ISSN No:-2456-2165

Investigation of Lung Cancer Prediction and Classification using CT-Scan Images by Employing Machine Learning & Population based Techniques

D. Kalaivani[1]
Research Scholar,
Department of Computer Science, PKR Arts College for Women, Tamilnadu, India

Dr.G.Dheepa[2]
Associate Professor,
Department of Computer Science, PKR Arts College for Women, Tamilnadu, India

Abstract:- According to the estimated reports of World Health Organization, with over 2.6 million new cases captured and diagnosed each year, lung cancer is the most prevalent cause of cancer-related deaths worldwide. Early detection and classification of LC is needed for effective analysis & treatments for better patient outcomes. Lung cancer prediction and classification at an early stage have shown significant potential for advanced ML algorithms, particularly DL models. Early detection of lung cancer facilitates patients to undergo timely and effective treatment, considerably improving their chances of survival. The purpose of this research is to put forward an ISBSSA (Improved Selection Based Squirrel Search Algorithm)-based machine learning approach for LC prediction and classification employing CT-SCAN illustrations. The suggested method makes use of a deep learning model called ISBSSA that has been trained on a substantial dataset of computed tomography (CT) images in order to accurately identify and classify lung cancer cells. For the experimental study, a Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis took from Cancer Imaging Archive (CIA) serves as the data source. The LC-CIA dataset which includes CT and PET-CT DICOM pictures of lung cancer patients as well as individuals who are healthy. The model is trained using appropriate machine learning algorithms along with ISBSSA such Naive Bayes Algorithm (NBA), Convolutional Neural Networks (CNNs), Support Vector Machines (SVMs), K-Nearest Neighbour (KNN) and Random Forests (RFs), to predict the presence and type of lung cancer cells in the CT & PET-CT DICOM images which was extracted. The findings of this study show that the proposed approach is successful in effectively predicting and classifying lung cancer cells in CT scans, which might have significant implications for the early detection and treatment of the disease. The comprehensive results show that 94.02% accuracy, 91.80% sensitivity, 92.76% specificity, 96% precision, 92% recall, 0.90 True Positive, 0.87 True Negative, and 96.13% F-Score are achieved to detect and classify the HD in an effective manner, which is the advantage of employing ML and DL approaches.

Keywords:- *Lung Cancer Prediction, Classification, Machine Learning, Deep Learning, Feature Selection, Data Mining, Image Processing.*

I. INTRODUCTION

Lung cancer is a one of the significant public health issue and one of the main causes of cancer-related fatalities globally. Lung cancer accounts for 11.9% of all cancer diagnoses and 18.4% of cancer deaths worldwide, according to the World Health Organization. Early detection and therapy [1] is essential for increasing the survival rates of lung cancer patients. In predicting the presence or absence of lung cancer, machine learning and deep learning models using advanced computational approaches have showed considerable potential. Machine learning techniques such as Support Vector Machines (SVM), Random Forest, K-Nearest Neighbours (KNN), and Naive Bayes have been employed with great accuracy for lung cancer prediction. Deep learning models, such as Convolutional Neural Networks (CNN), have also been utilised successfully for lung cancer prediction.

One of the main challenges in lung cancer prediction is the complexity and variability of the disease. Non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) are the two primary kinds of lung cancer [2]. NSCLC is the most prevalent kind of lung cancer, accounting for around 88% of all occurrences. SCLC is less frequent, but it is more aggressive and can spread to other regions of the body quickly. To properly forecast the existence of lung cancer, particular traits or biomarkers must be identified that can discriminate between the two kinds and reliably predict the presence of cancer [3].

To discover these biomarkers and predict the existence of lung cancer, machine learning and deep learning algorithms may be utilised for feature extraction and selection [4-5]. Feature extraction is the process of identifying significant features or qualities in raw data, whereas feature selection is the process of picking the most critical features for effective prediction. For example, in lung cancer prediction using CNN, the CT scans of the lungs are used as input [5], and the deep learning model learns to extract relevant features and classify the image as cancerous or non-cancerous. In SVM, the algorithm

separates the data into two classes based on a hyper plane that maximally separates the data points. SVM features can be chosen using approaches like Recursive Feature Elimination (RFE) or Principal Component Analysis (PCA). Similarly, Random Forest builds an ensemble of decision trees, and the features are chosen according on their relevance in the classification process [6]. SVM features can be chosen using approaches like Recursive Feature Elimination (RFE) or Principal Component Analysis (PCA). Similarly, Random Forest builds an ensemble of decision trees, and the features are chosen according on their relevance in the classification process [7]. Lung cancer is a serious health concern that requires early diagnosis and treatment. Machine learning and deep learning algorithms have shown great promise in predicting the presence or absence of lung cancer. These algorithms can be used for feature extraction and selection, data preprocessing, and data augmentation to identify relevant features and accurately predict the presence of lung cancer. With further research and development, machine learning and deep learning can provide an important tool for early diagnosis and treatment of lung cancer [8]. Various researchers are trying to predict the LC at early stage with the help of DL and ML models. This paper also aims to classify the LC and early prediction with the help of ML and DL approach and also a new approach called ISBSSA to improve the accuracy of LC prediction and classification at early occurrence.

II. BACKGROUND STUDY

Shimazaki et al. (2022) describe a deep learning-based segmentation approach for identifying lung cancer on chest radiographs. According to the authors, lung cancer is the largest cause of cancer deaths globally, and early identification is critical for successful treatment. Low-dose computed tomography (LDCT) is the current gold standard for lung cancer screening, although it is costly and has a high false positive rate. The algorithm produced an area under the receiver operating characteristic curve (AUC) of 0.67, which was much higher than the average AUC of 0.77 achieved by radiologists. The algorithm's sensitivity was 0.80 and its specificity was 0.74. Kumar et al. (2022) recommend employing machine learning to predict lung cancer using text datasets. A dataset of 2000 patient records with demographic information, medical history, symptoms, and test findings was employed. The dataset was preprocessed by transforming it to a bag-of-words format, which is a typical approach for text-based machine learning applications. In this model, early detection was not accomplished. Binson et al. (2021) use machine learning methods to investigate the potential of electronic nose (e-nose) technology for forecasting pulmonary illnesses. The authors point out those traditional diagnostic procedures for lung disorders can be intrusive and time-consuming, resulting in diagnostic and treatment delays. E-nose technology is a non-invasive and quick method of detecting volatile organic compounds (VOCs) in exhaled breath, which can provide important information regarding pulmonary illnesses. SVM and XGBoost both performed well in this model, with SVM attaining an accuracy of 85.4% and XGBoost reaching an accuracy of 86.2%. Nanglia et al. (2021) present a hybrid lung cancer classification system that combines support vector machines (SVM) with neural networks. Lung cancer is one of the major causes of cancer-related deaths globally, according to the authors, and early diagnosis and proper categorization of lung nodules can dramatically improve patient outcomes. However, due to their complexity and heterogeneity, lung nodules are difficult to classify. Based on the collected characteristics, the scientists utilized SVM and neural networks individually to categorize the lung nodules. The drawback of this model is the LC classification accuracy is not upto the mark. Yakar et al. (2021) conducted a pilot research to evaluate the application of machine learning algorithms for the prediction of radiation Pneumonitis in stage III lung cancer patients. The researchers gathered clinical and dissymmetric data from 54 individuals with stage III lung cancer who had radiation treatment. They developed models for the prediction of radiation Pneumonitis using machine learning methods, notably decision trees, random forests, and support vector machines. The disadvantage of their study is that it is constrained by the small sample size and single-center design, and they recommend that future studies confirm their findings on bigger and more diversified datasets. Using a genetic approach for feature selection, Maleki et al. (2021) suggested a k-NN technique for predicting the prognosis of lung cancer. To create their k-NN model, the scientists gathered clinical and demographic information from 879 lung cancer patients. Age, gender, smoking status, tumor size, and histology were just a few of the 29 potential features they employed a genetic algorithm to narrow down to the most important ones. They then classified patients into low-, intermediate-, or high-risk categories depending on their prognosis using a k-NN method. Their model had good performance in differentiating between low-, intermediate-, and high-risk individuals, with an accuracy of 77.8% and an AUC of 0.78. A hybrid strategy for lung tumor segmentation utilizing the support vector machine (SVM) and marker-controlled watershed algorithms is suggested by Vijh et al. (2021). The goal of the study was to precisely segment the lung tumor area, a crucial step in the early detection and management of lung cancer. The form, texture, and statistical data that are retrieved from the segmented tumor patches are used to train the SVM algorithm. 42 CT scan pictures were utilized in the investigation, of which 21 had benign tumors and the remaining 21 had malignant tumors. However, more validation and testing on a bigger dataset are required to determine whether the suggested strategy is useful in clinical settings. A pathway-based method is used by Nancy Lan Guo and Ying-Wooi Wan (2020) to find a smoking-associated 6-gene signature that is indicative of lung cancer risk and survival. The scientists found gene expression profiles that potentially predict lung cancer risk and survival using microarray data from the Gene Expression Omnibus database. To find the gene signature, they applied a pathway-based method, which groups genes according to how they operate in biological networks. To confirm the gene signature, the scientists also employed machine learning techniques including support vector machines (SVM) and random forests. Further study and development

ISSN No:-2456-2165

of the found 6-gene signature might result in a valuable tool for determining lung cancer risk and developing a tailored treatment strategy**.**

The purpose of this research study is to study and introduce a ML with GA method to classify the LC disease; make early-stage LC detections using CT scan images; maximize the accuracy level.

III. PROPOSED METHODOLOGY

➢ *Issbssa:*

The proposed method for LC prediction is carried out by using CNN, NBA, SVM, RF, KNN and ISBSSA. The ML and DL models used to feature selection and extraction. Here the ISBSSA is portrayed for the LC prediction and it is basically inspired by the food forging behavior of the flying squirrels in the real life. Depends on the weather condition the flying squirrel search for food and store them for future use. If the climate is hot the squirrels will fall down from a tree. It will move out and rapidly search for food for daily needs. It will eat acorns which are available in the hot climate. Once they consumed it will search food for winter. During bad weather the hickory nuts will help squirrels to satisfy the needs. So the process will be continuously depends on the weather condition in the area. Let's take this into mathematical model, the following hypothesis are taken into account.

- *In a forest, the flying squirrel can be counted as 1, it will stay on the tree for whole year*
- *The foraging behavior of FS is dynamic depends on the climatic condition and resource available*
- *Only 3 types of trees in forest hickory trees, oak trees and normal trees.*
- *The 'n' is the number of squirrels searching for food 'f' in the trees 't' to jump and forage. If winter s will jump to t1 and vice versa.*
- *Using JSM the squirrels will jump into one dimensional and two dimensional search spaces.*
- *It has ability to change the locations 'l' where it will be recorded for measure.*
- *Position is calculated and later the best solution is measured.*
- *Increment the counter and calculate the fitness value for new solution.*

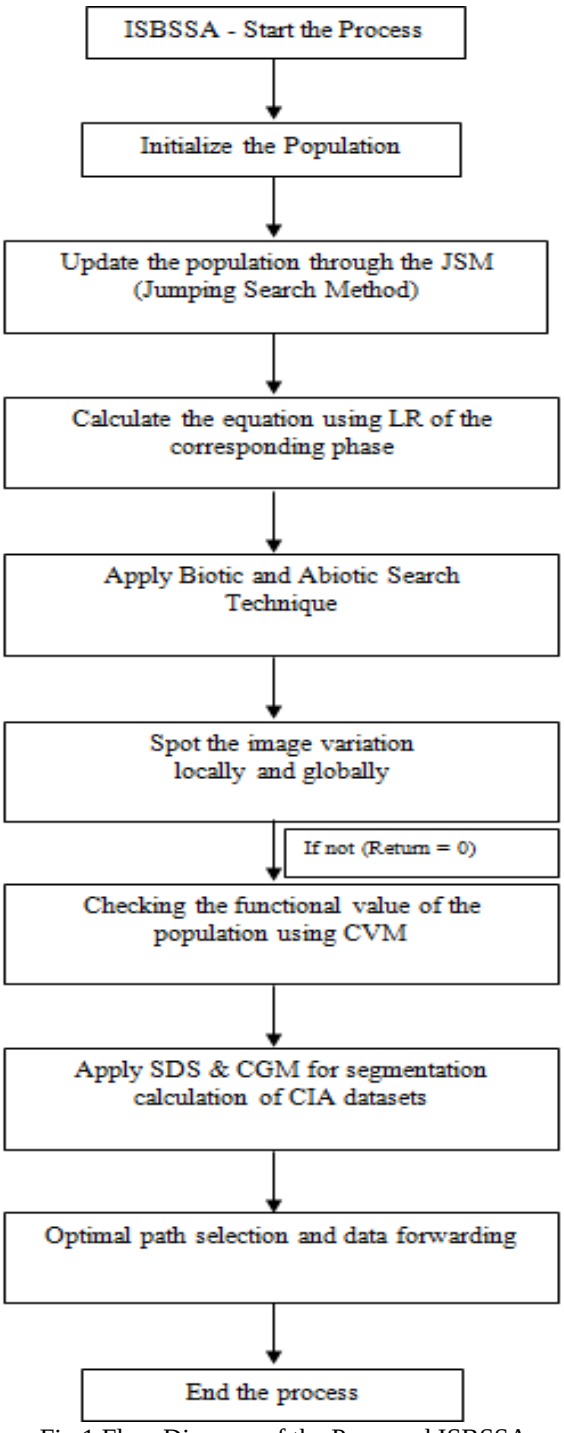

Fig 1 Flow Diagram of the Proposed ISBSSA

In real time scenario, Applying the biotic and abiotic search in image processing that helps to identify the LC in the CT-PET scan images loaded in the system and classify the disease and generates the output. As the ISASSB generates the fitness value, in real time the CGM (Conjugate Gradient Method) and ECS (Edge Colour Segmentation) is used to identify the spot to detect the LC at early stage.

➢ *NBA:*

To examine the wide range of supervised and non supervised learning NBA is used. Suppose in a CIA dataset with only two features (age and smoking history) and a binary target variable indicating whether or not the patient

has lung cancer. We can train a Naive Bayes model on this data by calculating the probability of each feature given the class (i.e. the probability of each feature given lung cancer and the probability of each feature given no lung cancer), and then combining these probabilities using Bayes' theorem. For example, suppose we calculate the following probabilities:

- P(age=50|lung cancer) = 0.2
- P(age=50|no lung cancer) = 0.3
- P(smoking history=heavy|lung cancer) = 0.8
- P(smoking history=heavy|no lung cancer) = 0.4

In observing a new patient with age 50 and a heavy smoking history, Naive Bayes' theorem is used to calculate the probability of each class:

P(lung cancer | age=50, smoking history = heavy) ∝ P(age=50 | lung cancer) * P(smoking history = heavy | lung cancer) * P(lung cancer) = **0.2 * 0.8 * P(lung cancer)**

P(no lung cancer | age=50, smoking history = heavy) ∝ P(age=50 | no lung cancer) * P(smoking history = heavy | no lung cancer) * P(no lung cancer) = **0.3 * 0.4 * P(no lung cancer)**

Compare these probabilities to make a LC prediction. If the probability of lung cancer is higher than the probability of no lung cancer, we predict that the patient has lung cancer. Otherwise, we predict that the patient does not have lung cancer. In NBAs the prediction level and accuracy is shown in the findings.

➢ *CNN:*

CNNs are used to analyze medical images loaded in the dataset (CIA Images), such as chest X-rays or CT scans, to detect signs of lung cancer in all stages. After preprocessing, feature selection and extraction the system start functioning to identify the prediction levels. Here are the LC prediction steps when implementing through CNNs,

- Input layer: The input to the model is a CIA dataset medical image of the lungs, such as a PET/DICOM and CT scan images.
- Convolutional layer: The 1st layer applies a set of filters to the input image to extract features that are relevant for detecting LC.
- ReLUAF activation: A rectified linear unit (ReLUAF) activation function is applied to the output of the convolutional layer to introduce non-linearity into the CNN model.
- Pooling layer - PL: The output of the activation function is passed through a PL to down sample the image spotted in the CIA dataset.
- Dropout layer - DO: A dropout layer is added to prevent overfitting by randomly dropping out a fraction of the units in the layer during training.
- Fully connected layer - FCL: The pooled output is flattened and passed through one or more fully connected layers, which use the extracted features to make a final prediction.
- Output layer- OL: The final layer produces a binary classification output indicating whether or not the input image shows signs of lung cancer. Once the LC is predicted the binary values are measured and performance is calculated.

A loss function that calculates the difference between the expected result and the actual output is minimized during training by adjusting the model's weights via back propagation and gradient descent. Once the model has been trained, it may be used to determine from the characteristics it extracts from an image if a fresh medical imaging has indicators of lung cancer. CNN has shown more accuracy in terms of LC prediction.

➢ *SVM:*

SVM works on separating the classes in a HD feature space. In of LC prediction SVM learns the decision boundary that separates the medical CT scan images into 0 and 1 that indicates disease and non disease. Model training and model evaluation steps will be carried out after FE and FS.

➢ *Input:*

CIA dataset medical images {x1, x2, xn}, where each xi is a vector of extracted features, and corresponding labels {y1, y2, yn} indicating whether the image shows signs of lung cancer (1) or not (0).

➢ *Output*:

A trained SVM model that predicts the presence or absence of lung cancer in new medical images.

- Split the CIA data into training and test sets.
- Select a subset of features from the training data that are most relevant for lung cancer prediction, using a feature selection technique such as mutual information or L1 regularization.
- Train an SVM on the selected features using the training set, optimizing the hyperparameters using cross-validation.
- Evaluate the performance of the trained SVM on the test set, using PEM shown below.
- Adjust the hyperparameters or feature selection criteria and repeat steps 3 and 4 until an acceptable performance is reached if the SVM's performance is not adequate.
- By removing pertinent information and using the learnt decision boundary, the model may be used to predict the presence or absence of lung cancer in CIA medical images after training.
- The SVM shows the remarkable performance in LC prediction compares to NBA and CNN which is shown in the diagram.

➢ *KNN:*

In case of KNN, the following steps are followed for LC prediction, the steps are,

- Preprocessing
- Extracting from CIA dataset

ISSN No:-2456-2165

- Splitting datasets to train and test
- Choose value of k-distance metric by using Euclidian distance method
- Train the KNN model by storing feature vectors
- Predict the class
- Evaluate the performance
- Adjust the hyper parameters
- Find the K-Nearest Neighbour

As KNN is using the distance vector metric, the prediction accuracy is little high compares to CNN, NBA and SVM.

➢ *RF:*

By constructing several DTs (Decision Trees) and combining their projections to increase accuracy. The RF selects the number of trees and other hyperparameters after FS and FE: In order to use random forest, you must provide the number of trees to be built, the maximum depth of each tree, and additional hyperparameters like the minimum number of samples needed to divide a node. To boost the model's performance, these hyperparameters will be modified using methods like cross-validation.

➢ *Input:*

- *Dataset: CIA medical dataset (preprocessed)*
- *Number of trees: n_trees*
- *Maximum depth of each tree: max_depth*
- *Minimum number of samples required to split a node:* Min, Samples, Split

➢ *Output***:**

Random Forest model for lung cancer prediction

- Split the dataset into training and test sets
- For each tree i in n_trees
- Randomly select a subset of the training data
- Construct a decision tree using the selected data with a maximum depth of max_depth and minimum samples required to split a node of min_samples_split
- Store the decision tree in the ensemble
- Return the Random Forest model
- Use the Random Forest model to predict the lung cancer status for new patients:
- For each patient's medical data, traverse each decision tree in the ensemble
- Record the predicted class for each tree
- Assign the new patient to the class with the most votes among the trees
- As RF used decision tree method, the accuracy of LC prediction is high and shows the evident result and shown in the results.

IV. THE ISBSSA ALGORITHM

- Begin : Define the input
- Random positions for ‘n’ of floating squirrels (Sq)
- calculate the fitness of Sq
- Sort positions of Sq based on fitness
- Announce the floating Sq on 3 type of trees (hickory trees, acorn, nut tree)
- Elect random and float some Sq move from normal trees to hickory trees
- for t=1 to n1 (n=total Sq)
- Evaluate the season
- Update the foraging value
- Update the fitness
- Randomly reposition the Sq
- Update the lowest value
- The final position of h_tree considered as best solution
- End

Select the features that appear the most frequently: Following the completion of the SSA process, the frequency of each feature in the best solutions can be calculated. The most often occurring features can be chosen as the best feature subset for lung cancer prediction. Use the selected features to train a classifier: Once the optimal feature subset is selected, it can be used to train a classifier such as SVM or KNN for lung cancer prediction on the test set. Parameter tuning is essential to find the best result.

- *Search Process:*

✓ Initialize the best solution as the one with the highest fitness value - HFS.
✓ For each Sq in the population, calculate the movement distance based on its fitness and the fitness of the best solution.
✓ Update the location of the Sq by adding a random displacement vector to its current position.
✓ Assess the fitness of the new solution and compare it with the previous one.
✓ If the new solution is better than the previous one, update the best solution and the population.

V. MATLAB R2020A IMPLEMENTATION

The performance evaluation of ISBSSA and other ML models such as NBA, CNNs, SVM, KNN, and RF for LC prediction is done using the MATLAB R2020a tool. It is used to analyse the proposed work and showcase the graphical representation of the model's efficacy. For conducting the simulation, NS2.35 is used. The number of iterations was 800, but 1000 iterations are required to reach the termination condition. The CIA datasets (355 instances with 4 types of cancer support in 251,135 images) used for the implementation process are clearly shown in Table 1.

Table 1 CIA Datasets for LC prediction (CT Images)

Image Statistics	
Modalities	CT/PT DICOM Images
Number of Instances	355
Number of Studies	436
Number of Series	1295
Number of PE Images	251135
Attributes - Clinical LC	
1.Patient ID	2.Sex
2.Age	4.Weight
5.T-Stage	6.N-Stage
7.M-Stage	8.Histopathological Grading
9.Smoking History	10.Target Value

➢ *Performance Metrics*

The following metrics are used in this analysis to compare the performance of the proposed work ISBSSA along with NBA, CNNS, SVM, KNN and RF. The results are shown in the graphical format where the readers can compare the analysis. The metrics are,

- *Accuracy in LC prediction* - Shows the efficiency of ISBSSA and other ML models for LC prediction..
- *True Positive & True Negative* - Accuracy of LC prediction where the system actually does.
- *False Positive & False Negative* - Failure rate of the system towards LC prediction.
- *Precision* - Calculates the amount of TP rates among all positive predictions/cases made by system.
- *Recall* - Calculates the amount of TP out of actual positive cases/instances in the given CIA dataset.
- *Fault/Error Comparison* - It shows the error percentage of the proposed ISBSSA and other ML models.
- *Sensitivity and Specificity* - Detects accurately all the positive and negative instances in the image given in •CIA dataset for ISBSSA and other ML models.

VI. FINDINGS AND DISCUSSIONS

➢ *Accuracy in LC prediction*

Figure 2 shows the detailed analysis of LC prediction and its accuracy. Among all the ML and DL models, ISBSSA outperforms all other algorithms and stands at the top. The ISBSSA initiates the process at the beginning of the image search in the LC-CIA dataset, where the depth ratio is calculated and measured by its coherent value. The depth ratio and, due to optimisation of images, the accuracy rate is high in ISBSSA compared to other models. The rate is calculated as 0-1 and ISBSSA shows outstanding performance of 1 in terms of LC prediction.

Table 2 Accuracy Analysis of LC prediction

Accuracy in LC Prediction						
Rounds	**NBA**	**CNN**	**SVM**	**KNN**	**RF**	**ISBSSA**
0	0	0	0	0	0	0
200	0.2	0.2	0.2	0.2	0.2	0.5
400	0.3	0.4	0.4	0.4	0.4	0.7
600	0.5	0.5	0.5	0.4	0.6	1
800	0.5	0.6	0.6	0.5	0.7	1

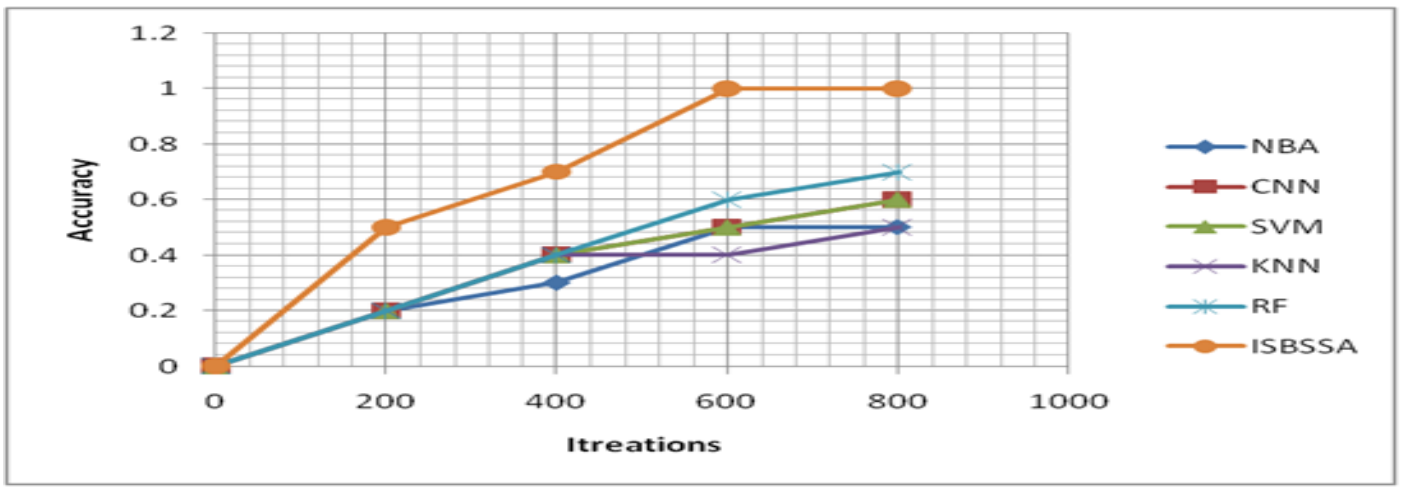

Fig 2 Accuracy in LC Prediction

ISSN No:-2456-2165

➢ *True Positive and True Negative*

Figure 3 shows a comparative analysis of DM and ML models in terms of TPR and TNR. As far as LC prediction is concerned, the TPR and TNR stand top in ISBSSA whereas NBA, CNN, SVM, KNN and RF models are relatively good. As the number of iterations increases the TPR and TNR rate is gradually increases in all the approaches. As ISBSSA is capable of working with both binary and normalized data, the complete search space is calculated and disease spot is identified in the search space.

Table 3 TP & TN Analysis of LC Prediction

True Positive & True Negative						
Rounds	**NBA**	**CNN**	**SVM**	**KNN**	**RF**	**ISBSSA**
0	0	0	0	0	0	0
200	0.3	0.4	0.5	0.5	0.6	0.7
400	0.3	0.4	0.6	0.6	0.6	0.9
600	0.5	0.6	0.6	0.7	0.9	1
800	0.5	0.6	0.7	0.8	0.9	1

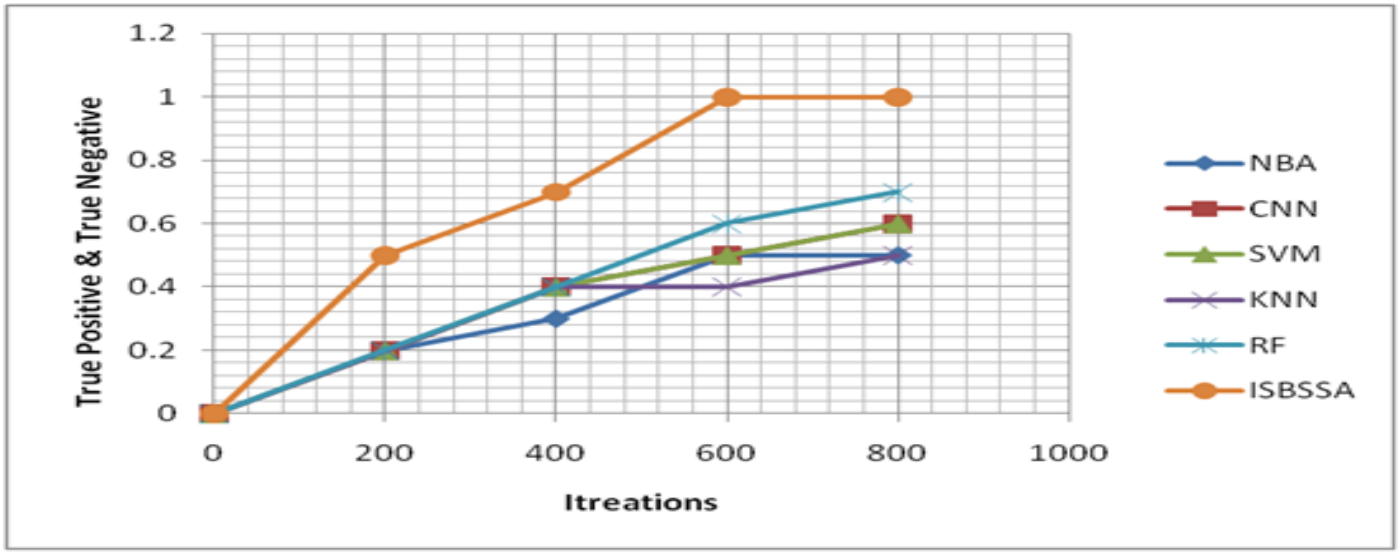

Fig 3 True Positive & True Negative

➢ *False Positive and False Negative*

Figure 4 portrays the detailed analysis of False Positive and False Negative of the proposed ML and DL approaches. ISBSSA showed the evident success in minimizing FPR and FNR. On the chosen datasets with features and instances, the new ISBSSA ML technique outperformed the other DL methods due to its high accuracy rate and measured the proportion of negative cases that are incorrectly identified as positive and vice versa. The new model ISBSSA performs well in prediction and classification of lung cancer due to the computational methods, and the results are given below. Minimum of 0.05 were plotted where the number of iterations is 800 in ISBSSA. The model works out well in all iterations/rounds for all the optimized images loaded in the datasets. The spotting of colours in CT-PET scan images is done by CGDM and ECSM method.

Table 4 FP & FN Analysis of LC Prediction

False Positive and False Negative						
Rounds	**NBA**	**CNN**	**SVM**	**KNN**	**RF**	**ISBSSA**
0	0	0	0	0	0	0
200	0.6	0.5	0.4	0.3	0.3	0.1
400	0.5	0.4	0.4	0.3	0.3	0.1
600	0.5	0.3	0.3	0.2	0.2	0.1
800	0.4	0.3	0.3	0.2	0.2	0.05

ISSN No:-2456-2165

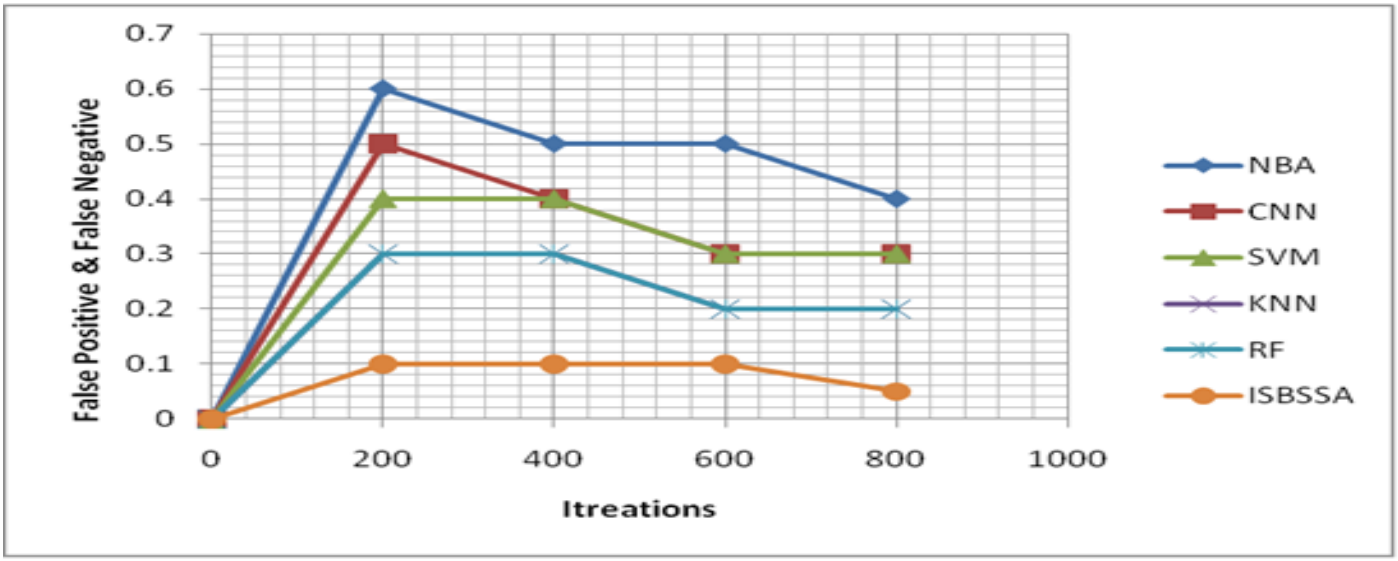

Fig 4 False Positive & False Negative

➢ *Precision and Recall*

Figure 5 present the precision and recall analysis of NBA, CNN, SVM, KNN, RF and ISBSSA. It shows the number of iterations and performance rate by calculating the accuracy of LC prediction at early stage. Despite the fact that the number of rounds increases, the P&R level is high in the proposed ML and DL approach. ISBSSA reached 1 performance/accuracy level where the other approaches are marginally low.

Table 5 Precision & Recall Analysis of LC Prediction

Precision and Recall						
Rounds	**NBA**	**CNN**	**SVM**	**KNN**	**RF**	**ISBSSA**
0	0	0	0	0	0	0
200	0.2	0.2	0.2	0.2	0.2	0.4
400	0.3	0.3	0.4	0.4	0.4	0.6
600	0.5	0.5	0.4	0.4	0.5	0.8
800	0.5	0.6	0.6	0.5	0.8	1

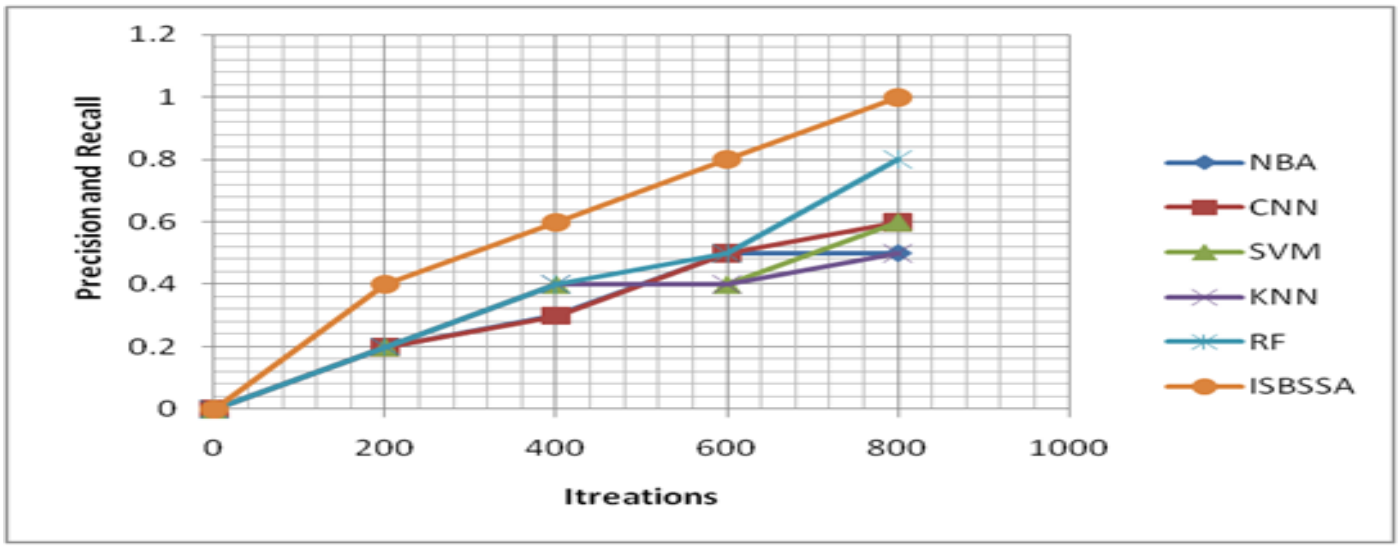

Fig 5 Precision & Recall

➢ *Fault/Error Comparison Rate*

Figure 6 illustrates the fault and error comparison rate. It shows that the error rate decreases as the number of iterations increases. Due to the optimisation of ISBSSA, the error or fault rate is low, and the approach outperforms other DL and ML models like NBA, CNN, SVM, KNN, RF, and ISBSSA. As the LC prediction is done at an early occurrence, the CT scan image optimisation is much needed for fast forwarding the process and getting quick results.

ISSN No:-2456-2165

Table 6 Error Comparison Rate Analysis of LC Prediction

Fault/Error Comparison Rate						
Rounds	**NBA**	**CNN**	**SVM**	**KNN**	**RF**	**ISBSSA**
0	0.8	0.7	0.7	0.6	0.6	0.3
200	0.8	0.7	0.6	0.5	0.5	0.3
400	0.7	0.7	0.5	0.4	0.4	0.2
600	0.6	0.6	0.5	0.4	0.3	0.1
800	0.4	0.5	0.5	0.4	0.3	0.1

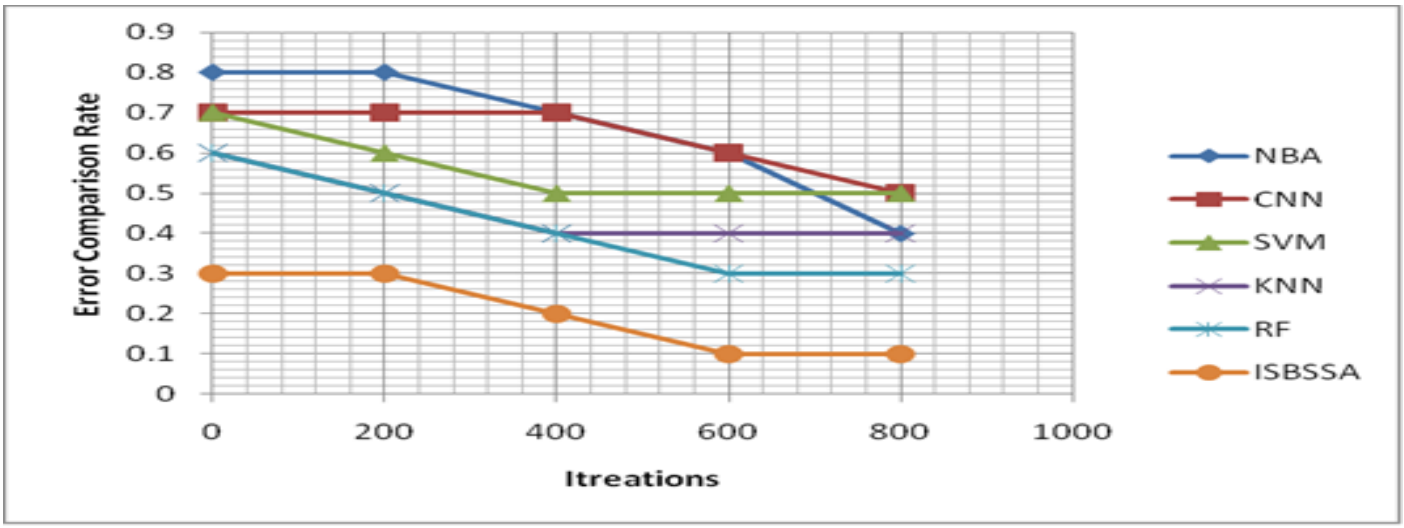

Fig 6 Error/Fault Comparison

➢ *Sensitivity and Specificity*

The performance analysis S&S of the ISBSSA and other ML and DL models towards LC prediction is shown in Figure 7. Figure 8 displays the evaluated outcomes. It is noted that the proposed ML technique produced remarkable performance with enhanced results. The new SDSM is employed in ISBSSA and provides better performance than any previous ML technique. Due to precise classification and extraction, the suggested ML techniques perform well up to maximum level 1 in terms of anticipating the LC at an early stage.

Table 7 Sensitivity & Specificity Analysis of LC prediction

Sensitivity and Specificity						
Rounds	**NBA**	**CNN**	**SVM**	**KNN**	**RF**	**ISBSSA**
0	0	0	0	0	0	0
200	0.2	0.2	0.2	0.4	0.5	0.7
400	0.3	0.3	0.3	0.5	0.6	0.9
600	0.4	0.4	0.4	0.5	0.7	1
800	0.5	0.55	0.5	0.5	0.7	1

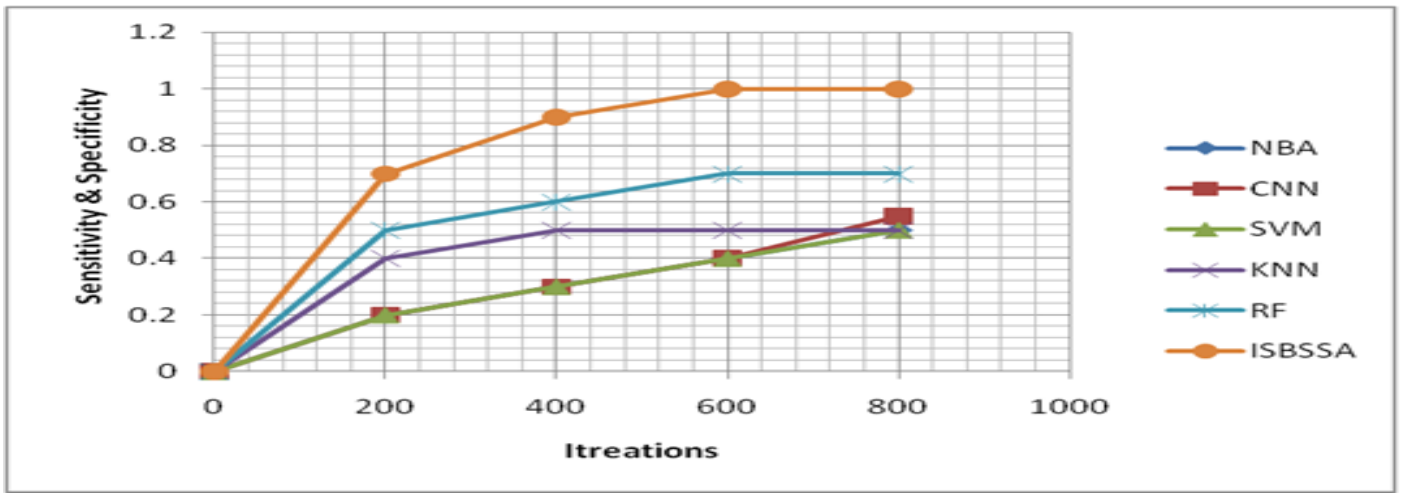

Fig 7 Sensitivity and Specificity

VII. CONCLUSION

In order to improve prediction accuracy and detect LC at an early stage, the research work offered a prediction and classification of lung cancer utilizing multiple ML techniques, including NBA, CNN, SVM, KNN, RF, and ISBSSA. The suggested method makes use of a dataset of CIA PT-CET CT scan images, and 355 cases with 15 or more features are utilised to choose the set of feature values using an ISB-SSA genetic algorithm (GA). The trained and test values are compared to the available datasets to make a prediction. 800+ iterations are carried out which contains patient information, such as medical history, lab results, and imaging data, to predict the likelihood of a patient having a specific type of habits like smoking, alcohol consumption etc. The results show that the ML ISBSSA technique prediction is more accurate and faster than the other ML, DL prediction models and successfully classified the LC. For the medical field, it is also a time-saving and user-friendly application. It is noted that ML decision support system is more suitable for LC prediction.

The suggested method has limitations since it only uses a set of attributes that have been set up, and prediction and accuracy levels may differ from instance to case. The algorithm may be improved in the future to forecast LC with skin disease accuracy rates even better and to make the system more adaptable based on the stability of the patients.

REFERENCES

[1] Li, P., Wang, S., Li, T., Lu, J., HuangFu, Y., & Wang, D. (2020). *A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis* [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/TCIA.2020.NNC2-0461

[2] Patra, R. (2020). Prediction of Lung Cancer Using Machine Learning Classifier. In: Chaubey, N., Parikh, S., Amin, K. (eds) Computing Science, Communication and Security. COMS2 2020. Communications in Computer and Information Science, vol 1235. Springer, Singapore. https://doi.org/10.1007/978-981-15-6648-6_11

[3] Kadir, T., & Gleeson, F. (2018). Lung cancer prediction using machine learning and advanced imaging techniques. *Translational Lung Cancer Research*, *7*(3), 304-312. https://doi.org/10.21037/tlcr.2018.05.15

[4] Nemlander E, Rosenblad A, Abedi E, Ekman S, Hasselström J, Eriksson LE, et al. (2022) Lung cancer prediction using machine learning on data from a symptom e-questionnaire for never smokers, formers smokers and current smokers. PLoS ONE 17(10): e0276703. https://doi.org/10.1371/journal.pone.0276703

[5] Nithyanandh S and Jaiganesh V. (2020), Reconnaissance Artificial Bee Colony Routing Protocol to Detect Dynamic Link Failure in Wireless Sensor Network. *International Journal of Scientific & Technology Research*, 10(10), 3244–3251. https://doi.org/10.35940/ijstr.b2271.0986231

[6] Ahsan Bin Tufail, Yong-Kui Ma, Mohammed K. A. Kaabar, Francisco Martínez, A. R. Junejo, Inam Ullah, Rahim Khan, "Deep Learning in Cancer Diagnosis and Prognosis Prediction: A Minireview on Challenges, Recent Trends, and Future Directions", *Computational and Mathematical Methods in Medicine*, vol. 2021, Article ID 9025470, 28 pages, 2021.https://doi.org/10.1155/2021/9025470

[7] Shimazaki, A., Ueda, D., Choppin, A. *et al.* Deep learning-based algorithm for lung cancer detection on chest radiographs using the segmentation method. *Sci Rep* 12, 727 (2022). https://doi.org/10.1038/s41598-021-04667-w

[8] C. Anil Kumar, S. Harish, Prabha Ravi, Murthy SVN, B. P. Pradeep Kumar, V. Mohanavel, Nouf M. Alyami, S. Shanmuga Priya, Amare Kebede Asfaw, "Lung Cancer Prediction from Text Datasets Using Machine Learning", *BioMed Research International*, vol. 2022, Article ID 6254177, 10 pages, 2022. https://doi.org/10.1155/2022/6254177

[9] V. A. Binson, M. Subramoniam, Y. Sunny and L. Mathew, "Prediction of Pulmonary Diseases With Electronic Nose Using SVM and XGBoost," in IEEE Sensors Journal, vol. 21, no. 18, pp. 20886-20895, 15 Sept.15, 2021, https://doi.org/10.1109/JSEN .2021. 3100390.

[10] Pankaj Nanglia, Sumit Kumar, Aparna N. Mahajan, Paramjit Singh, Davinder Rathee, A hybrid algorithm for lung cancer classification using SVM and Neural Networks, ICT Express, Volume 7, Issue 3, 2021, Pages 335-341. https://doi.org/10.1016/j.icte. 2020. 06.007.

[11] Yakar M, Etiz D, Metintas M, Ak G, Celik O. Prediction of Radiation Pneumonitis With Machine Learning in Stage III Lung Cancer: A Pilot Study. Technology in Cancer Research & Treatment. 2021; 20. https://doi.org/10.1177/15330338211016373

[12] Negar Maleki, Yasser Zeinali, Seyed Taghi Akhavan Niaki, A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection, Expert Systems with Applications, Vol 164,2021,113981. https://doi.org/10.1016/ j.eswa.2020. 113981.

[13] Vijh, S., Sarma, R., & Kumar, S. (2021). Lung Tumor Segmentation Using Marker-Controlled Watershed and Support Vector Machine. *International Journal of E-Health and Medical Communications (IJEHMC)*, *12*(2), 51-64. https://doi.org/10.4018/IJEHMC.2021030103

[14] Nancy Lan Guo, Ying-Wooi Wan, Pathway-based identification of a smoking associated 6-gene signature predictive of lung cancer risk and survival, Artificial Intelligence in Medicine, Volume 55, Issue 2, 2020, Pages 97-105. https://doi.org/10.1016/j. artmed. 2012. 01.001.

[15] Nithyanandh S and Jaiganesh V, (2020), Quality of service enabled intelligent water drop algorithm based routing protocol for dynamic link failure detection in wireless sensor network. *Indian Journal of Science and Technology*. 2020; 13(16), 1641-1647. https://doi.org/ 10.17485/IJST/v13i16.19

[16] Park, S., Lee, S.M., Lee, K.H. *et al.* Deep learning-based detection system for multiclass lesions on chest radiographs: comparison with observer readings. *European Radiology*, 30, 1359–1368 (2020). https://doi.org/10.1007/s00330-019-06532-x

[17] Yoo H, Kim KH, Singh R, Digumarthy SR, Kalra MK. Validation of a Deep Learning Algorithm for the Detection of Malignant Pulmonary Nodules in Chest Radiographs. *JAMA Netw Open.* 2020;3(9): e2017135.https://doi.org/10.1001/jamanetworkopen.2020.17135

[18] Hwang EJ, Park S, Jin K, et al. Development and Validation of a Deep Learning–Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs. *JAMA Netw Open.* 2019;2 (3):e191095.https://doi.org/10.1001/jamanetworkopen.2019.1095

[19] Nithyanandh S and Jaiganesh V, (2020), Dynamic Link Failure Detection using Robust Virus Swarm Routing Protocol in Wireless Sensor Network, *International Journal of Recent Technology and Engineering*, 8(2), 1574-1578. https://doi.org/10.35940/ijrte.b2271.078219

[20] Eui Jin Hwang, Sunggyun Park, Kwang-Nam Jin, Jung Im Kim, So Young Choi, Jong Hyuk Lee, Jin Mo Goo, Jaehong Aum, Jae-Joon Yim, Chang Min Park, Deep Learning-Based Automatic Detection Algorithm Development and Evaluation Group, Development and Validation of a Deep Learning–based Automatic Detection Algorithm for Active Pulmonary Tuberculosis on Chest Radiographs, *Clinical Infectious Diseases*, Volume 69, Issue 5, 1 September 2019, Pages 739–747, https://doi.org/10.1093/cid/ciy967

[21] Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas de Bel, Moira S.N. Berens, *et.al*, Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge, *Medical Image Analysis*, Volume 42,2017, Pages 1-13. https://doi.org/10.1016/j.media. 2017.06. 015.

[22] Dritsas, E.; Trigka, M. Lung Cancer Risk Prediction with Machine Learning Models. *Big Data Cogn. Comput.* 2022, *6*, 139. https:/doi.org/10.3390/ bdcc 6040139

[23] Marjolein A. Heuvelmans, Peter M.A. van Ooijen, Sarim Ather, Carlos Francisco Silva, Daiwei Han, Claus Peter Heussel, William Hickes, Hans-Ulrich Kauczor, et.al, Lung cancer prediction by Deep Learning to identify benign lung nodules, Lung Cancer,Volume 154,2021, Pages 1-4. https://doi.org/10.1016/j. lungcan. 2021.01.027.