

Diagnosis of Pulmonary Edema and Covid-19 from CT slices using Squirrel Search Algorithm, Support Vector Machine and Back Propagation Neural Network

R. Betshrine Rachel^a, H. Khanna Nehemiah^{a,*}, C.S. Marishanjunath^b and Rebecca Mercy Victoria Manoharan^c

^aRamanujan Computing Centre, College of Engineering Guindy, Anna University, Chennai, Tamil Nadu, India

^bDepartment of Information Science and Technology, College of Engineering Guindy, Anna University, Chennai, Tamil Nadu, India

^cDepartment of Computer Science and Engineering, College of Engineering Guindy, Anna University, Chennai, Tamil Nadu, India

Abstract. A Computer Aided Diagnosis (CAD) framework to diagnose Pulmonary Edema (PE) and covid-19 from the chest Computed Tomography (CT) slices were developed and implemented in this work. The lung tissues have been segmented using Otsu's thresholding method. The Regions of Interest (ROI) considered in this work were edema lesions and covid-19 lesions. For each ROI, the edema lesions and covid-19 lesions were elucidated by an expert radiologist, followed by texture and shape extraction. The extracted features were stored as feature vectors. The feature vectors were split into train and test set in the ratio of 80 : 20. A wrapper based feature selection approach using Squirrel Search Algorithm (SSA) with the Support Vector Machine (SVM) classifier's accuracy as the fitness function was used to select the optimal features. The selected features were trained using the Back Propagation Neural Network (BPNN) classifier. This framework was tested on a real-time PE and covid-19 dataset. The BPNN classifier's accuracy with SSA yielded 88.02%, whereas, without SSA it yielded 83.80%. Statistical analysis, namely Wilcoxon's test, Kendall's Rank Correlation Coefficient test and Mann Whitney U test were performed, which indicates that the proposed method has a significant impact on the accuracy, sensitivity and specificity of the novel dataset considered. Comparative experimentations of the proposed system with existing benchmark ML classifiers, namely Cat Boost, Ada Boost, XGBoost, RBF SVM, Poly SVM, Sigmoid SVM and Linear SVM classifiers demonstrate that the proposed system outperforms the benchmark classifiers' results.

Keywords: Pulmonary Edema, Covid-19, Squirrel Search Algorithm (SSA), Support Vector Machine (SVM), Back Propagation Neural Network (BPNN) classifiers

1. Introduction

Pulmonary Edema (PE) is the result of an imbalance between the formation and reflux of lung tissue

fluid, resulting in the absorption of vast amounts of tissue fluid by lymph and vein failure in the lungs. The fluid transudes into the interstitium of the lungs and then accumulates in the alveolars from the lung capillaries, causing a severe disturbance of pulmonary ventilation and gas exchange [1]. Two types of PE cases are acute and chronic.

Since December 2019, typical pneumonia termed covid 19 has been increasing exponentially world-

*Corresponding author. Khanna Nehemiah H., Professor, Ramanujan Computing Centre, College of Engineering Guindy, Anna University, Chennai 600025, Tamil Nadu, India. E-mail: nehemiah@annauniv.edu.

wide [3]. As of October 12th, 2022, a total of 619.7 million cases of Covid-19 had been diagnosed worldwide, with a total of 6.5 million people dying as an outcome of the condition [4]. It poses a significant threat and challenge to world health and the economy. Medical specialists face uncertainty in making decisions based on their judgment for covid 19.

The identification and differentiation of covid-19 and PE is a challenging task. There has been an increased rate of mortality due to the lack of a diagnosis system and the lack of expertise and difficulty in differentiating PE and covid-19 lesions through human observation. Moreover, early diagnosis of the diseases is not possible, which accelerates the severity of the symptoms due to the lack of treatment at the early stages and even runs the risk of being vulnerable to human errors. Hence, a clinical decision support system that can considerably reduce the time and increase the quality of prediction is highly imperative to address these problems. In precise, a Clinical Decision Support System (CDSS) [5] can effectively foster better communication, improve quality and expedite the diagnosis and help physicians in the betterment of treatment. The main objective of this study is to design a CAD that helps in decreasing the false negative rate in the diagnosis of pulmonary edema and covid-19 and assisting physicians as a second opinion system.

Computed Tomography (CT) is better than conventional radiography because it produces high-resolution pictures and can detect early-stage lesions [6, 7]. CT is also used to diagnose pneumoconiosis, pneumonia, pulmonary edema, and lung cancer. Treatment requires early disease identification. Due to large CT scan data volume, it is one of the radiologists' most challenging responsibilities. Thus, radiologists need CAD tools to analyse CT scans for pulmonary diseases [8, 9].

The motivation for designing a Computer Aided Diagnosis (CAD) system is to assist medical practitioners in differentiating PE and covid-19 lesions efficiently. The main goal of employing a CAD system is to identify even small lesions at an early stage. It serves as a supporting and decisive role in the interpretation of medical imaging diagnosis. The implementation of a CAD system has the potential to lead to an improved analysis using CT images and locate the regions of interest. The CAD serves as a second opinion system, whilst physicians are still the primary source for diagnosing. This can be a unique methodology to take up the challenge of differentiating PE and covid-19.

Critical reviews of the latest studies [5, 10–20] show that CAD system-based prediction models using clinical datasets can be significantly helpful in a timely, effective, and accurate diagnosis of the disease. The main advantage of using CAD systems is the quality services which can be used by medical practitioners to enhance the accuracy in clinical decision making. It also helps in achieving time efficiency, image diagnosis accuracy, interpretation time and sensitivity, which benefits the diagnosis of covid-19, PE and other pulmonary disorders. It not only classifies numerous pulmonary diseases but also provides an early and accurate diagnosis of the anomaly shown in the images. The present study aims to establish a CAD system with the following contributions.

- i. Segmentation was done using “Otsu’s Thresholding”.
- ii. ROIs were extracted, and the feature vectors were split as train and test sets in the ratio of 80 : 20.
- iii. A wrapper-based feature selection (FS) approaches using Squirrel Search Algorithm (SSA) with the Support Vector Machine (SVM) classifier’s accuracy as the fitness function has been used to select the optimal features.
- iv. The selected features were trained using the Back Propagation Neural Network (BPNN) classifier.
- v. Statistical Analysis namely: Wilcoxon’s test, Mean Whitney U test and Kendall’s Rank Correlation Coefficient were performed, which indicates that the proposed method has a significant impact on the accuracy, sensitivity and specificity of the novel dataset considered.
- vi. With the advantage of SSA combined with BPNN classifier, more accurate (88.02%) solutions are provided when compared with other traditional ML classifiers, namely Linear SVM, Ada Boost, Cat Boost, XG Boost, Poly SVM, RBF SVM and Sigmoid SVM.

The rest of the article is structured as described below. Section 2 presents the results of the literature review conducted in relation to the work done. In Section 3, the framework for the system is designed. The outcomes of the experiments are discussed in Section 4. Under Section 5, the description of the dataset and an analysis of its performance are presented. The conclusion is presented in Section 6, as well as the future scope of the work.

2. Literature survey

2.1. Classification algorithms and feature selection filter algorithms using medical data

Sweetlin et al. [18] used CAD to improve lung CT pulmonary bronchitis diagnosis. Optimal thresholding segmented the lungs. The size and texture of Pathology bearing Regions (PBRs) are retrieved. Relevant characteristics were identified using a hybrid Ant Colony Optimization (ACO) – cosine similarity – SVM classifier approach. Tandem run recruiting was employed to choose promising features. Simple performance measures evaluate the trained SVM classifier. ACO with tandem run strategy was 81.66% accurate, whereas, without tandem run, it was 77.5% accurate.

Mohammed et al. [21] developed adaptive wrapper FS techniques based on a new metaheuristic algorithm, the coronavirus herd immunity optimiser (CHIO), with and without a greedy crossover (GC) operator strategy to improve CHIO's search space exploration. 23 medical benchmark datasets and a real-time covid-19 dataset have been used to test CHIO and CHIO-GC. CHIO-GC surpassed CHIO in search capabilities, as measured by selection size, accuracy, standard deviation, convergence speed and F-measure. Two FS approaches and four filter-based methods have been compared to the proposed CHIO-GC. On all 23 medical datasets, CHIO-GC had the highest accuracy of 0.79.

Sanjay Agrawal et al. [22] developed a hybrid technique for analysing brain MRI (Magnetic Resonance Imaging) data. A layered thresholding technique for brain MRIs was developed using edge size information. Brain image slice Gray Level Co-Occurrence Matrix (GLCM) was used to compute edge magnitude. Cuckoo Search (CS), SSA, Adaptive Cuckoo Search (ACS), and ACS-SS algorithms discovered multiple optimal threshold values. This was evaluated with Harvard T2-w brain MRI images. The ACS-SS methodology is preferable for qualitative and statistical evaluations, according to the analysis. Hybridisation validates objective function value indices. ACS is CPU-efficient, and it improves fitness function and speed.

Mohit et al. [23] developed SSA. This research studies southern flying squirrel's dynamical hunting and gliding to optimise SSA. Both CEC 2014 and 33 standard functions were used to evaluate SSA. We used convergence rate analysis, statistics, Wilcoxon's

test, and ANOVA to compare nature-inspired optimisation techniques. Comparative statistical study reveals that SSA converges to global optimal solutions. SSA has shown accuracy and consistency for complex CEC 2014 benchmark functions, while alternative approaches are difficult. The heat Flow Experiment verified the algorithm's applicability and stability. In this real-time experiment, SSA's numerical optimisation results were competitive with other optimisers. It allegedly solves actual NP-hard combinatorial optimisation issues.

Parthiban et al. [24] introduced HFWE-based FS to select the best Chronic Kidney Disease (CKD) dataset characteristics. This technique minimises CKD analysis feature breadth, while SVM categorises features. The filter algorithm uses Relief, One-R, GR, and Gini Index (GI). The Improved Bat Algorithm (IBA) wrapper algorithm picks CKD dataset analytical features. The SVMt-based embedded algorithm picks CKD dataset analytical characteristics. SVM classifiers are used for low feature set validity. 400 patients, 250 with early CKD. UCI machine learning (ML) repository supplied CKD samples. UCI characteristics predict CKD. HWFFS algorithm features are tested to determine CKD or non-CKD. Classification performance depended on six parameters. SVMHFWE-FS outperforms (Artificial Neural Network) ANN-HFWE-FS and SVMHFWE-FS with 93.33% accuracy, 6.66%, and 8.1%. Noisy, missing data. Thus, missing and noisy inputs necessitate a categorization system for future work.

Elena Velichko et al. [25] proposed an EDECOVID-net which is a method based on ML to distinguish covid-19 symptoms from PE with the help of radiomic features using lung CT scans. With a simple structure and just a few math computations, the EDECOVID-net has been put together. The slices with edema and covid have been acquired and pre-processed and the texture features are extracted and then classified using EDECOVID-net. The dataset consisted of 13717 imaging patches, which included 5759 and 7958 covid-19 and edema images admittedly and extracted through CT incision. A comparison of this model with other competitive ML models has been made to prove its superiority. These competitive models are namely, VGG-16 with 94%, VGG-19 with 96%, Xception with 95%, ResNet101 with 97%, and DenseNet201 with 97% accuracy. The differentiation between covid-19 and PE classification has been 98% accurate by this EDECOVID-net computer-aided model. Moreover, it has achieved 0.99 precision, 0.97 recall, 0.98

specificity, sensitivity and F1 score each which is better in comparison to other models involved in the study.

Kumar et al. [26] used a Gabor filter and SVM to differentiate lung x-rays. The model accurately differentiated normal from PE chest x-rays. Most Gabor filter values were chosen empirically, and the angles were chosen to be within the range of normal rib cage incline. Model performance may be affected by pulmonary segmentation. All of the feature vectors in this study had an accuracy of 95.90 or higher and an AUC of 0.96 or higher. This may be attributed to the binary classification and the short number of test photos. Kurtosis, entropy, and skewness have the highest AUC of 1 and can be used for classification.

Liang Sun et al. [27] presented an Adaptive Feature Selection Guided Deep Forest (AFS-DF) for the problem of covid-19 classification. This has been achieved by first extracting location-specific features. CT slices have been involved in the study to capture these. Further, for extracting the high-level features where only a small data is available, a deep forest model has been implemented for learning the features' high-level representation. This model has helped in reducing the redundancy in features. Experimental results on covid-19 classification have shown that the proposed AFSDF-LR, AFSDF-SVM and AFSDF-RF outperformed their counterparts LR, SVM and RF.

2.2. *Spam detection using bidirectional transformers and machine learning classifier algorithms*

Yanhui et al. [28] in their study created an efficient spam detection approach using a pretrained bidirectional encoder representation from the transformer (BERT) and ML algorithms to classify ham or spam emails. Email texts were fed into the BERT, and features obtained from the BERT outputs were used to represent the texts. Four classifier algorithms in ML were employed to classify the features of the text into ham or spam categories. The proposed model was tested using two public datasets in the experiments. The results of the evaluation metrics demonstrate that the Logistic Regression (LR) algorithm achieved the best classification performance in both datasets. The experimental results demonstrate that LR achieved the best classification performance in two publicly available datasets.

Luis et al. [29] differentiated covid-19 pneumonia in chest X-rays. Using hand-crafted radiomic fea-

tures and ML, the study compared covid-19 chest X-ray images and pneumonia. Lung texture was segmented in two ways. Upper, middle, and lower lung zones are separated by the L-R method. First- and second-order texture features are derived from 80 : 20 training/testing data. 136 segmented chest X-rays from two public datasets were used. SVM, Ada Boost (AB), and LR classifiers are analysed. AB distinguished covid-19 from normal pneumonia better than other models. Lung zones outperform L-R in classification parameters (sensitivity, accuracy, AUC). The proposed approach using a model of lung segmentation in six separate zones achieved an AUC of 0.98.

Rongyao et al. [30] published a study that compares moderate and severe cases, differences, HDLSS data, and class imbalances. Chest CT slices using multi-task multi-modality SVM can detect covid-19 early. The system consisted of three-step (i) multi-modality handmade features were extracted to retain the difference between severe and moderate; (ii) Data augmentation (iii) Classification and regression. Comparison approaches include baseline, log-least, M3T, MvCCDA, CPM-Nets, and MVF. The suggested method performs better in classification and regression than six state-of-the-art methods on one real-time and two falsified covid-19 datasets. The proposed method obtained 89.47% and 83.20% on the two falsified datasets and 92.57% on the real-time dataset, better than other methods.

Mohammad H et al. [31] proposed an enhanced whale optimisation algorithm for medical FS (E-WOA) to solve the problems in global optimisation. It uses a pooling mechanism and three different searching strategies, namely migrating, preferential selecting, and enriched encircling prey. Comparative results have been shown where E-WOA outperforms the variants of traditional WOA. Further, a binary E-WOA named BE-WOA has been proposed for effectively selecting features precisely in medical datasets. This BE-WOA FS has been used to select features in covid-19 diagnosis. Experimental results and statistical results have been shown to demonstrate the efficiency of BE-WOA over other variants of WOA.

Makram et al. [32] suggested a prediction model that distinguished cases infected by covid-19 based on their symptoms and features. It uses a non-dominated sorting genetic algorithm (NSGA-II) for FS and to find the best trade-off between the two objectives namely, reducing the features and increasing the weights of those selected features. This has

been followed by AdaBoost classifier for classification. The results shown prove the efficiency of NSGA-II algorithm combined with AdaBoost, which has outperformed several other results through combinations of FS algorithms and classifiers.

Duchesne et al. [33] proposed the prediction of covid-19 radiological trajectory with the help of deep learning techniques on chest x-rays. The open CheXnet dataset has been used for training the deep learning algorithm, and patients' data has been collected from two open-source datasets. ROC analyses and Mann-Whitney tests have been performed. ROC-AUC curve determines the label learning, and Mann-Whitney tests compare the results obtained between different outcome groups.

2.3. *Situation awareness prediction technology based on improved RBF neural network algorithm*

Mostafa et al. [3] employed an ANN to diagnose covid-19. 400 Covid-19 cases were included in the dataset. Backpropagation trains ANN. MLP activation produces the best diagnostic output. Performance measures have determined the optimal ANN structure. 9-10-15-2 is the finest ANN (two hidden layers with 10 in the first and 15 in the second accounting for a total of 25 sigmoid nodes). Accuracy, specificity, sensitivity, and AUC-ROC were calculated for the system. The system has achieved competent results with 94% accuracy, 90.6% specificity, 96.4% sensitivity and 98.2% ROC.

Anuja et al. [34] developed an ML-based automatic covid-19 identification system. CXR and CT were used to detect coronavirus. First, quality enhancement and noise removal are applied to the slices. Using fuzzy c-means clustering, the pre-processed images are segmented. Third, the principal component analysis identified four features from the feature vector: gradient histogram, textural, statistical, and discrete wavelet transform. Normal, pneumonia, and covid-19 classifications use k-Nearest Neighbor (k-NN), Sparse Representation (SRC), ANN, and SVM classifiers. Nine datasets from diverse sources were analysed. k-NN, SRC, ANN, and SVM had covid-19 detection accuracies of 91.70, 94.40, 96.16, and 99.14%.

Adi et al. [35] developed a lung cancer detection technique using microscopic biopsies. Digital image processing includes texture feature extraction using GLCM and image classification using BPNN. Tex-

ture features are extracted using contrast, correlation, energy, and homogeneity, and lung biopsy images are identified using an ANN algorithm. This new system can classify images with a 95% training accuracy and 81.25 testing accuracy. These two studies suggest this technology can detect lung cancer.

According to the literature, the above-mentioned approaches can solve the FS problems in various applications. Based on these above inferences, the proposed CAD framework is developed to cater to an effective feature selection mechanism that can significantly impact the classifier's performance. Therefore, the SSA is used and is trained with SVM classifier as the fitness function to select features of PE and covid-19. With the advantage of SSA combined with BPNN classifier, more accurate (88.02%) solutions are provided when compared with other traditional ML classifiers namely, Linear SVM, Ada Boost, Cat Boost, XG Boost, Poly SVM, RBF SVM and Sigmoid SVM.

3. System framework

The present study aims to establish a CAD system with the following contributions. (a) Segmentation has been done using "Otsu's Thresholding". (b) ROIs were extracted, and the feature vectors were split as the train and test sets in the ratio of 80:20. (c) A wrapper based feature selection approach using SSA with the SVM classifier's accuracy as the fitness function has been used to select the optimal features. (d) The selected features were trained using the Back Propagation Neural Network (BPNN) classifier. (e) Statistical Analysis, namely Wilcoxon's test, Mean Whitney U test and Kendall's Rank Correlation Coefficient were performed, which indicates that the proposed method has significant impact on the accuracy, sensitivity and specificity on the novel dataset considered (f) Combining SSA with BPNN classifier provide a more accurate solution when compared to other seven benchmark classifiers. SVM classifier is less prone to the problem of overfitting and hence produces robust and accurate systems, whereas BPNN classifiers are very sensitive to noisy data and outliers. SSA yields the best feature selection subset to improve the accuracy of the classifier which in turn enhances the performance of SVM and BPNN classifiers with minimal computational cost and time [36]. The flowchart for the proposed CAD system framework is shown in Fig. 1.

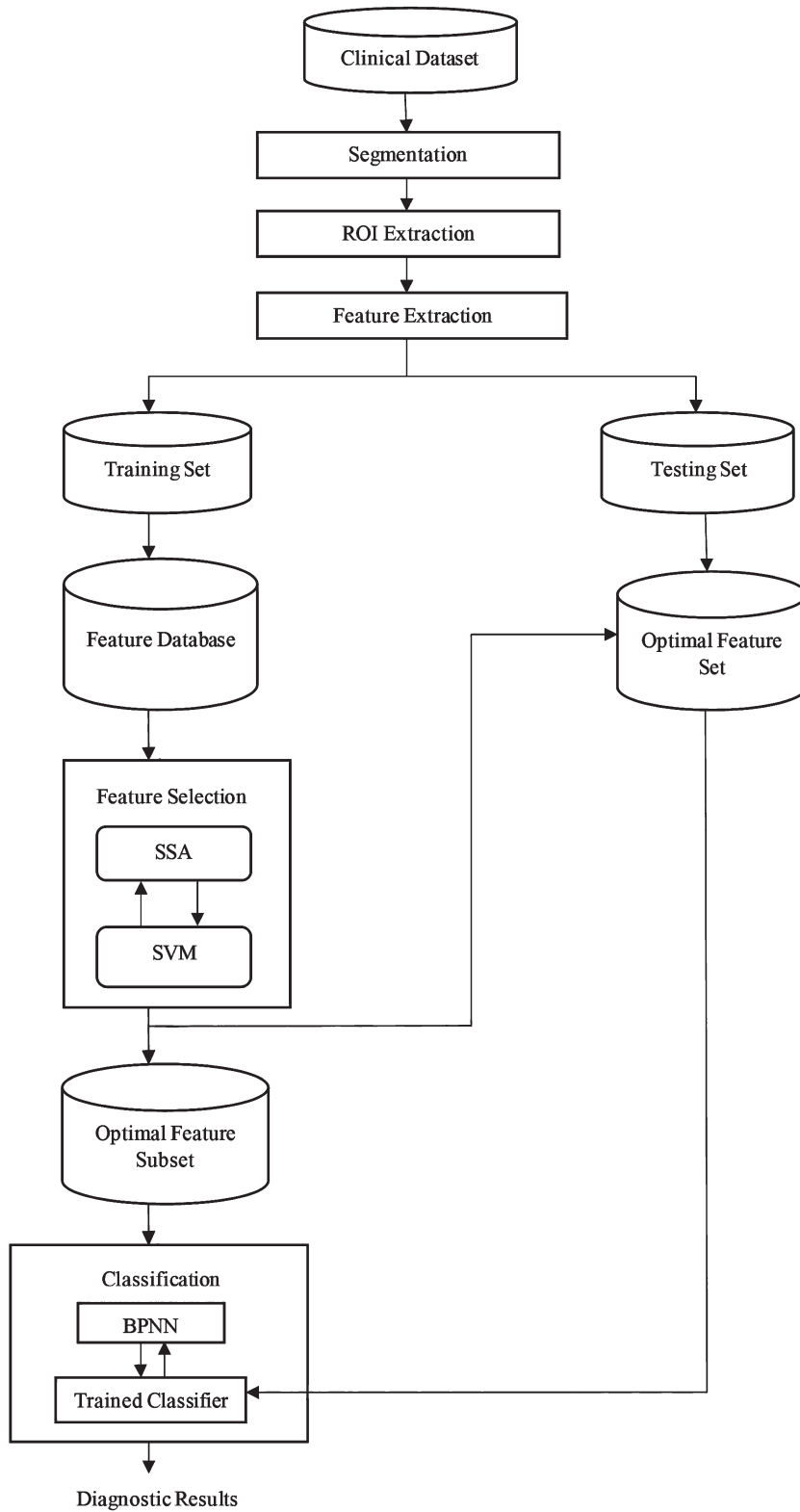


Fig. 1. Flowchart for Proposed CAD System Framework.

3.1. Segmentation

The procedure of segmentation involves separating the lung tissues from the lung CT slices in order to obtain the objects of interest for further analysis by removing the background of the lung tissues. This is done in order to perform segmentation [26]. The Otsu's thresholding technique was followed to accurately segment the lung tissues from their background.

The OTSU's thresholding method was put forward by Japanese scholars OTSU in 1979 [37]. It takes the maximum inter-class variance between the background and the foreground as a measuring standard. The Otsu's algorithm is based on a discriminant analysis that partitions the image based on the gray levels. According to the grayscale characteristics, the variance between the lung CT slices and the lung tissues is maximised to minimise the probability of misclassification, realising an accurate segmentation of the image.

The steps taken in Otsu's thresholding method [38] are outlined as follows.

Input: Lung CT slice.

Process:

Step 1: Convert the CT slices obtained into grayscale slices for further processing. The intensity of the gray level slice is expressed in L gray levels $[1, 2, \dots, L]$. The entire number of points in the CT slice can be expressed as $X = x_1 + x_2 + \dots + x_L$ where x_i denotes the number of points with gray level at i . The histogram of the slice is the possibility that the grey level i can occur, and it is represented by the equation (1).

$$p(i) = \frac{x_i}{X}, x_i \geq 0, \sum_{i=1}^L x_i \quad (1)$$

Step 2: Divide the slice into two classes C_0 and C_1 by a threshold t where C_0 denotes pixels within $[1, 2, \dots, t]$ levels and C_1 represents pixels within $[t + 1, \dots, L]$ levels. C_0 and C_1 corresponds to the lung CT slice (background) and lung tissues (foreground) respectively.

Step 3: Find the occurrence of probabilities, ω_0 for class C_0 and ω_1 for class C_1 respectively using the equations (2) and (3).

$$\omega_0 = \omega(t) = \sum_{i=1}^t p(i) \quad (2)$$

$$\omega_1 = 1 - \omega(t) = \sum_{i=t+1}^L p(i) \quad (3)$$

The average gray levels of these lung CT slice class and lung tissue class can be computed using the equations (4) and (5).

$$\mu_0 = \sum_{i=1}^t \frac{i * p(i)}{\omega_0} = \frac{1}{\omega(t)} \sum_{i=1}^t i * p(i) \quad (4)$$

$$\mu_1 = \sum_{i=t+1}^L \frac{i * p(i)}{\omega_1} = \frac{1}{1 - \omega(t)} \sum_{i=t+1}^L i * p(i) \quad (5)$$

Step 4: Calculate total mean μ_r for the whole CT slice using the equation (6).

$$\mu_r = \omega_0 \mu_0 + \omega_1 \mu_1 \quad (6)$$

Step 5: Calculate the between-class variance σ_B^2 of the two classes C_0 and C_1 using the Equation (7).

$$\sigma_B^2 = \omega_0(\mu_0 - \mu_r)^2 + \omega_1(\mu_1 - \mu_r)^2 \quad (7)$$

Step 6: Calculate the total variance σ_T^2 using the equation (8).

$$\sigma_T^2 = \sum_{i=0}^{L-1} (i - \mu_T)^2 \quad (8)$$

Step 7: In the discrimination analysis, compute the separable degree η of the class using the equation (9).

$$\eta = \max_{1 \leq t \leq L} \sigma_B^2 \quad (9)$$

Step 8: Maximise the inter-class variance using the equation (10) to minimize the probability of misclassification for choosing the optimal threshold to improve the performance of segmentation.

$$t^* = \arg \max_{1 \leq t \leq L} \sigma_B^2$$

The lung tissues were segmented from the obtained CT slice using morphological operations.

Output: Segmented lung tissues.

3.2. ROI extraction

Input: Segmented lung tissues

Process:

Step 1: Locate and extract all the suspicious lesions from the segmented lung tissues and categorise the ROIs [18, 39].

Step 1.1: ROIs that show the manifestation of Edema are categorised as Class 0.

Step 1.2: ROIs that show the manifestation of covid-19 are categorised as Class 1 and

Step 1.3: ROIs that do not show any disease manifestation are categorised as Class 2.

Step 2: Assign a class label to each ROI with special attention, based on the radiologist's descriptive information of each lesion.

Output: Extracted ROIs with class labels (0, 1, 2).

3.3. Feature extraction

A feature is an information extracted from the images in terms of numerical values that are difficult to understand and correlate by humans. Features contain information required to distinguish between classes. The pathology bearing regions (PBRs) are the ROIs in this work. The ROIs considered in this work are edema lesions, namely interlobular septal thickening, peri-bronchovascular interstitial thickening, pleural effusion, GGO and thick fissure and covid-19 lesions, namely GGO, Consolidation, Pleural effusion, Crazy paving appearance, air space consolidation, Broncho-vascular thickening and Traction Bronchiectasis (TB). Features in a medical image are categorised as Texture based features, Run-length features, Geometric features and haralick features. The features were generated using the GLCM. The feature vector is generated for all ROI by combining the 80 features, which includes 12 geometric features and 68 texture features which are given in Table 1. The feature vectors were split in the ratio of 80 : 20, with 80% used for training and 20% for testing the trained classifier [10, 13, 16, 18, 39].

3.4. Feature selection

In FS, the relevant feature subset was chosen using a wrapper-based FS approach based on the SSA and the SVM classifier's accuracy as the fitness function [23]. The algorithm starts by finding out where flying squirrels are at random. In a search space with d dimensions, the vector shows where the flying squirrels are. So, flying squirrels can glide in 1-D, 2-D, 3-D, or "hyper-dimensional" search space and also change their location vectors. The parameters used in SSA are given in Table 2. The four rules of SSA are as follows.

Rule 1: There are n number of flying squirrels in a deciduous forest and only one squirrel is assumed to be on one tree.

Table 1
Features Extracted

Extracted Features			
Geometric		Texture (0°, 45°, 90°, 135°)	
1.	Equivalent Diameter	1.	Sum of squares variance
2.	Convex Area	2.	Contrast
3.	Perimeter	3.	Sum Entropy
4.	Major axis length	4.	Autocorrelation
5.	Eccentricity	5.	Sum Average
6.	Euler number	6.	Correlation
7.	Minor axis length	7.	Cluster Prominence
8.	Orientation	8.	Maximum Probability
9.	Area	9.	Cluster Shade
10.	Solidity	10.	Dissimilarity
11.	Filled area	11.	Information Measure of correlation
12.	Extent	12.	Energy
		13.	Entropy
		14.	Homogeneity
		15.	Difference Variance
		16.	Inverse Difference
		17.	Difference Entropy

Table 2
Parameters used in SSA

Parameter	Value	Definition
n	50	Initial population
G_c	1.9	Gliding Constant
R_1, R_2, R_3	[0,1]	Random number
P_{dp}	0.1	Probability of the existence of a predator
sf	16 to 37	Scaling factor
d_g	9 to 20 m	Gliding Distance
$U(0,1)$	[0,1]	Random number which is Uniformly Distributed

Rule 2: By exhibiting a dynamic foraging behaviour, all the flying squirrel individually searches for food, and they optimally utilize the available food resources.

Rule 3: In the deciduous forest there are only three types of trees namely normal tree, oak tree for acorn nuts food source and hickory tree for hickory nuts food source.

Rule 4: The deciduous forest region under consideration is assumed to contain three oak trees and one hickory tree.

The SSA is depicted as follows:

Input: Feature Vectors.

Process:

Step 1: Randomization of the data is when the algorithm gets started. The location of the i^{th} flying

squirrel is described by a vector, and there is a total of n flying squirrels. The following matrix depicts the location of the flying squirrels.

$$FS = \begin{bmatrix} FS_{1,1} & FS_{1,2} & \dots & FS_{1,d} \\ FS_{2,1} & FS_{2,2} & \dots & FS_{2,d} \\ \vdots & \vdots & \dots & \vdots \\ FS_{n,1} & FS_{n,2} & \dots & FS_{n,d} \end{bmatrix} \quad (11)$$

where $FS_{i,j}$ represents the j^{th} dimension of i^{th} flying squirrel

The initial location of each flying squirrel in the forest is allocated using a uniform distribution as follows,

$$FS_i = FS_L + U(0, 1) \times (FS_U - FS_L) \quad (12)$$

where FS_L and FS_U are the lower and upper bounds of the i^{th} squirrel on the j^{th} dimension.

Step 2: The fitness of each flying squirrel's location is evaluated. The fitness depicts the quality of the food searched by the flying squirrel. The fitness of location of the squirrel is calculated by using SVM classifier and the corresponding values were stored in the array as follows.

$$f = \begin{bmatrix} f_1([FS_{1,1} \quad FS_{1,2} \quad \dots \quad FS_{1,d}]) \\ f_2([FS_{2,1} \quad FS_{2,2} \quad \dots \quad FS_{2,d}]) \\ \vdots \\ f_n([FS_{n,1} \quad FS_{n,2} \quad \dots \quad FS_{n,d}]) \end{bmatrix} \quad (13)$$

Step 3: It is possible to determine the location of flying squirrels in an ascending order based on the level of fitness they possess. The squirrel with the lowest value is said to be on the hickory nut tree. The next three best squirrels are said to be on the acorn nut trees, and it is anticipated that they will make their way to the hickory nut tree. It is expected that all of the remaining flying squirrels will be found on the ordinary trees. In addition, the random selection causes some of the squirrels to be assumed to walk towards the hickory nut tree. This is done under the assumption that the squirrels have met the energy requirements that were set for them. The remaining squirrels will make their way to the acorn nut trees in order to satisfy their need for daily energy consumption. The only thing that ever has any effect on the behaviour of flying squirrels when it comes to looking for food is the presence of other animals that could harm them. Therefore, the position updating

method along with the predator presence probability is denoted by the symbol P_{dp} .

Step 4: The new locations are generated based on three cases. During the foraging process of flying squirrels, three situations may arise. In all three situations, it is assumed that in the absence of predator the flying squirrel glides and searches efficiently throughout the forest for its favourite food, while the presence of the predator makes it cautious, and it is forced to search a nearby hiding location. The new location can be modelled as given below.

Step 4.1: The Flying Squirrels which are on the acorn nut trees (FS_{at}) move towards the hickory nut tree. In this case, the new location of the flying squirrel is found using the equation (14).

$$FS_{at}^{t+1} = \begin{cases} FS_{at}^t + dgxG_Cx (FS_{ht}^t - FS_{at}^t) & R_1 \geq P_{dp} \\ RandomLocation & Otherwise \end{cases} \quad (14)$$

Step 4.2: The Flying squirrels on normal trees (FS_{nt}) may move towards acorn nut trees to fulfil their energy needs. In this situation, a new location of flying squirrels is found using the equation (15).

$$FS_{nt}^{t+1} = \begin{cases} FS_{nt}^t + dgxG_Cx (FS_{at}^t - FS_{nt}^t) & R_2 \geq P_{dp} \\ RandomLocation & Otherwise \end{cases} \quad (15)$$

Step 4.3: Some flying squirrels on normal tree, which already have consumed the acorn nuts may move towards the hickory tree. In this case, the new location is found using the equation (16).

$$FS_{nt}^{t+1} = \begin{cases} FS_{nt}^t + dgxG_Cx (FS_{ht}^t - FS_{nt}^t) & R_3 \geq P_{dp} \\ RandomLocation & Otherwise \end{cases} \quad (16)$$

Step 5: Calculate the seasonal constant (S_c). The seasonal monitoring condition is modelled using the following steps.

- a. Calculate the seasonal constant (S_c) using the equation (17).

$$S_c^t = \sqrt{\sum_{k=1}^d (FS_{at,k}^t - FS_{ht,k}^t)^2} \quad (17)$$

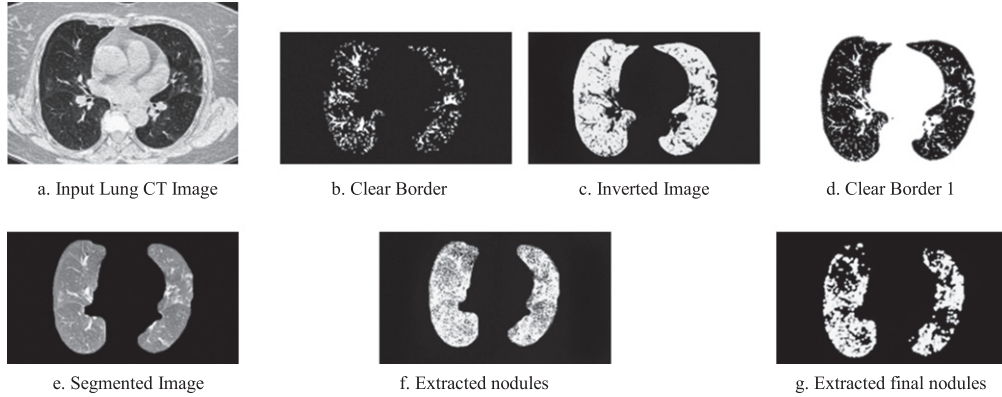


Fig. 2. a–2g. Experimental images obtained for Pulmonary Edema.

- b. Find the minimal value of seasonal constant S_{min} and check the seasonal monitoring condition, i.e., ($S_c^t < S_{min}$) using the Equation (18).

$$S_{min} = \frac{10E^{-6}}{(365)^{t/\left(\frac{m}{2.5}\right)}} \quad (18)$$

- c. If the seasonal condition is true, then relocate the flying squirrels.

Step 6: Random relocation is done for the flying squirrels which could not search the hickory nuts food source but still survived. The relocation of the flying squirrels is modelled through the Levy Distribution using the equation (19).

$$FS_{nt}^{new} = FS_L + Levy(n) \times (FS_U - FS_L) \quad (19)$$

Step 7: Repeat steps from 4 to 6 until convergence is met. The location of the flying squirrel on the hickory nut tree is the final optimal solution, which is stored in the feature database [22, 23].

Output: Optimal Feature Subset.

3.5. Classification subsystem

The dataset is divided in the ratio of training-testing sets as (80-20). The BPNN is the classifier used and it has three layers. The input layer has 80 features, and the hidden layer has 21 selected features, and the output layer represents 3 classes namely, class 0 for edema, class 1 for covid-19 and class 2 for the normal dataset [10, 13].

Input: Optimal Feature Subset

Process:

Step 1: Initiate the neural network (NN) by setting its biases and weights.

Step 2: The input to the NN is the 21 features selected by the FS.

Step 3: In the hidden layer, the sigmoid function is utilised as the activation function for the hidden layer.

Step 4: Compute Error rate by subtracting predicted output from the actual output.

Step 5: Update the weights and biases in a backward pass using Gradient Descent function.

Step 6: Steps 2 to 5 are repeated until the learning is done via minimising the error rate or when the maximum number of iterations is reached.

Output: Diagnostic Measures.

4.. Experimental results

The real-time PE and covid-19 datasets were used in the experiment. The results from different stages of the process obtained for a slice containing PE lesions are displayed in Figs. 2a – 2g. Figure 2a shows the input CT lung slice, and Figs. 2b – 2e shows the stages involved in segmenting the input slice. Figures 2f and 2g show the nodules (Interlobular Septal Thickening and Ground Glass Opacity (GGOs)) which were extracted. Similarly, the experimental results obtained for a healthy lung slice are given in the series of Figs. 3a – 3g. The performance of the proposed system was compared with the existing ML classifiers, namely Linear SVM (78%), Ada Boost (80%), RBF SVM (76%), XGBoost (81%), Poly SVM (75%), Cat Boost (82%) and Sigmoid SVM (65%) classifiers. From the results, it is inferred that the proposed system outperforms the existing ML classifiers. The comparison graph of the classifiers is given in Fig. 4.

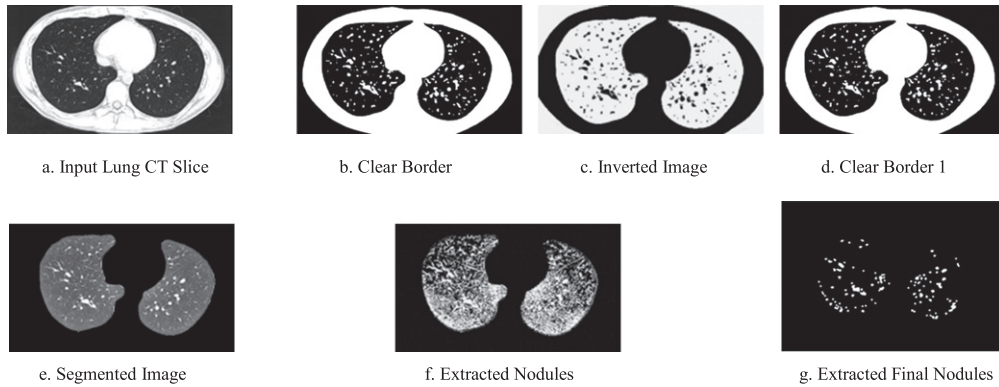


Fig. 3. a – 3 g. Experimental images obtained for Normal Lung Slices.

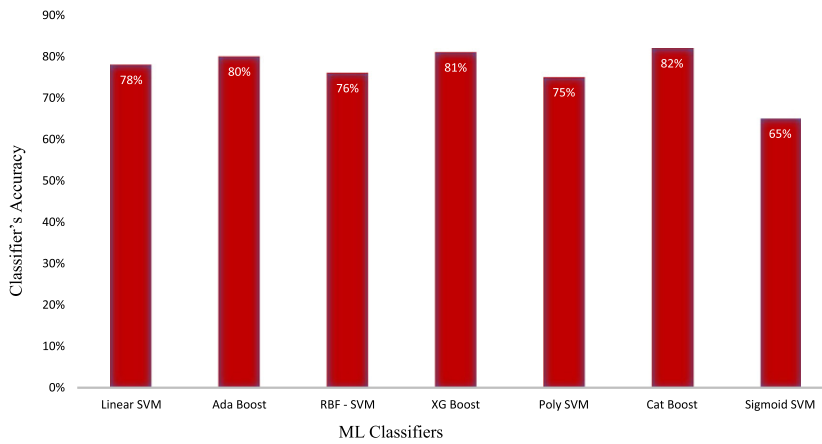


Fig. 4. Comparison of the classifiers.

5. Dataset description and performance evaluation

The real-time dataset has been collected from Bharat Scan Centre, Chennai, India. There was a total of 50 patient datasets obtained, which included 15 sets of healthy people, 9 patients diagnosed with PE disease, and 26 patients diagnosed with covid-19. Training has been carried out for PE using 735 Computed Tomography scan slices which include 270 slices with PE, 240 slices with covid-19 and 225 normal slices. Feature vectors of ROIs obtained from the same patient are present either in the training or testing set and not in both sets. To protect privacy, we deleted the Personally Identifiable Information (PII) from all CT scans. Although the covid-19 CT slices include enough data, proper annotated labels are also required. Each ROI was identified using the expert radiologist's opinion, manually labelled, and made clear by the radiologist. The dataset description

Table 3
Dataset description

Patient cases	Number of patients	Total slices	Number of ROIs	Training set ROIs	Testing set ROIs
Covid-19 disease	26	240	235	188	47
Pulmonary edema disease	9	270	247	197	50
Normal	15	225	224	179	45
Total	50	735	706	564	142

for the PE disease and covid-19 disease are given in Table 3.

The performance measures, namely accuracy, specificity, precision, and sensitivity is calculated using Equations (20) through (23).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (20)$$

Table 4
Performance evaluation

Feature selection using SSA	No. of features	Accuracy (%)	Precision (%)	Specificity (%)	Sensitivity (%)
With FS	21	88.02	88.12	84.78	83.4
Without FS	80	83.80	84.23	81.21	81.59

Table 5
Confusion matrix

Actual and predicted	Correctly identified	Incorrectly identified
Correctly identified	39	10
Incorrectly identified	7	39

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{21}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{22}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{23}$$

Where, true positive (TP) refers to the number of instances correctly classified as the presence of disease.

A true negative (TN) refers to the number of instances correctly classified as the absence of disease.

A false positive (FP) refers to the instances classified as the presence of disease in the absence of disease.

A false negative (FN) refers to the instances which are classified to be the absence of disease when it is actually present. The performance metrics are given in Table 4.

While using the SSA, 21 optimal features are selected and a maximum of 88.02% accuracy is achieved. The precision value achieved with FS is 88.12%, the specificity value is 84.78% and the sensitivity value is 83.4%. The confusion matrix is given in Table 5.

Statistical significance of the proposed training method is established using three statistical analysis methods namely, Wilcoxon’s test, Kendall’s Rank Correlation Coefficient test and Mean Whitney U test. Wilcoxon Test showed that the difference was statistically significant. The results in p value is $< .001$, which is below the specified significance level of $< .05$. The result of the Wilcoxon test is therefore significant for the present data and then the null hypothesis is rejected. A Mann-Whitney U-Test showed that the difference with respect to the dependent variable was statistically significant, i.e., $p = < .001, r = 0.8$. Thus,

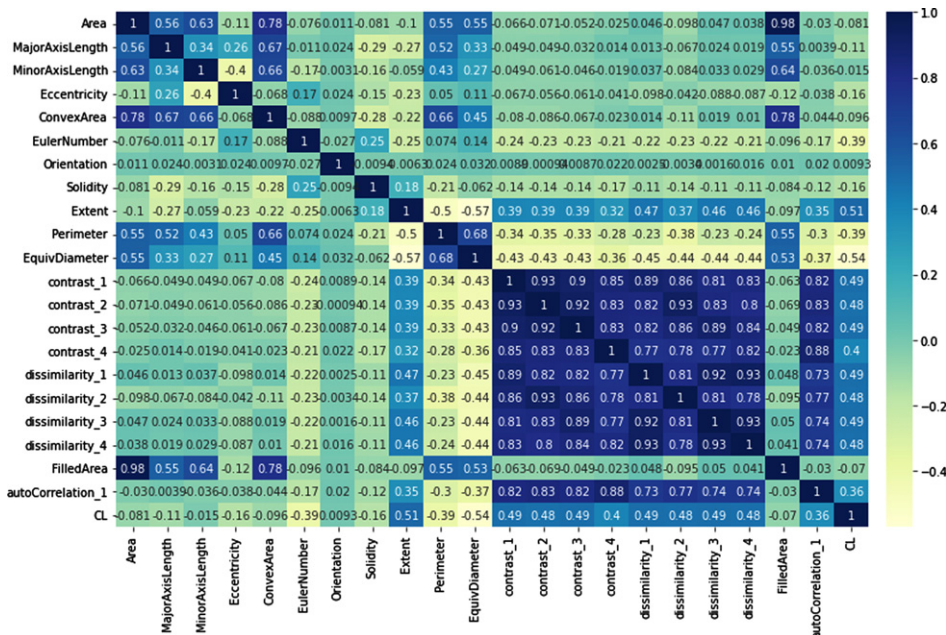


Fig. 5. Kendall’s Rank Correlation Coefficient graph.

the null hypothesis value is rejected. Kendall's Rank Correlation Coefficient graph is given in Fig. 5.

6. Conclusion

The covid-19 pandemic has led to reduced seriousness in the identification and treatment of PE, which also has significant impact on the mortality rate of patients. Hence an effective CAD system which can address the presence of covid-19 and PE has been designed and implemented. Through this research work, the false negative rates can be considerably reduced.

From the input chest CT slices, the lung parenchyma was segmented, and the ROIs were extracted. A list of 80 features was obtained from each ROI to formulate feature vectors. The wrapper-based FS was performed using SSA with the accuracy of the SVM classifier as the fitness function. The optimal feature subset obtained from the FS subsystem was trained using the BPNN classifier, and the performance was evaluated using the performance metrics. The proposed CAD system yielded better accuracy when Squirrel Search Algorithm have been used with the fitness function of SVM classifier and BPNN classifiers whereas it has the limitation of not being trained with big data and does not indicate the level of severity of the disease.

The system was examined using the real-time covid dataset and the experimental results showed that the framework with SSA has shown a greater accuracy of 88.02% compared to that without SSA with 83.80% accuracy. Statistical analysis namely, Wilcoxon's test, Kendall's Rank Correlation Coefficient test and Mean Whitney U test were performed which indicates that the proposed method has significant impact on the accuracy, sensitivity and specificity on the novel dataset considered.

In future, this work can be extended by using the Linguistic Neutrosophic Cubic Set (LNCS). By using the linguistic features, future work can be extended to the theory of LNCS by developing aggregation operators based on the properties of the LCNS.

References

- [1] N.C. Staub, Pulmonary Edema, *Physiol Rev* **54**(3) (xxxx), 678–811.
- [2] National Guideline Centre (UK), Chronic Heart Failure in Adults: Diagnosis and Management, *London: National Institute for Health and Care Excellence (UK)*, 2018.
- [3] M. Shanbehzadeh, R. Nopour and H. Hazemi-Arpanahi, Developing an Artificial Neural Network for detecting covid-19 disease, *Journal of Education and Health Promotion*, 2022.
- [4] World Health Organisation Coronavirus (COVID-19) Dashboard. <https://covid19.who.int/>
- [5] R. Sammout, K.B. Salah, K. Ghedira et al., A Proposal of Clinical Decision Support System Using Ensemble Learning for Coronary Artery Disease Diagnosis, *Wireless Mobile Communications and Healthcare, Proceedings from 9th EAI Conference*, 2020.
- [6] Y. Fang, H. Zhang, J. Xie et al., Sensitivity of chest CT for COVID-19: Comparison to RT-PCR, *Radiology* **296**(2) (2020), E11-E117.
- [7] C. Long, H. Xu, Q. Shen, et al., Diagnosis of the Coronavirus disease (COVID-19): rRT-PCR or CT? *European Journal of Radiology* **126** (2020), 108961.
- [8] M. Yen Ng, E.Y.P. Lee, J. Yang et al., Imaging Profile of the COVID-19 infection: Radiologic findings and literature review, *Radiology Cardiothoracic Imaging* **2**(1) (2020).
- [9] N. Hollings, P. Shaw, Diagnostic imaging of lung cancer, *European Respiratory Journal* **19**(4) (xxxx), 722–742.
- [10] A. Isaac, H. Nehemiah, A. Isaac et al., Computer-Aided Diagnosis system for diagnosis of pulmonary emphysema using bio-inspired algorithms, *Computers in Biology and Medicine* **124** (2020), 103940.
- [11] M. Sethi, S. Rani, A. Singh, A CAD System for Alzheimer's Disease Classification Using Neuroimaging MRI 2D Slices. *Computational and Mathematical Methods in Medicine*, 2022.
- [12] S. Haggag, F. Khalifa, H. Abdeltawab et al., Automated CAD System for Intermediate Uveitis Grading Using Optical Coherence Tomography Images, *IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–4, 2022.
- [13] A. Isaac, H. Nehemiah, S.D. Dunston et al., Feature selection using competitive coevolution of bio-inspired algorithms for the diagnosis of pulmonary emphysema, *Biomedical Signal Processing and Control*, pp. 103340, 2021.
- [14] D. Napoleon and I. Kalaiarasi, Classifying Lung Cancer as Benign and Malignant Nodule Using ANN of Back-Propagation Algorithm and GLCM Feature Extraction on Chest X-Ray Images, *Wireless Personal Communications* **126**(1) (xxxx), 167–195.
- [15] K.A.A. Aziz, M.I. Saripan, F.F.A. Saadand, CT Lung Images Segmentation Using Image Processing And Markov Random Field, *Malays J Med Health Sci*, pp.31–35, 2022.
- [16] A. Isaac, H. Nehemiah and A. Kannan, Computer-Aided Diagnosis System for Diagnosis of Cavitory and Miliary Tuberculosis Using Improved Artificial Bee Colony Optimization, *IETE Journal of Research*, pp. 1966440, 2021
- [17] N. AlHinai "Introduction to biomedical signal processing and artificial intelligence. In *Biomedical signal processing and artificial intelligence in healthcare*, pp. 1–28, 2020.
- [18] D. Sweetlin, H.K. Nehemiah and A. Kannan, Feature selection using ant colony optimization with tandem-run recruitment to diagnose bronchitis from CT scan images, *Comput Meth Progr Biomed* **145** (2017), 115–125.
- [19] S. Sreejith, H.K. Nehemiah and A. Kannan, A clinical decision support system for polycystic ovarian syndrome using red deer algorithm and random forest classifier, *Healthcare Analytics*, p. 100102, 2022.

- [20] S. Sreejith, H.K. Nehemiah and A. Kannan, A Framework to Classify Clinical Data Using a Genetic Algorithm and Artificial Flora-Optimized Neural Network, *International Journal of Swarm Intelligence Research (IJSIR)* **13**(1) (xxxx), 1–22.
- [21] M. Alweshah, S. Alkhalailah, M. Azmi Al-Betar et al., Coronavirus herd immunity optimizer with greedy crossover for feature selection in medical diagnosis, *Knowledge-Based Systems* **235**(107629).
- [22] S. Agrawal, L. Samantaray, R. Panda et al., A New Hybrid Adaptive Cuckoo Search-Squirrel Search Algorithm for Brain MR Image Analysis, *Hybrid Machine Intelligence for Medical Image Analysis*, pp. 85–117, 2020.
- [23] M. Jain, V. Singh and A. Rani, A novel nature-inspired algorithm for optimization: Squirrel search algorithm, *Swarm and Evolutionary Computation*, **44** (2019), 148–175.
- [24] R. Parthiban, S. Usharani, D. Saravanan, D. Jayakumar, U. Palani, D. Stalin David, D. Raghuraman, Prognosis of chronic kidney disease (CKD) using hybrid filter wrapper embedded feature selection method, *European Journal of Molecular and Clinical Medicine* **07**(09), (2020).
- [25] Elena Velichko, Fariddodin Shariaty, Mahdi Orooji et al., Development of computer-aided model to differentiate covid-19 from pulmonary edema in lung CT scan: EDECOVID-net, **141**(105172) (2022).
- [26] A. Kumar, Y. Wang, K. Liu et al., Distinguishing normal and pulmonary edema chest X-Ray using Gabor filter and SVM, *IEEE International Symposium on Bioelectronics and Bioinformatics*, no. 14334305, 2014.
- [27] L. Sun, Z. Mo, F. Yan et al., Adaptive feature selection guided deep forest for covid-19 classification with chest ct, *IEEE Journal of Biomedical and Health Informatics* **24**(10) (xxxx), 2798–2805.
- [28] Y. Guo, Z. Mustafaoglu and D. Koundal, Spam Detection Using Bidirectional Transformers and Machine Learning Classifier Algorithms, *Journal of Computational and Cognitive Engineering* **00** (2022), 1–5.
- [29] L.V.D. Moura, C.M. Dartora, D. Oliveira et al., A Novel Approach to Differentiate COVID-19 Pneumonia in Chest X-ray, *IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 446–451, 2020.
- [30] R. Hu, J. Gan, X. Zhu et al., Multi-task multi-modality SVM for early COVID-19 Diagnosis using chest CT data, *National Library of Medicine* **59**(102782), (2021).
- [31] M.H.N. Shahraki, H. Zamani and S. Mirjalili, Enhanced whale optimization algorithm for medical feature selection: A COVID-19 case study, *Computers in Biology and Medicine* **148**(105858) (2022).
- [32] M. Soui, N. Mansouri, R. Alhamad et al., NSGA-II as feature selection technique and AdaBoost classifier for COVID-19 prediction using patient's symptoms. *Nonlinear Dynamics* **106**(2) (2021), 1453–1475.
- [33] S. Duchesne, D. Gourdeau, P. Archambault et al., Tracking and predicting covid-19 radiological trajectory using deep learning on chest X rays: initial accuracy testing, *medRxiv*, 2020.
- [34] A. Bhargava, A. Bansal and V. Goyal, Machine learning-based automatic detection of novel coronavirus (COVID-19) disease, *Multimedia Tools and Applications* **81** (2022), 13731–13750.
- [35] K. Adi, C.E. Widodo, A.P. Widodo et al., Detection of lung cancer using Gray Level Co-Occurrence Matrix (GLCM) and back propagation neural network classification, *Journal of Engineering Science and Technology* **11**(2) (2018), 8–12.
- [36] K. Pavya, Classification Algorithms-Support Vector Machine, Back Propagation Neural Network and K-Nearest Neighbor: A Review. *International Journal of Innovative Research in Science, Engineering and Technology* **7**(7) (2018), 8110–8115.
- [37] N. Otsu, A Threshold selection method from grey level histograms, *IEEE Trans. Systems, Man and Cybernetics* **9**(1) (1979), 62–66.
- [38] C.R. Gonzalez and E.R. Woods, Digital Image Processing, Third Edition, *Pearson India Education Services*, 2017.
- [39] D. Sweetlin, H.K. Nehemiah, A. Kannan, Patient specific model based segmentation of lung computed tomographic images, *Journal of Information Science and Engineering* **32**(5) (2016), 1373–1394.