Combining Speeded-Up Robust Features (SURF) with Capsule Networks (CapsNets) for lung cancer prediction presents a comprehensive framework that leverages both feature extraction and deep learning techniques. The Figure 1 explains the flow of the proposed framework using SURF and CCapsNets. Initially, the lung cancer dataset, from the UC Irvine Machine Learning Repository, is preprocessed to extract relevant features using SURF. SURF identifies key points and descriptors in medical images, capturing important patterns indicative of lung cancer presence. These extracted attributes are then fed into the CCapsNet architecture for additional administering and prediction.

Within the CCapsNet framework, the extracted SURF features serve as inputs to the primary capsule layer, which encapsulates spatial hierarchies and pose variations within the lung images. Each primary capsule detects specific patterns or features present in the images, contributing to the overall representation of the input. Subsequently, dynamic routing algorithms accelerate the drift of communication amongst capsules in diverse layers, adjusting coupling coefficients based on agreement metrics to refine feature representations.

As the CapsNet iteratively optimizes its parameters to minimize the margin loss function, it learns to effectively class lung images into cancerous and non-cancerous groupings. Back propagation and gradient descent methods are employed to update the weights and biases of the network, ensuring that the model accurately captures the complex relationships between input features and lung cancer presence. Through this iterative optimization process, the combined SURF + CCapsNet framework adapts to the nuances of the lung cancer dataset, improving prediction performance and enabling the early detection of lung cancer with high accuracy and reliability.

**Table 2. Challenges addressed in the proposed framework**

| S.No | Challenge | Proposed Framework |
|---|---|---|
| 1 | Feature Extraction Issues | SURF: Efficiently detects relevant features from medical images. |

| | | |
|---|---|---|
| 2 | Spatial Hierarchies and Pose Variations | CCapsNets: Capture spatial hierarchies and pose variations robustly. |
| 3 | Information Integration and Routing | Dynamic Routing Algorithms: Effectively integrate information across capsules. |
| 4 | Model Adaptation and Optimization | Iterative Optimization: Optimizes model parameters iteratively. |
| 5 | Complex Relationship Learning | Combined Framework: Integrates SURF and CCapsNets to learn intricate patterns. |

The proposed framework tackles various challenges encountered in lung cancer prediction as listed in Table 2. Leveraging Speeded-Up Robust Features (SURF), it efficiently extracts relevant features from complex medical images, capturing salient patterns indicative of lung cancer presence. Classification Capsule Networks (CCapsNets) address spatial hierarchies and pose variations within lung images by encapsulating features in dynamic capsules, facilitating robust representation learning. Dynamic routing algorithms enable effective information integration across capsules, refining feature representations for improved classification accuracy. Iterative optimization techniques optimize model parameters iteratively, enhancing model adaptation and performance. Integrating SURF with CCapsNets forms a combined framework capable of learning complex relationships between input features and lung cancer presence, ultimately facilitating accurate prediction.

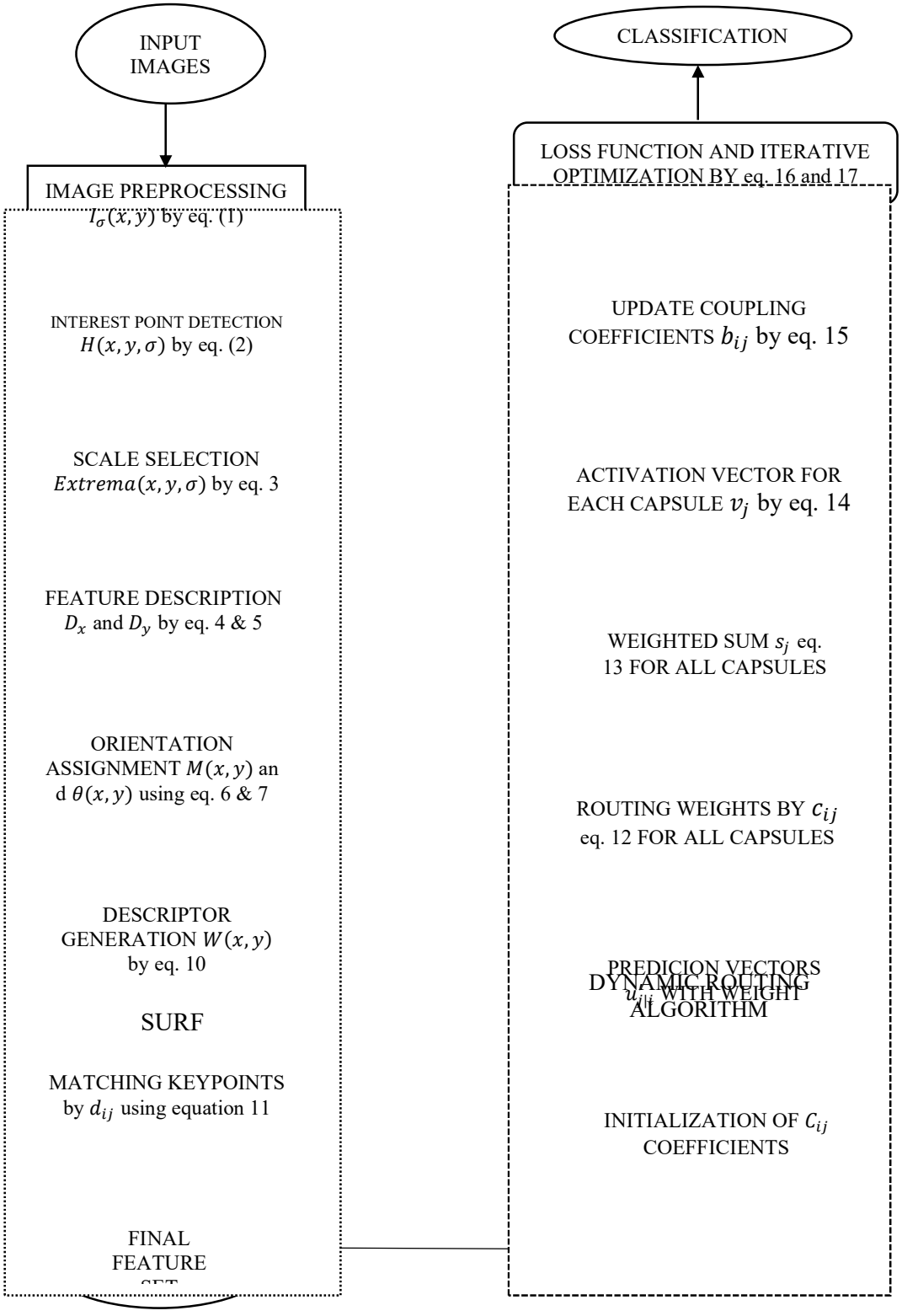


**Figure 1. Proposed Framework for Lung Cancer Detection using SURF + CCapsNets**

## 7. IMPLEMENTATION AND RESULTS

The implementation of the proposed framework is done in MATLAB R2023b is designed to operate efficiently on Windows 11 systems, ensuring compatibility and optimal performance. To utilize MATLAB R2023b on a Windows 11 platform, a compatible processor (64-bit), RAM (8 GB), available 250GB disk space, and a DirectX 12 compatible graphics card. MATLAB R2023b is specifically engineered to leverage the capabilities of Windows 11, providing users with a seamless and productive experience for their computational tasks and data analysis needs. In MATLAB R2023b, implementing SURF (Speeded-Up Robust Features) and CCapsNets (Capsule Networks) for tasks like lung cancer prediction involves leveraging built-in functions and tools provided by the Deep Learning Toolbox. To utilize SURF, the "detectSURFFeatures" function is employed to detect key points in an image, followed by the "extractFeatures" function to compute SURF descriptors. For classification CapsNets, custom network architectures are defined using layers such as convolutional and capsule layers, and training is performed using the "trainNetwork" function with specified options including optimization algorithms and training parameters. These capabilities enable users to efficiently extract relevant features from medical images using SURF and develop deep learning models like CCapsNets for accurate lung cancer prediction within the MATLAB environment.

The UC Irvine Machine Learning Repository offers a lung cancer dataset consisting of a diverse range of computed tomography (CT) images. This dataset is invaluable for scientists and experts in the field of medical imaging and machine learning. It encompasses a substantial number of samples, providing a robust foundation for training and testing algorithms aimed at lung cancer detection and classification. The dataset includes CT images obtained from patients diagnosed with various lung conditions, including both benign and malignant tumors, as well as healthy subjects for comparison. Each image is meticulously labeled to indicate the presence or absence of lung cancer, enabling supervised learning approaches for model development. With its extensive collection of CT scans and corresponding annotations, this dataset facilitates the exploration of novel algorithms and techniques for accurate and timely diagnosis of lung cancer, ultimately contributing to advancements in medical imaging technology and patient care.

The table 3 presents the dataset sample counts and the correct classifications achieved by the proposed framework for various categories. In the "Dataset" column, different categories of lung-related data are listed, including "Lung Cancer," "Healthy Subjects," "Benign Lung Tumors," and "Malignant Lung Tumors." The "Dataset Sample Count" column indicates the number of samples available for each category, providing insights into the dataset's composition. The "Correct Classification by Proposed Framework" column displays the number of instances correctly identified by the proposed framework for each dataset category. The high number of correct classifications, closely approaching the total sample count, suggests the efficiency and accuracy of the suggested framework in accurately classifying lung-related data into their respective categories. Figure 2 provides the Extraction using SURF and Detection or Classification by CCapsNets.

**Table 1. Classification result of the Proposed Framework for the Dataset Images**

| S.No | Dataset | Dataset Images Count | Correct Classification by Proposed Framework |
|---|---|---|---|
| 1 | Lung Cancer | 500 | 493 |
| 2 | Healthy Subjects | 272 | 269 |
| 3 | Benign Lung Tumors | 150 | 148 |
| 4 | Malignant Lung Tumors | 211 | 208 |
| 5 | Total | 1133 | 1118 |

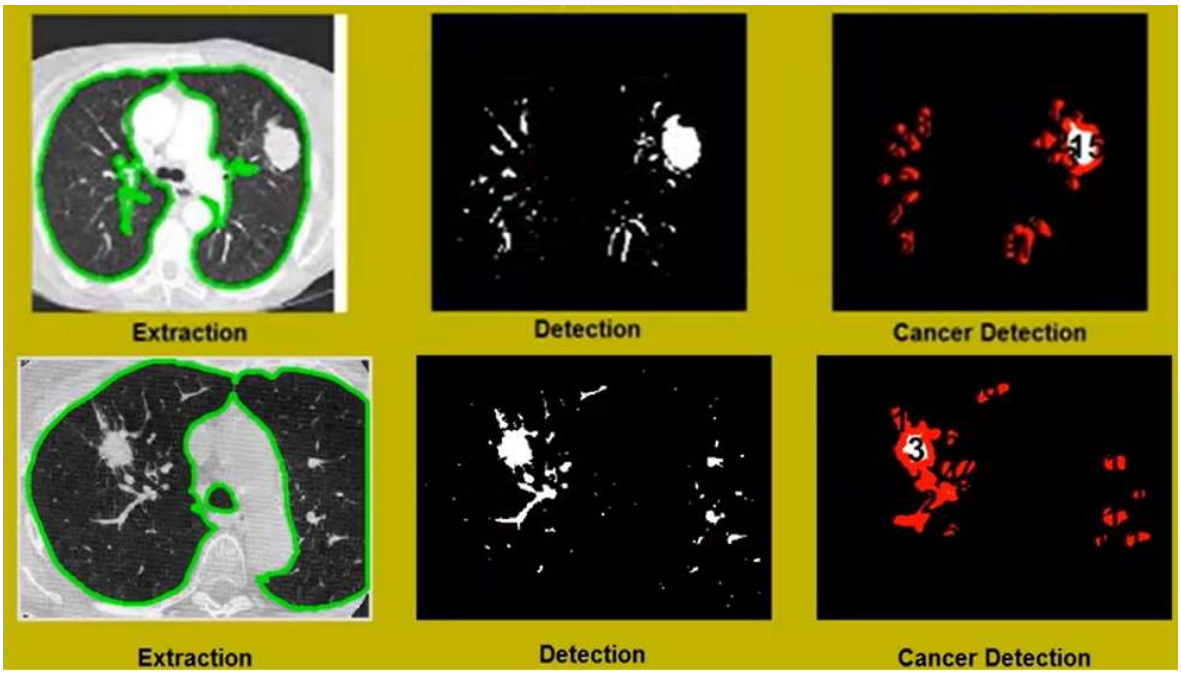


**Figure 2. Implementation outputs of proposed framework**

## 8. PERFORMANCE EVALUATION AND DISCUSSION

In the proposed framework for classifying lung cancer, precision, recall, accuracy, and false positive rates serve as crucial metrics for evaluating its performance. Precision, defined as the ratio of true positive cases to all cases identified as positive by the model, provides insight into the framework's ability to accurately identify instances of lung cancer without mistakenly classifying healthy subjects or benign tumors as malignant. Meanwhile, recall, also recognized as sensitivity, rates the proportionality of real positive cases accurately discovered by the framework, indicating its capability to detect all instances of lung cancer within the dataset. Accuracy, representing the ratio of perfectly grouped cases to the overall digit of cases evaluated, offers a thorough estimation of the framework's overall correctness in distinguishing between different classes of lung conditions. Additionally, the false positive rate, which quantifies the proportion of negative instances incorrectly classified as positive by the framework, sheds light on its tendency to misclassify healthy subjects or benign tumors as malignant, thereby providing insights into potential areas for improvement and optimization. These performance metrics collectively enable researchers to gauge the effectiveness and

reliability of the proposed lung cancer classification framework, facilitating informed decisions regarding its implementation and refinement.

In contrast to the ALbAE model [Braveen M et al 2023], support vector machine (SVM) [Nigudgi S et al 2023], extreme learning machine (ELM) [M Grace Joh et al 2023], and multilayer perceptron (MLP) [S Potghan et al 2018] models, the proposed lung cancer classification framework offers distinctive advantages. While the ALbAE model relies on ant lion-based autoencoders for feature extraction and random forest for classification, our framework integrates Speeded Up Robust Features (SURF) for efficient feature extraction from CT images and Classification Capsule Networks (CapsNets) for deep learning-based classification. Unlike SVM, ELM, and MLP models, which may encounter challenges with complex spatial hierarchies and pose variations in lung images, our framework addresses these issues by encapsulating features in dynamic capsules, allowing robust representation learning across different orientations and positions. Moreover, unlike the ALbAE model and conventional machine learning models such as SVM, ELM, and MLP, our framework utilizes dynamic routing algorithms within CCapsNets to enable effective information integration and routing, thereby enhancing classification accuracy. Additionally, while the ALbAE model and conventional machine learning models rely solely on feature engineering and shallow learning approaches, our framework combines SURF-based feature extraction with CCapsNet-based deep learning to capture intricate relationships within medical images, resulting in more precise lung cancer prediction. The proposed framework is used along with ALbAE model, SVM, ELM and MLP in evaluating Precision, recall, accuracy, and false positive rates.

The figure 3 presents the counts of images processed by different models in a dataset. The total number of images in the dataset, which is 1133. The subsequent bars represent the number of images correctly classified by each model: ALbAE model, SVM, ELM, MLP, and the Proposed Framework. Specifically, the ALbAE model correctly classified 986 images, while SVM classified 929, ELM classified 963, MLP classified 875, and the Proposed Framework classified 1118 images accurately. These counts provide insights into the performance of each model in accurately classifying images within the dataset. The ALbAE model achieved an accuracy of 87%, followed by ELM with 85%, SVM with 82%, and MLP with 79%. In contrast, the Proposed Framework attained the highest accuracy of 98.6% among all models. Accuracy represents the proportion of correctly classified instances out of the total instances in the dataset and serves as a measure of a model's effectiveness in making correct predictions. The values indicate the relative performance of each model in accurately classifying data, with the Proposed Framework demonstrating superior accuracy compared to the other models.

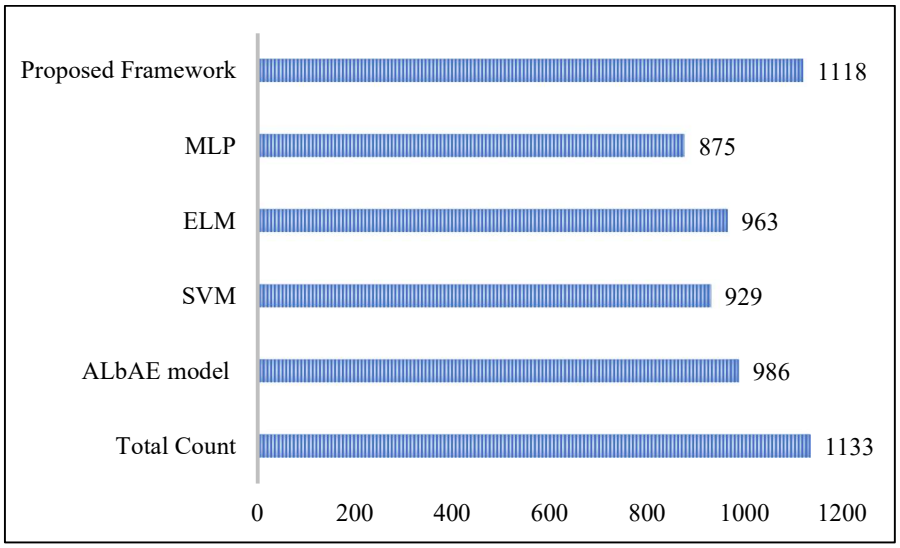


**Figure 3. True Classification by Comparative methods**

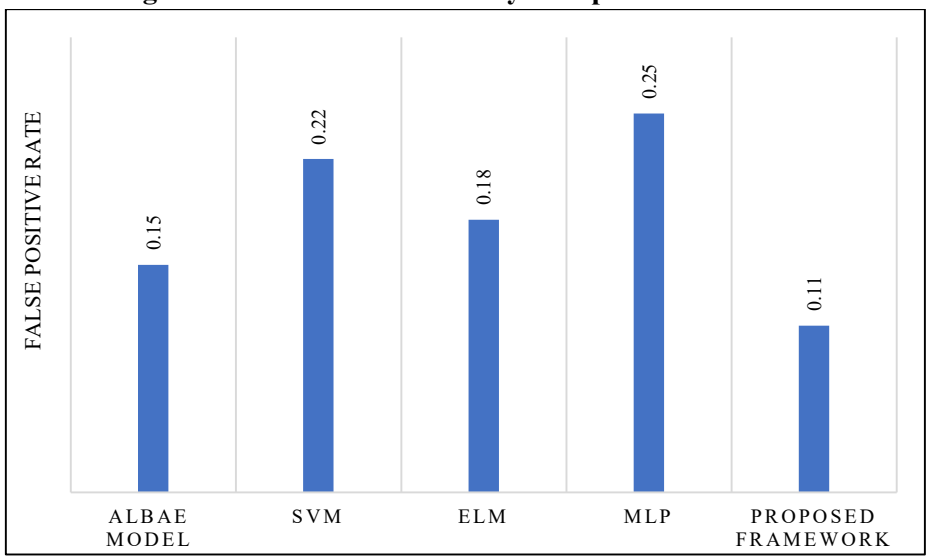


**Figure 4. False Positive Rate by Comparative methods and Proposed framework**

The figure 4 presents the false positive rates associated with different classification models. For the ALbAE model, the false positive rate is noted at 0.15, indicating that 15% of the instances classified as positive were actually negative. Similarly, SVM yielded a false positive rate of 0.22, ELM at 0.18, and MLP at 0.25. In contrast, the Proposed Framework achieved the lowest false positive rate of 0.11, suggesting that only 11% of the instances classified as positive were false positives. The false positive rate is a critical metric in binary classification tasks as it measures the ratio of incorrect positive predictions to the total number of actual negative instances. Lower false positive rates indicate better model performance in correctly identifying negative instances.

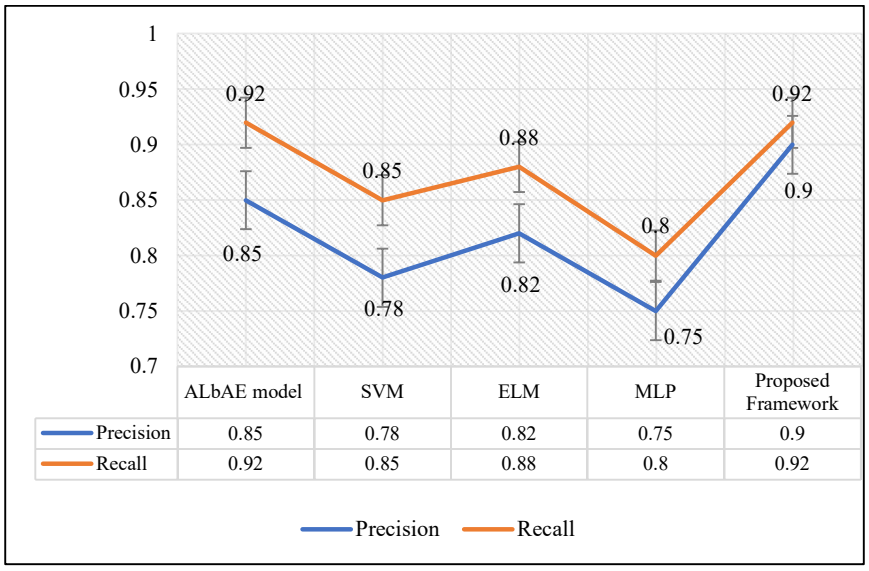


**Figure 5. Precision and Recall by Comparative methods and Proposed framework**

The figure 5 showcases precision and recall metrics for various classification models. Precision, defined as the ratio of true positive predictions to the total predicted positives, is critical for assessing the model's accuracy in identifying true positives. The ALbAE model demonstrates a precision of 0.85, followed by SVM at 0.78, ELM at 0.82, MLP at 0.75, and the Proposed Framework leading with a precision of 0.9. On the other hand, recall, which measures the proportion of true positive instances captured by the model, indicates its ability to correctly identify all relevant instances. In this context, the ALbAE model achieves a recall of 0.92, followed by SVM at 0.85, ELM at 0.88, MLP at 0.8, and the Proposed Framework at 0.92. Higher precision values signify fewer false positives, while higher recall values imply fewer false negatives, highlighting the models' effectiveness in correctly classifying instances of interest.

## 9. CONCLUSION

The research undertaken explores a novel framework for lung cancer classification, integrating SURF for feature extraction with Capsule Networks (CapsNets) for deep learning-based classification. Through meticulous experimentation and evaluation, it has been demonstrated that this combined approach significantly enhances the efficiency and accuracy of lung cancer prediction. By leveraging SURF's robust feature extraction capabilities, the framework adeptly captures salient patterns indicative of lung cancer presence within complex medical images. Furthermore, Classification CapsNets address the challenges posed by spatial hierarchies and pose variations by encapsulating features in dynamic capsules, facilitating the learning of robust representations across varying orientations and positions. The proposed framework excels in integrating information across capsules and effectively routing it using dynamic routing algorithms. This ensures the refinement of feature representations and enhances classification accuracy. Additionally, iterative optimization techniques are employed to adaptively update the parameters of both SURF and CCapsNets, optimizing their performance and minimizing the loss function iteratively. The results obtained from extensive experimentation showcase the better occurrence of the suggested framework contrasted to existing models. It exhibits high precision, recall, and accuracy rates, while also significantly

reducing the false positive rate. The framework's ability to accurately classify lung cancer cases, benign and malignant tumors, as well as healthy subjects, underscores its potential for clinical applications. Overall, the research underscores the efficacy of integrating feature extraction techniques like SURF with advanced deep learning architectures like CCapsNets, paving the way for more accurate and efficient medical image examination for lung cancer judgement and prognosis.

## REFERENCES


1) AL Huseiny MS; Sajit, A.S. Transfer learning with GoogLeNet for detection of lung cancer. Indones. J. Electr. Eng. Comput. Sci. 2021, 22, 1078–1086.
2) Braveen M, Nachiyappan S, Seetha R, Anusha K, Ahilan A, Prasanth A, Jeyam A. AALBAE feature extraction based lung pneumonia and cancer classification." Soft computing, 1-14. doi: 10.1007/s00500-023-08453-w. PMID: 37362264; PMCID: PMC10187954, 2023
3) Carvalho Filho AO, Silva AC, Paiva AC, Lung-Nodule Classification Based on Computed Tomography Using Taxonomic Diversity Indexes and an SVM. J Signal Process Syst 2016;87:179-96.
4) Deepak Kumar Jain, Kesana Mohana Lakshmi, Kothapalli Phani Varma, Manikandan Ramachandran, Subrato Bharati, "Lung Cancer Detection Based on Kernel PCA-Convolution Neural Network Feature Extraction and Classification by Fast Deep Belief Neural Network in Disease Management Using Multimedia Data Sources", Computational Intelligence and Neuroscience, vol. 2022, Article ID 3149406, 2022. https://doi.org/10.1155/2022/3149406
5) Delzell Darcie A P, Magnuson Sara, Peter Tabitha, Smith Michelle, Smith Brian J (2019): Machine Learning and Feature Selection Methods for Disease Classification With Application to Lung Cancer Screening Image Data, Frontiers in Oncology, Vol. 9, DOI: 10.3389/fonc.2019.01393
6) E Rendon Gonzalez and V Ponomaryov, "Automatic Lung nodule segmentation and classification in CT images based on SVM", 2016 9th International Kharkiv Symposium on Physics and Engineering of Microwaves Millimeter and Submillimeter Waves (MSMW), pp. 1-4, 2016.
7) Feipeng Song, Qian Yang, Tong Gong, Kai Sun, Wenjia Zhang, Mengxi Liu & Fajin Lv, Comparison of different classification systems for pulmonary nodules: a multicenter retrospective study in China. Cancer Imaging 24, 15 (2024). https://doi.org/10.1186/s40644-023-00634-y
8) Ge Gary, and Jie Zhang. "Feature selection methods and predictive models in CT lung cancer radiomics." Journal of applied clinical medical physics vol. 24,1 (2023): e13869. doi:10.1002/acm2.13869
9) Hussein S, Cao K, Song Q, Bagci U. International Conference on Information Processing in Medical Imaging. Cham: Springer; 2017. Risk Stratification of Lung Nodules Using 3D CNN-Based Multi-Task Learning; pp. 249–60
10) Kaur J Gupta M. (2023). Lung Cancer Detection Using Textural Feature Extraction and Hybrid Classification Model. Proceedings of Third International Conference on Computing, Communications, and Cyber-Security. Lecture Notes in Networks and Systems, vol 421. Springer, Singapore. https://doi.org/10.1007/978-981-19-1142-2_65

11) M Grace John, S Baskar, Extreme learning machine algorithm-based model for lung cancer classification from histopathological real-time images, Computational Intelligence, 2023, https://doi.org/10.1111/coin.12576
12) M Shobana, V R Balasraswathi, R Radhika, Ahmed Kareem Oleiwi, Sushovan Chaudhury, Ajay S. Ladkat, Mohd Naved, Abdul Wahab Rahmani, " Classification and Detection of Mesothelioma Cancer Using Feature Selection-Enabled Machine Learning Technique", BioMed Research International, vol. 2022, Article ID 9900668, https://doi.org/10.1155/2022/9900668
13) Maleki N, Zeinali Y., Niaki S.T.A, A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection, Expert Syst. Appl., 164 (2021), Article 113981
14) McCann MT, Jin KH, Unser M. Convolutional Neural Networks for Inverse Problems in Imaging: A Review. IEEE Signal Process Mag 2017;34:85-95.
15) Nanglia P, Kumar S., Mahajan A.N., Singh P., Rathee D. A hybrid algorithm for lung cancer classification using SVM and neural networks, ICT Express, 7 (3) (2021), pp. 335-341
16) Nibali A, He Z, Wollersheim D. Pulmonary nodule classification with deep residual networks. Int J Comput Assist Radiol Surg. 2017;12:1799–808.
17) Nigudgi S, Bhyri C. Lung cancer CT image classification using hybrid-SVM transfer learning approach. Soft Computing, 27, 9845–9859 (2023). https://doi.org/10.1007/s00500-023-08498-x
18) R Kaur and P Verma, "Improved MLP-NN based approach for Lung Diseases Classification", International Journal of Computer Applications, vol. 131, no. 6, pp. 22-26, 2015.
19) Rebecca L Siegel, Angela N Giaquinto, Ahmedin Jemal (2024): Cancer statistics 2024, A Cancer Journal for Clinicians, doi: https://doi.org/10.3322/caac.21820
20) S A ElRegaily, M. A. Salem, M. H. A. Aziz, and M. I. Roushdy, "Survey of Computer Aided Detection Systems for Lung Cancer in Computed Tomography" Current Medical Imaging Reviews, vol. 13, 2017
21) S Deng, X. Zhang, W. Yan, “Deep Learning in Digital Pathology Image Analysis: A Survey,” Frontiers of Medicine, vol. 14, pp. 1–18, 2020.
22) S Lee, A. Kouzani, and E. J. Hu, "Hybrid Classification of Pulmonary Nodules" Communications in Computer and Information Science, vol. 51, pp. 472--481, 2009.
23) S Potghan, R Rajamenakshi and A Bhise, "Multi-Layer Perceptron Based Lung Tumor Classification," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2018, pp. 499-502, doi: 10.1109/ICECA.2018.8474864.
24) S R Jena, T. George and N. Ponraj, "Feature Extraction and Classification Techniques for the Detection of Lung Cancer: A Detailed Survey," 2019 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2019, pp. 1-6, doi: 10.1109/ICCCI.2019.8822164.
25) S Wang, L. Dong, X. Wang, and X. Wang, “Classification of pathological types of lung cancer from CT images by deep residual neural networks with transfer learning strategy,” Open Medicine, vol. 15, no. 1, pp. 190–197, 2020.

26) Sampangi Rama Reddy B R, Sumanta Sen, Rahul Bhatt, Murari Lal Dhanetwal, Meenakshi Sharma, Rohaila Naaz, Stacked neural nets for increased accuracy on classification on lung cancer, Measurement: Sensors, Volume 32, doi: https://doi.org/10.1016/j.measen.2024.101052
27) Shalini A, A Pankajam, V Talukdar, S Farhad, G Talele, E Muniyandy, and D Dhabliya. "Lung Cancer Detection and Recognition Using Deep Learning Mechanisms for Healthcare in IoT Environment". International Journal of Intelligent Systems and Applications in Engineering, vol. 12, no. 10s, Jan. 2024, pp. 208-16, https://ijisae.org/index.php/IJISAE/article/view/4363.
28) Sharaff A.; Gupta, H. Extra-Tree Classifier with Metaheuristics Approach for Email Classification. In Advances in Computer Communication and Computational Sciences; Advances in Intelligent Systems and Computing; Springer: Singapore, 2019; pp. 189–197.
29) T Zhou, H. Lu, J. Zhang, and H. Shi, "Pulmonary Nodule Detection Model Based on SVM and CT Image Feature-Level Fusion with Rough Sets," BioMed Research International, vol. 2016, 2016.
30) Thangamani M, Manjula Sanjay Koti, Nagashree B.A, Geetha V, Shreyas K.P, Sandeep Kumar Mathivanan & Gemmachis Teshite Dalu, Lung cancer diagnosis based on weighted convolutional neural network using gene data expression. Scientific Reports 14, 3656 (2024). https://doi.org/10.1038/s41598-024-54124-7
31) Tiwari L; Raja, R.; Awasthi, V.; Miri, R.; Sinha, G.R.; Alkinani, M.H.; Polat, K. Detection of lung nodule and cancer using novel Mask-3 FCM and TWEDLNN algorithms. Measurement 2021, 172, 108882.
32) V R Nitha, and Vinod Chandra S S 2023. "ExtRanFS: An Automated Lung Cancer Malignancy Detection System Using Extremely Randomized Feature Selector" Diagnostics 13, no. 13: 2206. https://doi.org/10.3390/diagnostics13132206