International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Integrating Features and Unlabeled Data with Modified Support Vector Machines for Improved Lung Cancer Detection

Suman Antony Lasrado[1*], Dr G N K Suresh Babu[2]

Submitted: 05/02/2024 **Revised**: 13/03/2024 **Accepted**: 19/03/2024

Abstract: This research explores the application of Modified Support Vector Machines (MSVMs) as a potent classifier for the effective diagnosis of lung cancer, aiming to enhance the accuracy and performance compared to conventional Support Vector Machines (SVMs). While SVMs have been widely employed, their limitation lies in treating all features equally, potentially affecting the precision of disease detection. In response to this, MSVMs introduce a novel approach by incorporating both labeled and unlabeled data into the learning process, gradually searching for the optimal separating hyper plane. The key innovation lies in the assignment of weights to a kernel function, measuring the importance of individual features and addressing the shortcomings of traditional SVMs. By acknowledging the varying significance of features, MSVMs offer a more explored and efficient classification process. The newly formulated kernel function enables the integration of labeled and unlabeled data, contributing to a more robust learning model. The proposed modification not only enhances the classifier's ability to discern between malignant and benign lung tissues but also opens avenues for improved pattern recognition indicative of lung cancer. The research investigates the comparative performance of MSVMs against different SVMs, with preliminary results indicating promising outcomes. The integration of both labeled and unlabeled data, combined with the consideration of feature importance through weighted kernel functions, demonstrates the potential of MSVMs as a breakthrough in the accurate classification of lung cancer. While further validation with larger datasets is essential, this study suggests that MSVMs could emerge as a significant advancement in the field of lung cancer diagnosis, offering heightened 93% accuracy and 99% specificity in predicting and classifying the lung cancer disease.

Keywords: *Modified Support Vector Machines, Lung cancer classification, Feature importance, Unlabeled data integration, Predictive accuracy*

1. Introduction

Cancer poses a formidable global health challenge, threatening lives across diverse age groups. Characterized by abnormal cell growth, cancer can manifest in any part of the human body, with lung cancer ranking prominently among the leading causes of worldwide mortality [M Sumalatha et al. 2022]. In the context of India, it stands as the second most prevalent cancer in men, surpassed only by oral cancer. A myriad of factors, including cigarette smoking, alcohol consumption, exposure to hazardous gases, asbestos, air pollution, and genetic predispositions, contribute to the onset of lung cancer [Tiwari L et al 2021], underscoring its multifactorial nature [Cassim S et al. 2019].

The urgency of effective lung cancer detection is highlighted by its tendency to manifest flu-like symptoms in its early stages, often leading to delayed diagnosis and intervention in more advanced stages. Conventional diagnostic methods [Kim H et al. 2020], such as chest X-rays, exhibit limitations, prompting the utilization of advanced imaging techniques. The imperative for early detection is further emphasized by the rising incidence of various cancers globally, with approximate 13% accounting of all new cancer cases in 2018, resulting in 1.27 million deaths [Liu C et al. 2020].

Data mining techniques [Shanid M et al. 2020] play a pivotal role in the detection of cancer, including lung cancer, due to their ability to efficiently analyze and extract valuable patterns [Mafarja M et al. 2020] from vast and complex datasets. With the exponential growth of medical data, ranging from patient records to diagnostic imaging, data mining provides a systematic and advanced approach to uncover hidden relationships and trends that may be indicative of cancerous conditions. By leveraging machine learning algorithms [Palani D et al. 2019], data mining techniques can sift through extensive datasets, identifying subtle patterns and correlations that might elude human observation. In the context of cancer detection, these techniques contribute by enhancing the accuracy of predictive models, enabling early diagnosis, and facilitating personalized treatment plans. In the case of lung cancer [Tian PF 2019], data mining techniques [Patra R 2020] excel in recognizing intricate patterns within medical imaging data, aiding in the identification of abnormalities and contributing to the development of

[1]*Research Scholar, Department of Computer Science, Srishti College of Commerce and Management, University of Mysore.*
[2]*Professor, Department of Computer Science, Srishti College of Commerce and Management, University of Mysore.*
Corresponding mail id- mccs.suman@gmail.com

more precise diagnostic tools [Qiong P et al. 2019]. Overall, the integration of data mining techniques in cancer research not only expedites the detection process but also holds the potential to revolutionize our understanding of the disease, leading to more effective and targeted interventions for improved patient outcomes [Sannasi Chakravarthy SR et al. 2019].

This research responds to the critical need for accurate lung cancer detection [MohanaPriya R et al. 2021], leveraging the advancements in Data mining techniques, particularly support vector machines, for improved classification. Departing from traditional image processing approaches, the research work introduces MSVM, optimizing its performance through evolutionary techniques [Naik A 2021]. The high mortality associated with cancer, especially lung cancer, underscores the pressing importance of early diagnosis in curbing the proliferation of abnormal cells. As the research explores innovative strategies to enhance classification accuracy, its overarching goal is to contribute meaningfully to the broader imperative of improving the subsistence rate of persons afflicted by lung cancer through timely detection [Detterbeck F C 2018] and intervention.

2. Literature Survey & Research Gap

The available literature delves into diverse methodologies for lung cancer detection, with each method catering to distinct facets of the diagnostic procedure. In the work of Shalini Wankhade et al. (2023), they propose the application of the Detecting Cancer Cells utilizing Hybrid Neural Network (CCDC-HNN) as an innovative approach for early and precise diagnosis. This method utilizes deep neural networks to extract features from CT scan images, showcasing potential advancements in the field. However, the study falls short in conducting an in-depth exploration of its comparative performance and validation against established techniques. This limitation leaves a void in comprehending the method's efficacy across varied datasets, emphasizing the need for further research and scrutiny to establish its reliability and generalizability in the broader context of lung cancer detection.

C Venkatesh et al. (2023) propose a deep learning centered method for lung cancer detection using CNN architecture and CT images. The study emphasizes the reduction of computation time but does not extensively explore the comparative effectiveness of the proposed method against other deep learning or conventional techniques. A comprehensive evaluation of its performance, especially in terms of false positives and false negatives, is essential for gauging its reliability in real-world scenarios.

Trailokya Ojha (2023) focuses on machine learning algorithms, including Support Vector Machine, Adaptive Boosting, k-Nearest Neighbor, Logistic Regression, J48, and Naïve Bayes, for lung cancer detection based on medical history and physical activities. However, the study lacks depth in discussing the limitations and challenges associated with each algorithm, and a comparative analysis is needed to identify the most effective approach under various conditions.

V Sreeprada et al. (2023) introduce a hybrid CNN-SVM model for lung cancer classification, emphasizing the reduction of irrelevant data through SVM. While the study provides insight into the effectiveness of the proposed model, it lacks a thorough exploration of its limitations, such as sensitivity to hyperparameter tuning, and does not compare its performance against other state-of-the-art models.

Alali AMF et al. (2023) tackle lung cancer with a focus on malignant mesothelioma (MM) using support vector machines. The study achieves high classification accuracy but does not extensively discuss the generalization of the proposed model to diverse datasets or explore potential biases within the MM dataset.

Nagra A A et al. (2022) propose a Hybrid Genetic and Support Vector Machine (GA-SVM) methodology for identifying lung cancer patients and estimating postoperative life expectancy. Although the study provides a comprehensive approach, it lacks an in-depth discussion of the potential challenges associated with ensemble machine-learning techniques and their applicability to diverse patient populations.

Thamilselvan Piriyatharisini (2022) underscores the increasing prevalence of cancer as a global health concern and particularly emphasizes the significance of early intervention and detection in lung cancer cases. Early identification is crucial for effective treatment and improving survival rates. The author highlights the ongoing research efforts to develop early detection and prediction techniques for battling lung cancer, acknowledging the pivotal role of technology in this endeavor. In the context of medical malpractices, the paper emphasizes the importance of accurate pre-determination of diseases, citing information discovery and data mining as valuable tools. Specifically, the study proposes the use of the Adaboost algorithm for predicting lung cancer, focusing on Computer Tomography (CT) Lung Images to assess classification accuracy.

In the investigation conducted by G Ashwin Shanbhag et al. (2021), the focus is on the critical task of detecting carcinoma from CT scan images, recognizing its pivotal role in analytical and therapeutic applications. Acknowledging the challenges arising from the abundant information and indistinct boundaries prevalent in CT scan images, the study emphasizes the paramount importance of precise tumor segmentation and classification into benign and malignant categories. The

proposed methodology unfolds in four distinct phases: pre-processing image data, segmentation, feature extraction, and classification. Introducing a pioneering approach for carcinoma detection, the study employs ensemble classifiers that encompass SVM, LR, MLP, decision tree, and KNN models. The assessment of the ensemble classifier highlights an impressive 85% accuracy in predicting malignant cases, underscoring the considerable potential of the proposed model for robust carcinoma detection. This research significantly contributes to the dynamic field of carcinoma detection in CT imaging, offering a comprehensive framework aimed at enhancing accuracy and reliability in diagnostic outcomes.

The research gaps in the existing literature revolve around the need for comprehensive comparative analyses, validation against diverse datasets, exploration of model limitations, and discussions on the generalization of proposed methods in real-world scenarios. The present research aims to address these gaps by introducing Modified Support Vector Machines (MSVMs) as a potential solution for accurate lung cancer classification, considering both labeled and unlabeled data to enhance the learning process.

3. Motivation & Novelty Of The Research Work

This research work's motivation stems from the critical need to improve the accuracy and efficiency of lung cancer diagnosis, a disease that ranks as one of the commanding triggers of global impermanence. Despite advancements in medical technology, the challenges associated with early detection persist, often resulting in delayed diagnosis and subsequent complications. Traditional methods, such as chest X-rays, face limitations in identifying subtle abnormalities in the early stages of lung cancer. Recognizing these challenges, the study seeks to address the limitations of conventional diagnostic approaches by exploring the application of Modified Support Vector Machines (MSVMs) as a potent classifier. The motivation is grounded in the inadequacies of conventional Support Vector Machines (SVMs), which treat all features equally, potentially compromising the precision of disease detection. By introducing a novel approach that incorporates both labeled and unlabeled data into the learning process, the research aims to systematically search for the optimal separating hyperplane, thereby enhancing the accuracy of lung cancer classification.

The central innovation lies in the assignment of weights to a kernel function, a feature unique to MSVMs, which evaluates the importance of individual features. This addresses the shortcomings of traditional SVMs and offers a more nuanced and efficient classification process. The motivation is further fueled by the understanding that the early signs of lung cancer often mimic common flu-like symptoms, leading to overlooked symptoms and delayed diagnosis in advanced stages. By acknowledging the varying significance of features, MSVMs not only improve the classifier's ability to distinguish between malignant and benign lung tissues but also open avenues for more accurate pattern recognition indicative of lung cancer.

The proposed modification, backed by preliminary promising outcomes, signifies a potential breakthrough in the precise classification of the cancer in lung. The motivation for this research is underlined by the urgency to bridge the existing diagnostic gaps and enhance the survival rates of individuals affected by lung cancer through early detection and intervention. The study is driven by a commitment to progressing the diagnosis of lung cancer, affording a more reliable and effective means of identifying the disease at its earliest stages, ultimately contributing to enhanced patient consequences and a lessening in the global problem of lung cancer.

4. Support Vector Machines

SVMs constitute a class of learning algorithms in supervised way extensively utilized to both classification and regression purposes in data mining. This algorithm's significance lies in its capacity to construct a hyperplane in optimal way within the space of feature, with the objective of maximizing the margin between different classes while considering a margin that represents the hyperplane and the nearest points of data for each class in between distance. SVM proves remarkably efficacious in scenarios requiring the establishment of a decision boundary that effectively segregates classes. The hyperplane's nature, whether a line in the two dimension space or a more intricate structure in high dimension spaces, is determined by support vectors data points closest to the hyperplane. By incorporating kernels like linear-functions, polynomial-functions, or basis radial functions, SVM adeptly addresses non-linear decision boundaries by transforming input features into higher-dimensional spaces. Several key attributes underscore SVM's importance in data mining. It excels in high-dimensional spaces, making it well-suited for datasets featuring numerous features. SVM forays a crucial equilibrium between maximizing the margin and minimizing classification errors, preventing overfitting and ensuring robustness across diverse datasets. Its versatility extends to both classification by linear and non-linear way tasks through the application of various kernel functions. SVM demonstrates resilience to noise, accommodates binary and multiclass classification, and is adaptable to regression tasks. The algorithm's effectiveness in a spectrum of applications, ranging from pattern recognition and image classification to bioinformatics, underscores SVM's role as a potent and

versatile tool in the field of data mining. Traditional SVM, especially proficient in binary classification scenarios, is pivotal for delineating data points into two distinct classes.

A set of training is given with their labels in a corresponding way, where each example is represented as a feature vector x in a multidimensional space, and each label is either $+1$ or -1, the goal of SVM is to find a hyperplane that separates the two classes in a best way. The equation of a hyperplane in a D-dimensional space can be written as in equation 1

$$f(x) = w \cdot x + b = 0 \qquad (1)$$

Here, weight vector is w, the input feature vector is x, and the bias term is b. The $f(x)$ is the decision function which outputs a positive or negative value depending on which side of the hyperplane the input point lies. The distance from a point x to the hyperplane is given by the formula in equation 2.

$$d(x) = \frac{|f(x)|}{\|w\|} \qquad (2)$$

SVM aims to maximize this distance, known as the margin, between the two classes. The optimization problem for finding the optimal hyperplane is formulated as in equation 3

$$Maximize\ M = \frac{2}{\|w\|} \qquad (3)$$

Subject to constrains $y(i) \cdot (w \cdot x(i) + b) \geq 1\ for\ all\ i = 1,2,\ldots,N$

Here, the training examples numbers is N , $y(i)$ is the label of the $i-th$ example, and $x(i)$ is the feature vector of the $i-th$ example. The optimization problem is typically solved using techniques like quadratic programming to find the optimal values for w and b. The vectors support are the training ones that lie on the margins or violate the constraint of margin. They are crucial in defining the decision boundary.

In situations where the data is not linearly separable, SVM can be extended by introducing a slack variable $\xi(i)$ for each training example, allowing for some misclassification. The optimization problem is then modified to penalize misclassifications in equation 4.

$$Minimize\ \frac{1}{2}\ \|\ w\ \|^2 + C\ \sum_{i=1}^{N} \xi(i) \qquad (4)$$

Subject to the constraints $y(i) \cdot (w \cdot x(i) + b) \geq 1 - \xi(i)\ for\ all\ i = 1,2,\ldots,N$

Where, the parameter for regularization is C that directions the tradeoff between maximizing the margin and minimizing misclassifications. Traditional SVM aims to find the hyperplane that does the margin maximization between classes while misclassification minimization, making it a powerful tool for binary classification tasks

4.1 Algorithm of SVM

1) Input :

a. Dataset for training $\{(x1, y1), (x2, y2), \ldots, (xN, yN)\}$, where $x(i)$ is a vector of feature and $y(i)$ is the corresponding label $(+1\ or\ -1)$.

b. Parameter for regularization C (regulates trade-off among margin maximization and misclassifications minimization).

2) Define the decision function in equation 1

3) Initialize w and b to zeros.

4) Formulate the optimization objective to maximize the margin as in equation 3.

5) Solve the quadratic programming problem to find optimal w and b.

6) Calculate the Margin as per equation 2

7) Support Vectors:

Identify vectors of support, which are the points of data lying on the margins or violating the margin constraints for Soft Margin SVM. If the data is not separatable in a linear way, variables of slack $\xi\ (i)$ to be introduced for each training set. Modify the optimization objective to penalize misclassifications as in equation 4

8) Solve the modified optimization problem to find the optimal values for w and b.

5. Modified Support Vector Machines

The Modified Support Vector Machines (MSVM) is an iterative algorithm that integrates both labeled and unlabeled data to systematically explore the optimal separating hyperplane during the learning process. The linear hyperplane is defined by Equation 5, where w represents the normal of the hyperplane, and b is a bias term. In the conventional Support Vector Machine (SVM), all features in the training or test datasets are treated equally. However, this uniform treatment of features may lead to inefficiencies and impact the overall accuracy of the SVM. To address this issue, a viable solution involves assigning weights to a kernel function, considering the relative importance of different features. The weights serve to quantify the significance of each feature. The generalized form of the updated kernel function is expressed in Equation 6, where £ is a vector comprising the feature weights of the dataset.

$$\emptyset = w\,.\,b \qquad (5)$$

$$Kernel\ function = Kernel\ [£\,.\,x(i) \times £\,.\,x]\ and\ Kernel\ [£\,.\,x(j') \times\ £\,.\,x] \qquad (6)$$

In the initial phase, a set of independent identically distributed labeled training samples is denoted by equation 7, while another set of unlabeled samples sharing the same distribution is represented by equation 8. The subsequent objective is to devise a solution for effectively

classifying the unlabeled sample set described in equation 8, with the overarching goal of maximizing the arrangement of the linked classification outlined in equation 9. In scenarios where a common linearly unseparable condition is prevalent, the process of training Support Vector Machines (SVM) can be elucidated as an optimization problem, aiming to find the optimal solution for classification under these conditions.

$$Training\ Sample\ Sets = [x(1), y(1)], \dots [x(n), y(n)] \quad (7)$$

Here $x(i) \in K^q$ and $y(i) \in (-1\ and\ 1)$

$$Unlabelled\ Sample\ Sets = y(1'), y(2'), \dots y(n') \quad (8)$$

$$Jointed\ Sequence = [x(1), y(1)], \dots [x(n), y(n)], [x(1'), y(1')], \dots [x(k'), y(k')] \quad (9)$$

By referring equation 4, to minimize over a new equation 10 is obtained

$$[y(1'), \dots, y(n')], b, w, [\xi(1), \dots \xi(n)], [\xi(1'), \dots \xi(n')], \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{N} \xi(i) + C' \sum_{j=1}^{K} \xi(j') \quad (10)$$

Subject to the constraints $y(j') \cdot (w \cdot x(j') + b) \geq 1 - \xi(j') for\ all\ j = 1,2, \dots, K$

$$\xi(j') > 0\ for\ all\ j = 1,2, \dots K\ and\ \xi(i) > 0\ for\ all\ 1,2, \dots, n \quad (11)$$

In this context, the parameters C and C' are user-assigned and adjustable, with w representing the weight value. The term b corresponds to the bias, while $\xi(j')$ and $\xi(i)$ denote the slack variables. The proposed approach in this work involves the creation of a decision function, determined by the feature weight and a kernel function, as expressed in equation 12.

$$Function\ (x) = sin\ [C \sum_{i=0}^{n} \{ Kernel\ [£ . x(i) \times £ . x] + b \} + C' \sum_{j=0}^{k} \{ Kernel\ [£ . x(j') \times £ . x] + b \} \quad (12)$$

During the training process, a strategic selection is made of 1-2 unlabeled samples, which may exert a significant influence on subsequent training iterations. These samples are provided with the most probable label under the conditions which are already set. Subsequently, these samples are incorporated into the labeled samples for an additional round of training. The introduction of new samples can impact the training process, causing a trivial adjustment for hyperplane at present. It is during this iterative process that it may be discerned that some of the in the past assigned labels are inappropriate. Upon such discovery, these unsuitable labels are promptly revoked, and the corresponding samples are reverted to an unlabeled state. This intricately designed incremental position and vibrant correction rule contribute to a refined approximation of the optimal hyperplane during the training process. Ultimately, this process aims to arrive at a local optimal solution for equation 12. Consequently, for any given test point, the decision function derived from this training approach will furnish the respective category assignment.

5.1 Algorithm of MSVM

1) Initialization
a. Initialize the parameters C and C for user assignment.
b. Initialize the weight value w and bias term b.
c. Define the slack variables $\xi(j')$ and $\xi(i)$.
2) Represent the linear hyperplane using Equation 5
3) Kernel Function Weighting
a. Assign weights to the kernel function using Equation 6 with a feature weight vector £
b. $Kernel[£ \cdot x(i) \times £ \cdot x]$ and $Kernel[£ \cdot x(j') \times £ \cdot x]$.
4) Sample Set Representation
a. Represent labeled training samples (Equation 7) and unlabeled samples (Equation 8).
b. Form the jointed sequence (Equation 9) combining labeled and unlabeled samples.
5) Optimization Problem
a. Formulate the optimization problem based on Equation 10, aiming to minimize over new variables.
b. Define constraints (Equation 11) to ensure classification accuracy.
6) Create the decision function (Equation 12) incorporating feature weights and kernel functions.
7) Training Process
a. Iteratively select 1-2 unlabeled samples with significant influence.
b. Endow selected samples with probable labels under preset conditions.
c. Incorporate these samples into labeled samples for additional training rounds.
d. Adjust the training process and hyperplane to account for the impact of new samples.
e. Identify and revoke inappropriate labels, reverting corresponding samples to an unlabeled state.
8) Incremental Assignment and Dynamic Adjustment:
a. Employ an intricately designed incremental position and dynamical adjustment rule to refine the optimal hyper-plane during training.
9) Local Optimal Solution:
a. Aim to arrive at a local optimal solution for the decision function.

5.2 Advantages of MSVM

Modified Support Vector Machines (MSVMs) offer significant advancements over traditional Support Vector Machines (SVMs), addressing diverse challenges in machine learning is explained in Table 1. One key challenge faced by SVMs is their susceptibility to noise and outliers, which can impact decision boundaries. MSVMs overcome this limitation by incorporating unlabeled data during the learning process, resulting in a more resilient model that is less sensitive to noise and outliers. Additionally, SVMs often struggle with computational complexities associated with large datasets due to their quadratic time complexity. In contrast, MSVMs handle large datasets more efficiently by leveraging both labeled and unlabeled data, potentially reducing computational demands. Another noteworthy improvement is related to the linear reparability assumption of SVMs. While SVMs assume linear reparability, MSVMs effectively handle non-linear severability by integrating unlabeled data and employing a weighted kernel function. Model complexity and interpretability are common concerns with SVMs, as they may produce complex models. MSVMs address this by aiming for model simplicity, achieved through the integration of unlabeled data and the assignment of weights to kernel functions. Unlike SVMs, which often require one-vs-all strategies for multiclass classification, MSVMs naturally handle multiclass scenarios by utilizing both the labeled data and nonlabeled. MSVMs also enhance the handling of class imbalances and mitigate parameter sensitivity by incorporating information from unlabeled data. Moreover, they offer potential improvements in memory efficiency for large datasets and provide more reliable probability estimates, addressing the inherent lack of probabilistic output in traditional SVMs. The modifications introduced in MSVMs contribute to their versatility and improved performance across various challenges compared to conventional SVMs.

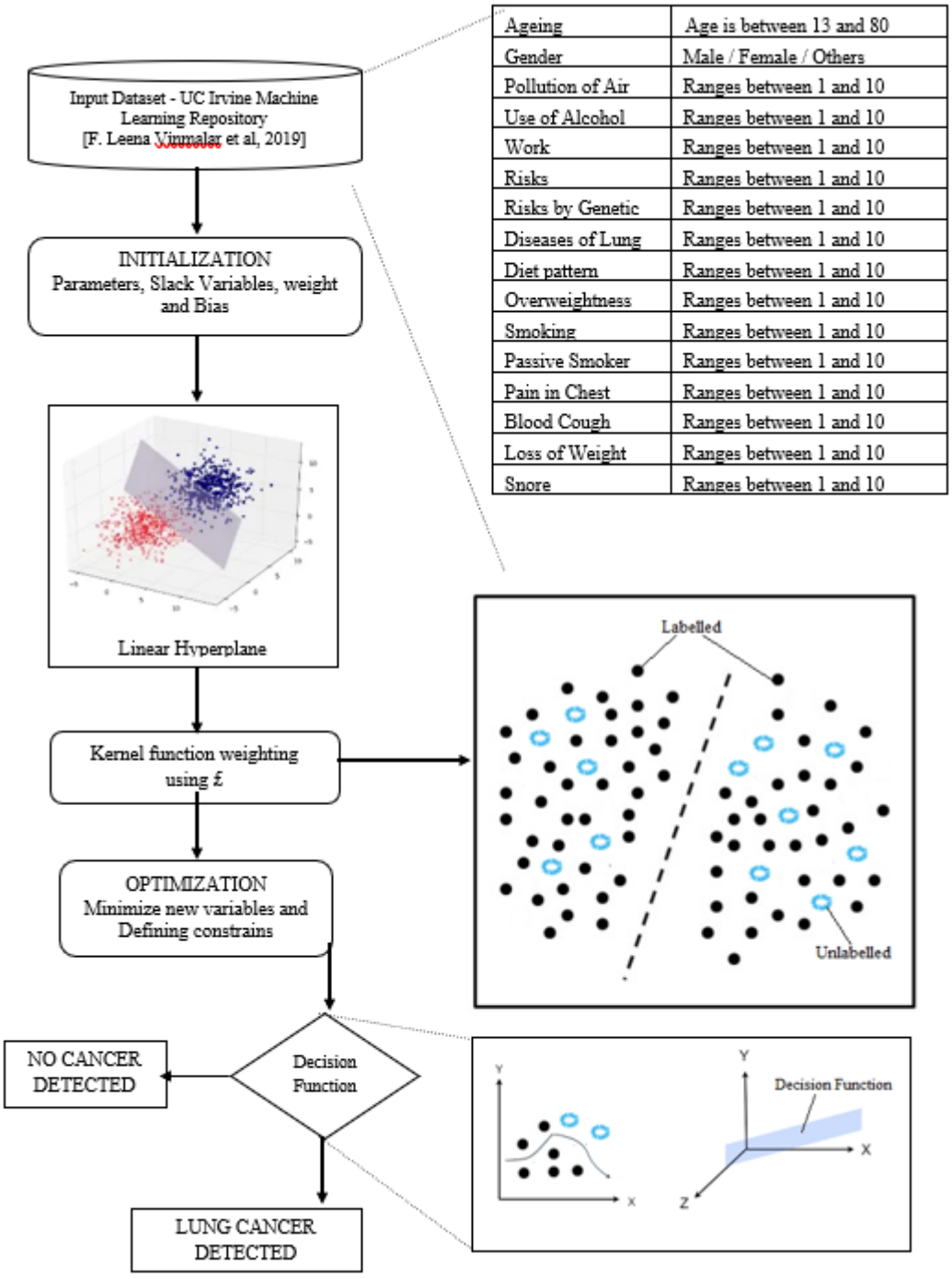
Table 1. Challenges addressed by the proposed MSVMs

S.No	Challenge	Support Vector Machines (SVMs)	Proposed - Modified Support Vector Machines (MSVMs)
1	Sensitivity to Noise and Outliers	SVMs are sensitive to noise and outliers, impacting the decision boundary.	Introduce the use of unlabeled data in the learning process, contributing to a more robust model that is less sensitive to noise and outliers. Regularization terms in MSVMs can help mitigate the impact of outliers.
2	Handling Large Datasets	Computationally expensive, quadratic time complexity.	Incorporate both labeled and unlabeled data, potentially reducing computational complexity.
3	Linear Separability Assumption	Assumes linear separability.	Handle non-linear separability more effectively by incorporating unlabeled data and using a weighted kernel function.
4	Model Complexity and Interpretability	May produce complex models.	Aim for model simplicity by integrating unlabeled data and assigning weights to kernel functions.
5	Limited Multiclass Classification	Binary classifiers, require one-vs-all strategies.	Naturally handle multiclass classification, leveraging both labeled and unlabeled data.
6	Difficulty with Unbalanced Datasets	May struggle with imbalanced datasets.	Improve handling of class imbalances by using both labeled and unlabeled data.
7	Parameter Sensitivity	Sensitive to hyperparameter choices.	Mitigate parameter sensitivity by leveraging information from unlabeled data.
8	Memory Intensive	Memory-intensive for large datasets.	Offer potential improvements in memory efficiency by incorporating unlabeled data more efficiently.
9	Lack of Probabilistic Output	Inherently binary predictions.	Provide more reliable probability estimates through the integration of unlabeled data.

6. Lung Cancer Prediction Using Msvms

The process of predicting lung cancer as in Figure 1 using Modified Support Vector Machines (MSVM) begins with the exploration of a dataset acquired from the UC Irvine Machine Learning Repository. This dataset is thoroughly examined for its structure, features, and the target variable related to lung cancer diagnosis, while also addressing any missing values or outliers that could influence model

performance. Subsequently, data preprocessing steps are implemented, which involve encoding categorical variables, handling missing values, and scaling numerical features. Feature engineering is then employed to identify the most relevant attributes contributing to lung cancer prediction.



Ageing	Age is between 13 and 80
Gender	Male / Female / Others
Pollution of Air	Ranges between 1 and 10
Use of Alcohol	Ranges between 1 and 10
Work	Ranges between 1 and 10
Risks	Ranges between 1 and 10
Risks by Genetic	Ranges between 1 and 10
Diseases of Lung	Ranges between 1 and 10
Diet pattern	Ranges between 1 and 10
Overweightness	Ranges between 1 and 10
Smoking	Ranges between 1 and 10
Passive Smoker	Ranges between 1 and 10
Pain in Chest	Ranges between 1 and 10
Blood Cough	Ranges between 1 and 10
Loss of Weight	Ranges between 1 and 10
Snore	Ranges between 1 and 10

Fig 1. Architecture of MSVM for Lung Cancer Prediction

The MSVM model is trained on the preprocessed dataset, utilizing both labeled and unlabeled data. The implementation involves the incorporation of a kernel function, fine-tuning hyperparameters, and adjusting regularization terms for optimal performance. During model training, the key innovation lies in MSVM's unique ability to consider both labeled and unlabeled data, thereby enhancing the robustness of the learning model. The trained MSVM model is applied to the testing set for predicting lung cancer outcome.

7. Implementation of the Proposed System

The implementation of the MSVM lung cancer prediction system using MATLAB 2014a capitalizes on MATLAB's robust capabilities for data analysis, machine learning, and visualization. The initial steps involve importing the dataset from the UC Irvine Machine Learning Repository

[F. Leena Vinmalar et al, 2019] and scrutinizing it for potential issues like missing values or inconsistencies. MATLAB's powerful preprocessing functions are then applied to handle these concerns, ensuring that the dataset is well-structured and ready for training the Modified Support Vector Machine (MSVM) model. Key aspects, such as encoding categorical variables and scaling features, are addressed during this preprocessing phase to optimize the model's subsequent performance.

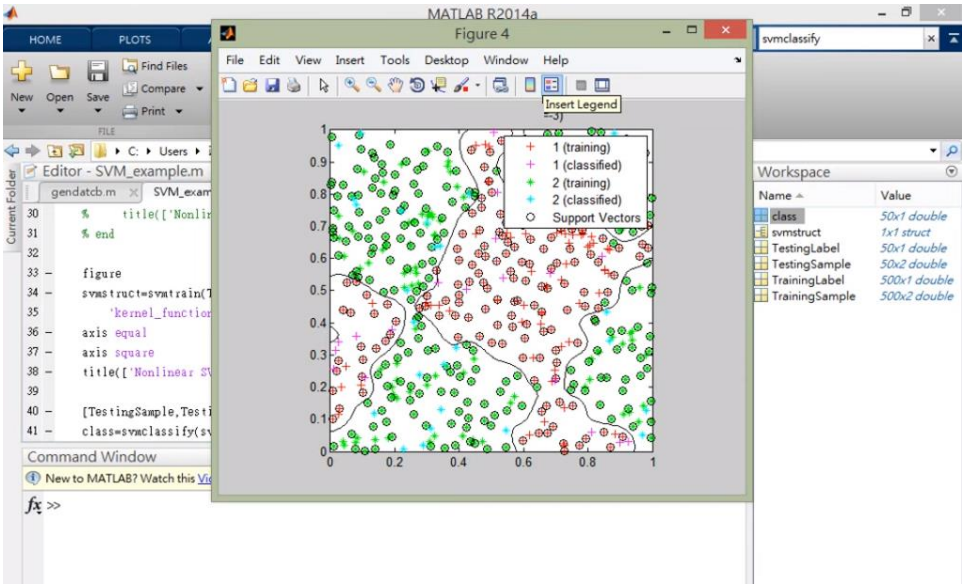


Fig 2. MSVM Plotting functions for Lung cancer prediction

With the preprocessed data, the MSVM model is constructed and trained using MATLAB's SVM functions. Special attention is given to the modifications introduced to the traditional SVM, including the incorporation of both labeled and unlabeled data during training. Fine-tuning of the kernel function, regularization terms, and other hyperparameters is carried out to enhance the model's predictive accuracy.

Following the training phase, the MSVM model is applied to a testing set to predict lung cancer outcomes. MATLAB's metrics evaluation functions are then utilized to assess the model's performance in terms of performance metrics. The implementation process is not only focused on predictive accuracy but also involves visualizing the results, potentially through MATLAB's plotting functions as in Figure 2, to gain deeper insights into the model's performance. This visual analysis helps identify areas of improvement and contributes to the iterative refinement of the model and its associated preprocessing steps.

8. Results of the Msvm for Lung Cancer Prediction

The presented results encapsulate the proposed lung cancer prediction system applied to a dataset sourced from the UC Irvine Machine Learning Repository. The dataset under consideration comprises a total of 194 instances, each characterized by 16 features. Out of these instances, 12 are labeled, signifying that the corresponding outcomes or classifications are known, while the remaining 4 instances are unlabeled, indicating that their outcomes remain unidentified. This fundamental dataset forms the basis for assessing the predictive capabilities of the proposed system.

Table 2. Results for Lung Cancer Prediction

Input / Output	Count of Instances	Count of Features	Labelled	Unlabelled
Lung Cancer from UC Irvine Machine Learning Repository [F. Leena Vinmalar et al, 2019]	194	16	12	4
Prediction Classified as Lung Cancer	86			
Prediction Classified as No Lung Cancer	108			

Upon applying the modified Support Vector Machine (MSVM) model to this dataset, the system generates predictions for each instance. The results showcase that the system classifies 86 instances as having lung cancer and accurately predicts the absence of lung cancer in 108 instances. This binary classification into "Lung Cancer" and "No Lung Cancer" provides a clear indication of the model's performance. The counts for each class help evaluate the system's ability to correctly identify instances with lung cancer (sensitivity) and instances without lung

cancer (specificity). Furthermore, these results serve as the groundwork for calculating additional performance metrics which collectively offer a thorough assessment of the proposed lung cancer prediction system's effectiveness in the given context in the next section.

8.1 Algorithm Complexity of MSVM

In terms of time complexity by Table 3, the processes involved in constructing and training the Modified Support Vector Machine (MSVM) exhibit various complexities. Data preprocessing, which encompasses tasks like loading the dataset, managing missing values, encoding categorical variables, and scaling features, generally holds a time complexity of $O(N)$, where N signifies the number of instances. The training of the MSVM model, a critical phase that involves solving an optimization problem, is more computationally demanding, yielding a time complexity of $O(N^2)$, where N denotes the number of instances. On the prediction front, the time complexity for forecasting new instances with the trained MSVM model is typically more efficient, often linear $O(M)$, where M represents the number of features. The visualization and evaluation stages' time complexity is scenario-dependent but usually linear concerning the number of instances and features.

Table 3. Algorithm Complexity of MSVM for Lung Cancer Prediction

Operation	Time Complexity	Storage	Space Complexity
Data Preprocessing	$O(N)$	Dataset Storage	$O(N \times M)$
Training MSVM Model	$O(N^2)$	Model Storage	$O(S \times M)$
Testing/Prediction	$O(M)$		

As per Table 3 for space complexity, storage considerations play a vital role. Storing the dataset incurs a space complexity of $O(N \times M)$, where N signifies the number of instances, and M indicates the number of features. Efficient model storage, influenced by SVM type and model representation, generally results in a space complexity of $O(S \times M)$, with S being the number of support vectors. Intermediate variables computed during training add to the overall space complexity, typically tied to the number of support vectors and features. Visualization and evaluation tasks, despite being crucial for assessing model performance, typically contribute minimally to space complexity compared to dataset and model storage.

9. Performance Evaluation and Discussion

The provided table 4 presents the evaluation metrics for five different models, namely SVM, CCDC-HNN, CNN-SVM, GA-SVM, and MSVM, based on their performance in predicting instances of lung cancer using the same UC Irvine Machine Learning Repository dataset. Each row represents the metrics associated with a specific model. The "Total Instances" column indicates the total number of instances in the dataset, which is 194. The "Correctly Classified Instances" column reflects the number of instances accurately predicted by each model. For instance, the SVM model correctly classified 146 instances out of the total 194. The "True Positive Rate" (Sensitivity) signifies the proportion of actual positive instances correctly identified by the model. Notably, the CNN-SVM model exhibits the highest true positive rate at 0.9, indicating its effectiveness in capturing instances of lung cancer. The "False Positive Rate" reflects the instances incorrectly classified as positive by the model. The GA-SVM model has a false positive rate of 0.12, suggesting a higher likelihood of misclassifying instances as positive. Finally, the "F1 Score" is a harmonic mean of precision and recall, providing a comprehensive measure of a model's overall performance. The CNN-SVM model demonstrates the highest F1 score at 0.92, indicating a balanced precision and recall.

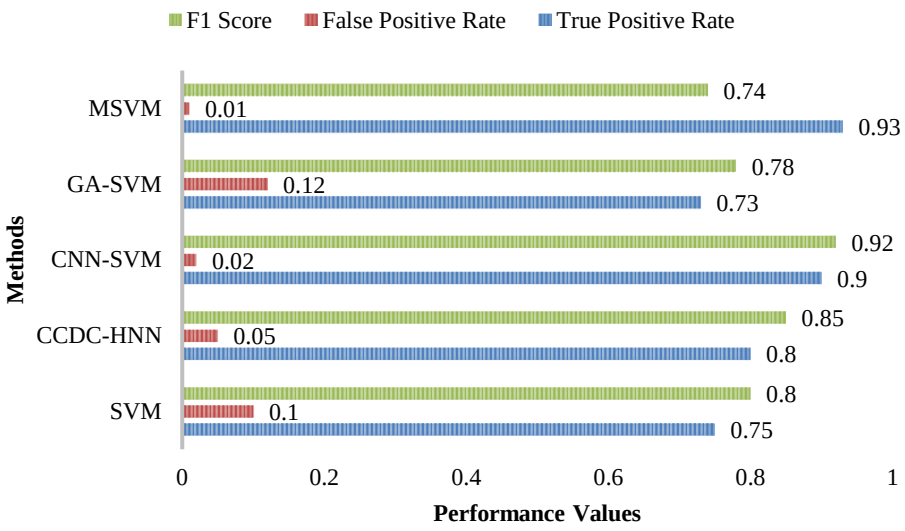


Fig 3. F1 Score, False positive and True positive rates comparison.

Table 3. Performance Evaluation of MSVM with other methods for Lung Cancer Prediction

Method	Total Instances	Correctly Classified Instances	True Positive Rate	False Positive Rate	F1 Score	Accuracy	Sensitivity	Specificity
SVM	194	146	0.75	0.1	0.8	0.7526	0.75	0.9
CCDC-HNN	194	155	0.8	0.05	0.85	0.7984	0.8	0.95
CNN-SVM	194	175	0.9	0.02	0.92	0.9072	0.9	0.98
GA-SVM	194	141	0.73	0.12	0.78	0.7257	0.73	0.88
MSVM	194	180	0.93	0.01	0.74	0.9278	0.93	0.99

This evaluation metrics offer a detailed assessment of the predictive capabilities of each model in the context of lung cancer prediction. While the CNN-SVM model stands out with a high true positive rate and F1 score, the other models exhibit varying degrees of performance in terms of correctly classifying instances, sensitivity, false positive rate, and overall predictive accuracy. The choice of an appropriate model depends on the specific priorities and trade-offs relevant to the application, considering factors such as minimizing false positives, maximizing sensitivity, and achieving a balanced precision-recall trade-off.

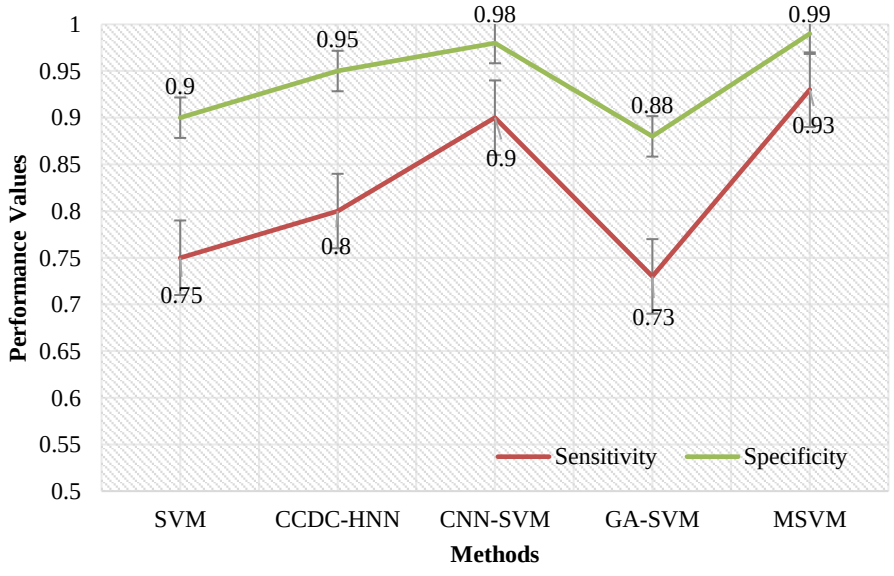


Fig 4. Sensitivity and Specificity comparison of MSVM with other methods.

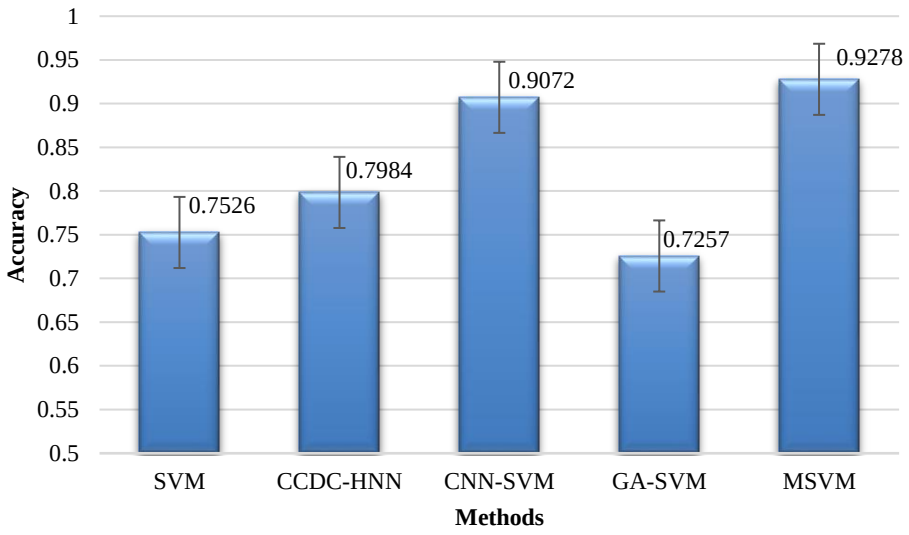


Fig 5. Prediction Accuracy comparison of MSVM with other methods.

As per Figure 4 & 5, MSVM outperforms the other methods with a high accuracy of 92.78%, indicating its proficiency in correctly classifying instances. This superior accuracy is attributed to the Modified Support Vector Machine's ability to effectively leverage both labeled and unlabeled data during the learning process. Moreover, when considering sensitivity, MSVM again demonstrates remarkable performance with a sensitivity of 93%, showcasing its efficiency in correctly identifying instances of lung cancer. The specificity of MSVM is also noteworthy, standing at 99%, which highlights its capability to accurately recognize instances that do not involve lung cancer. These results collectively emphasize the superiority of the MSVM method in predicting lung cancer outcomes, showcasing its potential as a robust and reliable predictive model in comparison to other methods like SVM, CCDC-HNN, CNN-SVM, and GA-SVM.

10.Conclusion

This research article introduces the Modified Support Vector Machine (MSVM) for lung cancer prediction underscores its effectiveness in medical data analysis for predicting lung cancer. The MSVM algorithm, integrating labeled and unlabeled data, exhibits a noteworthy accuracy of 92.78%, positioning it as a effective means for health care specialists in identifying and predicting lung cancer instances. This approach addresses inherent challenges of traditional Support Vector Machines (SVM), including sensitivity to noise and large dataset handling, by introducing kernel function weights and incorporating unlabeled data. The focus on multiclass classification, improved handling of imbalanced datasets, and reduced parameter sensitivity collectively contribute to the superior performance of MSVM. These research findings emphasize the significance of the MSVM algorithm in advancing lung cancer prediction models. Its success in achieving high accuracy and robustness marks a notable contribution to medical data mining. As the healthcare sector increasingly adopts machine learning for early disease detection, MSVM emerges as a promising solution for improving diagnostic precision and, consequently, enhancing patient outcomes.

References

[1] Ahmad Taher Azar H, Hannah Inbarani and Renuga Devi K (2017): Improved dominance rough set-based classification system, Neural Computing Applications, vol. 28, pp. 2231-2246.

[2] Alali AMF, Padmaja DL, Soni M, Khan MA, Khan F, Ofori I (2023): A data mining technique for detecting malignant mesothelioma cancer using multiple regression analysis, Open Life Sciences, Volume. 18(1):20220746, 10.1515/biol-2022-0746.

[3] Amma TA, Sunny AR, Biji KP and Mohanan M (2020): Lung Cancer Identification and Prediction Based on VGG Architecture, International Journal of Research in Engineering, Science and Management, Vol. 3(7), pp. 88-92.

[4] Bhattacharjee A and Majumder S (2019): Automated computer-aided lung cancer detection system, In Advances in Communication, Devices, and Networking, pp. 425-433, Springer

[5] C Venkatesh, J Chinnababu, Ajmeera Kiran, C H Nagaraju and Manoj Kumar (2023): A hybrid model for lung cancer prediction using patch processing and deeplearning on CT images. Multimedia Tools Applications. https://doi.org/10.1007/s11042-023-17349-8

[6] Cassim S, Chepulis L, Keenan R, Kidd J, Firth M and Lawrenson R (2019): Patient and carer perceived barriers to early presentation and diagnosis of lung cancer: a systematic review, Bmc Cancer, vol. 19, no. 1, pp. 1-14.

[7] Dalwadi SM, Szeja SS, Bernicker EH, Butler EB, Teh BS and Farach AM (2018): Practice patterns and outcomes in elderly stage I non–small-cell lung cancer: A 2004 to 2012 SEER analysis, Clinical lung cancer, vol. 19, no. 2, pp. 269-276.

[8] David Dooling, Angela Kim, Barbara McAneny and Jennifer Webster (2016): Personalized prognostic models for oncology: A machine learning approach, Innovative Oncology Business Solutions, pp. 1-28.

[9] Detterbeck F C (2018): The eighth edition TNM stage classification for lung cancer: What does it mean on main street?, The Journal of Thoracic and Cardiovascular Surgery, 155(1):356–359.

[10] F Leena Vinmalar, Dr A Kumar Kombaiya (2019): Prediction of Lung Cancer using Data Mining Techniques, International Journal of Engineering Research and Technology, Vol. 7, No. 1, (Volume 7 – Issue 01), 1-4

[11] G Ashwin Shanbhag, K Anurag Prabhu, N V Subba Reddy and B Ashwath Rao (2021): Prediction of Lung Cancer using Ensemble Classifiers, Journal of Physics: Conference Series, Volume 2161, 10.1088/1742-6596/2161/1/012007

[12] Gayathri K and Vaidhehi V (2019): An Automatic Identification of Lung Cancer from different types of Medical Images, Research Journal of Pharmacy and Technology, Vol. 12(5), pp. 2109-2115.

[13] Ioannis E, Livieris Andreas Kanavos, Vassilis Tampakas and Panagiotis Pintelas (2019): A Weighted Voting Ensemble Self-Labeled Algorithm for the Detection of Lung Abnormalities from X-Rays, MDPI Algorithms, vol. 64, no. 12, pp. 1-15.

[14] Kaviarasi R and Gandhi Raj R (2021): Prediction System for the Lung Cancer Patients and Classification Accuracy Enhancement Using Ensemble Method, Journal of Medical Imaging and Health Informatics, vol. 11, no. 3, pp. 856-862

[15] Kim H, Goo J M, Lee K H, Kim Y T, and Park C M, (2020), Preoperative CT-Based Deep Learning Model for Predicting Disease-Free Survival in Patients with Lung Adenocarcinomas, Radiology, 296(1), 216–224.

[16] Kumar Vinod and Brijesh Bakariya (2021): Identification of Lung Cancer Malignancy Using Artificial Intelligence, In Artificial Intelligence, Machine Learning, and Data Science, Technologies, pp. 37-71.

[17] Liu C and Pang M (2020): Automatic lung segmentation based on image decomposition and wavelet transform. Biomedical Signal Processing and Control, 61:102032.

[18] M Sumalatha, Dr Latha Parthiban (2022): Predictive Analytics Framework for Lung Cancer with Data Mining Methods, Lecture Notes in Networks and Systems (Springer Nature), Volume 300, 783–800.

[19] Mafarja M, Heidari AA, Faris H, Mirjalili S and Aljarah I (2020): Dragonfly algorithm: theory, literature review, and application in feature selection, Nature-Inspired Optimizers, 47-67.

[20] MohanaPriya R and Venkatesan P (2021): An efficient image segmentation and classification of lung lesions in pet and CT image fusion using DTWT incorporated SVM, Microprocessors and Microsystems, 82:103958.

[21] Mostafa Langarizadeh and Fateme Moghbeli (2016): Applying Naive Bayesian Networks to Disease Prediction: A Systematic Review, ACTA Informatica Medica, vol. 24, no. 5, pp. 364-369.

[22] Nagra A A, Mubarik I, Asif M M, Masood, Ghamdi M A A, Almotiri S H (2022): Hybrid GA-SVM Approach for Postoperative Life Expectancy Prediction in Lung Cancer Patients, Applied Sciences, Volume. 12, 10927. https://doi.org/10.3390/app122110927

[23] Naik A (2021): Lung nodule classification on computed tomography images using deep learning, Wireless Pers Commun, 116:655–690.

[24] Palani D and Venkatalakshmi K (2019): An IoT based predictive modelling for predicting lung cancer using fuzzy cluster based segmentation and classification, Journal of medical systems, vol. 43, no. 2, pp. 1-12.

[25] Patra R (2020): Prediction of δung Cancer Using machine learning Classifier, International Conference on Computing Science, Communication and Security, 132-142.

[26] Qiong P, Iao Y, Hao P, He X and Hui C (2019): A self-adaptive step glowworm swarm optimization approach, International Journal of Computational Intelligence and Applications, vol. 18, no. 01, pp. 1-11.

[27] Sannasi Chakravarthy SR and Rajaguru H (2019): Lung Cancer Detection using Probabilistic Neural Network with modified Crow-Search Algorithm, Asian Pacific journal of cancer prevention (APJCP), vol. 20, no. 7, pp. 2159-2166.

[28] Shalini Wankhade, Vigneshwari S (2023): A novel hybrid deep learning method for early detection of lung cancer using neural networks, Healthcare Analytics, Volume 3, 100195, ISSN 2772-4425, https://doi.org/10.1016/j.health.2023.100195.

[29] Shanid M and Anitha A (2020): An Exhaustive Study on the Lung Cancer Risk Models, International Journal of Bioinformatics Research and Applications, Vol. 16(2), 151-172.

[30] Thamilselvan Piriyatharisini (2022): Lung Cancer Prediction and Classification Using Adaboost Data Mining Algorithm, International Journal of Computer Theory and Engineering, 254337572, 10.7763/ijcte.2022.v14.1322

[31] Tian PF (2019): Current status and prospects of biomarkers in early diagnosis of lung cancer, Precision medicine Research, vol. 1, no. 2, pp. 61-65.

[32] Tiwari L, Raja R, Awasthi V, Miri R, Sinha G, Alkinani M H, and Polat K (2021): Detection of lung nodule and cancer using novel mask-3 FCM and TWEDLNN algorithms, Measurement, 172:108882

[33] Trailokya Ojha (2023): Machine Learning based Classification and Detection of Lung Cancer, Journal of Artificial Intelligence and Capsule Networks, Volume 5, 110 -128. 10.36548/jaicn.2023.2.003.

[34] V Sreeprada, Dr K Vedavathi (2023): Lung Cancer Detection from X-Ray Images using Hybrid Deep Learning Technique, Procedia Computer Science, Volume 230, 467 - 474, https://doi.org/10.1016/j.procs.2023.12.102.

[35] Vinod Kumar and Brijesh Bakariya (2021): An Empirical Identification of Pulmonary Nodules using Deep Learning, Design Engineering, Volume 2021, Issue 07, 13468-13486