Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 6, pp. 4589–4603, DOI 10.12694/scpe.v25i6.3311

MACHINE LEARNING BASED LUNG CANCER DIAGNOSTIC SYSTEM USING OPTIMIZED FEATURE SUBSET SELECTION

RAMYA PERUMAL [*], YOGESH KUMARAN S [†], I.MANIMOZHI [‡], A.C.KALADEVI [§] AND C.ROHITH BHAT [¶]

Abstract. Lung is a vital organ that plays a major role in respiration. Without breathing, one may not survive in this world. Hence lung is an important organ that acts as filter to absorb oxygen and supply it to heart where pumping takes place through blood vessel in the circulatory system .The pumped blood takes oxygen and other nutrients to every other parts of the body. Hence one must take care of lung. There are various diseases associated with lungs. Lung Cancer is a deadly disease that spread across the countries all over the world. An early detection of lung cancer has been proved to improve the survival rate of human life. There are various resources are available to detect the lung cancer disease. They are low dose CT-scans, X-rays, blood-based screening, pathology slide reading, biopsy's test, survey data(clinical dataset) etc. helps to predict the disease well in advance. Our proposed work uses two clinical datasets that has various features to detect how likely the persons get affected from the lung disease. Dataset1 includes features such as age, gender, smoking, yellow fingers, anxiety, peer-pressure, chronic disease, fatigue, allergy, wheezing, alcohol, coughing, shortness of breath, swallowing difficulty, and chest pain. Also, the work has experimented with another dataset2 that represents causes of lung cancer due to exposure of pesticide. Our proposed diagnostic system consider all these features in total and perform feature selection to extract optimal feature subsets using cuckoo search algorithm then perform classification using machine learning algorithms such as Linear Support Vector Machine, Logistic Regression and Random Forest algorithm. It is observed that with the cuckoo search algorithm, dataset 1 achieves an accuracy of 100%, precision of 100%, recall of 100%, and F1-score of 100% by LR Classifier. The Linear SVC classifier achieves an accuracy of 90%, a precision of 88%, a recall of 86%, and an F1-score of 87%.The Random forest Classifier achieves an accuracy of 91%, precision of 86%, recall of 93%, and F1-score of 90%. For dataset 2, both the LR classifier and Linear SVC classifier outperform with an accuracy of 100%, precision of 100%, recall of 100%, and F1-score of 100%. Whereas Random Forest provides accuracy of 97%, precision of 97%, recall of 96%, and F1-score of 97%.

1. Introduction.

Lungs are important organs for breath control. Humans have two lungs in their chest one on the left side leaving space for the heart and the other on the right side. It prevents unwanted toxic gases from entering the parts of the body. The chest gets expanded during inhalation and shrinks during exhalation which supports widely in the process of respiration. It purifies the blood with oxygen and ensures every cell in the body gets a sufficient supply of oxygen. Air is an important substance that reaches the lungs through the nasal cavity, pharynx, larynx, trachea, and bronchi and end-up in the alveoli. The function of the capillaries in the alveoli is to absorb oxygen and leave out carbon dioxide [1].

There are various diseases associated with the lungs. Lungs get infected, inflamed even it may cause serious complications such as the growth of unwanted cancerous cells [22]. Lung cancer is the second most common cancer present in both men and women. The American Cancer Society estimates for lung cancer is about 2,38,340 new cases and the death toll raised to 1,27,070 in the US for the year 2023. Age and Smoking are the major factors that must be considered for lung cancer[20]. Lung cancer causes 1 in 5 people accounting for death. The women's risk is about 1 in 17. The demographic key statistics report that lung cancer accounts for 5.9% of all cancers and 8.1% of all cancer-related deaths. The main challenge in Lung Cancer is the late diagnosis of the disease resulting in a poor prognosis. Another challenge that exists in the detection of the disease is limited clinical parameters and the relevant population at risk. The accuracy of disease detection is highly dependent

[*] Department of Computer Science and Engineering, Sona College of Technology, Salem (ramyaperumal@sonatech.ac.in)

[†] Department of Computer Science and Engineering, School of Engineering and technology, Jain University (yogesh.ks@jainuniversity.ac.in)

[‡] Department of Computer science and Engineering East Point College of Engineering & Technology, Bangalore, India (drmanimozhi.i@eastpoint.ac.in)

[§] Department of CSE, Sona College of Technology, Salem (kaladeviac@sonatech.ac.in)

[¶] Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, India. (rohithbhat2000@gmail.com)

on the unavailability of the relevant population, systematic data gathering, and data preparation should always consider the clinical application and relevant population at risk. The research says that a stereotypical lung cancer patient is likely to be a 70years old smoker with a history of cardiovascular disease, a chronic obstructive pulmonary disorder, and blood analysis denoting inflammation, hyponatremia, and hypoalbuminemia. These are the risk factors associated with Lung Cancer disease. The integration of relevant clinical information with these associated risk factors characterizes a large risk cohort. The chances of a person getting lung cancer are 20-25% if he smokes a pack of cigarettes each day when compared to a non-smoker. Some of the symptoms of lung cancer include coughing, coughing up blood, chest discomfort, shortness of breath, etc [12]. The procedure for the detection of lung cancer disease is a chest x-ray, computed tomography, magnetic resonance imaging, sputum cytology, etc [9]. All these approaches are time-consuming and expensive. The treatment of lung cancer includes surgery, chemotherapy, radiation therapy, and immune therapy [11]. The diagnosis of lung cancer comes to know by the doctor at its advanced stage only and the survival rate highly relies on age, race, and health condition also it differs from person to person [21].

The evolution of machine learning algorithms finds its application in various healthcare analytics such as diabetes, cardiovascular disease, hypercholesterolemia, acute liver failure, stroke, etc. The machine learning algorithm replicates the human learning system without being explicitly programmed [19]. There are different types of learning algorithms. They are supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, the data and class labels are given as input to the system. In the training phase, the system learns the data with its associated class labels. In the testing phase, it uncovers the latent pattern and then classifies the data accordingly to its classes. This type of learning is termed supervised learning in which data and class labels are available. Another type is unsupervised learning in which data alone is present, the machine itself automatically groups the data instances based on similarities. There is another class of learning termed reinforcement learning in which the system brings into action to maximize the reward in a particular situation. The intelligent agent interacts with its environment and takes steps based on rewarding desired actions and punishing undesired ones [11].

The primary objective of this research work is build an effective diagnostic system to detect lung cancer with remarkable performance measure.

The main contribution of the proposed system is as follows,

1. Collection of datasets from an online repository then perform data pre-processing to standardize the features for further processing.
2. The optimal feature subset selection is obtained by using the Cuckoo Search Algorithm; a FS algorithm that eliminates irrelevant, redundant features and selects novel features to enhance the efficacy of the proposed work. It also overcomes the time and space complexity of data.
3. After selection of novel features from the dataset, then the proposed system is subjected to the use of machine learning algorithms namely Logistic regression, linear Support Vector Machine, and Random Forest for classifying the person who is infected with lung cancer disease or not.
4. Evaluate the proposed computer-assisted lung cancer diagnostic system by using performance measures such as accuracy, precision, recall, and f1-score that provide remarkable results.

2. Related Works. Venkatesh et al. used ensemble learning methods such as Adaboost, and Bagging and integrated three machine learning algorithms viz. K-Nearest Neighbour, Decision tree, and Neural Networks were evaluated on the SEER dataset. The Surveillance, Epidemiology, and End Results program of the National Cancer Institute is an authoritative repository of cancer statistics in the United States. It achieved accuracy with the ensemble method namely bagging in combination with KNN classifier 93.2%, Decision tree 97.3%, and neural network 91.2%. It also accomplished accuracy with the ensemble method namely boosting in combination with the KNN classifier 95.1%, decision tree 98.2%, and neural network 93.1% [2].

Vikas et al. experimented with a dataset collected from data world which consists of 1000 samples with 25 attributes. The author used two machine learning algorithms as Support Vector Machine and Random Forest and compared those algorithms with and without the Feature Selection technique namely Chi-Squared. It achieves an accuracy of 98%, precision of 100%, recall of 100%, and F1-score 100% with an execution time of 0.010 seconds [3].

Faisal et al. used various machine learning algorithms such as Neural Networks, Naïve Bayes, and Support

Vector Machines. The obtained results are compared with ensemble learning methods such as Random Forest and Gradient Boosted tree. It was observed that the ensemble learning method namely Gradient Boosted Tree outperformed with an accuracy of 90%, a precision of 87.8%, a recall of 83.7%, and an F1 score of 85.7% [4].

Puneet et al used a dataset gathered from Lanzhou University consisting of 277 patient blood indices details. He integrated machine learning algorithms such as XGBoost, Grid Search CV, Gaussian Naïve Bayes, Support Vector Machine, Decision tree, and K-Nearest Neighbour for lung cancer prediction. The experiment showed that XGBoost outperformed with an accuracy of 92.16% recall of 96.97% and AUC Area Under Curve of 95% [5].

Alsinglawi et al detected lung cancer patients by using machine learning algorithms such as Random Forest, XGBoost, and Logistic Regression. He analyzed by experimenting with the dataset MIMIC-III dataset. As the dataset is imbalanced, the used over-sampling technique (SMOTE) for the validation. Among the classifiers, the Random forest with SMOTE technique performed better with an accuracy of AUC 98% and recall of 98% [6].

Safiyari et al.used various ensemble learning methods such as Bagging, AdaBoost, MultiBoosting, Dagging, and RandomSubspace in combination with machine learning algorithms such as RIPPER, Decision Stump, C4.5, SMO, Bayes Net, Logistic Regression, and Random Forest. It has experimented with the SEER dataset that consists of 6,43,924 samples with 149 attributes. Among the classifiers, Adaboost outperformed with an accuracy of 88.98% and an AUC of 94.9% [7].

Patra et al.used several machine learning algorithms viz. Radial Basis function network(RBF), KNN classifier, J48, Support Vector Machine, Logistic Regression, Artificial Neural Network, Naïve Bayes, and Random forest were evaluated with the dataset collected from the UCI repository. It consists of 32 instances and 57 attributes. The results of different classifiers were compared and proved that RBF outperformed with an accuracy of 81.2%, a precision of 81.3%, an F1-score of 81.3%, and an AUC of 74.9% [8].

3. Proposed System. The proposed diagnostic system consists of modules such as Data Preprocessing, Feature Selection, and Classification. The block diagram of the proposed Lung cancer diagnostic system.

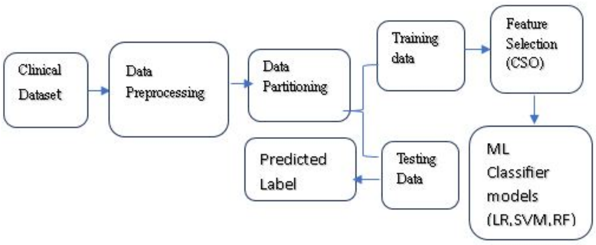

Fig. 3.1: Block Diagram of Lung Cancer Diagnostic System

3.1. Dataset Collection. Survey, or clinical dataset 12 features. They are of numerical features 11 features. Out of 12 categorical datatypes, one is of the class datatype that represents whether a person is affected by lung cancer or not [13]. It consists of 2000 data instances with 1000 data instances representing lung cancer-affected data instances and the remaining 1000 representing healthy persons. Another dataset includes lung cancer caused by exposure to pesticides which consists of 680 samples, 67 attributes, and 1 class attribute representing a patient with Lung cancer or not.

Table 3.1 represents the attributes that are majorly responsible for the cause of lung cancer. These 15 attributes in this dataset are of categorical type denoting its presence or absence in the data instances are the sources of lung cancer. Those 15 attributes are taken in to account in total may leads to computation time and space complexity. Hence, we primarily focus on feature selection that contributes the performance upgradation in predicting the lung cancer disease.

Table 3.1: Dataset I

Attribute	Description
Smoking	Yes = 1 No = 0
Yellow Fingers	Yes = 1 No = 0
Anxiety	Yes = 1 No = 0
Peer Pressure	Yes = 1 No = 0
Genetics	Yes = 1 No = 0
Attention Disorder	Yes = 1 No = 0
Born on even day	Yes = 1 No = 0
Car Accident	Yes = 1 No = 0
Fatigue	Yes = 1 No = 0
Allergy	Yes = 1 No = 0
Coughing	Yes = 1 No = 0
Cancer	Yes = 1 No = 0

Table 3.2: Dataset II

Attribute	Description
ID	Responder's identification number: 1.0, 2.0, 3.0, ... is ID of the case with lung cancer, 1.1, 1.2, 2.1, ... is ID of control without lung cancer
LungCA	Lung cancer status of responders: 0 refers to control without lung cancer, 1 refer to the case with lung cancer
Gender	Gender/sex of the responders, 1 refers to male, 0 refers to female
Age	Age in the year of each responder
Age_group	Responders in each group, 1 refers to those with an age less than or equal to 54, 2 refer to those with age 55-64 yr, 3 refer to those with age 65-74 yr, 4 refer to those with age 75 yr or more
Status	Marital status of the responders, 1 refers to those who are single, 2 refer to those who are married, 3 refers to those who are divorced/spouse passed away/separated
Education	Education level completed by the responders, 1 refers to those who are finished primary school, 2 refer to those who are finished high school, 3 refers to those who are finished their undergraduate or higher degree
Occupation	Occupation of the responders, 0 refers to those who are non-farmers, 1 refer to those who are a farmer
Residency	Living duration (years) in a community of the responders, 1 refers to those who have lived in a community for less than 21 years, 2 refer to those who have lived in a community for 21-30 years, 3 refers to those who have lived in a community for more than 30 years
Distances	Responders' distances between home and their nearest farmland, 1 refers to responders who have a distance less than 500 m, 2 refers to those who have distances 500-1,000 m, 3 refers to those who have distances more than 1,000 m
Cooking_fume	Cooking fume exposure, 1 refers to those who have ever exposure to cooking fume, 2 refer to those did not exposure to cooking fume,
Air_Pollution_ exposure	Responders' exposure to air pollution from various sources, e.g. working in a factory with air pollution (asbestos, diesel engine exhaust, silica, wood dust, painting, and welding exposure), 0 refer to responders who did not expose to air pollution, 1 refer to responders who exposure to air pollution
CigSmoke1	Tobacco use by responders, 0 refer to those who have never smoked a cigarette, 1 refer to current smoker or ex-smoker
CigSmoke2	Tobacco use by responders, 0 refers to those who have never smoked a cigarette, 1 refer to those who smoke less than 109500 cigarettes, 2 refers to those who smoke 109500 cigarettes or more
Cigarette_total	The number of cigarettes the study responders smoked in a lifetime.
Cigarette_year	Number of years the responders have smoked cigarette
Cigarette_number	Number of cigarettes responders smoked per day
CigSmoke_Status	Tobacco status of responders, 1 refers to those who have never smoked a cigarette, 1 refer to ex-smoker, 3 refers to a current smoker
Herbicide	Exposure to herbicides of responders, 0 refers to those who have never used herbicides, 1 refer to those who ever used herbicides
Herbicides_year	Number of years each responder used herbicides

Herbicide_year_group	Groups of years each responder using herbicides, 1 refers to those using the herbicides 1-10 years, 2 refer to those using the herbicides 11-30 years, 3 refers to those using the herbicides more than 30 years
Herbicides days	Number of days using the herbicides of each responder
Herbicides day group	Responders' quartile of days using the herbicides, 1 refers to those who have several days using the herbicides less than 160 days (Quartile 1), 2 refers to those who have several days using the herbicides between 160-500 days (Quartile 2), 3 refer to those who have several days using the herbicides between 500-960 days (Quartile 3), 4 refer to those who have several days using the herbicides more than 960 days (Quartile 4)
Insecticides	Exposure to insecticides of responders, 0 refers to those who do not use insecticides, 1 refer to those who are use insecticides
Insecticide year	Number of years using the insecticides of each responder
Insecticide year group	Groups of years each responder using insecticides, 1 refers to those using the herbicides 1-10 years, 2 refers to those using the herbicides 11-30 years, 3 refers to those using the herbicides more than 30 years
Insecticide days	Number of days using the insecticides of each responder
Insecticide day group	Responders' quartile of days using the herbicides, 1 refers to those who have several days using the insecticides less than 200 days (Quartile 1), 2 refers to those who have several days using the insecticides 200-480 days (Quartile 2), 3 refer to those who have several days using the insecticides 481-1,200 days (Quartile 3), 4 refer to those who have several days using the insecticides more than 1,200 days (Quartile 4)
Fungicides	Exposure to fungicides of responders, 0 refers to those who are not using fungicides, 1 refers to those who are using fungicides
Fungicide years	Number of years using fungicides of each responder
Fungicide year group	Groups of years each responder using fungicides, 1 refers to those using the fungicides 1-10 years, 2 refer to those using the fungicides 11-30 years, 3 refers to those using the fungicides more than 30 years
Fungicide days	Number of days using the fungicides of each responder
Fungicide day group	Responders' quartile of days using fungicides, 1 refers to those who have several days using fungicides less than 96 days (Quartile 1), 2 refers to those who have several days using fungicides between 96-160 days (Quartile 2), 3 refers to those who have several days using fungicides between 161-530 days (Quartile 3), 4 refer to those who have a number of days using fungicides more than 530 days (Quartile 4)
Glyphosate use	Exposure to Glyphosate herbicide (Roundup/ Touchdown/ Spark) of responders, 0 refers to those who did not use Glyphosate, 1 refers to those who used Glyphosate
Glyphosate days	Number of days using the glyphosate of each responder
Paraquat use	Exposure to Paraquat herbicide (Gramoxone/ Knockxone) of responders, 0 refers to those who did not use Paraquat, 1 refers to those who used Paraquat
Paraquat days	Number of days using the paraquat of each responder
2,4-Dichlorophenoxy use	Exposure to 2,4-Dichlorophenoxy herbicide of responders, 0 refers to those who did not use 2,4-Dichlorophenoxy, 1 refer to those who used 2,4-Dichlorophenoxy
2,4-Dichlorophenoxy days	Number of days using the 2, 4-Dichlorophenoxy of each responder
Butachlor use	Exposure to Butachlor herbicide of responders, 0 refers to those who did not use Butachlor, 1 refer to those who used Butachlor
Butachlor days	Number of days using the butachlor of each responder
Propanil use	Exposure to Propanil herbicide of responders, 0 refers to those who did not use Propanil, 1 refers to those who used Propanil
Propanil days	Number of days using the propanil of each responder
Alachlor use	Exposure to Alachlor herbicide of responders, 0 refers to those who did not use Alachlor, 1 refer to those who used Alachlor
Alachlor days	Number of days using the alachlor of each responder
Endosulfan use	Exposure to Endosulfan insecticide of responders, 0 refers to those who did not use Endosulfan, 1 refer to those who used Endosulfan
Endosulfan days	Number of days using the endosulfan of each responder
Dieldrin use	Exposure to Dieldrin insecticide of responders, 0 refer to those who did not use Dieldrin, 1 refer to those who used Dieldrin
Dieldrin days	Number of days using the dieldrin of each responder
DDT use	Exposure to DDT (Dichlorodiphenyltrichloroethane) insecticide of responders, 0 refers to those who did not use DDT, 1 refer to those who used DDT
DDT days	Number of days using the DDT of each responder
Chlorpyrifos use	Exposure to Chlorpyrifos insecticide of responders, 0 refers to those who did not use Chlorpyrifos, 1 refer to those who used Chlorpyrifos

Chlorpyrifos days	Number of days using the chlorpyrifos of each responder
Folidol use	Exposure to Folidol insecticide of responders, 0 refers to those who did not use Folidol, 1 refers to those who used Folidol
Folidol days	Number of days using the folidol of each responder
Mevinphos use	Exposure to Mevinphos insecticide of responders, 0 refers to those who did not use Mevinphos, 1 refers to those who used Mevinphos
Mevinphos days	Number of days using the mevinphos of each responder
Carbaryl/Savin use	Exposure to Carbaryl/Savin insecticide of responders, 0 refers to those who did not use Carbaryl/Savin, 1 refers to those who used Carbaryl/Savin
Carbaryl/Savin days	Number of days using the carbaryl/savin of each responder
Carbofuran use	Exposure to Carbofuran insecticide of responders, 0 refers to those who did not use Carbofuran, 1 refer to those who used Carbofuran
Carbofuran days	Number of days using the carbofuran of each responder
Abamectin use	Exposure to Abamectin insecticide of responders, 0 refers to those who did not use Abamectin, 1 refers to those who used Abamectin
Abamectin days	Number of days using the abamectin of each responder
Armure/Propiconazole use	Exposure to Armure/Propiconazole fungicide of responders, 0 refers to those who did not use Armure/Propiconazole, 1 refers to those who used Armure/Propiconazole
Armure/Propiconazole days	Number of days using the armure/propiconazole of each responder
Methyl aldehyde use	Exposure to Methyl aldehyde fungicide of responders, 0 refers to those who did not use Methyl aldehyde, 1 refer to those who used Methyl aldehyde
Methyl aldehyde days	Number of days using the Methyl aldehyde of each responder
Morphology Group	Morphology of lung cancer cases, 0 refers to control (not lung cancer), 1 refer to adenocarcinoma, 2 refers to squamous cell carcinoma, 3 refers to small cell carcinoma, 4 refers to large cell carcinoma, 5 refers to neoplasm, malignant, and 6 refer to other and unspecified

3.2. Data Preprocessing. Feature standardization is the conversion of numerical features to the same unit of measurement with zero mean and unit standard deviation. Data pre-processing technique includes data cleaning, missing values handling, and categorical variables transformation[1]. If missing values are omitted, we are getting a lesser number of data instances. To overcome this issue, we perform artificial data are included to have complete data instances in total. The missing data values can be filled with suitable data measures. For handling missing data, it is necessary to determine whether the median or mean value of the corresponding numerical attribute is updated in the missing entry. Mean represents the average value of the data attribute. The median is the center or middle value of the data attribute. These values can be interpreted by performing a statistical analysis of the data. Describe() is the method found in the Python library that provides a detailed description of the attribute in terms of mean, count, first quartile, median, third quartile, minimum, and maximum values. For handling categorical data attributes, the mode is the suitable measure to fill in the missing entry. Mode represents the highest frequency occurrence of the data attribute value.

3.3. Feature Selection using Cuckoo Search Algorithm. Our proposed lung cancer diagnostic system uses a bio-inspired algorithm namely Cuckoo Search Algorithm(CS). It mimics the reproduction strategy of cuckoo bird. Cuckoo bird lays eggs in another bird's nest for their reproduction. The host bird once found it is an alien egg either it throw away the alien egg or abandon the nest built for a new one for reproduction. If it does not notice the egg ,it hatches the alien egg .The cuckoo bird imitates the host bird and get more food for their survival. To overcome these issues, the CS algorithm is used in the proposed work and it has advantages as follows,

1. It has fewer parameters to find the optimal feature subset.
2. It guarantees global convergence.
3. It maintains a balanced combination of a random walk and a global explorative random walk controlled by switching parameter Pa.

These characteristics inspire us to use the algorithm. It supersedes the Genetic algorithm and Particle Swarm Optimization algorithm.

1. Each cuckoo bird egg represents a feature. Hence the first step is to randomly generate an initial population of n at the position $X = \{x_1^0, x_2^0, \ldots, x_n^0\}$ and then assess their objective values to find the current global best g_t^0. Here all the features for detecting lung cancer are considered in total that represents the initial population.

2. The best fitted eggs are responsible for next generation. The fitness of an egg or solution is determined by its objective value. The optimal solution with the lowest objective values is subjected to the next generation. Therefore update the new solutions/positions by,

$$x_i^{(t+1)} = x_i^{(t)} + \alpha \otimes L(\lambda) \tag{3.1}$$

3. The Pa=[0,1] is the probability that the host bird is noticing the alien bird's egg. In this way, the irrelevant and redundant features in the lung cancer diagnostic system are eradicated.
4. Here the stopping criterion is finding the best global solution otherwise it returns to the step2. host bird finds the alien bird's egg represents the worst solution which is far away from the optimal solution[24].

The local random walk is defined by

$$\boldsymbol{x}_i^{t+1} = \boldsymbol{x}_i^t + \alpha s \otimes H\left(p_a - \epsilon\right) \otimes \left(x_j^t - \boldsymbol{x}_k^t\right) \tag{3.2}$$

where x_j^t and x_k^t are two different solutions selected randomly by random permutation H(u) is a Heaviside function,α is the random number drawn from uniform distribution and s is step size and $\otimes$ is the entry-wise product.

The global random walk is described by using Levy Flights,

$$x_i^{t+1} = x_i^t + \alpha L(s, \lambda) \tag{3.3}$$

$$L(s, \lambda) \sim \frac{\lambda \Gamma(\lambda) \sin(\pi\lambda/2)}{\pi} \frac{1}{s^{1+\lambda}}, (s \gg 0) \tag{3.4}$$

The objective function is given by,

$$f\,(x) = \ \alpha * error + \beta * (\frac{No.\ of\ selectedfeatures}{Maximum\ number\ of\ features}) \tag{3.5}$$

where α=0.99, β=1-α=0.01

The error is calculated by considering the difference between the estimated value by the classifier model and the actual value of the observed data.

The fitness value of the cuckoo search algorithm is given by,

$$f_{x},_{y} = f(\min)_y + \gamma_{xi}(f(\max)_y - f(\min)_y) \tag{3.6}$$

We employ a greedy selection algorithm to find the right combination of optimal feature subsets for each iteration which maximizes the performance of detecting lung cancer disease.

```
Algorithm 1. Cuckoo Search Algorithm
Initial_population; initialize the population with Nst host nests;
Evaluate the Initial population;
Set the max iter;
iter=0;
while(iter<max)
C=select random cuckoo; //select a random cuckoo
C*=levy flights©; //apply levy flights on C to generate new solution
Fc*=Evaluatefitness(C*); //compute the fitness of C*
N=random nest (); //select a nest at random among Nst
$F_N$=Evaluatefitness©;
if ($F_{C*}$ >$F_N$)
N=C*; // Replace N by new generated solution C*
end if
Abandon the worst pa nest;//where pa is a fraction of nests
```

```
Construct new nests using levy_flights
Save the best nests;
Find the current best nests
**iter=iter+1**
**end while**
```

3.4. Classification. The machine learning algorithm consists of two phases. They are the training phase and testing phase. During the training phase, it replicates the human learning system. It learns data with associated class labels which means it learns by examples. If any unseen data are provided to the lung cancer diagnostic system during the testing phase, it predicts the class label by interpreting the hidden pattern of the learned data. In the testing phase, the machine learning algorithm evaluates the model building that is generated during the training phase.

3.4.1. Logistic Regression Algorithm. Logistic regression is widely used for both regression and classification. It uses the sigmoid function to classify data instances. The hypothesis function is given by,

$$Z = WX + B \tag{3.7}$$

$$h\Theta(x) = \text{sigmoid}(Z) \tag{3.8}$$

$$\text{Sigmoid}(Z) = 1/(1 + e^{-z}) \tag{3.9}$$

If the Z value goes ∞, then Y(Predicted) =1. Then the data point belongs to class 1.
If the Z value goes -∞, then Y(Predicted) =0. Then the data point belongs to class 0.

3.4.2. Support Vector Machine. It is widely used for binary and multi-class classification algorithms. It uses a decision line to separate two classes. It uses hyperplane for more than two class problems. Finding the optimal hyperplane is a challenging task[18]. The optimal hyperplane is the one that maximally separates the data points from its margin. The equation of the hyperplane is given by,

$$Y = wixi + b \tag{3.10}$$

Where Y is the output variable which is of categorical type, b is the bias parameter, xi is the input vector and wi is the weight vector.

If Y<1, then the data point belongs to the negative class.

If Y>=1, then the data point belongs to the positive class[10]. It also has capability that it automatically eliminates the noisy features in order to obtain optimal feature subsets [16].

3.4.3. Random Forest Classifier. It is widely used for both classification and regression problems. It combines several decision trees on different samples and takes the majority to predict the class of unknown data instances. It serves as ensemble method that facilitates for deeper understanding of data [17]. It is faster as it is working only on the subset of the features in this model. The number of decision trees constructed is between 64-128 trees as it balances the ROC-AUC and processing time. The advantage of random forest is that it is good at handling high-dimensional data. Its training speed is faster. It is robust to outliers and non-linear data. It can handle unbalanced data. The drawbacks of random forests are not interpretable. It consumes considerable memory for large datasets. It can tend to overfit so need to tune the hyperparameters.

4. Experimental Setup. All computations are performed on Intel (R) Core (TM) i5-8250U CPU @1.80GHz with 64bit Windows 10 is the operating system. All the experiments are performed using the Python software package. The proposed lung cancer diagnostic system uses two datasets dataset collected of which one is from the Kaggle repository and another dataset collected based on exposure to pesticides causes lung cancer. The datasets are subjected to stratified 10 kfold cross-validations to overcome biasing.

4.1. Performance measures. The performance of the classifier model is assessed by using a confusion matrix. It comprises True Positive, True Negative, False Positive, and False Negative[14, 15, 21].

True Positive represents the number of instances having lung cancer and it is also correctly predicted by the classifier model.

True Negative represents the number of instances having no lung cancer and it is also correctly predicted by the classifier model.

False Positive represents the number of instances having lung cancer but it is predicted as normal by the classifier model

False Negative represents the number of instances having no lung cancer but is predicted as a patient by the classifier model.

Accuracy. It is defined as the ability of the classifier that makes correct predictions about its classes out of the total number of data instances.

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \tag{4.1}$$

Precision. Precision is defined as the ability of the classifier that makes the correct prediction of lung cancer data instances from the total number of predictions. It is also known as Positive Predictive Value(PPV).

$$Precision = TP/(TP + FP) \tag{4.2}$$

Recall. Recall is defined as the ability of the classifier that makes the correct prediction of data instances having lung cancer out of correctly identified lung cancer data instances.

$$Recall = TP/(TP + FN) \tag{4.3}$$

Specificity. Specificity is defined as ability of the classifier that make correct prediction of negative samples .

$$Specificity = TN/(TN + FP) \tag{4.4}$$

F1-Score. It is the weighted average of precision and recall.

$$F1 - score = (2 * Precision * Recall)/(Precision + Recall) \tag{4.5}$$

4.2. Result. Our proposed lung cancer diagnostic system has experimented with two datasets from an online data repository. The Descriptive statistics of sampled datasets 1 and the are as in Table 4.1.

Table 4.2 represents the correlation matrix of all the features in the data and their relationship using the Pearson correlation coefficient. The correlation matrix represents the strength of the relationship that exists between features in the data. The value +1 represents features that are perfectly positively correlated,-1 represents the features that are perfectly negatively correlated and 0 represents uncorrelated features.

After applying the Cuckoo Search feature selection algorithm, the optimal feature subset is generated and the selected features are as follows for the sampled dataset1.For the sampled dataset1, among the 12 features, only 7 features such as age, anxiety, yellow_fingers, attention disorder, Born_an_Even_Day, Fatigue, and Coughing are considered by the CS algorithm that is optimally discriminate the data instances into their categories.

The Cuckoo Search algorithm is a metaheuristic algorithm in which the term heuristics represents the parameter settings are completely trial and error based. Whereas the term meta that contributes optimal solution beyond higher level. There are two components associated with metaheuristic algorithm. They are local search and global search. The global search is good at explore search space at global scale. The local search use information that is good at search in local region [23]. The proposed work uses parameter set-up for Cuckoo search algorithm of sampled dataset1 which is given below.

Table 4.1: The Descriptive statistics of sampled dataset 1

Column 1	**Smoking**	**Yellow Fingers**	**Anxiety**	**Peer Pressure**	**Genetics**	**Attention Disorder**	**Born an Even Day**	**Car Accident**	**Fatigue**	**Allergy**	**Coughing**	**Lung cancer**
Mean	0.753	0.782	0.6305	0.3415	0.1395	0.3225	0.4895	0.72	0.737	0.343	0.7005	0.722
Standard Error	0.01	0.009	0.0108	0.0106	0.0077	0.0105	0.011 18	0.01	0.0098	0.011	0.0102	0.01
Median	1	1	1	0	0	0	0	1	1	0	1	1
Mode	1	1	1	0	0	0	0	1	1	0	1	1
Standard Deviation	0.432	0.413	0.4828	0.4743	0.3466	0.4676	0.50001	0.45	0.4404	0.475	0.4582	0.448
Sample Variance	0.186	0.171	0.2331	0.225	0.1201	0.2186	0.25001	0.2	0.1939	0.225	0.2099	0.201
Kurtosis	-0.63	-0.131	-1.709	-1.554	2.3394	-1.424	-2.0002	-1	-0.84	-1.56	-1.234	-1.023
Skewness	-1.17	-1.367	-0.541	0.669	2.0826	0.76	0.04204	-1	-1.077	0.662	-0.876	-0.989
Range	1	1	1	1	1	1	1	1	1	1	1	1
Minimum	0	0	0	0	0	0	0	0	0	0	0	0
Maximum	1	1	1	1	1	1	1	1	1	1	1	1
Sum	1505	1564	1261	683	279	645	979	1446	1474	686	1401	1443
Count	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000

Table 4.2: Pearson Correlation Coefficient considering all features of Sample dataset1

0	**Smoking**	**Yellow Fingers**	**Anxiety**	**Peer Pressure**	**Genetics**	**Attention Disorder**	**Born an Even Day**	**Car Accident**	**Fatigue**	**Allergy**	**Coughing**	**Lung cancer**
Smoking	1	0	0	0	0	0	0	0	0	0	0	0
Yellow Fingers	0.775	1	0	0	0	0	0	0	0	0	0	0
Anxiety	0.401	0.308	1	0	0	0	0	0	0	0	0	0
Peer Pressure	0.149	0.115	0.003	1	0	0	0	0	0	0	0	0
Genetics	0.01	-0.004	0.0063	0.0205	1	0	0	0	0	0	0	0
Attention Disorder	0.004	0.007	-0.015	0.0152	0.2687	1	0	0	0	0	0	0
Born an Even Day	-0.02	-0.006	-0.038	-0.007	-0.0219	-0.021	1	0	0	0	0	0
Car Accident	0.051	0.049	0.0308	0.024	0.146	0.3028	-0.0309	1	0	0	0	0
Fatigue	0.163	0.125	0.0509	0.0279	0.0996	0.0331	-0.0285	0.46	1	0	0	0
Allergy	0.036	0.047	0.0185	-0.005	-0.0082	0.0198	-0.0333	0.04	0.0943	1	0	0
Coughing	0.262	0.207	0.1372	0.0496	0.1372	0.0541	-0.0279	0.21	0.4598	0.307	1	0
Lung cancer	0.491	0.377	0.1899	0.057	0.2276	0.0683	-0.0119	0.17	0.3687	-0.03	0.5167	1

Tables 4.3-4.9 represent with and without the Cuckoo Search feature selection algorithm for dataset1. For the sampled dataset1, among the 12 features, only 7 features such as age, anxiety, yellow fingers, attention disorder, Born an Even Day, Fatigue, and Coughing are considered by the CS algorithm that supports optimally discriminating the data instances into their categories.

Table 4.3: Parameter setup

Parameters	**Values**
Alpha	0.01
Beta	2
No. of iterations	100
MSE	0.11
Pa	0.25
Number of Features N	7 out of 12

Table 4.4: Performance of LR Classifier without CS algorithm

	precision	**recall**	**f1-score**	**support**
0	0.77	0.70	0.73	166
1	0.89	0.92	0.90	434
accuracy			0.86	600
macro avg	0.83	0.81	0.82	600
weighted avg	0.86	0.86	0.86	600

Table 4.5: Performance of Linear SVC Classifier without CS algorithm

	precision	**recall**	**f1-score**	**support**
0	0.77	0.69	0.73	166
1	0.88	0.92	0.90	434
accuracy			0.86	600
macro avg	0.83	0.80	0.81	600
weighted avg	0.85	0.86	0.85	600

Table 4.6: Performance of RF Classifier without CS algorithm

	precision	**recall**	**f1-score**	**support**
0	0.71	0.72	0.71	166
1	0.89	0.89	0.89	434
accuracy			0.84	600
macro avg	0.80	0.80	0.80	600
weighted avg	0.84	0.84	0.84	600

Table 4.7: Performance of LR Classifier with CS algorithm

	precision	**recall**	**f1-score**	**support**
0	1.00	1.00	1.00	45
1	1.00	1.00	1.00	23
accuracy			1.00	68
macro avg	1.00	1.00	1.00	68
weighted avg	1.00	1.00	1.00	68

Table 4.8: Performance of Linear SVC Classifier with CS algorithm

	precision	**recall**	**f1-score**	**support**
0	0.84	0.78	0.81	55
1	0.92	0.94	0.93	145
accuracy			0.90	200
macro avg	0.88	0.86	0.87	200
weighted avg	0.90	0.90	0.90	200

Table 4.9: Performance of RF Classifier with CS algorithm

	precision	**recall**	**f1-score**	**support**
0	0.90	0.88	0.92	55
1	0.83	0.98	0.90	145
accuracy			0.91	200
macro avg	0.86	0.93	0.90	200
weighted avg	0.86	0.90	0.90	200

Tables 4.10-4.12 and 4.13-4.15 represent with and without the Cuckoo Search feature selection algorithm for dataset2.

The diagram from Fig. 4.1 shows the number of iterations versus fitness scores by using LR Classifier with CS algorithms. The diagram from Fig. 4.2 shows the number of iterations versus fitness scores by using Linear SVC Classifier. The diagram from Fig. 4.3 shows the number of iterations versus fitness scores by using RF Classifier.

Fig. 4.4 represents with and without the Cuckoo Search feature selection algorithm for dataset 2. The diagram from Fig. 4.5 shows the number of iterations versus fitness scores by using LR Classifier. The diagram from Fig. 4.6 shows the number of iterations versus fitness scores by using Linear SVC Classifier.

Table 4.10: Performance of LR Classifier without CS algorithm of Dataset2

	precision	recall	f1-score	support
0	0.97	0.96	0.97	135
1	0.97	0.99	0.99	69
accuracy			0.98	204
macro avg	0.97	0.96	0.98	204
weighted avg	0.97	0.96	0.98	204

Table 4.11: Performance of Linear SVC Classifier without CS algorithm

	precision	recall	f1-score	support
0	0.99	0.99	0.99	135
1	0.99	0.98	0.98	69
accuracy			0.99	204
macro avg	0.99	0.98	0.98	204
weighted avg	0.99	0.99	0.98	204

Table 4.12: Performance of RF Classifier without CS algorithm

	precision	recall	f1-score	support
0	0.83	0.80	0.84	55
1	0.93	0.92	0.91	145
accuracy			0.91	200
macro avg	0.88	0.86	0.87	200
weighted avg	0.91	0.91	0.87	200

Table 4.13: Performance of LR Classifier with CS algorithm of dataset2

	precision	recall	f1-score	support
0	1.00	1.00	1.00	135
1	1.00	1.00	1.00	69
accuracy			1.00	204
macro avg	1.00	1.00	1.00	204
weighted avg	1.00	1.00	1.00	204

Table 4.14: Performance of Linear SVC Classifier with CS algorithm of dataset2

	precision	recall	f1-score	support
0	1.00	1.00	1.00	135
1	1.00	1.00	1.00	69
accuracy			1.00	204
macro avg	1.00	1.00	1.00	204
weighted avg	1.00	1.00	1.00	204

Table 4.15: Performance of RF Classifier with CS algorithm of dataset2

	precision	recall	f1-score	support
0	0.97	0.99	0.98	135
1	0.97	0.94	0.96	69
accuracy			0.97	204
macro avg	0.97	0.96	0.97	204
weighted avg	0.97	0.97	0.97	204

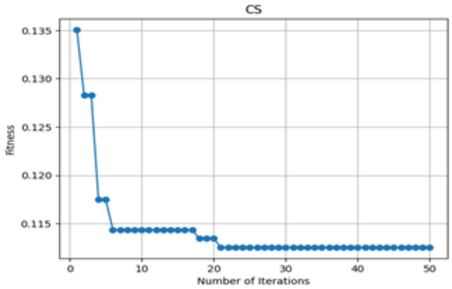

Fig. 4.1: Number of iteration Vs Fitness score of LR Classifier with CS algorithm of dataset1

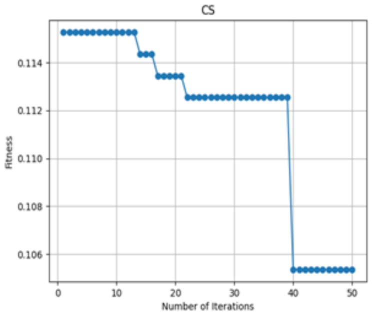

Fig. 4.2: Number of iteration Vs Fitness score of Linear SVC classifier with CS algorithm of dataset1

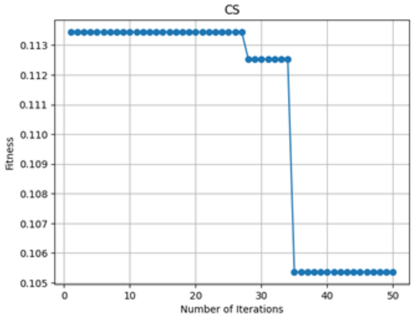

Fig. 4.3: Number of iteration Vs Fitness score of RF Classifier with CS algorithm of dataset1

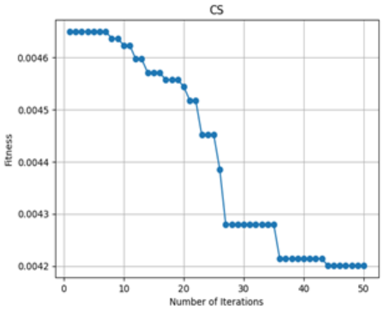

Fig. 4.4: Number of iterations Vs Fitness score of LR Classifier with CS of dataset 2

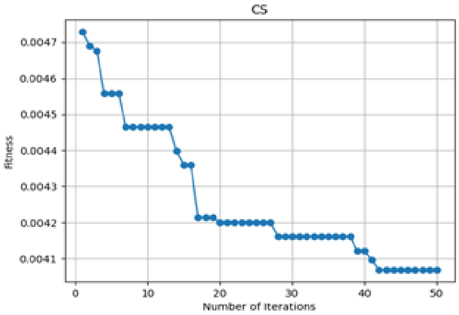

Fig. 4.5: Number of iterations Vs Fitness score of Linear SVC Classifier with CS algorithm of dataset2

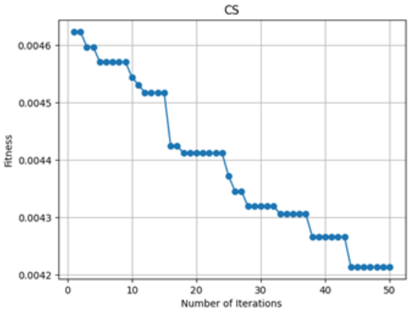

Fig. 4.6: Number of iterations Vs Fitness score of RF Classifier with CS algorithm of dataset2

Table 4.16: Performance of LR Classifier with GA algorithm

	precision	recall	f1-score	support
0	0.92	0.94	0.92	135
1	0.94	0.96	0.95	69
accuracy			0.93	204
macro avg	0.92	0.94	0.92	204
weighted avg	0.92	0.94	0.94	204
Specificity			0.93	

Table 4.17: Performance of Linear SVC Classifier with GA algorithm

	precision	recall	f1-score	support
0	0.93	0.95	0.94	135
1	0.92	0.94	0.93	69
accuracy			0.92	204
macro avg	0.93	0.95	0.94	204
weighted avg	0.93	0.95	0.94	204
Specificity			0.93	

Table 4.18: Performance of RF Classifier with GA algorithm

	precision	recall	f1-score	support
0	0.91	0.93	0.92	135
1	0.91	0.93	0.92	69
accuracy			0.93	204
macro avg	0.91	0.93	0.92	204
weighted avg	0.91	0.93	0.92	204
Specificity			0.91	

5. Conclusion. Lung cancer is the second most common cancer present in both men and women. It could be diagnosed at its advanced stage only by doctors. If it could be diagnosed at its early stage, the survival rate could be improved. To facilitate the process, our proposed lung cancer diagnosis system has experimented with two survey datasets and provides better results in terms of accuracy, precision, recall, and f1-score. There are two machine learning models namely LR classifier and Linear SVC classifier and one ensemble learning model namely random forest tree are used. The research work has been conducted with and without a cuckoo search algorithm as a feature selection technique to select the optimal feature subset for enhancing performance in lung cancer detection. It is observed that with the cuckoo search algorithm, dataset 1 achieves an accuracy of 100%, precision of 100%, recall of 100%, and F1-score of 100% by LR Classifier. The Linear SVC classifier achieves an accuracy of 90%, a precision of 88%, a recall of 86%, and an F1-score of 87%.The Random forest Classifier achieves an accuracy of precision of 86%, recall of 93%, F1-score of 90%, and accuracy of 91%. For dataset 2, both the LR classifier and Linear SVC classifier outperform with an accuracy of 100%, precision of 100%, recall of 100%, and F1-score of 100%. Whereas Random Forest provides accuracy of 97%, precision of 97%, recall of 96%, and F1-score of 97%.

REFERENCES

[1] Elias Dritsas * and Maria Trigka ,Lung Cancer Risk Prediction with Machine Learning Models, Big Data Cogn. Comput. 2022. https://doi.org/10.3390/bdcc6040139

[2] Venkatesh, S., & Raamesh, L. (2022). Predicting Lung Cancer Survivability: a Machine Learning Ensemble Method on Seer Data. doi:10.21203/rs.3.rs-1490914/v1

[3] Vikas et al. Lung Cancer Detection Using Chi-Square Feature Selection and Support Vector Machine Algorithm. (2021). International Journal Of Advanced Trends In Computer Science And Engineering, 10(3), 2050-2060 doi:10.30534/ijatcse/2021/801032021

[4] M. I. Faisal, S. Bashir, Z. S. Khan and F. Hassan Khan, ”An Evaluation of Machine Learning Classifiers and Ensembles for Early Stage Prediction of Lung Cancer,” 2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST),2018, pp. 1-4, doi: 10.1109/ICEEST.2018.8643311

[5] Puneet and A. Chauhan, ”Detection of Lung Cancer using Machine Learning Techniques Based on Routine Blood Indices,” 2020 IEEE International Conference for Innovation in Technology (INOCON), 2020, pp. 1-6, doi: 10.1109/INOCON50539.2020.9298407

[6] Alsinglawi, B., Alshari, O., Alorjani, M. et al. An explainable machine learning framework for lung cancer hospital length of stay prediction. Sci Rep 12, 607 (2022). https://doi.org/10.1038/s41598-021-04608-7

[7] A. Safiyari and R. Javidan, ”Predicting lung cancer survivability using ensemble learning methods,” 2017 Intelligent Systems Conference (IntelliSys), 2017, pp. 684-688, doi: 10.1109/IntelliSys.2017.8324368.

[8] 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

[9] Reem Nooreldeen and Horacio Bach Current and Future Development in Lung Cancer Diagnosis, International Journal of Molecular Science 2021, https://doi.org/10.3390/ijms22168661

[10] SurenMakajua,P.W.C.Prasad, AbeerAlsadoon a, A. K. Singh , A. Elchouemic,_Lung Cancer Detection using CT Scan Images,Elsevier Procedia Computer Science 2018, pp.107-114

[11] Hwa-Yen Chiu 1,2,3,4 , Heng-Sheng Chao 1,5,* and Yuh-Min Chen,” Application of Artificial Intelligence in Lung Cancer,2022, .https://doi.org/10.3390/cancers14061370

[12] Elinor Nemlander , Andreas RosenbladID1,3,4, Eliya Abedi, Simon Ekman5,Jan Hasselstro m, Lars E. ErikssonID, Axel C. Carlsson, Lung cancer prediction using machine learning on data from a symptom e-questionnaire for never smokers, formers smokers and current smokers, PLOS ONEhttps://doi.org/10.1371/journal.pone.0276703 , 2022

[13] Ibrahim M. Nasser, Samy S. Abu-Naser, Lung Cancer Detection Using Artificial Neural Network , International Journal of Engineering and Information Systems, Vol. 3 Issue 3, March 2019, Pages: 17-23

[14] E. Punarselvam, Mohamed Yacin Sikkandar, ohsen Bakouri, N. B. Prakash T.Jayasankar and S.Sudhakar,Different Loading Condition and Angle Measurement of Human Lumbar SpineMRI Image Using ANSYS”,Springer Journal of Ambient Intelligence and Humanized Computing (2020), ISSN: 1868-5137 , (Print) 1868-5145 (Online) https://doi.org/10.1007/s12652-020-01939-7.

[15] Ankur Gupta, Rahul Kumar 1, Harkirat Singh Arora, And Balasubramanian Raman 1, (Member,IEEE),”MIFH: A Machine Intelligence Framework For Heart Disease Diagnosis”, IEEEAccess,DigitalObjectIdentifier10.1109/ACCESS.2019.2962755vol.8, pp. 14659-14674, January 2020.

[16] Liaqat Ali, Awais Niamat, Javed Ali Khan, Noorbakhsh Amiri Golilarz, Xiong Xingzhong, AdeebNoor,RedhwanNour,and Syed Ahmad Chan Bukhari “ An Optimized Stacked Support Vector Machines Based Expert System For The Effective Prediction Of Heart Failure” L. Ali et al.: IEEE Access Digital Object Identifier 10.1109/ACCESS.2019.2909969, vol.7,pp.54007- 54013 March 2019

[17] Yogita Solanki,Sanjiv Sharma,” A Survey on Risk Assessments of Heart Attack” Using Data Mining Approaches I.J.Information Engineering and Electronic Business, 2019, DOI: 10.5815/ijieeb.2019.04.05 vol. 4, pp.43-51 July 2019.

[18] Roger Alan Steina , Patrıcia A. Jaquesa , Joao Francisco Valiatib,” An Analysis of Hierarchical Text Classification Using Word Embeddings” September 2018, https://doi.org/10.1016/j.ins.2018.09.001.

[19] Book,Tom M Mitchell, ”Machine learning” *(McGraw Hill Science*), ISBN 0070428077 2013.

[20] L.Maria Jenifer, T.Sathya,B. Sathiyabhama , “GSA Based Classification of Lung Nodules in CT Images” Proceedings of the International Conference on Intelligent Computing Systems” March 2018.

[21] T.Sathiya, R.Reenadevi, B.Sathiyabhama, “Lung nodule classification in CT images using Grey Wolf Optimization algorithm”, Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 6, 2021, Pages. 1495-1511,May 2021.

[22] Muntasir Mamun, Afia Farjana, Miraz Al Mamun, Md Salim Ahammed, Lung cancer prediction model using ensemble learning techniques and a systematic review analysis, IEEE Access 2022.

[23] Book Xin-She Yang, “Nature Inspired Optimization Algorithm, Elsevier 2014.

[24] Suchetha N K, Anupama Nikhil, Hrudya P, Comparing the Wrapper Feature Selection Evaluators on Twitter Sentiment Classification, IEEE Access 2019, Second International Conference on Computational Intelligence in Data Science (ICCIDS-2019).

Edited by: Polinpapilinho Katina
Special issue on: Scalable Dew Computing for future generation IoT systems
Received: Feb 2, 2024
Accepted: May 15, 2024