Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 21 No. 8 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v21i08.55121

PAPER

Comparative Analysis of Hybrid and Ensemble Learning in Lung Cancer Diagnosis

Manmath Nath Das[1], Niranjan Panda[1](✉), Rasmita Rautray[1], Jyotsnarani Tripathy[2]

[1]Department of Computer Science and Engineering, Siksha 'O' Anusandhan University, Bhubaneswar, Odisha, India

[2]Dept of CSE-AIML & IoT, Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology (VNRVJIET), Hyderabad, Telengana, India

niranjanpanda@soa.ac.in

ABSTRACT

Lung cancer remains one of the leading causes of cancer-related deaths globally, primarily due to delayed diagnosis. Early and accurate detection is critical for improving patient survival rates. However, microarray data used in cancer diagnosis pose a significant challenge due to the curse of dimensionality, where a small number of samples are associated with a large number of features, potentially reducing the accuracy (ACC) of prediction models. To address this, we propose a hybrid machine learning (ML) approach that integrates correlation-based feature selection (CFS) and the elephant search algorithm (ESA) for effective feature selection and optimization. Additionally, we introduce an ensemble deep learning (EDL) strategy, combining a deep neural network (DNN) with a bagging classifier and ensemble techniques such as weighted averaging (WA), soft voting (SV), and hard voting. Experiments conducted on a microarray dataset from the national center for biotechnology information (NCBI) evaluated performance using ten metrics. The hybrid approach (CFS+ESA+MLP) achieved an ACC of 97.18%, while the ensemble DL model with hard voting (HV) attained a superior ACC of 98.87%. These results demonstrate the effectiveness of the proposed methodologies in enhancing early lung cancer diagnosis.

KEYWORDS

ensemble learning (EL), machine learning (ML), deep learning (DL), lung cancer, classification and prediction

1 INTRODUCTION

Cancer is the abnormal proliferation of cells in the body. Every individual is born with the potential for cancer. Detection is challenging for one of us unless the individual is ill or exhibits symptoms of a cold. When a group of cells is impacted, growth may become unregulated. Lung cancer in humans occurs owing to many circumstances. Factors include chronic inflammation, tobacco smoke and dust exposure, radioactive chemicals, age, sex, race, and inheritance. However, one of the cited criteria may be uncontrollable. Lung cancer refers to the proliferation of malignant cells in the

Das, M.N., Panda, N., Rautray, R., Tripathy, J. (2025). Comparative Analysis of Hybrid and Ensemble Learning in Lung Cancer Diagnosis. *International Journal of Online and Biomedical Engineering (iJOE)*, 21(8), pp. 41–55. https://doi.org/10.3991/ijoe.v21i08.55121

Article submitted 2025-02-24. Revision uploaded 2025-04-29. Final acceptance 2025-04-29.

lungs, which may subsequently metastasize to lymph nodes and other organs in the body [1, 2]. Machine learning (ML) is a subset of artificial intelligence (AI) that utilizes a compilation of data and algorithms to enhance human learning processes, hence increasing accuracy (ACC) [3]. ML is an essential element of the expanding domain of data science, utilizing statistical approaches and trained algorithms to facilitate categorization, predictions, and a profound comprehension of data mining initiatives. ML has been extensively employed in the medical domain for studies aimed at detecting cardiac problems. The diverse applications of ML algorithms facilitated the identification process, resulting in a higher success rate in treatment and an improved patient survival rate. The implementation of ML improves the learning capacity of the datasets utilized during training. Various methodologies exist for the development of learning systems. The three types of ML are supervised, unsupervised, and reinforcement learning. The learning process for regression is unsupervised, whereas the learning process for classification is supervised. Moreover, deep learning (DL), an enhancement of ML, is important in early lung cancer diagnosis [4, 5].

1.1 Objectives

This study proposes a two-stage technique evaluation. A hybrid ML strategy is initially presented, including CFS for feature selection, ESA for optimization, and seven traditional techniques: logistic regression (LR), naïve bayes (NB), K-nearest neighbor (KNN), support vector machine (SVM), multi-layer perceptron (MLP), decision tree (DT), and random forest. A composite DL method incorporating a deep neural network (DNN) and a bagging classifier is implemented in the second stage, with weighted averaging (WA), soft voting (SV), and hard voting (HV) techniques. Experiments are conducted on a national center for biotechnology information (NCBI) microarray dataset utilizing 10 assessment measures. The contributions of this work are delineated as follows:

- Build a hybrid approach that provides us with enhanced accuracy.
- The proposed work was designed in such a way that it can easily detect lung cancer at an early stage.
- The proposed model would be available at a low cost.
- People should be aware of lung cancer so they don't hesitate to visit the doctor when facing some symptoms.

1.2 Paper structure

This proposed work is organized as follows: Section 2 outlines the current research projects in this field, along with a summary table. Section 3 outlines the utilization of the proposed dataset in this study. Section 4 pertains to the investigation's architectural dimension, encompassing the proposed project's design, flowchart, block diagrams, and operational idea. Section 5 outlines the analytical framework of the proposed inquiry, contrasted with the relevant findings assessed in this study. Section 6 finishes the examination with a viable extension of the proposed endeavor.

2 RELATED WORK

Cui et al. [6] utilized ML on PAH datasets to identify, via microarray, that low-expression genes may collectively affect the PAH sickness model with 99% ACC. Lai et al. [7] created a DNN-based NSCLC prognosis model employing E-MTAB-923

datasets with 75.44% ACC and 0.8163 area beneath the curve (AUC). Pawar [8] created a web tool that uses ML on Gene expression omnibus (GEO) microarray data to diagnose cancer types with 94.99% ACC. To construct a lung cancer diagnosis and relapse prediction model with 97.9% ACC, 97.2% sensitivity (SEN), 100% specificity, 100% precision (PRE), and 0.98 AUC, Abdu-Aljabar and Awad [9] evaluated ML techniques using the NCBI site microarray and NGS dataset.

Tabares-Soto et al. [10] classified cancer kinds using a microarray gene expression data model and ML and DL methods, including FNNs and CNNs, on 11_tumor datasets. 98% F1-score (FS), 100% PRE, 100% recall, and 90.6% ACC. Basavegowda and Dangnew's Deep Feed Forward model for microarray cancer data classification using PCA [11] has 100% ACC, PRE, recall, F-measure, and 1 AUC. Thallam et al. [12] predicted early-stage lung cancer with 99.5% ACC on the microarray dataset using ML approaches such as SVM, RF, KNN, ANN, and a voting classifier. Takahashi et al. [13] integrated survival subtypes in lung cancer models with 99.0% AUC using ML approaches such as XGBoost, LightGBM, and K-Means clustering using multi-omics LUAD and lung LUSC datasets from TCGA.

Dagnew and Shekar [14] classified a microarray cancer data tree-based features model with 100% classification ACC, PRE, recall, F-score, and AUC as 1 using ML methods such as LR, MLP, SVM, RF, DT, KNN, and voting ensemble learning (EL) on Shenzhen University, GEMS, and Elvira datasets. Doppalapudi et al. [5] used DL approaches such as ANN, CNN, and RNN on SEER datasets to predict lung cancer survival periods with 71.18% ACC, 77.70% PRE, and 83.36% recall, 79.96% F-score, and 0.92 AUC. Rezaee et al. [15] classified lymphoma, leukaemia and prostate microarray datasets with 97.51%, 99.6%, and 96.34% ACC using DNN DL. Khoirunnisa et al. [16] identified lung cancer with 91% ACC using the CRNN methodology based on DL techniques, CNN, and RNN on the lung cancer microarray dataset. An ML-based lung cancer diagnosis model by Rani and Prasad [17] used RF, AdaBoost, and Kernel PCA (KPCA) on GSE4115 from the GEO database with 81% ACC, 81.2% PRE, 78.9% recall, and 77.7% FS. Table 1 lists cutting-edge works.

Table 1. Summary of the considered state-of-the-art works

Ref	Techniques Employed	Dataset(s) Employed
Cui et al. [6]	LDA, ANN, SVM	PAH Dataset
Lai et al. [7]	DNN, RF	E-MTAB-923 Dataset
Pawar [8]	SVM	Microarray data from GEO
Abdu-Aljabar and Awad [9]	XGBoost, SVM, gcForest	Microarray and NGS datasets
Tabares-Soto et al. [10]	KNN, SVM, LR, LDA, NB, RF, DT, MLP, K-means, FNNs, CNNs	11_tumor database
Basavegowda and Dangnew [11]	Deep Feed Forward based on PCA	Microarray cancer dataset
Thallam et al. [12]	SVM, KNN, RF, ANN and Voting	Microarray dataset
Takahashi et al. [13]	XGBoost, LightGBM, K-Means	LUAD and LUSC from TCGA
Dagnew and Shekar [14]	LR, MLP, SVM, RF, DT, KNN and Voting	Microarray cancer Datasets from Shenzhen University data repositories, GEMS, and Elvira
Doppalapudi et al. [5]	ANN, CNN, and RNN	SEER Dataset
Rezaee et al. [15]	DNN	Lymphoma, leukemia, and prostate cancer Microarray Datasets
Khoirunnisa et al. [16]	CNN, RNN, and CRNN	Lung cancer microarray dataset

3 MATERIALS AND METHODS

This study provides a comprehensive overview of the methodologies and resources utilized. It details the dataset characteristics, pre-processing techniques, and the implementation of various ML, DL, and EL approaches. For clarity and reproducibility, each component is thoroughly explained in the respective sub-sections.

3.1 Dataset employed and pre-processing

The microarray lung cancer dataset, i.e., GSE30219, is used to test the proposed hybrid and ensemble ML methods for tumor diagnostic prediction in lung cancer. This dataset is available through the NCBI data repository. This microarray data set was collected as part of the French Ligue Nationale Contre le Cancer's Cartes d'Identite des Tumeurs (CIT) program [18], which analyzed 293 lung tumor samples and 14 non-tumoral lung samples. Table 2 briefly summarizes the microarray lung cancer dataset used in this study. The first phase of every ML algorithm is the pre-processing of data. Data collection procedures are sometimes not strictly regulated, resulting in inaccurate, incomplete, or unreliable information. Therefore, it must have pre-processed the datasets by deleting extraneous, inaccurate, or undesired information. There is a significant lack of complete clinical data in lung cancer databases. These records are deemed irrelevant and should be destroyed, as they lead to inappropriate learning outcomes. The choice category in a classification problem must be a number. However, the format is still nominal in some datasets. The data must be converted from nominal to numerical form. Cancer and normal are the nominal classifications used in this analysis. This calls for a conversion to the numerical forms 0 and 1.

Table 2. Dataset description

Dataset		Dimension		Class Distribution	
Name	Type	No. of Patients	No. of Features	Cancer	Normal
GSE30219	Microarray	307	54675	293	14

3.2 Feature extraction technique: correlation-based feature selection

The correlation between each feature and the target variable or between features is measured using the correlation-based feature selection (CFS) method used in ML and statistics [19]. The objective is to maintain the traits that are highly correlated with the dependent variable and eliminate the ones that are irrelevant or unhelpful. The most common metric used for measuring correlation is the Pearson correlation coefficient (Υ), which ranges from –1 to 1. Equation 1 shows the Pearson coefficient calculation. Let's say there are n features, including the target variable. Equation 2 shows the boundary values and the relation between the two features depending on the Υ.

$$\Upsilon(f_i, c) = \frac{\Sigma(f_i - \bar{f})(f_j - \bar{c})}{\sqrt{(f_i - \bar{f})^2 (f_j - \bar{c})^2}} \quad (1)$$

$$\Upsilon(f_i, c) = \begin{cases} 1, & \textit{Perfect positive linear correlation} \\ 0, & \textit{No correlation} \\ -1, & \textit{Perfect negative linear correlation} \end{cases} \quad (2)$$

Where f_i, f_j are the data points, and c is the target variable. $\overline{f}$, and $\overline{c}$ are the mean of the feature f_i *and target variable* c.

3.3 Optimization technique: elephant search algorithm

The elephant search algorithm (ESA) mimics elephant birth and death by following natural life cycles [20]. Regardless of their current placements, male and female elephants age and die at the same pace in the original plan. Mammalian lifespans can be optimized by natural selection, in which only the fittest survive, to reward successful search agents with prime locations over those that fail. ESA should offer healthy elephants with a history of high performance a longer lifespan than bad elephants, who will be taken out soon. Strong elephants that are becoming fitter may find a global answer. Since their premature death may arrive at the same speed as any other elephant, don't constrain them so they can't achieve their goals. Healthy elephants may find the world's best solution with more time. Ideally, the next generation of elephants would include healthy, long-lived elephants and newborn elephants reborn from ordinary elephants. A kid of the same gender is born after an elephant dies. To help the new young elephant remember all the essential facts so it may refine its current position and find new, possibly lucrative places near the best-known ones. Let's say the group of elephants is denoted as E; male and female elephants are denoted as E_m and E_f, respectively. E_{nb} denotes the new baby elephant. The characteristics of the elephants mentioned above can be described as follows:

- E_f is the group leader, and in the group, there are several E_m and E_f with a ratio of 1:4 (E_m: E_f). E_f is responsible for performing the local search with a radius $\mathcal{L}_f$.
- E_m is responsible for exploring the search space with a radius of $\mathcal{L}_f$ and finding the global solution. Depending upon the obtained solution, the entire group slowly migrates to that solution.
- E_o is the oldest female elephant in the group. The location of the E_{nb} can be determined by using equation 3.

$$Location(E_{nb}) = 0.6 * Location(E_{mother}) + 0.3 * Location(E_{father}) + 0.1 * Location(E_o) \quad (3)$$

3.4 Classification techniques

There are a number of classification schemes that can be used to estimate the risk of developing lung cancer. ML and statistical techniques classify patients' medical records, genetic profiles, and clinical data as benign, malignant, or relapsing/non-relapsing. The research employs a wide range of ML classification techniques: NB, LR, SVM, MLP, DT, RF, and KNN [21–23].

3.5 Deep neural network

Deep neural network: In an abstract sense, a specific type of artificial neural network consisting of many layers to present relations found in complex data, every layer contained in such an architecture implements mathematical transforms, namely the weighted sums as well as activation functions, which basically transmit the information progressively over the structure. Its huge depth through many of its hidden layers enables DNN to learn hierarchical features—starting from simple patterns at the first layers, and heading towards deeper abstract representations of them. Despite their impressive abilities, DNN often require considerable computational resources, large training datasets, and careful tuning of hyper-parameters to achieve peak performance [24, 25].

3.6 Ensemble learning

Ensemble learning enhances the effectiveness of ML and DL techniques by combining multiple models to make more accurate predictions than those produced by any single model. The basic approach is to use a collection of classifiers and combine their outputs, often using techniques such as voting. This paper discusses three different ensemble approaches, including WA, SV, and HV [26, 27]. A weighted average ensemble is one of the methods where multiple models have predictions averaged, with each model contributing by a weighted factor reflecting how well the model made a prediction. The highest score in the final prediction, the class, has it. Voting ensembles make up soft or hard voting: SV averages out the probabilities; HV is by majority class. Another approach is to aggregate the outputs of several models into a single dataset, train a meta-model on that dataset, and use the ensemble for efficient prediction of outcomes. Take λi as the initial prediction probability of various base learners or classifiers (Bi). In equation (4), the final prediction of the ensemble model is represented by ρ.

$$\rho = Max_i \sum_{k=1}^{B} \omega_k \lambda_k \tag{4}$$

The categorization problem-solving HV ensemble uses majority voting. This ensemble uses predictions from many datasets trained models. Let ρ be the projected class label using hard voting. Calculate this projected value using equation (5), where c is the class for the attribute a_1 of the dataset D.

$$\rho = mode\{c(a_1), c(a_2), \ldots\ldots, c(a_n)\} \tag{5}$$

4 PROPOSED MODEL

This study proposes a novel two-stage evaluation framework. The first stage involves utilizing CFS as a feature selection method, while the second stage employs the ESA as an optimization technique. Additionally, seven traditional ML classifiers—LR, NB, KNN, SVM, MLP, DT, and random forest (RF)—are included. Initially, the dataset undergoes preprocessing using a standard scaler.

Subsequently, CFS is applied to identify relevant features, and ESA is utilized to optimize the selected feature set further. The optimized features are then split into training and testing datasets with a test size of 20%. Finally, the performance of the dataset is evaluated using the seven classifiers. Algorithm 1 represents the pseudocode for the proposed hybrid ML model, and Figure 1 is a full process diagram for the study.

Algorithm 1: Pseudocode of Stage-1 of the Proposed Model

Require: Dataset D ($f_1, f_2, f_3 \ldots .., f_n$) Elephant population N (E_m, and E_f). $\mathcal{L}_m$, *and* $\mathcal{L}_f$ as the radius of E_m, E_f, EAge as the age limit of the elephant, T_{max} as the maximum iteration.

1. Preprocess (D)
2. Apply CFS to D to obtain the selected feature set D'
3. Invoke ESA() to D' to select the optimal number of feature sets
4. If ($t < T_{max}$ and G_{best} is not satisfactory)
   - **a)** For ($\forall E_{mi}, E_{mj} \in E_m$)
     - **i.** E_m = Location($E_{m,t-1}$))
     - **ii.** Calculate Euclidean distance (d) of (E_{mi}, E_{mj})
     - **iii.** If ($d < \mathcal{L}_m$)
       1. Abort
     - **iv.** Else
       1. Calculate the new fitness function $F_{new}(E_m)$
     - **v.** EndIf
     - **vi.** If ($F_{new}(E_m) < F(E_m)$)
       1. Update fitness function as $F_{new}(E_m)$.
       2. Update global best G_{best}
     - **vii.** EndIf
   - **b)** EndFor
   - **c)** For ($\forall E_{fk}, E_{fl} \in E_f$)
     - **i.** Calculate Euclidean distance (d) of (E_{fk}, E_{fl})
     - **ii.** If ($d < \mathcal{L}_f$)
       1. Abort
     - **iii.** Else
       1. Calculate the new fitness function $F_{new}(E_f)$
     - **iv.** EndIf
     - **v.** If ($F_{new}(E_f) < F(E_f)$)
       1. Update fitness function as $F_{new}(E_f)$.
       2. Update global best L_{best}
     - **vi.** EndIf
   - **d)** EndFor
   - **e)** For ($\forall E_m, E_f \in N$)
     - **i.** If ($F_{new}(E_m) < F_{new}(E_f)$)
       1. $G_{best} = L_{best}$
     - **ii.** EndIf
   - **f)** EndFor
   - **g)** $t = t + 1$
5. EndIf
6. Obtain D" as the optimal dataset.
7. Apply ML-classifiers to D."
8. Evaluate the performance metrics

In the second phase, we propose a DL ensemble using a DNN, a bagging classifier, and WA, SV, and hard voting. This study trains an approach using the Breast Cancer relapse dataset using DNN and simple EL approaches. The pre-processed dataset uses DNN, clustering ensemble classifier, and EL methods, including WA and voting. Algorithm 2 shows the pseudocode for the proposed model, and Figure 2 shows the workflow of the proposed model.

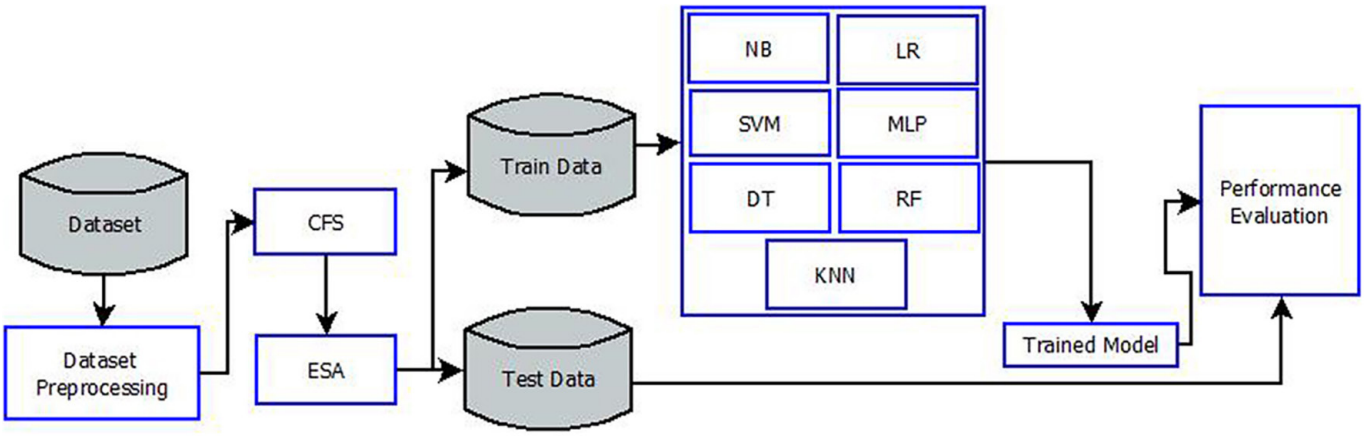

Fig. 1. Workflow of the proposed model (Stage-1)

Algorithm 2: Pseudocode of Stage 2 of the Proposed Model

1. Input raw cancer dataset.
2. Perform data pre-processing:
  a) Handle missing values.
  b) Normalize features.
  c) Split Dataset with distribution ratio (D) as 0.25.
3. Initialize SSA_Optimizer() for feature optimization.
4. Train DNN classifier on the feature set.
5. Initialize Ensemble_Learning_Approach():
  a) Set the number of iterations (I).
  b) Generate Initial predictions using ML classifiers.
  c) Apply ensemble methods: Weighted_Averaging(), Soft_Voting(), and Hard_Voting().

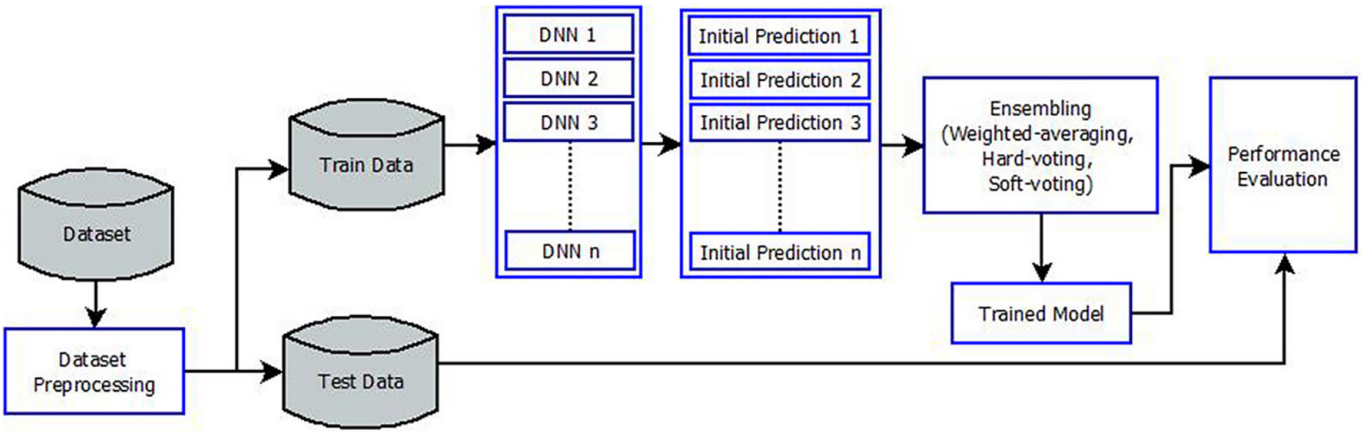

Fig. 2. Workflow of the proposed model (Stage-2)

5 RESULTS AND DISCUSSIONS

Assessment of ensemble models requires certain assumptions and methods. The study distinguishes between ML and DL, incorporates feature selection and optimization into ML, and discusses ensemble techniques such as WA, SV, and HV to improve prediction ACC. The experimental system has 16 GB RAM, 512 GB SSD, 2 TB HDD, 3.2 GHz AMD Ryzen 7 CPU, and Windows 11. Systematic experimental techniques make empirical analysis essential for determining findings. To forecast and

assess true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), a confusion matrix is created. Performance evaluation metrics include ACC, PRE, SEN, specificity (SPE), misclassification rate (MCR), FS, false negative rate (FNR), false positive rate (FPR), Matthews correlation coefficient (MCC), and balanced accuracy (BA) and their mathematical expressions in equations (7)–(15). The receiver operating characteristic (ROC) curve and AUC provide the true positive rate (TPR)-FPR balance [28, 29].

$$A_{CC} = \frac{T_A + T_B}{T_A + T_B + F_A + F_B} \tag{6}$$

$$M_{CR} = \frac{F_A + F_B}{T_A + T_B + F_A + F_B} \tag{7}$$

$$P_{RE} = \frac{T_A}{T_A + F_A} \tag{8}$$

$$S_{EN} = \frac{T_A}{T_A + F_B} \tag{9}$$

$$F - 1S = \frac{2 \times T_A}{2 \times T_A + F_B + F_B} \tag{10}$$

$$S_{PE} = \frac{T_B}{T_B + F_A} \tag{11}$$

$$F_{NR} = \frac{F_B}{T_A + F_B} \tag{12}$$

$$F_{PR} = \frac{F_A}{T_B + F_A} \tag{13}$$

$$M_{CC} = \frac{(T_A + T_B) - (F_A + F_B)}{\sqrt{(T_A + F_A)(T_A + F_B)(T_B + F_A)(T_B + F_B)}} \tag{14}$$

$$B_A = \frac{R_{EC} + S_{PE}}{2} \tag{15}$$

The research presents a two-stage method evaluation procedure. The hybrid ML technique employs CFS and ESA to pick and optimize features using seven traditional classification methods: LR, NB, KNN, SVM, MLP, DT, and RF. In the second phase, our ensemble DL ensemble deep learning (EDL) employs WA SV, and HV with a DNN and bagging classifier. Table 3 provides comprehensive hybrid ML performance evaluations. Figure 3 shows the ACC comparison of the proposed model. "CFS+ESA+MLP" and "CFS+ESA+RF" surpass all six other hybrid ML-based techniques with accuracy of 97.18%, a precision of 97.99%, a specificity of 89.66%, an FS of 98.32%, etc. As indicated in Table 3, the hybrid ML-based strategy "CFS+ESA+MLP" improves MCR, FNR, FPR, MCC, and BA, making it the suggested hybrid model for the microarray lung cancer dataset.

Table 3. Results obtained utilizing hybrid ML approaches

Methodology	ACC (%)	MCR (%)	PRE (%)	SEN (%)	FS (%)	SPE (%)	FNR (%)	FPR (%)	MCC (%)	BA (%)
CFS+ESA+NB	94.35	5.65	95.92	97.24	96.58	81.25	2.76	18.75	80.51	89.25
CFS+ESA+LR	95.48	4.52	97.32	97.32	97.32	85.71	2.68	14.29	83.03	91.52
CFS+ESA+SVM	93.79	6.21	94.04	98.61	96.27	72.73	1.39	27.27	78.48	85.67
CFS+ESA+MLP	97.18	2.82	97.99	98.65	98.32	89.66	1.35	10.34	89.56	94.16
CFS+ESA+DT	94.92	5.08	96.62	97.28	96.95	83.33	2.72	16.67	81.71	90.31
CFS+ESA+RF	96.05	3.95	96.73	98.67	97.69	81.48	1.33	18.52	84.17	90.08
CFS+ESA+KNN	93.22	6.78	96.71	95.45	96.08	78.26	4.55	21.74	71.17	86.86

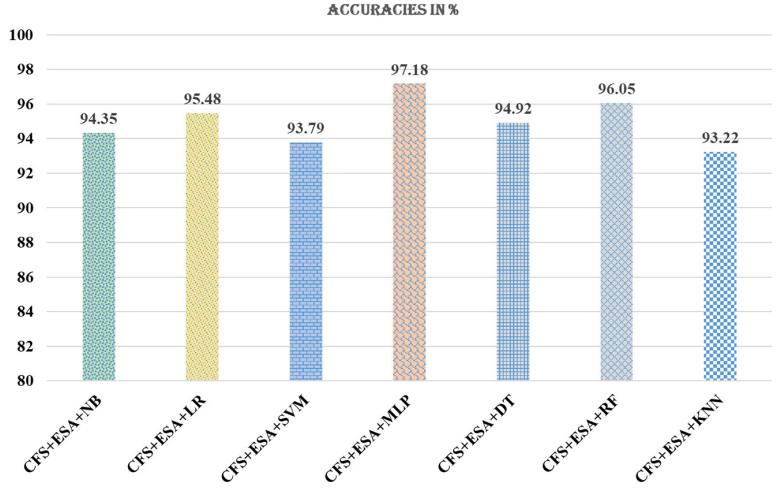

Fig. 3. Findings in % of accuracies for suggested hybrid ML-based approaches

Table 4 presents the findings from comprehensive assessments of the efficacy of the recommended EDL approaches. Figure 4 presents the findings as percentages for ACC, MCR, PRE, SNE, SPE, FS, FNR, FPR, MCC, and BA, respectively. The ensemble DL method "DNN+BC+MV," including a DNN, bagging classifier, and majority voting classifiers, surpasses all previously proposed methods, achieving an accuracy of 98.87% according to performance assessments. Moreover, the offered EDL methodology, exhibiting 99.33% in accuracy, SEN, and FS, with 96.43% specificity, surpasses other recommended ensemble DL methods; hence, this ensemble DL technique, termed "DNN+BC+MV," is regarded as the proposed ensemble DL strategy.

Figure 5 depicts the following calculation results: the ROC curve and the AUC value for the recommended ensemble DL technique known as "DNN+BC+MV." The AUC value of 0.99 that was obtained for this recommended EDL technique underlines the significance of its use on the microarray lung cancer dataset that was taken into consideration. Table 5 shows the performance evaluation comparison of the proposed model with existing literature.

Table 4. Results obtained utilizing ensemble DL approaches

Methodology	ACC (%)	MCR (%)	PRE (%)	SEN (%)	FS (%)	SPE (%)	FNR (%)	FPR (%)	MCC (%)	BA (%)
DNN+BC+WA	97.74	2.26	98.04	99.34	98.69	88.46	0.66	11.54	90.79	93.9
DNN+BC+SV	97.18	2.82	98.01	98.67	98.34	88.89	1.33	11.11	88.93	93.78
DNN+BC+MV	98.87	1.13	99.33	99.33	99.33	96.43	0.67	3.57	95.76	97.88

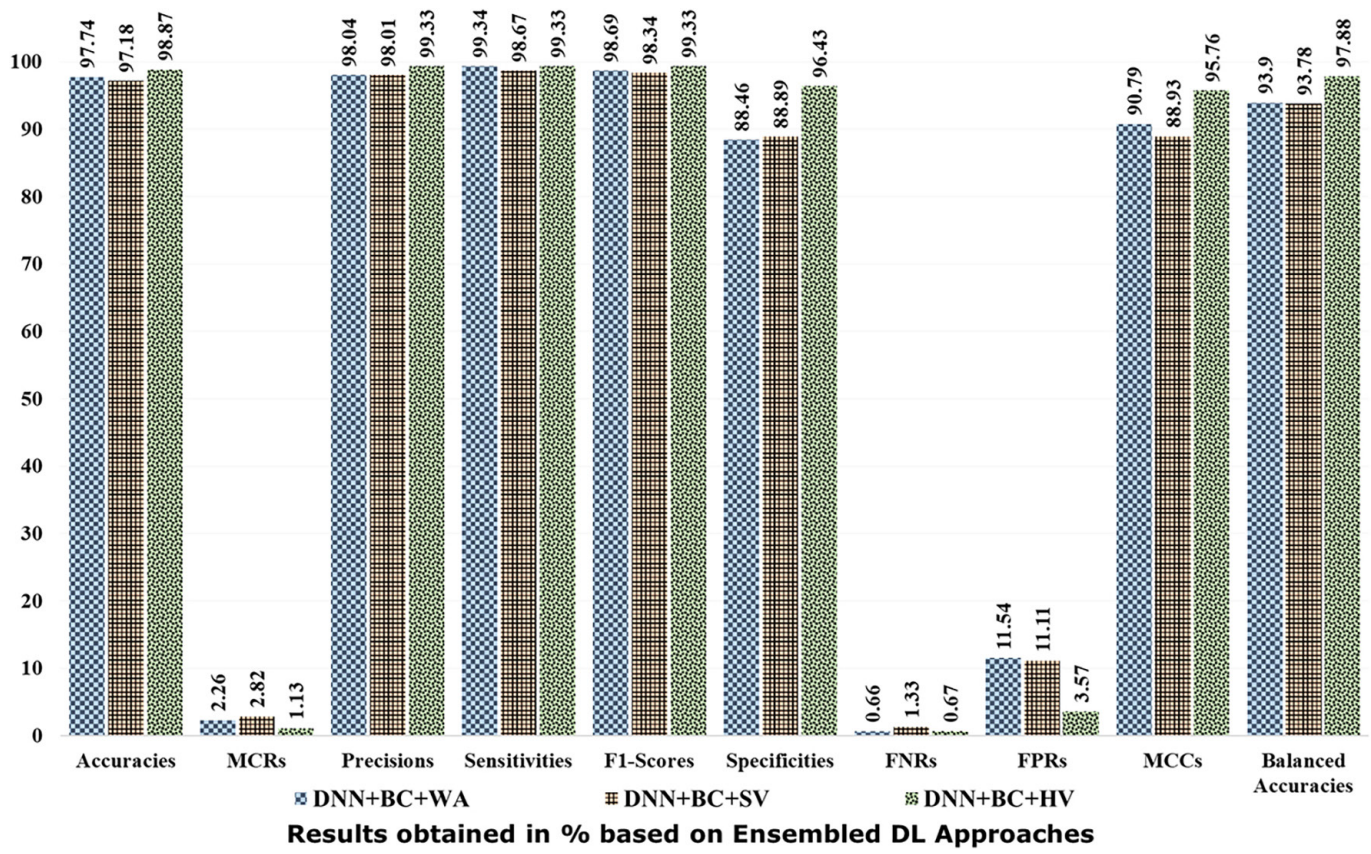

Fig. 4. Findings in % for suggested ensemble deep learning approaches

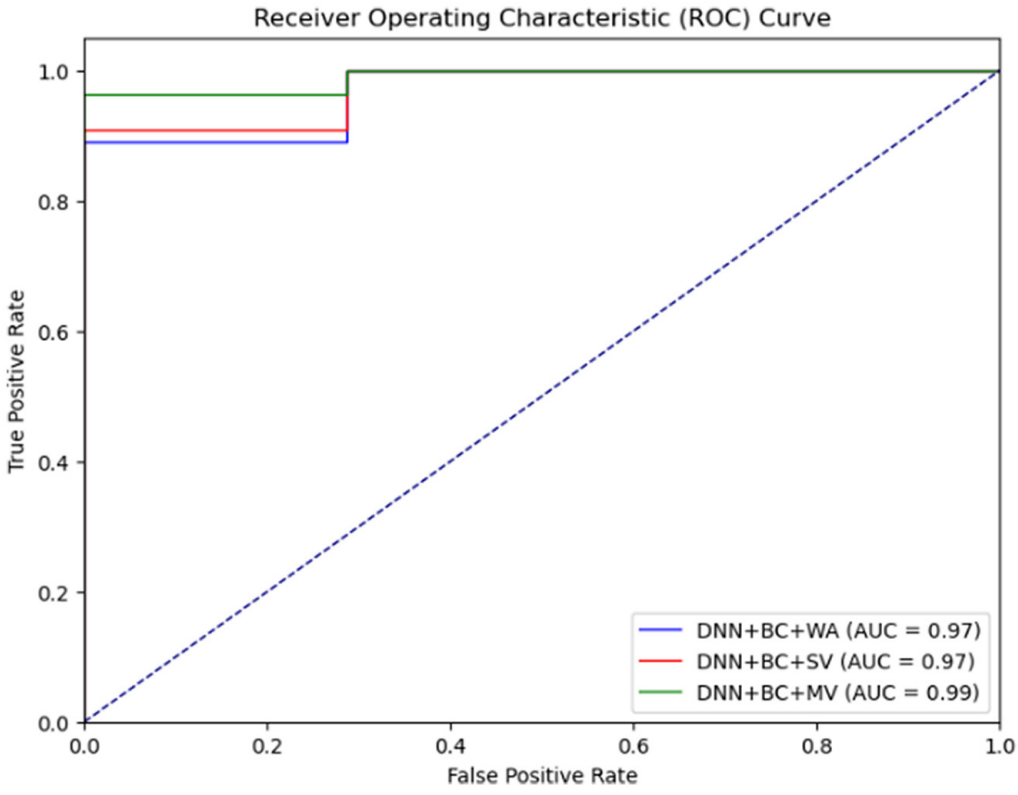

Fig. 5. AUCs obtained from ROCs plotted for suggested ensemble deep learning approaches

Table 5. Comparative analysis of the proposed hybrid approach to the considered state-of-the-art works

Ref	ACC (%)	PRE (%)	SEN (%)	SPE (%)	FS (%)	AUC
Lai et al. [7]	75.44	–	–	–	–	0.8163
Pawar [8]	94.99	–	–	–	–	–
Abdu-Aljabar and Awad [9]	97.9	100	97.2	100	–	0.98
Tabares-Soto et al. [10]	–	100	100	–	100	–
Basavegowda and Dangnew [11]	100	100	100	–	100	1
Thallam et al. [12]	99.5	–	–	–	–	–
Dagnew and Shekar [14]	100	100	100	–	100	1
Doppalapudi et al. [5]	71.18	77.70	83.36	–	79.96	0.92
Rani and Prasad [17]	81	81.2	78.9	–	77.7	–
Proposed Work	98.87	99.33	99.33	96.43	99.33	0.99

6 CONCLUSION AND FUTURE SCOPE

Lung cancer has a high mortality rate due to late diagnosis. Thus, early detection of this condition is crucial. The most problematic aspect of microarrays is the curse of dimensionality. Due to its small sample size and many characteristics, microarray data pose this problem. This this study evaluates prior efforts in two phases to give a new approach. A hybrid ML approach uses CFS and ESA for feature selection and optimization, and seven traditional classification methods: LR, NB, KNN, SVM, MLP, DT, and RF. The second phase combines EL with DNNs and bagging classifiers to create an ensemble DL method. These strategies include weighted average, gentle, and hard voting. Ten metrics are used to run tests on the NCBI microarray dataset. A 97.18% accuracy was attained using CFS, ESA, and a multilayer perceptron. DNN with a bagging classifier with HV yielded 98.87% accuracy.

The proposed models were evaluated on a limited microarray dataset, which may affect their generalizability to broader or more heterogeneous populations. Additionally, while highly accurate, the EDL approach may require substantial computational resources, posing challenges for deployment in real-time or resource-constrained clinical settings. Further validation on large, real-world datasets is necessary to confirm the model's robustness and practical applicability. This approach improves outcomes and helps construct an early planning method to build a foundation for lung cancer patients' medical, financial, and resource demands upon diagnosis. We want to find the best dataset preparation approaches to reduce noise and enhance the model's detection and prediction. We also plan to categorize cancer photos using DL in the future.

7 REFERENCES

[1] H. A. Miller, V. H. van Berkel, and H. B. Frieboes, "Lung cancer survival prediction and biomarker identification with an ensemble machine learning analysis of tumor core biopsy metabolomic data," *Metabolomics*, vol. 18, 2022. https://doi.org/10.1007/s11306-022-01918-3

[2] M. N. Das, N. Panda, and R. Rautray, "Enhancing lung cancer disease diagnosis by employing ensemble deep learning approaches," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 32, no. 3, pp. 1766–1773, 2023. https://doi.org/10.11591/ijeecs.v32.i3.pp1766-1773

[3] M. Saminathan, M. Ramachandran, A. Kumar, K. Rajkumar, A. Khanna, and P. Singh, "A study on specific learning algorithms pertaining to classify lung cancer disease," *Expert Systems*, vol. 39, no. 3, p. e12797, 2021. https://doi.org/10.1111/exsy.12797

[4] Y. Xie *et al.*, "Early lung cancer diagnostic biomarker discovery by machine learning methods," *Translational Oncology*, vol. 14, no. 1, p. 100907, 2021. https://doi.org/10.1016/j.tranon.2020.100907

[5] S. Doppalapudi, R. G. Qiu, and Y. Badr, "Lung cancer survival period prediction and understanding: Deep learning approaches," *International Journal of Medical Informatics*, vol. 148, p. 104371, 2021. https://doi.org/10.1016/j.ijmedinf.2020.104371

[6] S. Cui, Q. Wu, J. West, and J. Bai, "Machine learning-based microarray analyses indicate low-expression genes might collectively influence PAH disease," *PLOS Computational Biology*, vol. 15, no. 8, p. e1007264, 2019. https://doi.org/10.1371/journal.pcbi.1007264

[7] Y.-H. Lai, W.-N. Chen, T.-C. Hsu, C. Lin, Y. Tsao, and S. Wu, "Predicting the prognosis of non-small cell lung cancer by integrating microarray and clinical data with deep learning," *bioRxiv*, 2019. https://doi.org/10.1101/656140

[8] S. Pawar, "Web-based application for accurately classifying cancer type from microarray gene expression data using a Support Vector Machine (SVM) learning algorithm," in *Bioinformatics and Biomedical Engineering*, in Lecture Notes in Computer Science, I. Rojas, O. Valenzuela, F. Rojas, and F. Ortuño, Eds., vol. 11466, 2019, pp. 149–154. https://doi.org/10.1007/978-3-030-17935-9_14

[9] R. D. Abdu-Aljabar and O. A. Awad, "A comparative analysis study of lung cancer detection and relapse prediction using XGBoost classifier," *IOP Conference Series: Materials Science and Engineering*, vol. 1076, no. 1, p. 012048, 2021. https://doi.org/10.1088/1757-899X/1076/1/012048

[10] R. Tabares-Soto, S. Orozco-Arias, V. Romero-Cano, V. S. Bucheli, J. L. Rodríguez-Sotelo, and C. F. Jiménez-Varón, "A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data," *PeerJ Computer Science*, vol. 6, p. e270, 2020. https://doi.org/10.7717/peerj-cs.270

[11] H. S. Basavegowda and G. Dagnew, "Deep learning approach for microarray cancer data classification," *CAAI Transactions on Intelligence Technology*, vol. 5, no. 1, pp. 22–33, 2020. https://doi.org/10.1049/trit.2019.0028

[12] C. Thallam, A. Peruboyina, S. S. T. Raju, and N. Sampath, "Early stage lung cancer prediction using various machine learning techniques," in *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2020, pp. 1285–1292. https://doi.org/10.1109/ICECA49313.2020.9297576

[13] S. Takahashi *et al.*, "Predicting deep learning based multi-omics parallel integration survival subtypes in lung cancer using reverse phase protein array data," *Biomolecules*, vol. 10, no. 10, p. 1460, 2020. https://doi.org/10.3390/biom10101460

[14] G. Dagnew and B. H. Shekar, "Ensemble learning-based classification of microarray cancer data on tree-based features," *Cognitive Computation and Systems*, vol. 3, no. 1, pp. 48–60, 2021. https://doi.org/10.1049/ccs2.12003

[15] K. Rezaee, G. Jeon, M. R. Khosravi, H. H. Attar, and A. Sabzevari, "Deep learning-based microarray cancer classification and ensemble gene selection approach," *IET Systems Biology*, vol. 16, nos. 3–4, pp. 120–131, 2022. https://doi.org/10.1049/syb2.12044

[16] A. Khoirunnisa, A. Adiwijaya, and D. Adytia, "Implementation of CRNN method for lung cancer detection based on microarray data," *JOIV: International Journal on Informatics Visualization*, vol. 7, no. 2, pp. 600–605, 2023. https://doi.org/10.30630/joiv.7.2.1339

[17] K. M. S. Rani and V. K. Prasad, "Identification of lung cancer using ensemble methods based on gene expression data," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, no. 10s, pp. 257–266, 2023. https://ijisae.org/index.php/IJISAE/article/view/3249

[18] J. Hou *et al.*, "Gene expression-based classification of non-small cell lung carcinomas and survival prediction," *PLoS ONE*, vol. 5, no. 4, p. e10312, 2010. https://doi.org/10.1371/journal.pone.0010312

[19] T. P. Sahu and Ankita, "Correlation Feature Selection (CFS) and Feature Weighting (CFW) based improved BPSO for gene selection and cancer classification," in *Proceedings of the 2023 6th International Conference on Information Science and Systems*, 2023, pp. 208–214. https://doi.org/10.1145/3625156.3625199

[20] S. Deb, S. Fong, and Z. Tian, "Elephant search algorithm for optimization problems," in *2015 Tenth International Conference on Digital Information Management (ICDIM)*, 2015, pp. 249–255. https://doi.org/10.1109/ICDIM.2015.7381893

[21] P. Chaturvedi, A. Jhamb, M. Vanani, and V. Nemade, "Prediction and classification of lung cancer using machine learning techniques," *IOP Conference Series: Materials Science and Engineering*, vol. 1099, no. 1, p. 012059, 2021. https://doi.org/10.1088/1757-899X/1099/1/012059

[22] M. Kh. Khazaaleh *et al.*, "Handling DNA malfunctions by unsupervised machine learning model," *Journal of Pathology Informatics*, vol. 14, p. 100340, 2023. https://doi.org/10.1016/j.jpi.2023.100340

[23] O. M. Al-Hazaimeh, A. Abu-Ein, N. Tahat, M. Al-Smadi, and M. Al-Nawashi, "Combining artificial intelligence and image processing for diagnosing diabetic retinopathy in retinal fundus images," *International Journal of Online and Biomedical Engineering*, vol. 18, no. 13, pp. 131–151, 2022. https://doi.org/10.3991/ijoe.v18i13.33985

[24] S. Shandilya and S. R. Nayak, "Analysis of lung cancer by using deep neural network," in *Innovation in Electrical Power Engineering, Communication, and Computing Technology*, in Lecture Notes in Electrical Engineering, M. Mishra, R. Sharma, A. Kumar Rathore, J. Nayak, and B. Naik, Eds., vol. 814, 2021, pp. 427–436. https://doi.org/10.1007/978-981-16-7076-3_37

[25] S. Taha Ahmed and S. Malallah Kadhem, "Using machine learning via deep learning algorithms to diagnose the lung disease based on chest imaging: A survey," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 15, no. 16, pp. 95–112, 2021. https://doi.org/10.3991/ijim.v15i16.24191

[26] A. Safiyari and R. Javidan, "Predicting lung cancer survivability using ensemble learning methods," in *2017 Intelligent Systems Conference (IntelliSys)*, 2017, pp. 684–688. https://doi.org/10.1109/IntelliSys.2017.8324368

[27] B. Sahu *et al.*, "Novel hybrid feature selection using binary portia spider optimization algorithm and fast mRMR," *Bioengineering*, vol. 12, no. 3, p. 291, 2025. https://doi.org/10.3390/bioengineering12030291

[28] N. Gharaibeh, A. A. Abu-Ein, O. M. Al-hazaimeh, Khalid M. O. Nahar, W. A. Abu-Ain, and M. M. Al-Nawashi, "Swin transformer-based segmentation and multi-scale feature pyramid fusion module for alzheimer's disease with machine learning," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 19, no. 4, pp. 22–50, 2023. https://doi.org/10.3991/ijoe.v19i04.37677

[29] M. M. Al-Nawashi, O. M. Al-Hazaimeh, and M. Kh. Khazaaleh, "New approach for breast cancer detection-based machine learning technique," *Applied Computer Science*, vol. 20, no. 1, pp. 1–16, 2024. https://doi.org/10.35784/acs-2024-01

8 AUTHORS

Mr. Manmath Nath Das is currently pursuing Ph.D. as a research scholar at Shiksha 'O' Anusandhan University, Bhubaneswar, Odisha. Mr. Das completed his M.Tech. in CSE in 2011 from BPUT, Odisha. His research interest includes machine learning, deep learning, and cloud computing. He can be contacted at e-mail: manmathnath.das@gmail.com.

Dr. Niranjan Panda is working as an Associate Professor in the CSE Department, Siksha 'O' Anusandhan University, Bhubaneswar, India since 2011. He was awarded his Ph.D. degree from Siksha 'O' Anusandhan University, Bhubaneswar, India, in 2019. He got his M.Tech. Degree from KIIT University, Bhubaneswar, India, and B.Tech. from Utkal University, Bhubaneswar, India, in 2010 and 2005, respectively. His research interests include ad hoc networks, computer security, intelligent systems, and machine learning. He can be contacted at e-mail: niranjanpanda@soa.ac.in.

Dr. Rasmita Rautray received her BE degree and M.Tech. degree in CSE from Utkal University, Bhubaneswar, India, in 2003 and 2006, respectively, and the Ph.D. degree in CSE from Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, India, in 2017. She is currently working as an Associate Professor in the Department of CSE, FET (ITER), Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, India. Her research interests include data mining, soft computing, and information retrieval. She has contributed over 30 research papers in many National and International journals and conferences. She can be contacted at e-mail: rashmitarautray@soa.ac.in.

Ms. Jyotsnarani Tripathy is presently working as an Assistant Professor in CSE-AIML & IoT, VNR Vignana Jyothi Institute of Engineering & Technology, Hyderabad, Telangana, India since 2022, and also has 16 years of teaching experience. She is pursuing a Ph.D degree from Maharaja Sriram Chandra Bhanjadeo University, Baripada, Odisha, India. She was awarded an M.E in Computer Science from Utkal University, Bhubaneswar, Odisha, in 2008. She got a B.E. degree in Computer Science from Ajay Binay Institute of Technology (ABIT), Cuttack, Odisha, in 2006. Her research focuses on Image Processing, and her work has been published in over 10 international and national journals and conferences. She is also a member of ISTE and OITS. Her research interests include Image Processing, Machine Learning, and Deep Learning. She can be contacted at e-mail: jtjyotsna@gmail.com.