Bio-inspired swarm-intelligent with machine learning framework for prediction and classification of lung cancer

Naveen Kumar Singh¹ | Rahul Gupta² | Satyaapir Sahu³

¹Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India, Professor, Department of General Surgery.
²Jaipur National University, Jaipur, India, Assistant Professor, Department of Resp Medicine.
³Sanskriti University, Mathura, Uttar Pradesh, India, Assistant Professor, Department of Ayurveda.

Abstract Early detection of lung cancer nodules, which is often dependent on a tomography scan filmic examination, greatly increases the probability of survival. Earlier tumor detection reduces lung cancer mortality by increasing the likelihood of successful therapy. Traditionally, radiologists have analyzed medical pictures for signs of lung cancer using a laborious and inaccurate systematic approach. The confidentiality and integrity of medical data have become a serious challenge for healthcare applications as a result of the tremendous improvement in the transmission of medical data in the healthcare sector. This research makes use of computational intelligence methods. In this research, a brand-new Enhanced vortex search algorithm optimized Support Vector Machine (EVSAO-SVM) is developed for detection and classification. Preprocessing using the Gaussian Filter, segmentation with Otsu thresholding, feature extraction with local binary patterns (LPB), and classification and prediction with the EVSAO-SVM are the processes that are simulated. Regarding performance criteria including accuracy, sensitivity, specificity, and F1-measure, this study indicates the suggested framework's superiority over the traditional approaches. The results of the tests conducted demonstrate that the suggested model can reach up to 95.42 percent sensitivity, 96.24 percent accuracy, 98.92 percent specificity, and 94.26 percent F1 measure.

Keywords: lung cancer, Gaussian filter, Otsu thresholding, LPB, EVSAO-SVM

1. Introduction

A malignant growth that develops in the lungs’ cells is called lung cancer. One of the most prevalent and deadly cancers in the world, it is generally brought on by prolonged exposure to tobacco smoke, environmental causes, or a genetic predisposition (Alzubi et al 2019). Adenocarcinoma, squamous cell carcinoma, and giant cell carcinoma are other subtypes of NSCLC. The characteristics, methods of treatment, and prognoses of the many lung cancer types and subtypes vary (Chen et al 2021). Early lung cancer symptoms may not be apparent, which might delay diagnosis until it is too late. Long-lasting coughs, chest discomfort, shortness of breath, exhaustion, unintentional weight loss, and recurring respiratory infections are among the symptoms of lung cancer that are often seen. Imaging examinations like X-rays, computed tomography (CT) scans, or magnetic resonance imaging (MRI) are among the diagnostic methods for lung cancer. To confirm the diagnosis and identify the kind of malignancy, tissue samples taken by bronchoscopy, needle aspiration, or surgical procedures are required (Radhika et al 2019). The patient’s general health, cancer stage, and kind are only a few of the variables that affect how lung cancer is treated. Aside from surgery, treatment options also include radiation, chemotherapy, targeted therapy, and immunotherapy.

Chemotherapy employs medications to eradicate cancer cells across the body, while targeted treatment and immunotherapy focus on certain molecular markers or strengthen the body's defenses against cancer (Hyun et al 2019). Compared to many other malignancies, lung cancer has a relatively poor overall survival rate. For lung cancer, the five-year survival rate varies from roughly 10–20% for SCLC to 25–40% for early-stage NSCLC. But improvements in treatment choices, early detection methods, and campaigns to stop smoking have led to a modest increase in survival rates (Bonavita et al 2020). The majority of instances of lung cancer are significantly correlated with cigarette use. Avoiding or quitting smoking is the most efficient approach to stopping lung cancer from developing. Adopting a healthy lifestyle, quitting smoking, and limiting exposure to environmental toxins may all help lessen the risk of lung cancer (Lai et al 2020). The high incidence of lung cancer highlights the need for early detection, effective treatment options, and preventative measures. Research is still concentrated on developing targeted medications, advancing screening procedures, and comprehending the genetic and molecular elements of lung cancer to significantly enhance patients (Ibrahim et al 2021). X-rays of the chest or CT scans are
two common imaging techniques used in diagnosis, along with tissue biopsies to look for cancer cells. The stage of lung cancer is established after a diagnosis to help with treatment choices (Haragan et al 2019). Therapy is to eliminate or reduce the tumor, kill cancer cells, and stop or control the disease's spread (He et al 2020). It also affects the hue they show up on CT scans. The difficulty of characterizing them is exacerbated by these factors. The remaining segments are as follows: Part 2 covers the related works, Part 3 explains the proposed method, Part 4 evaluates the results, and Part 5 provides a conclusion.

Computing-based methods of intelligence are used in this investigation. Consequently, this research proposes a unique Hybrid model for detection and data transfer. The suggested approach has two stages: first, various image processing techniques are employed to identify cancer in MATLAB, and then, data is sent to authorized individuals through the IoT cloud. Preprocessing, segmentation using Otsu thresholding and the swarm intelligence technique, feature extraction using a local binary pattern, and classification using an SVM are all part of the simulated workflow. Through a variety of performance criteria including training time, this study proves that the swarm intelligent framework is superior to traditional algorithms (Venkatesh and Bojj 2020). Classifying lung illness, heart rhythm, brain activity, forest cover, and other medical images, among other uses, is only one example. The ability of deep learning models to efficiently choose features is crucial for all of these uses. Iterative window-based feature processing is often used by these models to assess various feature combinations. This means attempting to account for an infinite number of feature combinations to properly categorize a limited set of characteristics. The models cease iterating when the error rate between the current iteration and the one before it becomes too large to ignore (Baviskar et al 2021). Mandave demonstrated how these methods are often used to diagnose illnesses including cancer, anemia, Alzheimer's, renal disease, and skin disease by resolving issues like feature selection, classification, optimization, and picture segmentation. The current state of bio-inspired algorithms for diagnosing dementia is also discussed. A bio-inspired approach was shown to be significant in this study's diagnosis of dementia. They use an evolutionary machine learning strategy in this chapter to concurrently tune the parameters of SVM and conduct feature weighting to solve common diagnostic issues. The salp swarm method, a novel, and strong metaheuristic, is paired with SVM to accomplish this. When compared to alternative SVM-based frameworks using highly recognized algorithms like the genetic algorithm (GA) and the particle swarm optimization, the SVM method demonstrates several advantages (Al-Zoubi et al 2020).

Most cancer datasets include features measuring different levels of gene expression, but they often have fewer samples overall, therefore it is important to first remove duplicate features so that prediction algorithms can converge more quickly. These characteristics (genes) allow us to diagnose cancer, choose an effective treatment plan, and find methodological differences. The next step is to use Deep Learning (DL) to categorize the genes into groups that are associated with certain cancer types (Joshi et al 2023). Han et al (2021) concluded that combining the strengths of several swarm intelligence (SI) -based feature selection methods yield a superior feature subset compared to using the algorithms alone. The chosen characteristics were improved via the dynamic recursive feature removal framework after being voted. Priyadharshini and Kanimozhi (2020) proposed using a mean-based GA to select optimal subsets from the raw dataset based on the mean value of the features, and then we use an SVM classifier to evaluate the subset's accuracy, thereby reducing the computational cost and size of the model. Results comparing the suggested technique to the ant colony optimization (ACO) algorithm demonstrate the proposed method's superior accuracy. Pal et al (2022) proposed an attention Net to connect the encoder and decoder layers with a new intermediate layer. Since micro-califications all seem the same to the naked eye and segmenting breast pictures is a difficult procedure, radiologists have tried and failed to do it manually. As a result, automated techniques for early-stage identification of the illness are required to aid radiologists in making accurate diagnoses and subsequent treatment choices for their patients. SI plays a crucial role in illness classification and prediction by helping to get the best possible answer with little training. The effectiveness of SI in disease prediction modeling for breast cancer is reviewed and analyzed by (Kumar et al 2020). Hariprasath et al (2021) exploit the challenge of breast cancer detection to demonstrate the efficacy of several NIC-based methods. Separate and combined methods for diagnosis are also briefly discussed in this section. In the end, the experimental findings and additional statistical evaluations are presented. The goal of this chapter is to help the upcoming researcher in the field draw motivation from the established canon of works and develop a novel, cutting-edge strategy for improving upon existing methods.

2. Material and Methods

Images from CT scans of the lungs are sent into a system with pre-processing, segmentation, feature extraction, and classification steps.

The input picture is first process using a Gaussian filter to remove noise. Segmentation of the images is followed by the extraction of characteristics from the identified tumor, including average intensity, perimeter, area, and eccentricity. The tumor is binarized to determine whether or not it is malignant. If a malignant tumor is found, its stage may be determined as well.

2.1. The IQ-OTH lung cancer dataset

https://www.malque.pub/ojs/index.php/msj
Three months of data were collected in the fall of 2019 on lung cancer cases at the Iraq-Oncology Teaching Hospital (IQ-OTH). It consists of computed tomography (CT) images from both patients with various stages of lung cancer and healthy controls. Oncologists and radiologists from these institutions annotated IQ-OTH slides. As can be seen in Figure 1, the collection includes a thousand hundred and ninety pictures representing CT scan slices from 110 patients. Normal, benign, and malignant describe the three categories into which these situations fall. Forty have cancerous diagnoses, fifteen are benign, and the remaining fifty-five are considered to be typical. Initially, the CT scans were gathered in DICOM format. Siemens's SOMATOM is the name of the scanner being utilized. Breathe hold at full inspiration and a CT protocol including one hundred and twenty kV, one mm slices, three hundred fifty-thousand two hundred HU windows, and fifty-six and hundred HU centers were employed for readings. Before any examination, all photos were scrubbed of all identifying information. The institutional review boards of the hospitals involved in the research decided to forgo obtaining written permission from the patients. There are several "slices" in each scan. There might be anywhere from eighty to two hundred of these slices, and each one is a picture of a human chest seen from a different perspective. There is a wide range of ages, sexes, levels of education, regions of residency, and socioeconomic positions among the hundred and ten cases.
2.2. Gaussian Filter

Image smoothing is the initial step in this pre-processing stage for images. On the input picture, the Gaussian filter is used for smoothing. Noise may be eliminated quite well with Gaussian smoothing. The image's high-frequency components are eliminated with Gaussian. Smoothing eliminates noise and produces a more precise intensity surface. Equation (1) contains the numerical formula for the Gaussian filter.

\[ H(w, z) = \frac{1}{2\pi\sigma^2} e^{-\frac{(w^2+z^2)}{2\sigma^2}} \]  

Where \( w \) represents the horizontal distance from the center, \( z \) is the vertical distance and represents the standard deviation of a Gaussian distribution.

2.3. Segmentation

Segmentation typically divides the image's pixels into regions that correspond to the image's subjects. For most computer vision systems, this is the starting point. Pixel intensities are often used by segmentation algorithms. Setting a threshold parameter is required for all algorithms. Better segmentation occurs when the threshold is set correctly. Given the concentration measurements, a threshold is determined. In this study, the Otsu thresholding methods are applied to get the best possible threshold worth.

2.3.1. Otsu Thresholding

It is predicated on the principle of finding an entrance that minimizes the subjective discrepancy in the class, which is equivalent to finding a threshold that maximizes the variation across classes. Grayscale bimodal histograms are processed directly. There is no need to describe the structure of any additional entities or ensure regional consistency. It uses standard numerical values that may be adjusted regionally.

2.4. Feature extraction by LBP

The LBP approach has been utilised in several contexts. This operator describes textures by using symbols representing differences between centers and neighboring pixels. By thresholding the pixels around the central pixel with the threshold pixel, a binary cypher of each pixel may be obtained. If the value of the neighboring pixel is greater than or equal to the threshold value, then 1 is assigned to that location. At first, a histogram is built to analyze the frequency values of binary patterns. Characteristics of the image's texture stand in for the probability that a binary pattern is present. The following is the LBP equation (2).

\[ LBP(n_d, m_d) = \sum_{i=1}^{K-1} 2^i e(hl - h(n_d, m_d)) \]  

Where \( f(x) \) is a purpose whose result is 0 if \( h < 0 \) and 1 if \( h \geq 0 \); where \( g(n_d, m_d) \) is the gray value of the center and adjacent pixels. The center pixel \((n_d, m_d)\) is the LBP value.
2.5. **Enhanced vortex search algorithm optimized Support Vector Machine Methodology**

2.5.1. **Initial guess solution**

A vortex pattern may be used to describe the VSO algorithm's operation. Figure 3 depicts a 2-dimensional nested loop model of the VSO.

2.5.2. **Candidate solutions**

In Equation (3), \( r \) represents the random number used to produce the neighbor solutions around the center:

\[
q(\zeta |\mu, v) = \frac{1}{\sqrt{(2\pi)^d \mu}} \exp\left[-\frac{1}{2} \frac{(\zeta - \mu)^T (\zeta - \mu)}{v}\right]
\]

To determine the value of, we use equation (4). The literature provides further in-depth details.

\[
v = T^2 \cdot I_{d \times d}
\]

Where \( T^2 \) is the distribution's variance and \( I \) is the \( d \times d \) identity matrix. Using Eq. (5), we can get the standard deviation at time zero (\( T_0 \)):

\[
T_0 = \frac{\max(\text{upper \_lim}) - \min(\text{lower \_lim})}{2}
\]

Where \( T_0 \) is the starting radius (\( P_o \)), which is set to a high amount to ensure that the SS is completely covered.

2.5.3. **Substitution of the current solution**

During the selection process, \( C_0(X) \) is consulted to find a new center for the circle, and the optimal answer, \( (i=0) \), is chosen and stored in memory. As indicated in equation (6), before proceeding to the collection phase, the contender solutions must be contained inside the exploration limits.

\[
lower\_lim^{e} \leq T_i^{e} \leq upper\_lim^{e}
\]

Where \( k = 1, 2, \ldots, n \), and \( e \) is the number of dimensions. A new set of solutions \( C_i(X) \) is generated around the new center by decreasing the effective radius \( (r_i) \) of the previous circle, which was assigned the learned best solution \( X \) in the previous iteration.

It’s important to notice that as the radius becomes smaller in the next step, the region of the found national solutions improves. In the second stage, solutions \( C_i(X) \) are picked at random and analyzed to see which one is the best; if it is better than the previous best solutions; it is designated as the new finest explanation and stored. Then, after the minimal function evaluation termination requirement is met, the designated middle of the third phrase is remembered as the best answer up to that point. Figure 3 is an illustrative case study of the search procedure. The resultant pattern takes on a vortex structure after the algorithmic rule is broken, with the optimal location being smack in the center of the smallest circle. The interpreter pattern for a 2-dimensional issue with upper and lower bounds between \([-10, 10]\) is also shown in Figure 4.

---

*Figure 3 Two-dimensional nested-circle VSO model representing the active search process.*
2.5.4. The radius step-down process

The strategy for performing this adjustment is particularly crucial to the performance of the VSO algorithm; the radius step-down procedure is an example of an understanding step size alteration approach that is also used in random search algorithms. This procedure has to be carried out to guarantee that it enables the algorithmic program to act exploratory in the first stages and exploitative in the latter ones. Proper tuning of the radius value throughout the search procedure is necessary to accomplish such a process. Flow charts are used to offer a full explanation of the VSO method (see Figure 5), and more information about this procedure can be found in the reference.

Figure 5 is a working flowchart for the VSO algorithm.
2.5.5. **Support Vector Machine**

SVM is the most effective classification machine learning algorithm. It can take action on sets if there are only two possible categories. Classification problems for multi-class datasets may be simplified to binary classification problems by splitting the sets into subsets. The partitioning problem is not as severe since fewer sets are required for training. The outcome is a simple SVM yes/no comparison. It functions by generating the optimal hyperplane that divides points into their respective categories. It represents a plane or set of planes in a huge or multidimensional space. Aircraft further from the nearest training data have higher classification results. A larger margin means less room for error. The primary goal of SVM is to estimate the gaps between groups. The following figure 6 illustrates these points:

![Figure 6 SVM Hyperplane](https://www.malque.pub/ojs/index.php/msj)

Several approaches may be used to improve regression and classification. SVM has earned acclaim for its ability to find answers to issues both those with salient features and those without. Fast feature extraction from well-known sets and the generation of decision planes for splitting datasets are also part of this method. Results were obtained by constructing the line distant from the closest training data. The linear plane excels at dividing the world into two groups. For multi-class situations involving nonlinear planes, we use a variety of kernels from the toolbox.

SVMs may also be used to categorize cancer stages. When training a binary SVM classifier, it doesn’t matter whether each document belongs to that category or not. A medical report may not contain information about a single operation. The toolbox features allow the SVMs to perform as expected. The system is taught by being exposed to a huge dataset of possible outcomes across a range of difficulties. After being applied to test data, it gives a score that may be used as a threshold to identify whether something of the same kind has changed. After training the code to enhance the quality of the data, we run the validation step to ensure that it is functioning correctly. The initial step is to divide the data into smaller chunks, each of which may be processed by the trained machine. To use the computer to distinguish between stages depending on the degree of tissue development, feature extraction is followed by training an SVM to integrate the features and get familiar with points of data. If data is cross-checked, it may be possible to understand a system’s operation with less validation. Tolerable results are achieved, but when the number of features exceeds the number of samples, it becomes impractical to apply the essential functions promptly.

The difficulty of this approach ranges from doubles the number of features to three times the number of features, depending on the toolkit. Based on statistical theory, SVM is a one-of-a-kind process that utilizes sparse data and many attributes to get around the two-class classification problem and arrive at a judgment line. To divide non-linear data, functions may be used to create a copy of the data in a highly-featured space. All classification tasks on all datasets may have their accuracy and other metrics calculated using the aforementioned equations. The precision of a classifier refers to how well it reliably divides information. Accuracy in data classification is also represented by the value. The quality of the categorization is proportional to the amount of information utilised for training and testing. More aspects affect accuracy that includes kernel functions, box limits, etc.
3. Results and Discussion

The lung cancer detection process in this study begins in MATLAB and continues with the communication of the resulting characteristics to medical professionals. Algorithms were developed and implemented in this study using MATLAB on a 1.80 GHz Intel Core i5 PC with 8 GB of RAM. The results of this study show that the suggested framework is superior to the traditional algorithms in terms of many performance criteria. The results of the tests reveal that the suggested model's maximum sensitivity is 95.42%, accuracy is 96.24%, specificity is 98.19%, and F1-measure is 94.26%.

The percentage of samples for which the suggested approach accurately predicted results is used to assess how accurate the system. The accuracy is calculated using the equation (8).

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)
\]

Figure 7 displays comparable values for the accuracy metrics and makes it evident that the suggested technique can generate performance results that are better than those generated by the current research approaches. The suggested method’s accuracy of 96% outperforms the results of the current ones, which include K-nearest neighbors (KNN) having a 88% accuracy rate, RF at 90%, AlexNet at 92%, and EVSAO-SVM at 98.27%. When classifying the data, the suggested technique outperformed existing techniques.

\[
\text{sensitivity} = \frac{TP}{TP + FN} \quad (9)
\]

In Figure 8 Sensitivity-wise, KNN scores 82%, RF 85%, AlexNet 88%, and EVSAO-SVM 95.42%.
The harmonic mean of the proposed model is computed to merge "recall and precision" into a single component called the f1-measure. The f1-score is calculated using equation (10).

$$F1 - \text{measure} = \frac{(\text{precision}) \times (\text{recall}) \times 2}{\text{precision} + \text{recall}}$$

(10)

In Figure 9, EVSAO-SVM scored (94.26%) on the f1-score, followed by KNN (86%), RF (88%), and AlexNet (94.26%).

![Figure 9 Results of f1-measure.](image)

Specificity is the ability to assess the substance in the condition of elements that could be predicted to be present. The ratio between the value of TNs and the sum of TNs and FPs is a general understanding of specificity. The specificity is calculated using equation (11).

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

(11)

In Figure 10, the Specificity for KNN is 88%, RF is 92%, AlexNet is 94%, and EVSAO-SVM 98.92%.

![Figure 10 Results of specificity.](image)

4. Conclusions

Lung cancer is a malignant tumor that develops in the cells of the lungs. It is one of the majority common and deadliest cancers in the globe, and its main causes include chronic tobacco use, environmental factors, or a genetic predisposition. Computational intelligence approaches are used in this investigation. This research proposes a unique EVSAO-SVM for detection and classification. The confidentiality and honesty of medical information have grown to be significant problems for healthcare applications as a consequence of the considerable development in the spread of medical information in the healthcare industry. The accuracy offered by the suggested method is superior, at 96.5 percent. Due to the
algorithm’s local optima’s tendency to converge, the accuracy and specificity parameters that may be used to interpret the findings are constrained. Shortly, these findings may also be more efficiently examined by utilizing deep learning methods and cutting-edge hardware processors.

**Ethical considerations**

Not applicable.

**Declaration of interest**

The authors declare no conflicts of interest.

**Funding**

This research did not receive any financial support.

**References**


Joshi AA, Aziz RM (2023) Metaheuristic Model of Gene Selection for Deep Learning Early Prediction of Cancer Disease Using Gene Expression Data. DOI: 10.21203/rs.3.rs-84630/v2


