A novel machine learning based hybrid approach for breast cancer relapse prediction

Ghanashyam Sahoo[1], Ajit Kumar Nayak[2], Pradyumna Kumar Tripathy[3], Jyotsnarani Tripathy[4]
[1]Department of Computer Science and Engineering, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, India
[2]Department of Computer Science and Information Technology, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, India
[3]Department of Computer Science and Engineering, Silicon Institute of Technology, Bhubaneswar, India
[4]Department of CSE-AIML and IoT, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India

Article Info

Article history:

Received Jul 10, 2023
Revised Sep 14, 2023
Accepted Sep 25, 2023

Keywords:

Breast cancer relapse
Classification approaches
Correlation-based feature selection
Flower pollination algorithm
Machine learning

ABSTRACT

The second leading cause of death for women is breast cancer, which is growing. Some cancer cells may remain in the body, so relapse is possible even if treatment begins soon after diagnosis. Since there are now many machine learning (ML) approaches to recurrence prediction in breast cancer, it is important to compare and contrast them to find the most effective one. Datasets with many features often lead to incorrect predictions because of this. In this study, correlation-based feature selection (CFS) and the flower pollination algorithm (FPA) are used to improve the quality of the wisconsin prognostic breast cancer (WPBC) and University Medical Centre, Institute of Oncology (UMCIO) breast cancer relapse datasets respectively. Data imputation, scaling, pre-process raw data. The second stage uses CFS to select discriminative features based on important feature correlations. The FPA chose the optimum attribute combination for the most precise answer. We tested the approach using 10-fold cross-validation stratification. Various trials show 84.85% and 83.92% accuracy on the WPBC and UMCIO breast cancer relapse datasets, respectively. The hybrid method performed well in feature selection, increasing the accuracy of the relapse classification for breast cancer.

Corresponding Author:

Ghanashyam Sahoo
Department of Computer Science and Engineering, Siksha 'O' Anusandhan (Deemed to be University)
Bhubaneswar, Odisha, India
Email: ghanarvind@gmail.com

1. INTRODUCTION

The World Health Organization (WHO) reports that breast cancer is a major cause of death worldwide. This disorder is more common in older women of all races [1]. And it is the leading killer of female cancer patients. The most lethal aspects of breast cancer are metastasis and recurrence. Breast cancer that has been treated may come back years or decades later. Early detection and prediction are seen as helpful tools in the battle against this aggressive cancer by many researchers [2]. The use of machine learning (ML) and data mining (DM) for relapse prediction in breast cancer is thus an important area of research. In order to glean meaningful insights from large data sets, data miners employ statistical, probabilistic, and ML techniques [3], [4]. The ability to accurately forecast breast cancer helps doctors tailor treatments to the needs of each patient, improving care and survival rates. The utilization of healthcare resources for these individuals is also enhanced. Why breast cancer recurs so rapidly is still a mystery to clinical and statistical researchers [4]. Researchers are urged to use ML's learning and forecasting powers to diagnose and treat several diseases. ML methods employ mathematical models that link features to find patterns in large datasets. The use of ML algorithms in medical studies has grown in recent years. Several ML methods have been designed specifically for breast cancer classification in medical datasets [5].

Prediction models were developed by Ahmad *et al.* [6] using decision tree (DT) C4.5, support vector machine (SVM), and artificial neural network (ANN). In order to determine how sensitive, specific, and accurate these three popular algorithms are, we conduct these evaluations. To enhance recurrence prediction in breast cancer, Vazifehdan *et al.* [7] created a hybrid imputation approach that takes into account attribute reliance and incomplete attribute type. Using the wisconsin prognostic breast cancer dataset (WPBC), Chakradeo *et al.* [8] compared several ML models for detecting and predicting breast cancer recurrences. To evaluate susceptibility, risk, recurrence, and regeneration after resolution and survival, Aavula and Bhramaramba developed the extensible breast cancer prognosis framework (XBPF) [9]. The authors propose combining representative feature subset selection (RFSS) with SVMs to better forecast surveillance, epidemiology, and end results (SEER) dataset outcomes. Prediction strategies for recurrence of breast cancer within ten years of surgery were investigated by Lou *et al.* [10]. Using data from pathology reports and patient progress notes, Wang *et al.* [11] constructed an electronic health record-based distant recurrence prediction model for 6,447 breast cancer patients who were treated at Northwestern Memorial Hospital between 2001 and 2015. Breast cancer recurrence prediction motivated Momenzadeh *et al.* [12] development of a hidden markov model (HMM). The HMM produced sequential patterns from gene expression data to distinguish expression profiles, using ranked gene sets as observation sequences and hidden states as gene set priorities. Scaling only, scaling with principle component analysis (PCA), scaling with PCA and minority class oversampling, and scaling with selected characteristics were the four prominent ML models that Magboo and Magboo [13] tested for categorizing breast cancer recurrences on the WPBC dataset. Using ML techniques, Alzu'bi *et al.* [14] predicted the return of breast cancer. The medical staff at King Abdullah University Hospital (KAUH) agreed with the prognosis. Doctors and researchers validated the content of the medical dictionary. Alwohaibi *et al.* [15] used statistics and a brainstorming optimization algorithm (BSO) to refine their study of two breast cancer recurrence datasets. As the first step in the multi-phase process, statistical feature selection (SFM) is where we start. Using importance and correlation, SFM chooses discriminative features. Predictions based on class variables are used to rank traits. Second, the approach is analyzed by a multi-classifier (MC) using a combination of two SFMs and three classifiers. The BSO algorithm prioritized the most important factors. Ebrahim *et al.* [16] analyzed 1.7 million NIH datasets. They employed DT, linear discriminant analysis (LDA), logistic regression (LR), SVM, and ensemble techniques (ET). The features impacted precision as well as probabilistic neural network (PNN), deep neural network (DNN), and recurrent neural network (RNN) methods were employed for comparison study.

Dataset feature quality influences classification algorithms' instance classification. Noise hinders classification, although some data factors aid. These are why feature selection approaches exist. Good classification models can be computed cheaper with fewer features. Reducing unnecessary details makes healthcare diagnostic testing cheaper. Few recurrent breast cancer databases exist. Medical datasets often have missing data, duplicates, missing values, noise, and biases from non-representative events. Pretreatment enhances data classification and analysis. Specific missing data approaches. Second, these immense databases are heterogeneous. There are various ways to customize features. Wrapper strategies pick subsets of features using optimization algorithms, while discriminative methods select features. Finding the right dataset feature selection method might be tricky. Many high-dimensional breast cancer recurrence datasets have rich features. Third, dependency is uncertain. Determine the feature-target variable association by calculating the feature-class correlation coefficient. Predictable characteristics are highly connected to the class but not strongly associated with other features. Features improve prediction; hence, feature reliance is significant. Both independent and dependent factors predict. These challenges have no dataset-wide solution. The subset used can also affect classification. This study used a hybrid strategy to predict breast cancer relapse. The WPBC and University Medical Centre, Institute of Oncology (UMCIO) breast cancer datasets were obtained from the University of California Irvine machine learning (UCI-ML) Repository. Three phases make up this hybrid strategy: raw data is pre-processed via data imputation, scaling, and other methods. The correlation-based feature selection (CFS) uses important feature correlations to pick discriminative features. The flower pollination algorithm (FPA) found the best input properties for an exact answer. The feature selection methods and algorithms are employed for each dataset for optimal outcomes.

The rest of the paper is structured as follows: in section 2 depicts the data and methods used in the current research endeavors and the proposed work. The findings analysis is summed up in section 3. In section 4 concludes this study and suggests future research directions.

2. METHOD

2.1. Datasets description and pre-processing

This research investigates two breast cancer relapse datasets acquired from the UCI-ML Repository: the WPBC cancer relapse dataset and the UMCIO breast cancer relapse dataset. William, Wolberg, Nick Street, and Mangasarian from the University of Wisconsin developed the WPBC dataset [17]. The collection of data for the UMCIO dataset by the University Medical Center's Institute of Oncology was done in Zwitter and Soklic [18]. In this data collection, each individual is an individual who underwent cancer surgery. A brief description of the datasets employed in this research is depicted in Table 1. The data in both datasets is unbalanced. To maintain the original distribution's non-recurrence to recurrence ratio across folds, we employed stratified 10-fold cross-validation. Some datasets may benefit from preprocessing steps like data imputation and scaling. Eliminating missing values is simple when you use data imputation. Some missing values are usually not a big deal. Loss of statistical power and informative value results from discarding many cases. There must be imputation in the case of missing data. The missing numerical characteristics were reconstructed using a linear regression model. Sample median imputation was used for missing categories. Since the WPBC dataset was nearly complete except for four lymph node status variables, their associations with other factors were investigated. Missing feature values might be predicted using standard linear regression. However, the data range could be quite different for each attribute. To achieve this, we normalize or scale the feature data we have. Data preparation for ML relies on it. The absolute value of the WPBC feature data is divided by its standard deviation to achieve normalization.

Table 1. Brief description of the datasets employed

Dataset	Instances	Features	Non-relapse counts	Relapse counts
WPBC	198	35	151	47
UMCIO	286	10	201	85

2.2. Correlation-based feature selection

CFS is a feature selection algorithm to determine the correlation between two variables. To calculate the correlation, two different approaches are available [19]. One is to find out the linear coefficient, and the other is based on the information theory. In the first approach, the linear correlation coefficient ($\boldsymbol{\delta}$) can be calculated for the data point (fi, fj) as in the (1).

$$\boldsymbol{\delta} = \frac{\Sigma_i(f_i - \overline{f}_i)(f_j - \overline{f_j})}{\sqrt{(f_i - \overline{f}_i)^2}\sqrt{(f_j - \overline{f_j})^2}} \tag{1}$$

$\overline{f_\iota}, \overline{f_J}$ are the mean of data point f_i, and f_j respectively. $\boldsymbol{\delta}$ value lies in between the-1 and 1. The value of $\boldsymbol{\delta}$ shows the degree of correlation between them. If $\boldsymbol{\delta}$ becomes 0, the features are considered fully independent [20].

2.3. Flower pollination algorithm

FPA is an optimization technique that makes use of flowers' pollination behavior. Depending on the conditions, pollen may be transferred between flowers either biotically or abiotically [21]. The process of pollen being carried across large distances by animals and insects is called biotic pollination [22]. In abiotic pollination, pollen is dispersed over a short distance by the diffusion of water or air [23]. The four basic rules of FPA can be stated as follows:

- R1: biotic pollination is treated as global pollination following the levy flights.
- R2: abiotic pollination is called the local pollination process.
- R3: flower constancy feature of FPA assumes reproduction likelihood is correlated to flower resemblance.
- R4: local and global pollination switching can be controlled by a probability Ƥ in <0,1>.

The global pollination process of the FPA can be represented as (2) and (3) with G^* as the current global optimal solution and L as the levy flight step size with a constant scaling factor η.

$$\mathfrak{x}_m^{i+1} = \mathfrak{x}_m^i + \eta.\mathrm{L}.(G^* - \mathfrak{x}_m^i) \tag{2}$$

$$L \sim \frac{c\,\gamma(c)Sin(\frac{\pi c}{2})}{\pi} \cdot \frac{1}{\zeta^{c+1}} \tag{3}$$

where $\mathfrak{x}_m^i$ denotes the current flower $\mathfrak{x}_m$ at ith iteration with $\gamma()$ is the standard gamma function with a constant c. Similarly, the next position of a flower in the case of local search can be defined as (4).

$$x_m^{i+1} = x_m^i + \omega(x_n^i - x_o^i) \quad (4)$$

x_n^i and x_o^i are the positions of two different flowers used for pollination. ω is the uniform distribution of <0, 1>.

2.4. Proposed model

Several classification methods can be used to make predictions about the chance of relapse in breast cancer patients. Patients are separated into relapse and non-relapse groups using variables collected from their medical histories, genetic profiles, and clinical data through the use of ML and statistical methods. In addition to the CFS and FPA, several ML classification algorithms are applied to the dataset under consideration: SVM, random forest (RF), DT, LR, Naive Bayes (NB), K-nearest neighbours (KNN), and LDA [24]–[26]. The reported model uses two types of breast cancer relapse datasets. Figure 1 shows the workflow of the proposed model. Initially, the datasets undergo a preprocessing step to handle the outliers present in them. The CFS feature selection algorithm is applied to the processed dataset to identify the correlated features. The FPA optimization algorithm is then applied to the featured dataset to bring the optimized number of features into the front without hampering the dataset’s utility. Finally, seven different ML classifiers are applied to evaluate the performance of the proposed model with six different evaluative measures. Algorithm 1 shows the pseudocode of the proposed model.

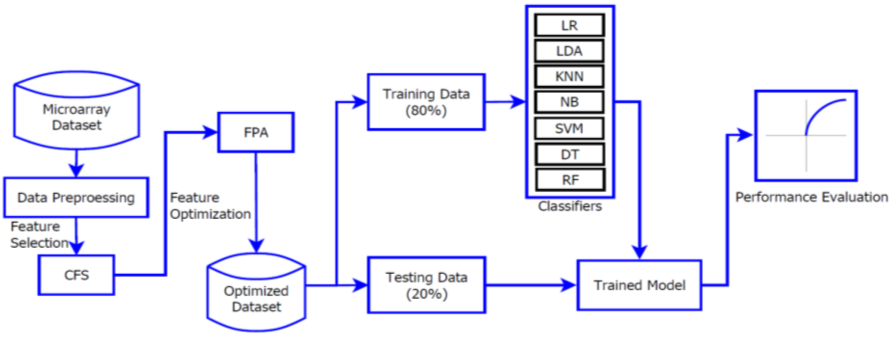

Figure 1. Workflow of the proposed model

Algorithm 1. Algorithm for proposed work

```
REQUIRE: Datasets D←{D1, D2}, Feature set (F)←{$f_1$, $f_2$,…..., $f_n$}
OUTPUT: Performance measures←{Ac,Pr,Sn,Sp,Er,Fv}
$D_i$∊D, apply data scaling
for j, k←1 to n
        $f_{scaled} = \frac{f_j - avg(F)}{\max(F) - \min(F)}$
        $D_i \leftarrow f_{scaled}$
end for
for $D_i$∊D, apply CFS
        for j ←1 to n
        calculate δ with respect to $f_c$ ($f_c$ is the class label of $D_i$)
        if ($\delta > 0.5$)
                for k ← 1 to m
                        $D_{featured} \leftarrow f_j$
                end for
        end if
        end for
        Apply FPA () to $D_{featured}$
                Initialize flower population $m \leftarrow \{m_1, m_2, \ldots m_k\}$
                Define objective function $F_{min}$ (), switch probability P
                Find $G^*$ in $n$
                while (i<max_iteration)
                        for j ← 1 to n
                                if (rand<P)
                                        Find out $x_m^{i+1}$ for global search
                                else
                                        Find out $x_m^{i+1}$ for local search
                                end if
                        end for
                end while
end for
```

3. RESULTS AND DISCUSSION

This research work has been carried out considering two breast cancer relapse datasets, namely, WPBC and UMCIO, taken from the UCI-ML repository and employing some performance parameters, namely, accuracy (Ac), precision (Pr), sensitivity (Sn), specificity (Sp), error rate (Er), and F-value (Fv). These performance parameters are determined, as in (6)-(10), based on the obtained confusion matrix (CM) from the experiments on Jupiter Python Notebook. The 2×2 CM consists of four parameters, namely, true positive (P_1), true negative (N_1), false positive (P_0), and false negative (N_0).

$$Ac = \frac{P_1 + N_1}{P_1 + N_1 + P_0 + N_0} \tag{5}$$

$$Pr = \frac{P_1}{P_1 + N_1} \tag{6}$$

$$Sn = \frac{P_1}{P_1 + N_0} \tag{7}$$

$$Sp = \frac{N_1}{N_1 + P_0} \tag{8}$$

$$Er = \frac{P_0 + N_0}{P_1 + N_1 + P_0 + N_0} \tag{9}$$

$$Fv = \frac{2 \times P_1}{2 \times P_1 + N_1 + N_0} \tag{10}$$

The experiments are performed in four ways. First, the seven traditional ML approaches are applied to the WPBC relapse dataset and recorded outcomes, as depicted in Table 2. Second, the hybrid approaches, i.e., the seven traditional ML approaches along with the CFS and FPA, and responses are recorded as depicted in Table 3. It can be observed that DT, with 74.75% accuracy, as depicted in Figure 2, outperforms others while considering only seven conventional ML approaches. In contrast, SVM, along with CFS and FPA, outperforms other hybrid approaches concerning the obtained accuracy of 84.85% in the case of the WPBC dataset, as depicted in Figure 3. In the third case, the seven traditional ML approaches are applied to the UMCIO breast cancer relapse dataset and recorded outcomes, as depicted in Table 4. Fourth, the hybrid approaches, i.e., the seven traditional ML approaches along with the CFS and FPA, and responses are recorded as depicted in Table 5. It can be observed that NB, with 80.07% accuracy, as depicted in Figure 4, outperforms others while considering only seven conventional ML approaches. In contrast, DT, along with CFS and FPA, outperforms other hybrid approaches concerning the obtained accuracy of 83.92% in the case of the UMCIO breast cancer relapse dataset, as depicted in Figure 5. Table 6 describes a comparative study of the proposed hybrid approaches with some considered state-of-the-art works. It is quite difficult to overcome the previous results in different datasets. At the same time, this recommended work outperforms other state-of-the-art works in the case of employing the same datasets, as shown in Table 6.

Table 2. Recorded responses employing ML approaches on the WPBC dataset

ML approaches	Ac (%)	Pr (%)	Sn (%)	Sp (%)	Er (%)	Fv (%)
LR	74.24	76.36	77.06	70.79	25.76	76.71
LDA	73.74	74.56	78.71	67.78	26.26	76.58
KNN	72.22	74.26	72.12	72.34	27.78	73.17
NB	71.72	77.88	73.95	68.35	28.28	75.86
SVM	75.76	81.91	74.78	77.11	24.24	78.18
DT	74.75	82.18	72.17	78.31	25.25	76.85
RF	72.73	76.64	73.87	71.26	27.27	75.23

Table 3. Recorded responses employing hybrid approaches on the WPBC dataset

Hybrid approaches	Ac (%)	Pr (%)	Sn (%)	Sp (%)	Er (%)	Fv (%)
CFS+FPA+LR	83.33	89.66	83.21	86.56	16.67	86.31
CFS+FPA+LDA	81.31	82.93	86.44	73.75	18.69	84.65
CFS+FPA+KNN	79.29	82.57	80.36	77.91	20.71	81.45
CFS+FPA+NB	80.31	86.07	82.68	76.06	19.69	84.34
CFS+FPA+SVM	84.85	90.59	84.81	84.93	15.15	84.81
CFS+FPA+DT	82.83	91.15	81.09	85.92	17.17	81.09
CFS+FPA+RF	80.81	85.22	82.35	78.48	19.19	83.48

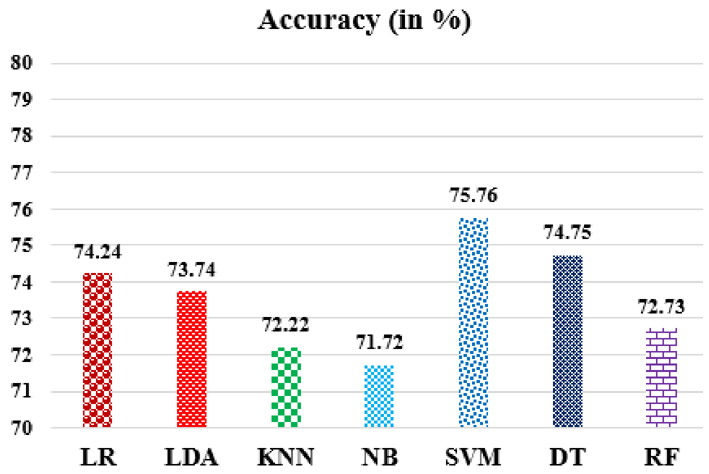

Figure 2. Obtained accuracies utilizing ML approaches on the WPBC dataset

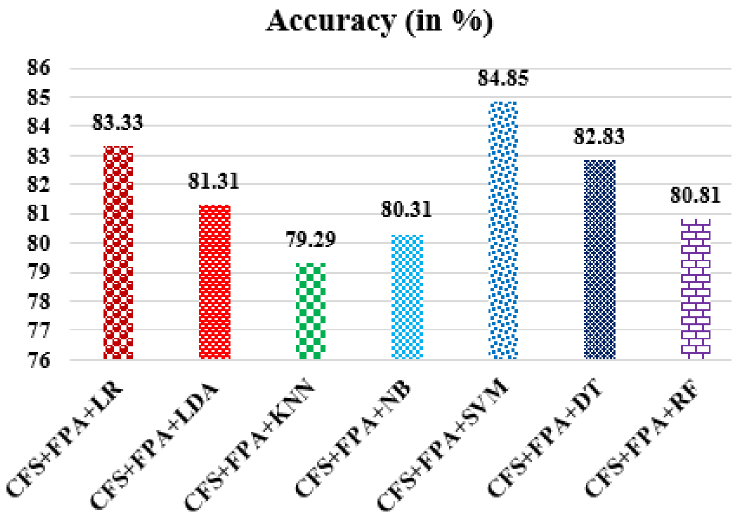

Figure 3. Obtained accuracies utilizing hybrid approaches on the WPBC dataset

Table 4. Recorded responses employing ML approaches on UMCIO dataset

ML approaches	Ac (%)	Pr (%)	Sn (%)	Sp (%)	Er (%)	Fv (%)
LR	77.62	81.87	82.78	68.87	22.38	82.32
LDA	77.27	80.65	83.81	66.36	22.73	82.19
KNN	76.57	81.49	80.11	70.91	23.43	80.79
NB	80.07	86.59	82.45	75.51	19.93	84.47
SVM	78.67	84.92	81.72	73.01	21.33	83.29
DT	79.37	86.11	82.01	74.23	20.63	84.01
RF	76.57	82.12	80.77	69.23	23.43	81.44

Table 5. Recorded responses employing hybrid approaches on UMCIO dataset

Hybrid approaches	Ac (%)	Pr (%)	Sn (%)	Sp (%)	Er (%)	Fv (%)
CFS+FPA+LR	79.72	85.25	83.42	72.73	20.28	84.32
CFS+FPA+LDA	79.37	85.03	83.68	70.83	20.63	84.35
CFS+FPA+KNN	78.67	84.18	81.87	73.08	21.33	83.01
CFS+FPA+NB	82.52	88.52	84.82	77.89	17.48	86.63
CFS+FPA+SVM	80.77	87.29	83.16	76.04	19.28	85.18
CFS+FPA+DT	83.92	90.81	85.28	80.91	16.08	87.96
CFS+FPA+RF	80.07	86.81	82.72	74.74	19.93	84.72

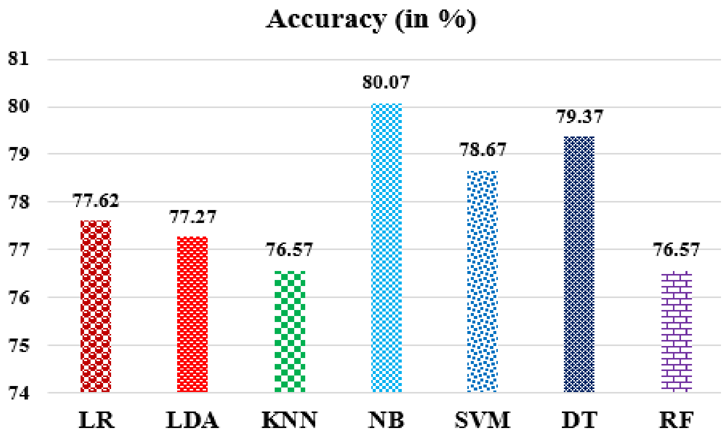

Figure 4. Obtained accuracies utilizing ML approaches on the UMCIO dataset

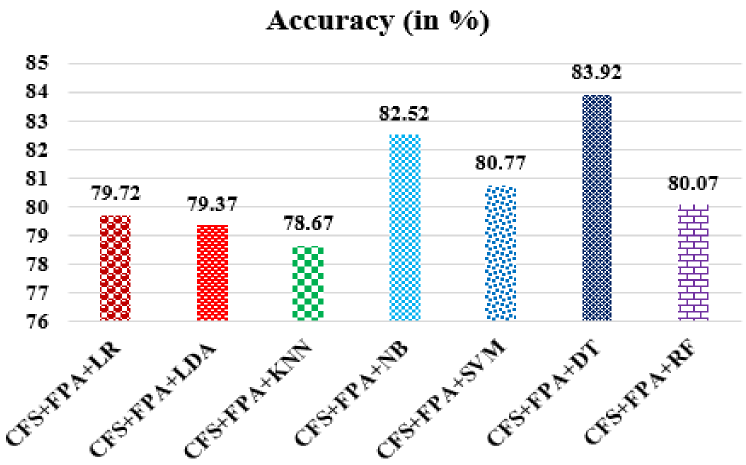

Figure 5. Obtained accuracies utilizing hybrid approaches on the UMCIO dataset

Table 6. Comparison of the proposed hybrid approach with some considered state-of-the-art works

Reference	Methodologies	Dataset (s)	Findings (%)				
			Ac	Pr	Sn	Sp	Fv
Ahmad *et al.* [6]	C4.5 DT, SVM, ANN	ICBC	95.7	--	97.1	94.5	--
Vazifehdan *et al.* [7]	DT, KNN, SVM	Omid Hospital	89.29	--	78.55	92.83	--
Chakradeo *et al.* [8]	LR, SVM, DT	WPBC	97.93	93.36	91.00	--	--
Aavula and Bhramaramba [9]	XBPF, RFSS	SEER	98.90	--	57.54	59.86	--
Lou *et al.* [10]	ANN, KNN, SVM, NB, COX	Questionary	98.87	97.90	95.89	99.54	--
Wang *et al.* [11]	K-CNN	NMEDW	--	55.7	46.8	98.1	50.0
Momenzadeh *et al.* [12]	HMM	7 micro arrays	--	--	--	--	--
Magboo and Magboo [13]	LR, NB, KNN, and SVM	WPBC	74	67	62	--	74
Alzu'bi *et al.* [14]	J48, NB, Bag, LR, SVM-SMO, KNN, MLP, OneR, PART,	KAUH	92.25	--	92.3	88.7	--
Alwohaibi *et al.* [15]	SVM, LR, LDA, SFS, CB, BSO, GBSO	WPBC, UMCIO	82 76	81 81	82 82	96 96	-- --
Ebrahim *et al.* [16]	DT, LDA, LR, SVM, ET, DNN, RNN, and PNN	NIH, USA	98.7	97.4	76.4	--	85.2
Proposed hybrid approach	CFS+FPA+7 ML approaches	WPBC UMCIO	84.85 83.92	91.15 90.81	86.44 85.28	86.56 80.91	86.31 87.96

4. CONCLUSION

This proposed hybrid approach includes two breast cancer relapse datasets, WPBC and UMCIO breast cancer datasets. These raw data were further refined in this study using CFS and FPA. Data imputation, scaling, and other forms of pre-processing are applied to raw data first. Second, crucial feature correlations are used in a CFS to choose the discriminative features. The FPA identified the optimal combination of the chosen characteristics as the means to an accurate answer. We tested the efficacy of the suggested technique using a 10-fold cross-validation stratified. As a result, the suggested hybrid strategy successfully selected characteristics and increased classification accuracy for breast cancer recurrence. When looking at just seven traditional ML methods, DT comes out on top with an accuracy of 74.75%, whereas in the case of the WPBC dataset, SVM combined with CFS and FPA achieves the highest accuracy of any hybrid method (84.85%). It is clear that NB performs best with an 80.07% accuracy when just seven standard ML techniques are included. In contrast, in conjunction with CFS and FPA, DT performs best when considering the accuracy of 83.92% achieved on the UMCIO breast cancer dataset. This study can further be implemented using other various breast cancer relapse datasets with different attributes and applying the ensemble approaches to enhance the outcomes.

REFERENCES

[1] S. R. Gupta, "Prediction time of breast cancer tumor recurrence using machine learning," *Cancer Treatment and Research Communications*, vol. 32, p. 100602, 2022, doi: 10.1016/j.ctarc.2022.100602.

[2] A. Bokhare and P. Jha, "Machine learning models applied in analyzing breast cancer classification accuracy," *IAES International Journal of Artificial Intelligence (IJAI)*, vol. 12, no. 3, pp. 1370–1377, Sep. 2023, doi: 10.11591/ijai.v12.i3.pp1370-1377.

[3] M. M. Hossin, F. M. J. M. Shamrat, M. R. Bhuiyan, R. A. Hira, T. Khan, and S. Molla, "Breast cancer detection: an effective comparison of different machine learning algorithms on the Wisconsin dataset," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 4, pp. 2446–2456, Aug. 2023, doi: 10.11591/eei.v12i4.4448.

[4] B. Thakur and N. Kumar, "Prediction, detection and recurrence of breast cancer using machine learning based on image and gene datasets," in *Lecture Notes in Electrical Engineering*, vol. 832, 2022, pp. 263–273.

[5] R. R. Kadhim and M. Y. Kamil, "Comparison of breast cancer classification models on wisconsin dataset," *International Journal of Reconfigurable and Embedded Systems (IJRES)*, vol. 11, no. 2, p. 166, Jul. 2022, doi: 10.11591/ijres.v11.i2.pp166-174.

[6] L. G. Ahmad, A. T. Eshlaghy, A. Poorebrahimi, M. Ebrahimi, and A. R. Razavi "Using three machine learning techniques for predicting breast cancer recurrence," *Journal of Health & Medical Informatics*, vol. 04, no. 02, 2013, doi: 10.4172/2157-7420.1000124.

[7] M. Vazifehdan, M. H. Moattar, and M. Jalali, "A hybrid Bayesian network and tensor factorization approach for missing value imputation to improve breast cancer recurrence prediction," *Journal of King Saud University - Computer and Information Sciences*, vol. 31, no. 2, pp. 175–184, 2019, doi: 10.1016/j.jksuci.2018.01.002.

[8] K. Chakradeo, S. Vyawahare, and P. Pawar, "Breast cancer recurrence prediction using machine learning," in *2019 IEEE Conference on Information and Communication Technology*, Dec. 2019, pp. 1–7, doi: 10.1109/CICT48419.2019.9066248.

[9] R. Aavula and R. Bhramaramba, "XBPF: an extensible breast cancer prognosis framework for predicting susceptibility, recurrence and survivability," *International Journal of Engineering and Advanced Technology*, vol. 8, no. 5, pp. 159–166, 2019.

[10] S. J. Lou *et al.*, "Machine learning algorithms to predict recurrence within 10 years after breast cancer surgery: a prospective cohort study," *Cancers*, vol. 12, no. 12, pp. 1–15, Dec. 2020, doi: 10.3390/cancers12123817.

[11] H. Wang, Y. Li, S. A. Khan, and Y. Luo, "Prediction of breast cancer distant recurrence using natural language processing and knowledge-guided convolutional neural network," *Artificial Intelligence in Medicine*, vol. 110, 2020, doi: 10.1016/j.artmed.2020.101977.

[12] M. Momenzadeh, M. Sehhati, and H. Rabbani, "Using hidden markov model to predict recurrence of breast cancer based on sequential patterns in gene expression profiles," *Journal of Biomedical Informatics*, vol. 111, 2020, doi: 10.1016/j.jbi.2020.103570.

[13] V. P. C. Magboo and M. S. A. Magboo, "Machine learning classifiers on breast cancer recurrences," *Procedia Computer Science*, vol. 192, pp. 2742–2752, 2021, doi: 10.1016/j.procs.2021.09.044.

[14] A. Alzu'bi, H. Najadat, W. Doulat, O. Al-Shari, and L. Zhou, "Predicting the recurrence of breast cancer using machine learning algorithms," *Multimedia Tools and Applications*, vol. 80, no. 9, pp. 13787–13800, 2021, doi: 10.1007/s11042-020-10448-w.

[15] M. Alwohaibi, M. Alzaqebah, N. M. Alotaibi, A. M. Alzahrani, and M. Zouch, "A hybrid multi-stage learning technique based on brain storming optimization algorithm for breast cancer recurrence prediction," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, pp. 5192–5203, Sep. 2022, doi: 10.1016/j.jksuci.2021.05.004.

[16] M. Ebrahim, A. A. H. Sedky, and S. Mesbah, "Accuracy assessment of machine learning algorithms used to predict breast cancer," *Data*, vol. 8, no. 2, p. 35, Feb. 2023, doi: 10.3390/data8020035.

[17] D. Dua and C. Graff, "UCI machine learning repository," *The University of California, School of Information and Computer Science*, 2010. http://archive.ics.uci.edu/ml.

[18] M. Zwitter and M. Soklic, "Breast cancer data," Institute of Oncology, University Medical Centre Ljubljana, 1988.

[19] M. Shobana *et al.*, "Classification and detection of mesothelioma cancer using feature selection-enabled machine learning technique," *BioMed Research International*, vol. 2022, pp. 1–6, Jul. 2022, doi: 10.1155/2022/9900668.

[20] R. Zhao, Y. Mu, L. Zou, and X. Wen, "A hybrid intrusion detection system based on feature selection and weighted stacking classifier," *IEEE Access*, vol. 10, pp. 71414–71426, 2022, doi: 10.1109/ACCESS.2022.3186975.

[21] X. S. Yang, M. Karamanoglu, and X. He, "Flower pollination algorithm: a novel approach for multiobjective optimization," *Engineering Optimization*, vol. 46, no. 9, pp. 1222–1237, Sep. 2014, doi: 10.1080/0305215X.2013.832237.

[22] S. A. F. Sayed, E. Nabil, and A. Badr, "A binary clonal flower pollination algorithm for feature selection," *Pattern Recognition Letters*, vol. 77, pp. 21–27, Jul. 2016, doi: 10.1016/j.patrec.2016.03.014.

[23] M. Abdel-Basset and L. A. Shawky, "Flower pollination algorithm: a comprehensive review," *Artificial Intelligence Review*, vol. 52, no. 4, pp. 2533–2557, Dec. 2019, doi: 10.1007/s10462-018-9624-4.

[24] A. Pati, A. Panigrahi, D. S. K. Nayak, G. Sahoo, and D. Singh, "Predicting pediatric appendicitis using ensemble learning techniques," *Procedia Computer Science*, vol. 218, pp. 1166–1175, 2022, doi: 10.1016/j.procs.2023.01.095.

[25] A. Pati *et al.*, "FOHC: firefly optimizer enabled hybrid approach for cancer classification," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 7s, pp. 118–125, Jul. 2023, doi: 10.17762/ijritcc.v11i7s.6983.

[26] M. A. Elsadig, A. Altigani, and H. T. Elshoush, "Breast cancer detection using machine learning approaches: a comparative study," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 1, pp. 736–745, Feb. 2023, doi: 10.11591/ijece.v13i1.pp736-745.

BIOGRAPHIES OF AUTHORS

Mr. Ghanashyam Sahoo is currently a research scholar in the Department of Computer Science and Engineering, ITER, Siksha 'O' Anusandhan (Deemed to be University), Odisha, India. He completed his M. Tech. in CSE from KIIT University Odisha, India 2008. His research interests include machine learning, statistical computing, deep learning, wireless sensor networks, IoT. He has authored a book, "data structure using C," and has three publications. He can be contacted at email: ghanarvind@gmail.com.

Dr. Ajit Kumar Nayak is the Professor and HoD of the Department of Computer Science and Information Technology, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha. He graduated in Electrical Engineering from the Institution of Engineers, India, in 1994, M. Tech. and Ph.D. in Computer Science from Utkal University in 2001 and 2010, respectively. His research interests include computer networking, ad hoc and sensor networks, ML, natural language computing, speech and image processing. He has published about 70 research papers in various journals and conferences. He can be contacted at email: ajitnayak@soa.ac.in.

Dr. Pradyumna Kumar Tripathy completed his M.Tech. and Ph.D. in Computer Science from Utkal University, India, in 2007 and 2015 respectively. He is an Associate Professor in the Computer Science and Engineering Department at Silicon Institute of Technology, Bhubaneswar, India. His research interests include reliability analysis of interconnection networks, parallel distributed systems, topological optimization of interconnection networks, IoT, and machine learning. He can be contacted at email: pradyumnatripathy@gmail.com.

Ms. Jyotsnarani Tripathy has been an Assistant Professor in CSE-AIML and IoT, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, Telangana, India since August 2022. She is pursuing her Ph.D. from Maharaja Sriram Chandra Bhanjadeo University, Baripada, Odisha, India. She was awarded an M.E degree from Utkal University, Bhubaneswar, Odisha, 2008 and got a B.E. Degree from ABIT, Cuttack, Odisha, in 2006. Her research interest lies in image processing, machine learning. She can be contacted at email: jtjyotsna@gmail.com.