ISSN 2063-5346

ARTIFICIAL BEE COLONY BASED SVM FOR LUNG CANCER CLASSIFICATION

A.Anupriya[1], Arunkumar Thangavelu[2]

Article History: **Received:** 01.02.2023 **Revised:** 07.03.2023 **Accepted:** 10.04.2023

Abstract

Lung cancer kills more people than any other type of cancer, and this is likely to stay true for a long time. Lung cancer can be treated if the signs are found early. If lung cancer symptoms are found early, the latest advances in artificial intelligence can be used to make an experimental diagnosis plan that will work. In this study, optimized support vector machines (SVMs) with an artificial bee colony were used to process the detection from the lung cancer dataset. An SVM classifier is used to categories lung cancer patients based on their symptoms. We examined our ABC-SVM model's balanced accuracy, F1 score, Mathew's correlation coefficient, Sensitivity and specificity to see how well it worked. The evaluated model was trained and tested using benchmark cancer datasets. Irvine. Patients with lung cancer can receive real-time treatment from any location and at any time, with the smallest amount of effort and latency. The suggested model was compared using SVM (linear), SVM (radial basis), GA-based SVM, and PSO-based SVM. When compared to existing methods, the proposed method is 95.65% as accurate.

Keywords: Artificial Bee Colony, Support Vector Machine, Lung Cancer, Predictive system .

[1]Research Scholar, Department of Computer Science, Bharathiyar University, Coimbatore, Tamilnadu. anupriya.mtech@gmail.com

[1]Professor, School of Computer Science and Engineering, VIT University, Vellore, Tamilnadu. arunkumar.thangavelu@gmail.com

DOI: 10.31838/ecb/2023.12.4.116

1. Introduction

Lung cancer is a type of cancer that starts in the lungs and is most common in people who smoke. There are two primary subtypes of lung cancer: non-small cell and small cell. Lung cancer is caused by smoking, being around people who smoke, being exposed to certain chemicals, and having it in your family [1]. Some of the signs are coughing (sometimes with blood), chest pain, wheezing, and weight loss. Most of the time, these signs don't show up until the cancer has gotten worse. Some of the different kinds of treatments are surgery, chemotherapy, radiation therapy, targeted drug therapy, and immunotherapy. As a result, early-stage lung nodules must be carefully checked and monitored. In this work, we explored the formation and advancement of cancer using machine learning and soft computing techniques for predicting whether the given data belongs to Lung cancer (LC) or Non-Lung Cancer (N-LC). Nowadays, various supervised machine learning techniques used for classification along with meta-heuristic soft computing approaches are used for selecting significant input features from the sample set and constructing the prediction models.

In this study, we try to improve the process of diagnosing lung cancer by using an SVM-based machine learning model. There are numerous diagnostic procedures available for various types of tumors. There are, however, only a few distinct methods for calculating their populations. In addition to the diagnosis, this paper will provide a method for determining the important causing factors of lung cancer. Thus, besides the fact that malignancies can be detected, their type can also be easily determined through adding up and gaining and the appropriate treatment guidelines can be calculated. Optimization is the process or act of discovering a means or alternative that is both cost-effective and yields the maximum performance for the employed approach [2]. Using a variety of optimization techniques, optimizing the findings in the medical area will aid in the early detection of cancer and other diseases. Other applications for optimization include computer modeling and the illustration of business-related difficulties. To achieve precise optimization results, various optimization algorithms such as the Genetic Algorithm, Ant Colony Optimization, Bees Algorithm, Particle Swarm Optimization, and Multi Swarm Optimization can be used [13], [14], [15] in different research works.

The specific objectives of this research work are

- To design a predictor model for lung cancer disease using soft computing and machine learning approaches.
- To compare the analysis of performance for the predictor models using the Matthews correlation coefficient and balanced accuracy instead of normal approaches.
- The most significant result of this study is the development of a reliable model for early lung cancer diagnosis.

The efficiency of the proposed approach is measured using Mathews Correlation Coefficient [4] and Balanced Accuracy. This paper's structure continues: Section 2 reviews literature. Methods are in Section 3. Section 4 describes the ABC-SVM technique, while Section 5 shows Results and Discussions. Section 6 concludes and progresses.

2. Literature Survey

This section discusses about the previous state-of-the-art approaches employed for lung cancer prediction using machine learning approaches. For the purpose of accurately predicting lung cancer, this study makes use of machine learning and image processing. The study includes a total of 83 CT images taken from 70 different participants [3]. During the pre-processing stage of a picture, the geometric mean is utilized. The quality of the image

gets better. The photos are divided up using K-means. A section of the image can be located using this segmentation. Following this step, techniques of classification based on machine learning are utilized. For the purpose of categorization, ANN, KNN, and RF were utilized. According to the findings of the study, the ANN model provides more accurate predictions of lung cancer.

SVM-based machine learning [7] was utilized to optimize lung cancer diagnosis. Patients with lung cancer are classified with an SVM classifier and Python for model implementation. The SVM model was evaluated based on multiple criteria. The tested model utilized cancer information from the collection at UC Irvine. This study will help smart cities improve their healthcare delivery to their citizens. Lung cancer patients can access high-quality care quickly and affordably, no matter where they are or what time of day it is. The proposed technique outperforms SVM and SMOTE by a factor of 98.8 when compared to the required model.

In this study, the authors look at different machine learning [7] classifier algorithms to sort data from the UCI machine learning repository into benign and malignant lung cancer. Before figuring out whether a set of data is cancerous or not, the WEKA gets the input data and changes it to binary format. Then, well-known classification techniques were used to decide whether the data set was cancerous or not. The comparison method shows that the proposed RBF classifier is accurate 81.25% and they propose that RBF is the best classifier technique for predicting lung cancer.

Due to the rising frequency of cancer, both the male and female death rates have risen [4]. Lung cancer is a disease characterized by uncontrollable cell division in the lungs. It is impossible to avoid lung cancer, although its risk can be diminished. Therefore, early identification of lung cancer is essential for patient survival. There is a direct correlation between the number of chain smokers and the incidence of lung cancer. The prediction of lung cancer was analyzed using classification methods such as Logistic Regression, SVM, Decision Trees, and Naïve Bayes. By analyzing the performance of classification algorithms, the main purpose of this work is the early diagnosis of lung cancer.

AI can automate cancer detection, allowing us to evaluate more patients in less time and at less expense [10], [11]. In this study, histopathology images of the lung and colon are classified using deep learning (DL) and digital image processing (DIP). With the proposed framework, cancerous tissues can be identified with 96.3% accuracy. This model will assist medical practitioners in constructing an automatic and reliable lung and colon cancer detection system.

From the above-mentioned analysis, it is clear that there is room for improvement in designing an effective predictor model for lung cancer. Hence, in this work, we have taken the artificial bee colony approach for selecting the significant features from the lung cancer dataset that will improve the classification accuracy of the support vector machine approach.

3. Methodology

This section discusses about the approaches employed in the study. The following sub sections give the working of support vector machine, genetic algorithm and particle swarm optimization.

3.1 Support Vector Machine (SVM)

SVM is a prominent supervised learning classification and regression method [12]. The SVM algorithm finds the optimum line or decision boundary that divides n-dimensional space into classes to classify the following data points. A "hyperplane" is the best way to decide between two options. SVM finds the extreme points and vectors needed to make the hyperplane. The SVM handles these extreme examples, called support vectors.

3.2 Genetic Algorithm

The genetic algorithm (GA) is a method for addressing both constrained and unconstrained optimization issues. Its theoretical underpinnings can be traced back to natural selection, the process by which all living things evolve. The genetic algorithm iteratively improves upon a pool of individual solutions. The genetic algorithm picks potential parents from the current population and uses them to generate offspring. As each generation passes, the population "evolves" toward the best solution. The genetic algorithm can be used to solve optimization problems that can't be solved with other methods, such as those where the goal function is discontinuous, non differentiable, random, or very nonlinear. Mixed-integer programming problems whose components can take integer values can be solved using the evolutionary approach [9]. In order to generate a new generation from the existing population, the genetic algorithm employs three distinct sets of rules at each stage:

- Most of the time, the selection is random and can depend on how people scored.
- Children are the offspring of a cross between two parents, as stipulated by the crossover regulations.
- In order to make an infant, random mutations must happen to each parent according to the rules of mutation.

3.3 Particle Swarm Optimization

In 1995, Kennedy and Eberhart invented particle swarm optimization (PSO) [12]. According to the original research, a school of fish or flock of birds travelling together "may benefit from all other individuals". In other words, all birds in the flock can share their discoveries and help the whole flock get the best hunt while a bird is flying and haphazardly looking for food. While it is possible to mimic a flock of birds' movements, it is also possible to assume that each bird is intended to aid in the search for the best solution in a high-dimensional problem space and that the solution found by the flock is also the best solution in the space. This is a heuristic approach because, in most cases, it is impossible to prove that an absolutely optimal global solution can be found. However, we frequently find that the PSO solution is close to the overall ideal.

4. ARTIFICIAL BEE COLONY based Support Vector Machine (ABC-SVM)

This section discusses about the proposed ABC-SVM approach. Figure 1 depicts the overall flow of ABC-SVM approach and the architecture is given in Figure 2. Before giving input into the model, proper scaling is done before starting the process.

Scaling: Initially every feature value in the lung cancer dataset is scaled between [0-1] in order to avoid larger numerical value domination and to avoid the possibilities of overflow.

$$v^s = \frac{v - min_f}{max_f - min_f} \qquad (1)$$

Where $v^s$the scaled value, v is is the original value of the feature, max_f is the upper bound of features and min_f is the lower bound of features. The initial values (IV) are chosen randomly based on

$$\text{IV} = Irange_{min} + (Irange_{max} - Irange_{min}) * r \qquad (2)$$

Where $Irange_{min}$ and $Irange_{max}$ are lower and upper bounds of Initial values and ‘r’ is a random number between [0 -1].

In this work Radial Basis Function was used as the kernel. The classification accuracy of SVM mainly depends upon ɣ and the weighting factor C. These two parameters are optimized with bee’s algorithm to improve the classification accuracy. The fitness function generally consists of three criteria namely (i) accuracy, (ii) selected features and (iii) features cost. In this work fitness function is represented as

$$Fitness\ Function = W_{accuracy} * SVM_{Accuracy} + W_{feature\ cost} * \left\{ \sum_{i=1}^{n_{features}} C_i * f_i \right\}^{-1} \quad (3)$$

Where $W_{accuracy}$ represents weight of classification accuracy, $W_{feature\ cost}$ represents weight of feature cost, $SVM_{Accuracy}$ is the SVM classification accuracy, $n_{features}$ represents the number of features, C_i be the cost of features i and f_i be the number of features.

(i) Feature Selection Procedure

1. Start with a small number of bees.
2. Use the training data set to figure out how much each bee's error function is worth.
3. Based on the error value found in step 2, make a new group of bees made up of the best bees in the chosen neighborhoods and scout bees that are placed at random.
4. Stop if the error function's value has dropped below a certain threshold or after a certain number of iterations.
5. If not, go back to step 2.

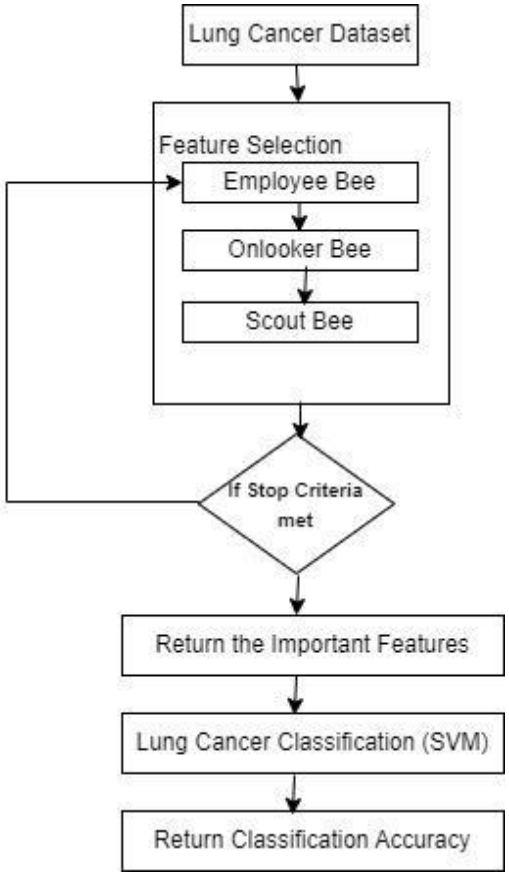

Figure 1 Flow Chart of ABC-SVM

(ii) Parameter Optimization

This section explains the algorithm's evaluation using computing experiments. Using a lung cancer dataset and several benchmark datasets, the experiments are undertaken. To evaluate the effectiveness of the proposed ABC–SVM, it is compared to two other algorithms, GA–SVM and PSO–SVM. These two algorithms were created using the same structure as ABC–SVM. Only the method for finding the optimal kernel parameters and feature selection differs. Also, the suggested framework is compared with the basic SVM (linear) and SVM (radial basis) algorithms to see if it gives better results.

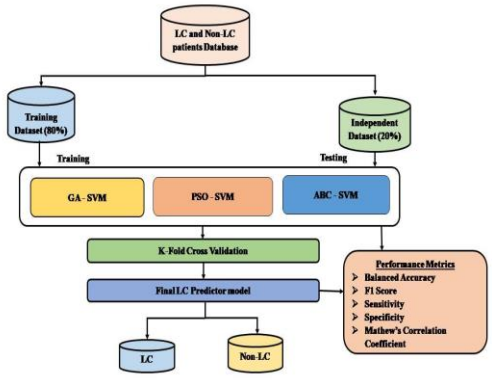

Figure 2 Overall Architecture of ABC-SVM

5. Experimentation Setup

The experiments are run in Python on an Intel(R) Core(TM) i7-2640 M CPU with 16 GB of RAM. Ten-fold cross-validation (CV) is used to get results that can be used to classify things without bias. Due to the random way the dataset was split up, a single 10-fold CV can't be used to make a strong classification. Because the results of metaheuristics are hard to predict, the

experiment is done ten times, and the final result is the average of the ten results.

5.1 Dataset Description

This dataset is taken from [6] consists of 16 attributes and 284 instances. The attribute information is tabulated in Table 1.

Table 1 Dataset Description

S.No	Attribute Name	Values
1	A1(Gender)	M(male), F(female)
2	(A2)Age	Age of the patient
3	(A3)Smoking	T=1, F=0
4	(A4)Yellow fingers	T=1, F=0
5	(A5)Anxiety	T=1, F=0
6	(A6)Peer pressure	T=1, F=0
7	(A7)Chronic Disease	T=1, F=0
8	(A8)Fatigue	T=1, F=0
9	(A9)Allergy	T=1, F=0
10	(A 10) Wheezing	T=1, F=0
11	(A 11)Alcohol	T=1, F=0
12	(A 12) Coughing	T=1, F=0
13	(A 14) Shortness of Breath	T=1, F=0
14	(A 15) Swallowing Difficulty	T=1, F=0
15	(A 16) Chest pain	T=1, F=0
16	(Label)Lung Cancer	T=1, F=0

The parameters considered for designing a predictor model of lung cancer using the methods GA, PSO, and ABC are tabulated in Table 2.

Table 2 Parameters for Optimization

Method	Factors	Parameters
GA	Cross Over	0.8
	Mutation Rate	0.02
	Number of chromosomes	15
PSO	Inertia(w)	0.4
	Learning Rate (L1)	0.5
	Learning Rate (L2)	0.5
ABC	Number of food sources	15

5.2 Evaluation Metrics

The proposed approach is evaluated based on the following metrics: Balanced Accuracy, Sensitivity, Specificity, F1 Score and Matthews Correlation Coefficient. Based on the values taken from the confusion matrix sensitivity, specificity, and MCC are calculated. Balanced accuracy is calculated based on the average of sensitivity and specificity.

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2} * 100\% \quad (4)$$

$$Sensitivity = \frac{TP}{TP + FN} * 100\% \quad (5)$$

$$Specificity = \frac{TN}{TP + FP} * 100\% \quad (6)$$

$$Matthews\ Correlation\ Coefficient(MCC) = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FN)}} \quad (7)$$

6. Results and Discussion

The effectiveness of SVM (linear) and SVM (radial basis) was initially examined using the original feature space, and the results are displayed in Table 4. SVM (Radial Basis) achieved 84.47% balanced accuracy, a F1 score of 0.8694, 93.51% sensitivity, 75.42% specificity, and a MCC of 0.6854. Due to the unsatisfactory performance of the SVM (Radial Basis) classifier, the ABC-SVM method was applied to the same dataset. ABC-SVM achieved 94.90% of balanced accuracy, 0.98 of the F1 score, 98.90% of sensitivity, 90.1% of specificity, and 0.800 of MCC, as shown in Table 5. Interestingly, ABC-SVM greatly outperformed SVM (Radial Basis) in terms of performance improvement. The suggested ABC-SVM outperforms the individual SVM (Radial Basis) classifier in terms of balanced accuracy (11.51%), F1 score (0.1151), sensitivity (5.39%), specificity (15.48%), and MCC (0.1146).

6.1 Training results

To validate the proposed ABC-SVM, it is compared to other meta-heuristic-based SVM approaches, such as PSO-SVM, GA-SVM, and individual SVM with linear and radial bases. The mean balanced accuracy, F1 score, sensitivity, specificity, and MCC are evaluated for each method as shown in Fig. 3. According to Table 3, the average balanced accuracy of ABC-SVM for ten iterations is 94.90%, which is higher than 11.27% for SVM (Linear), 10.43% for SVM (RB), 9.79% for GA-SVM, and 7.08% for PSO-SVM. For ten iterations, the average F1 Score of ABC-SVM is 0.98, which is higher than 0.1256 for SVM (Linear), 0.1151 for SVM (RB), 0.0859 for GA-SVM, and 0.0174 for PSO-SVM, respectively. For ten iterations, the average sensitivity of ABC-SVM is 98.90%, which is higher than the 6.25% of SVM (Linear), 5.39% of SVM (RB), 8.69% of GA-SVM, and 3.53% of PSO-SVM, respectively. Similarly, the average specificity of ABC-SVM for ten iterations is 90.9%, which is higher than 16.03% of SVM (linear), 15.48% of SVM (RB), 10.90% of GA-SVM, and 10.64% of PSO-SVM, respectively. The average MCC of ABC-SVM for ten iterations is 80.00%, which is higher than 20.36% of SVM (linear), 11.46% of SVM (RB), 13.00% of GA-SVM, and 10.28% of PSO-SVM, respectively.

Table 3 Training Set Comparison

Method	Balanced Accuracy	F1 Score	Sensitivity	Specificity	MCC
SVM(Linear)	0.8363	0.8544	0.9265	0.746	0.5964
SVM(RB)	0.8447	0.8649	0.9351	0.7542	0.6854
GA-SVM	0.8511	0.8941	0.9021	0.8	0.67
PSO-SVM	0.8782	0.9626	0.9537	0.8026	0.6972
ABC-SVM	0.9490	0.98	0.989	0.909	0.8

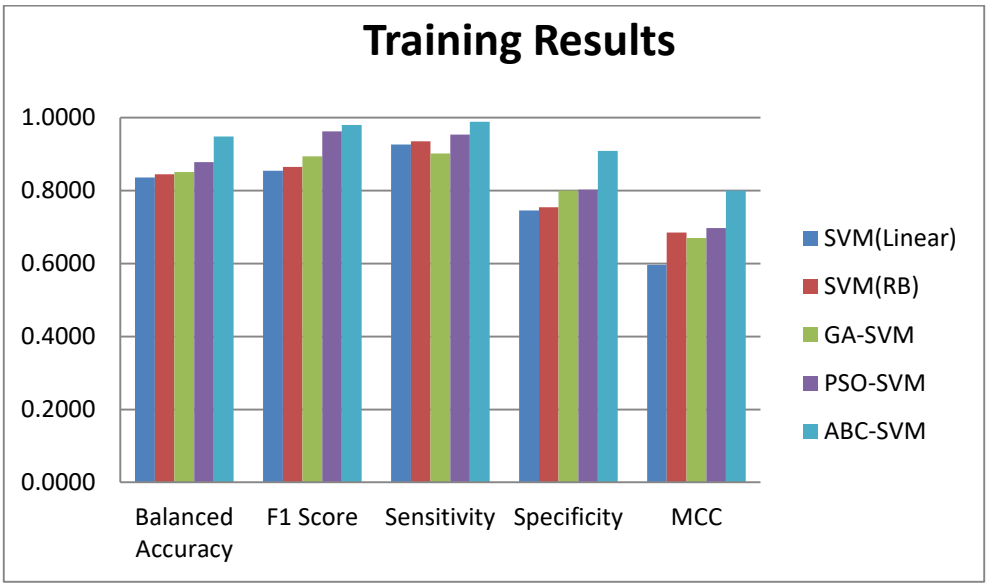

Figure 3 Training Results

6.2 Testing results

ABC-SVM is deemed an effective prediction model based on a number of performance metrics. To evaluate the transferability or robustness of the suggested model, however, it must be tested with an independent dataset. The experimental outcomes of the proposed models are shown in Table 4. Based on Table 4, the balanced accuracy of ABC-SVM is 95.65%, which is greater than the balanced accuracy of SVM (linear), SVM (RB), GA-SVM, and PSO-SVM, respectively, which are 11.08%, 11.25 %, 9.22%, and 8.19%. The F1-Score of ABC-SVM is 0.9801, which is greater than SVM (linear), SVM (RB), GA-SVM, and PSO-SVM, respectively, at 0.1159, 0.1047, 0.076, and 0.0175. ABC-SVM has a sensitivity of 99.40%, which is greater than SVM (linear), SVM (RB), GA-SVM, and PSO-SVM, respectively, at 5.46, 7.19, 6.19, and 5.04%. Similarly, ABC-SVM has a specificity of 91.9%, which is greater than 16.70% for SVM (linear), 17.32% for SVM (RB), 12.25% for GA-SVM, and 11.34% for PSO-SVM, respectively. ABC-SVM has an MCC of 0.86, which is more than SVM (linear), SVM (RB), GA-SVM, and PSO-SVM, each of which has an MCC of 0.14. Fig. 4 is a graphical representation of the evaluation of the independent dataset's performance.

Table 4 Testing Set Comparison

Method	Balanced Accuracy	F1 Score	Sensitivity	Specificity	MCC
SVM(Linear)	0.8457	0.8622	0.9394	0.752	0.601
SVM(RB)	0.8340	0.8754	0.9221	0.7458	0.605
GA-SVM	0.8643	0.9041	0.9321	0.7965	0.69
PSO-SVM	0.8746	0.9626	0.9436	0.8056	0.72
ABC-SVM	0.9565	0.9801	0.994	0.919	0.86

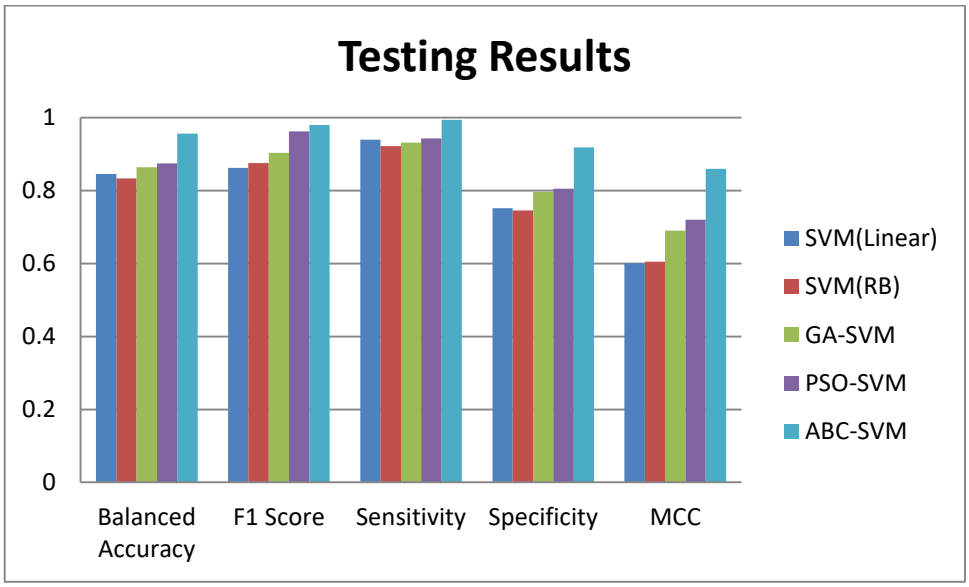

Figure 4 Testing Results

6.3 Comparison with other approaches

Using the same dataset, the proposed ABC-SVM approach is compared to state-of-the-art classifiers like SVM (linear), SVM (RB), GWO-SVM, Adaboost, and Gaussian Naive Bayes. Fig. 5 is a diagram that shows the results. The results show that the ABC-SVM did a better job than the most advanced classifiers. When compared to other methods, ABC-SVM provides the most noticeable improvements. It got higher accuracy with SVM (linear) of 7%, SVM (RB) of 8%, GWO-SVM (linear) of 5%, AdaBoost (Santos, 2021) of 15.5%, and Gaussian NB (Santos, 2021) of 8%. But the features chosen by each method are different, and it's surprising that all of the models focused on predicting lung cancer. From the analysis, it's clear that ABC-SVM has done a better job than other classifiers at telling the difference between LC and non-LC.

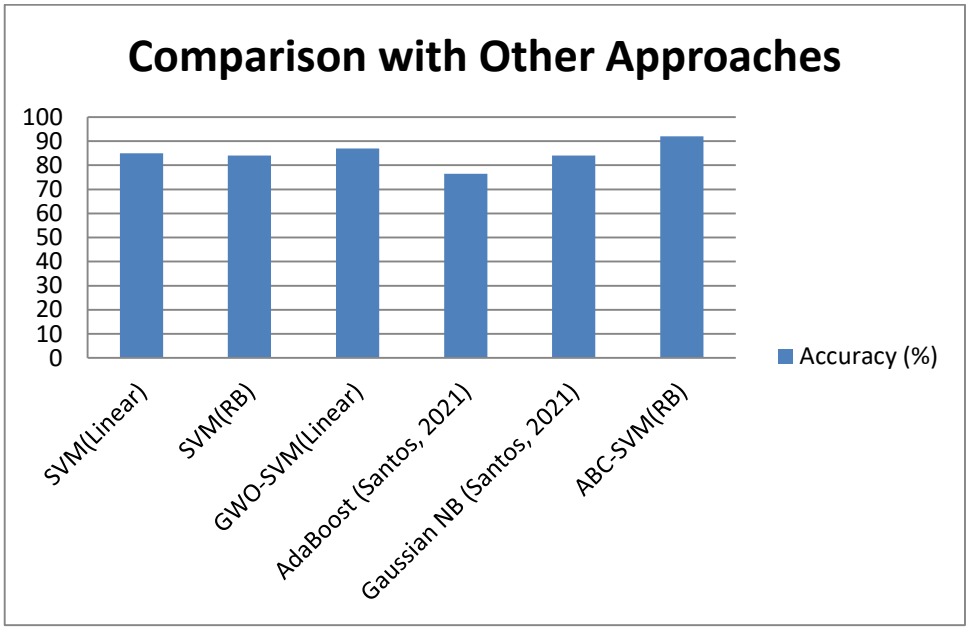

Figure 5 Comparison with other approaches

6.4 Time Taken

The running times of the approaches used in this work are given in Table 5. From the results, it is evident that our proposed ABC-SVM approach produces results in 0.001ms. GA based SVM and PSO based SVM attain results in 0.002ms. SVM on a linear and radial basis produces results in 0.006ms. The results are diagrammatically shown in Fig. 6.

Table 5 Time taken Comparison

Method	Time Taken(ms)
SVM(Linear)	0.006
SVM(RB)	0.006
GA-SVM	0.002
PSO-SVM	0.002
ABC-SVM	0.001

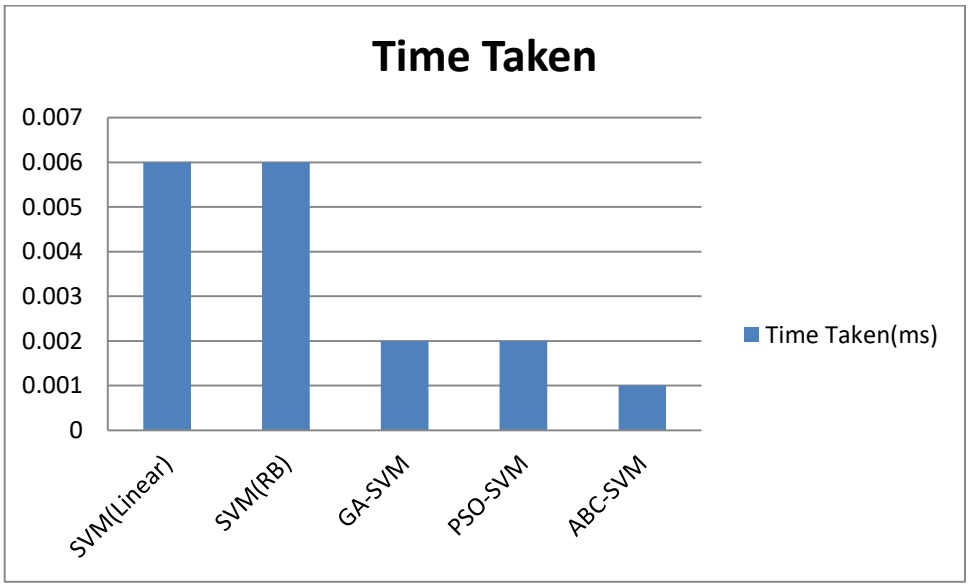

Figure 6 Time taken comparison

7. Conclusion and Future Enhancement

Due to the complex architecture of cancer cells, diagnosing lung cancer might be one of the most challenging medical undertakings. There are more than a hundred additional malignancies that should also be avoided like the plague. Delaying therapy for lung cancer greatly increases the likelihood of the patient dying from the disease. If caught and treated quickly enough, cancer can be cured. Researchers in this work employ ABC-SVM to foresee cases of lung cancer. The main goal of this system is to alert individuals from cancer in advance so they can save time and money. Positive results from evaluating the proposed method's performance indicate that an improved SVM can be used to aid in the diagnosis of lung cancer by oncologists. If the prognosis is correct, the physician may be able to prescribe a more effective treatment and make an earlier diagnosis. To create the most precise lung cancer prediction model, future work will broaden the scope of the current proposal to include other meta-heuristic methodologies and use a real-time lung cancer dataset.

REFERENCES

[1] Anil Kumar, C., Harish, S., Ravi, P., Svn, M., Kumar, B. P., Mohanavel, V & Asfaw, A. K. (2022). Lung Cancer Prediction from Text Datasets Using Machine Learning. *BioMed Research International*, *2022*.

[2] Banerjee, N., & Das, S. (2020, March). Prediction lung cancer–in machine learning perspective. In *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)* (pp. 1-5). IEEE.

[3] Burki, T. K. (2016). Predicting lung cancer prognosis using machine learning. *The Lancet Oncology*, *17*(10), e421.

[4] Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*(1), 1-13.

[5] Du, J., Liu, Y., Yu, Y., & Yan, W. (2017). A prediction of precipitation data based on support vector machine and particle swarm optimization (PSO-SVM) algorithms. Algorithms, 10(2), 57.

[6] https://data.world/sta427ceyin/survey-lung-cancer

[7] Kadir, T., & Gleeson, F. (2018). Lung cancer prediction using machine learning and advanced imaging techniques. *Translational lung cancer research*, *7*(3), 304.

[8] Masud, M., Sikder, N., Nahid, A. A., Bairagi, A. K., & AlZain, M. A. (2021). A machine learning approach to diagnosing lung and colon cancer using a deep learning-based classification framework. *Sensors*, *21*(3), 748.

[9] Mirjalili, S. (2019). Genetic algorithm. In Evolutionary algorithms and neural networks (pp. 43-55). Springer, Cham.

[10] Nageswaran, S., Arunkumar, G., Bisht, A. K., Mewada, S., Kumar, J. N. V. R., Jawarneh, M., & Asenso, E. (2022). Lung Cancer Classification and Prediction Using Machine Learning and Image Processing. *BioMed Research International*, *2022*.

[11] Palanisamy, S., & Kanmani, S. (2012). Artificial bee colony approach for optimizing feature selection. International Journal of Computer Science Issues (IJCSI), 9(3), 432.

[12] Patra, R. (2020, March). Prediction of lung cancer using machine learning classifier. In *International Conference on Computing Science, Communication and Security* (pp. 132-142). Springer, Singapore.

[13] Radhika, P. R., Nair, R. A., & Veena, G. (2019, February). A comparative study of lung cancer detection using machine learning

algorithms. In *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)* (pp. 1-4). IEEE.

[14] Shanmugam, S., Sugumaran, V., Thangavelu, A., & Sekaran, K. Predicting rheumatoid arthritis from the biomarkers of clinical trials using improved harmony search optimization with adaptive neuro-fuzzy inference system. *Journal of Intelligent & Fuzzy Systems*, (Preprint), 1-13.

[15] Xie, Y., Meng, W. Y., Li, R. Z., Wang, Y. W., Qian, X., Chan, C., ... & Leung, E. L. H. (2021). Early lung cancer diagnostic biomarker discovery by machine learning methods. *Translational oncology*, *14*(1), 100907.