International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7419
***Article Received**: 05 May 2023 **Revised:** 28 June 2023 **Accepted:** 23 July 2023*

En-PaFlower: An Ensemble Approach using PSO and Flower Pollination Algorithm for Cancer Diagnosis

Sudhir Kumar Senapati[1], Manish Shrivastava[2], Satyasundara Mahapatra[3]
[1]Department of Computer Science and Engineering
Vivekananda Global University
Jaipur, India
sudhir.aricent@gmail.com
[2]Department of Computer Science and Engineering
Vivekananda Global University
Jaipur, India
manish.shrivastava@vgu.ac.in
[3]Department of Computer Science and Engineering
PSIT
Kanpur, India
satyasundara123@gmail.com

Abstract— Machine learning now is used across many sectors and provides consistently precise predictions. The machine learning system is able to learn effectively because the training dataset contains examples of previously completed tasks. After learning how to process the necessary data, researchers have proven that machine learning algorithms can carry out the whole work autonomously. In recent years, cancer has become a major cause of the worldwide increase in mortality. Therefore, early detection of cancer improves the chance of a complete recovery, and Machine Learning (ML) plays a significant role in this perspective. Cancer diagnostic and prognosis microarray dataset is available with the biopsy dataset. Because of its importance in making diagnoses and classifying cancer diseases, the microarray data represents a massive amount. It may be challenging to do an analysis on a large number of datasets, though. As a result, feature selection is crucial, and machine learning provides classification techniques. These algorithms choose the relevant features that help build a more precise categorization model. Accurately classifying diseases is facilitated as a result, which aids in disease prevention. This work aims to synthesize existing knowledge on cancer diagnosis using machine learning techniques into a compact report. Current research work aims to propose an ensemble-based machine learning model En-PaFlower using Particle Swarm Optimization (PSO) as the feature selection algorithm, Flower Pollination algorithm (FPA) as the optimization algorithm with the majority voting algorithm. Finally, the performance of the proposed algorithm is evaluated over three different types of cancer disease datasets with accuracy, precision, recall, specificity, and F-1 Score etc as the evaluation parameters. The empirical analysis shows that the proposed methodology shows highest accuracy as 95.65%.

Keywords- Cancer; Machine Learning; Particle Swarm Optimization (PSO); Flower Pollination algorithm (FPA); majority voting.

I. INTRODUCTION

One of the medical industry's most cutthroat and competitive disciplines is cancer research. It's a disease characterized by the uncontrolled proliferation of cells with the ability to invade other parts of the body. Cancer is a disease in which the patient's DNA may undergo changes. Cancer cells often display much more genetic modifications than normal cells do; nonetheless, malignant tumors typically display unique, explicitly combining genetic aberrations in different people [1]. Despite the fact that malignant tumors have more genetic alterations than healthy cells, this is the case. In 2012, there were around 12 million reported cases, and of them, approximately 7 million persons had died due to the lack of a suitable diagnosis process, as determined by statistical analysis of the NCI and WHO data sheet. Over the next two decades, the patient population grew by around 70%, and roughly one in six of them died due to an insufficient diagnostic model. In 2015, the number of deaths is predicted to have reached 8.8 million, and by 2018, that number is expected to have risen to over 9.6 million [2]. The development of this death troll is aided by the lack of a reliable diagnostic paradigm. This has led to the categorization of cancers becoming an increasingly important area of research in the medical community. This is because there is a significant possibility of patient survival if cancer is properly classified, detected, and diagnosed at an early stage. However, with the help of accurate diagnosis and effective treatment, cancer mortality rates may be lowered.

There are two sorts of data sets that may be used in cancer diagnosis. The microarray data set and the biopsy fall under this

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7419
***Article Received**: 05 May 2023 **Revised:** 28 June 2023 **Accepted:** 23 July 2023*

category. The patient's genetic information is not included in the Biopsy dataset, but the patient's laboratory test results are. Because genetic information plays such a crucial role in cancer diagnosis, biopsy data does influence cancer diagnosis in a meaningful way [3].

The MT presents a fantastic chance to learn more about the genetic basis of illness. Gene expression data, on the other hand, include high dimensionalities that are unimportant while looking for diseases. We cannot trust the high-dimensional gene expression data since it contains redundant information and is thus useless [4]. The inability to process all data at once, as well as the possibility that its subset data processing may lead to the reasons behind over-fitting, information loss, and other such issues, can all be briefly described as limitations of microarray data that are directly affecting the accuracy of classification [5,6].

Both the detection and outlook for cancer improve greatly. However, when the model is given the data from the biopsy, it does not offer an appropriate answer due to the lack of genetic information in the test result. In contrast, microarray data may be quickly analysed using ML for better diagnostics. Microarray data presents challenges because to its high dimensional nature, which must be surmounted before the data can be used for categorization [7]. Because microarray data includes a large quantity of genetic information but a limited number of samples, applying machine learning algorithms to it presents a unique challenge known as the small sample size problem. This problem will impact the quality of the microarray data. The most critical component in generating reliable diagnostic findings is having a correct interpretation of the microarray data, which is notoriously difficult to deal with. Researchers have shown that dimensionality reduction, a technique through which high-dimensional data may be reduced to a considerable dimension [8], is the key to solving this issue. Researchers have identified dimensionality reduction as the answer to this issue. It is possible to further divide the dimensionality reduction process into two sub-processes. Selecting the most useful and relevant features to enhance classification is the goal of feature selection, also known as a variable selection; feature extraction, on the other hand, aims to reduce the number of features by combining them in a way that does not affect the classification model's accuracy. The process of choosing which features to use is also known as "variable selection."

In the realm of machine learning, the ideas of classification and prediction rely on many different components of datasets, each of which is crucial to the classification and prediction process [9]. The features are the distinguishing traits of anything. The vast amount of data points in the actual dataset has been a major focus throughout the testing and training phases. Dimensionality reduction is often used in the microarray dataset to get around the problem of a small sample size [10]. Dimensionality reduction is being utilized to reduce the feature space prior to applying the classification algorithm. It's a method for transforming a high-dimensional dataset into one with fewer dimensions. The massive increase in the size of the dataset has resulted in the widespread use of many DRTs across a wide range of industries. In addition, there is constant innovation in the form of novel procedures. The DRT method may be used to extract the most relevant information from high-dimensional data for classification or prediction while disregarding less important information [11]. Using low-dimensional data may help address the issue of high-dimensional data, which hinders the effectiveness of any machine learning model. DRT plays a crucial role in microarray data processing because of the ease with which low-dimensional information may be generated, investigated, and evaluated [12].

A. Motivation

With only partial information about the cancer tissues available from the biopsy data set, the conventional diagnostic approach poses a barrier to successful treatment. As a result, the micro-array data will contain genetic information, which is essential for improving cancer detection. The fundamental benefit of using the data from the microarrays as input to the diagnostic model is that a large number of the genetic behaviors of the tissues may be taken into account with the aim of producing a more accurate diagnosis. Data from a DNA microarray will have high dimensionality since it will have a wealth of information about the expression of genes involved in cancer but will be lacking in particulars about individual samples. Analyzing the data from microarrays is a never-ending effort because of the large number of variables and the noise that is typically connected with the observations. This method generates hundreds of variables because of its high throughput, and each one has to be thoroughly evaluated in order to deliver useful results.

B. Objective

The objectives of this work are listed below:

- To analyze the concept of microarray data
- To implement the Particle Swarm Optimization (PSO) algorithm as a feature selection algorithm to deal with the high dimensionality issue of the microarray dataset.
- To employ the Flower Pollination Algorithm (FPA) as the optimization algorithm to select more relevant features.
- To develop an ML-based ensemble method using a majority voting technique for effective cancer diagnosis.
- To evaluate the performance of the proposed model over 3 different microarray datasets with 7 different evaluation parameters.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7419
***Article Received**: 05 May 2023 **Revised:** 28 June 2023 **Accepted:** 23 July 2023*

C. Structure of the Paper

The structure of the paper can be summarized as follows. Section 1 shows the introduction of this work along with the motivation and objective of the work. Section 2 shows the literature review of the work. Section 3 shows the proposed work in addition to the dataset description and methodologies used in the work. Section 4 shows the empirical analysis of the research work. Finally, the overall conclusion and future scope is presented in Section 5.

II. RELATED WORK

Zhu and Hastie [13] have introduced penalized logistic regression (PLR) using RFE and univariate ranking (UR). From experiments, it is observed that the proposed model outperforms others in terms of feature selection, test samples, and cross-validation on Microarray datasets.

Zhou et al [14] have mentioned a logistic regression-oriented Bayesian technique for classification and feature selection. Some experiments are performed on various Microarray datasets namely, acute leukemia, small round blue cell tumors, and hereditary breast cancer etc. It is shown from the results that effectiveness in identifying important features and classification accuracy by this proposed model.

In order to classify and predict microarray data in a prostate cancer dataset, Zhao et al. [15] used the PLR method in conjunction with the top score pair (TSP) method. They then compared the results with those obtained using Lasso, fisher discriminative analysis (FDA), and support vector machines (SVMs). The authors claim that their suggested method is superior to others in terms of both classification and prediction performance.

Morais-Rodrigues et al [16] have mentioned a technique using logistic regression for breast cancer classification on gene expression omnibus (GEO) data series. The authors considered all the features, without reducing any, in this proposed model and claimed that of achieving better performance in comparison with others.

Fan et al [17] have presented a naïve Bayes-based sequential feature extraction model for Microarray data classification. Some experiments are performed on 5 Microarray datasets and claimed that of achieving comparatively an enhanced performance by this proposed model.

Wu et al [18] have proposed a Laplace naïve Bayes with mean shrinkage (LNB-MS) model to identify biomarkers and to classify cancer based on Microarray data. From experiments, it is shown as the guarantee of gene selection by this proposed model.

Nagi and Bhattacharya [19] have proposed an ensemble method termed as SD-EnClass for the classification of Microarray cancer data. The enhanced classification accuracy achieved through this proposed approach is combined with stacking, bagging, and boosting methods to obtain more improved performance in terms of classification accuracy.

Mahfouz et al [20] have introduced a decisive ensemble classifier, derived from KNN, to overcome the problems associated with imbalanced small-sized datasets. The authors claimed that of achieving enhanced accuracy in comparison with base classifiers on the datasets namely, CNS, Leukemia, Notterman, GDS3257, and Kentridge etc.

Maulik and Chakraborty [21] have introduced a predictive model based on the combination of semi-supervised SVMs and fuzzy preference-based rough set (FPRS) for the selecting genes or features. The authors claimed of achieving significant success in the discovery of drugs and diagnosing cancer in comparison with consistency-based feature selection (CBFS) and signal-to-noise ratio (SNR) methods considering 6 – Microarray datasets in both multiclass and binary classification problems.

Huo et al [22] have proposed an SGL-SVM model based on Sparse Group Lasso and Support Vector Machine. The authors validated this method with Microarray and NGS datasets. From experiments, it is concluded that achieving good classification accuracy on selected featured genes for high dimensional and small tumor datasets classification.

A novel bio-marker gene selection algorithm has been developed in [23]. In order to develop the model, the firefly algorithm (FF) in addition to the SVM is used. F-score filter method is used to enhance the performance level of the FF-SVM model. The proposed model F-score FF-SVM model is evaluated over 5 different cancer microarray datasets.

III. PROPOSED WORK

For the current work Particle Swarm Optimization (PSO) algorithm has been applied as the feature selection algorithm to deal with the high dimensionality issue of the microarray data. To select more relevant features the Flower Pollination Algorithm (FPA) is used as the optimization algorithm. For the selected features the SVM, MLP, and RF have been used as the base learners. Finally, the majority voting technique is applied as an ensemble learning technique to enhance the performance of the proposed En-PaFlower. The workflow of this proposed CorPaSVM is reflected in Figure 1.

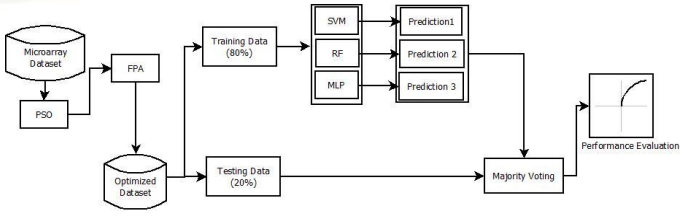

Figure 1. Workflow of proposed work

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7419
***Article Received**: 05 May 2023 **Revised:** 28 June 2023 **Accepted:** 23 July 2023*

A. Support Vector Machine (SVM)

SVM is a supervised ML algorithm with the objective to find a hyperplane in a multidimensional space. The support vector machine's goal is to provide the best possible hyperplane such that the two classes can be distinguished from one another. This ideal hyperplane not only differentiates the two groups but also increases the gap that exists between them to its greatest extent. The gap that exists between the hyperplane and the SVs is what is referred to as the margin. It is very popular due to its primary advantage such as it can be a very effective one even in high dimensional space but the primary issue present in this approach is that it does not provide probabilistic estimations. High accuracy can be achieved by tuning the hyperparameters such as gamma, coat, and kernel level but in reality, it becomes very hard to define the exact hyperparameters which directly enhances the computational cost and overhead [24].

B. Random Forest (RF)

Multiple decision trees form the basis of the random forest algorithm. Bagging and bootstrap aggregation are used to train the 'forest' produced by the random forest algorithm. Bagging is a meta-algorithm for improving the performance of machine learning algorithms by working as an ensemble. Based on the decision trees' predictions, the (random forest) algorithm determines the result. Predictions are made by averaging the results from several trees. A more accurate result may be achieved by using a larger number of trees. When compared to a random forest, a decision tree method becomes obsolete. It improves accuracy by decreasing the likelihood of a dataset overfitting [25].

C. Multi-Layer Perception (MLP)

MLP is a feedforward network that takes in many inputs and generates multiple outputs. Each node in the input, hidden, and output layers of a traditional MLP is connected to every other node in the layer below and the layer above it. Each node's outputs are weighted units that are subsequently sent through a nonlinear activation function in order to distinguish between data that cannot be linearly separated [26,27].

D. Particle Swarm Optimization (PSO)

Particle Swarm Optimization, sometimes known as PSO, is an approach for stochastic optimization that is population-based. This replicated the physical motions that the members in the swarm would make in order to look for something. It is helpful to think of each particle as a point inside a D-dimensional searching region; these points have two qualities, the first of which is their location, and the second of which is their velocity. In this strategy, the goal function is responsible for making adjustments to the location and velocity of each and every particle in the population [28]. This is done in order to get satisfactory outcomes from the simulation. In a broad sense, this algorithm takes use of social interaction that occurs between different groups. The speeds at which the particles go through the search region are constantly shifting due to the dynamic nature of the search. Each particle will update two absolute values during each iteration. These values are the best position that the particle has achieved to date (it's personal best), as well as the best position that the swarm has achieved during that time period (global best) [29].

E. Flower Pollination Algorithm (FPA)

For worldwide optimum performance, the flower pollination algorithm (FPA) is presented. The natural phenomena of wind pollination of moving plants served as inspiration for this novel metaheuristic algorithm. Pollination is the process by which pollen is transferred from the flagellum of an individual flower to the pistil stigma of that bloom or another flower of the same plant species. Both vegetative and reproductive cells may be found in pollen. The vegetative cell divides and grows into a pollen tube once pollen is placed on the stigma of the pistil. A reproductive cell may split in half along its patch and send out two separate pollen tubes to the ovary. A zygote is created when an egg cell fertilises one of the reproductive cells. The resulting zygote is the beginning of a whole new plant. Based on the mechanisms used to spread pollen, we may classify pollination as either biotic or abiotic. Insects and other animals provide biotic pollination for the vast majority of blooming plants. Abiotic pollination, on the other hand, does not rely on the movement of live creatures to spread pollen. Instead, pollination is accomplished without the use of insects. Cross-pollination happens when pollens from one plant are transferred to another of the same species, whereas self-pollination occurs when pollen is transferred to the same flower or blooms of the same species [30].

Domain optimization defines and incorporates the techniques of biotic pollination, cross-pollination, abiotic pollination, and self-pollination into the flower pollination algorithm. There is a complicated set of processes at work in plant production strategies, and pollination is one of them. The optimisation challenge may be solved by the combination of a flower and the gametes it produces. The flower's consistency as a custom fit is obvious [31]. With global pollination, bees and other insects travel great distances to disperse pollen from one flower to another. Contrarily, local pollination occurs inside a restricted region of a particular bloom when it is shaded by wind or water. Switch probability describes the likelihood of global pollination occurring. When this process is skipped, local pollination takes its place [32].

Global pollination and local pollination are the two most important phases of this algorithm. In global pollination insects play an important role as pollinators, and the pollen they collect

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7419
***Article Received**: 05 May 2023 **Revised:** 28 June 2023 **Accepted:** 23 July 2023*

from flowers may travel far away due to the insects' wings and mobility. It can be represented by equation 1.

$$f_i^{t+1} = f_i^t + \mathcal{L}(f_i^t + \rho) \quad (1)$$

f_i^t is the pollen i from the solution vector f at t[th] iteration. ρ is the current best global solution. $\mathcal{L}$ is the pollination strength so that $\mathcal{L} > 0$.

F. Dataset Description

For the current research work the 3 Cancer microarray datasets from the UCI machine learning data repository have been considered. Table 1 shows the dataset description. Figure 1, 2, and 3 shows the class count, head, and tail description of the dataset. Figure 2-4 shows the class count for Breast Cancer, Lung Cancer, and Ovarian Cancer respectively.

TABLE I. DATASET DESCRIPTION

Dataset	Feature	Sample	Distribution
Breast Cancer	24481	97	Relapse- 46 Non-relapse- 51
Lung Cancer	12533	203	Cancer-139 Noncancer-64
Ovarian Cancer	15154	253	Normal- 91 Cancer-161

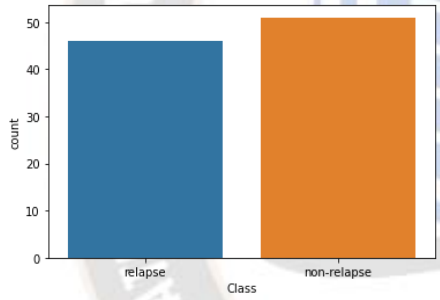

Figure 2. Class count for the Breast Cancer Dataset.

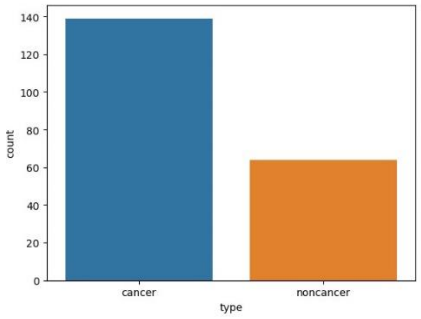

Figure 3. Class count for the Lung Cancer Dataset.

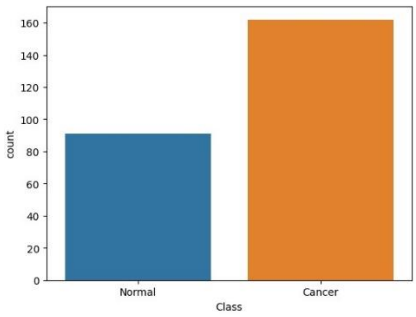

Figure 4. Class count for the Lung Cancer Dataset.

IV. RESULT AND DISCUSION

The proposed system has been implemented with a system having 8GB RAM, Windows 10 OS, Intel i5 processor with 2.3GHz clock speed, and 1TB HDD, SSD. The performance of the proposed system has been evaluated based on the evaluating parameter such as accuracy (A_c), mis-classification rate (M_r) precision (P_r), recall (R_e), F-1 Score ($F1_s$), and Mathew's corelation coefficient (M_c) with true positive (TRP), true negative (TRN), false positive (FLP), and false negative (FLN) [33-35]. The above-said parameters can be calculated by using equations 2-8.

$$A_c = \frac{TRP+TRN}{TRP+TRN+FLP+FLN} \quad (2)$$

$$M_r = (1 - A_c) * 100\% \quad (3)$$

$$P_r = \frac{TRP}{TRP+FLP} \quad (4)$$

$$R_e = \frac{TRP}{TRP+FLN} \quad (5)$$

$$F1_s = \frac{2*\frac{TRP}{TRP+FLP}*\frac{TRP}{TRP+FLN}}{\frac{TRP}{TRP+FLP}+\frac{TRP}{TRP+FLN}} \quad (6)$$

$$SP_e = \frac{FLN}{TRN+FLP} \quad (7)$$

$$M_c = \frac{(TRP*TRN)-(FLP*FLN)}{\sqrt{(TRp+FLP)(TRP+FLN)(TRN+FLP)(TRN+FLN)}} \quad (8)$$

TABLE II. PERFORMANCE EVALUATION OF VARIOUS PROPOSED HYBRID MODELS

Dataset	Methodology	A_c	M_r	P_r	R_e	$F1_s$	SP_e	Mc
Breast	PSO+FPA+SVM	77.32	22.68	71.15	84.09	77.08	71.70	55.69
	PSO+FPA+MLP	80.41	19.59	79.25	84.00	81.55	76.60	60.83
	PSO+FPA+RF	79.38	20.62	82.69	79.63	81.13	79.07	58.47
Lung	PSO+FPA+SVM	78.82	21.18	73.04	87.50	79.62	71.03	58.97
	PSO+FPA+MLP	77.34	22.66	77.31	82.88	80.00	70.65	54.11
	PSO+FPA+RF	79.31	20.69	78.87	90.32	84.21	62.03	55.67
Ovarian	PSO+FPA+SVM	76.28	23.72	82.68	83.62	83.15	59.21	43.16
	PSO+FPA+MLP	79.45	20.55	82.39	87.35	84.80	64.37	53.39
	PSO+FPA+RF	78.26	21.74	82.53	84.05	83.28	67.78	52.23

TABLE III. PERFORMANCE EVALUATION OF PROPOSED EN-PAFLOWER

Dataset	A_c	M_r	P_r	R_e	$F1_s$	SP_e	Mc
Breast	93.81	6.19	92.73	96.23	94.44	90.91	87.55
Lung	93.10	6.90	91.92	93.81	92.86	92.45	86.21
Ovarian	95.65	4.35	96.41	96.99	96.70	93.10	90.34

Table 2 and 3 quantifies the evaluative parameters for different proposed hybrid models including PSO+FPA+SVM, PSO+FPA+MLP, and PSO+FPA+RF along with the proposed En-PaFlower respectively. 5-10 shows the performance comparison of the proposed work in contrast to different classifiers.

The empirical analysis shows that the performance of proposed En-PaFlower shows accuracy for Breast Cancer, Lung Cancer, and Ovarian Cancer as 93.81%, 93.10%, and 95.65%

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7419
***Article Received**: 05 May 2023 **Revised:** 28 June 2023 **Accepted:** 23 July 2023*

respective. It can be clearly depicted that the En-PaFlower outperforms all other proposed hybrid models. For Ovarian Cancer the proposed model shows highest accuracy with precision, recall, F-1 Score, Specificity as 96.41%, 96.99%, 96.7%, 93.1% respectively.

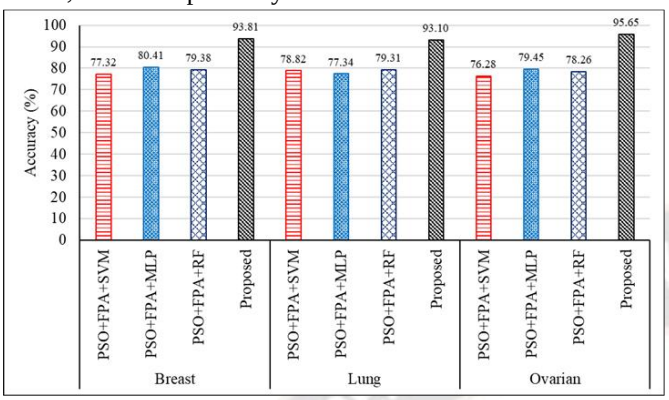

Figure 5. Accuracy Comparison of proposed system in contrast to different hybrid methods.

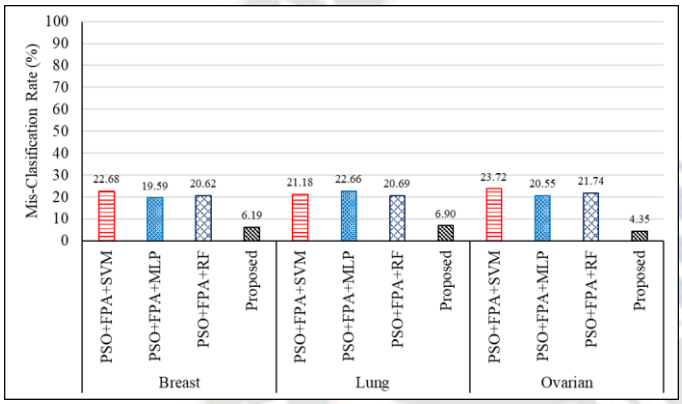

Figure 6. Mis-Classification rate (Mc) Comparison of the proposed system in contrast to different hybrid methods.

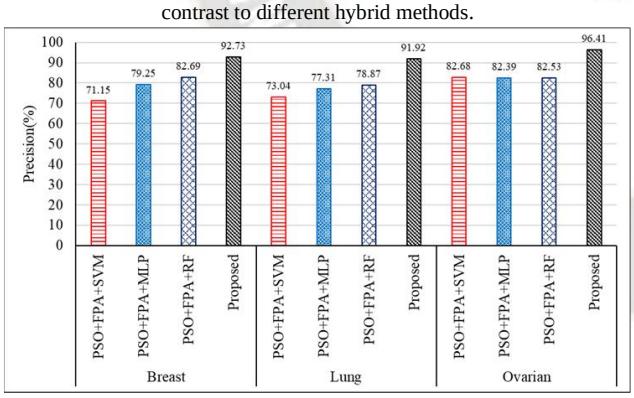

Figure 7. Precision Comparison of the proposed system in contrast to different hybrid models.

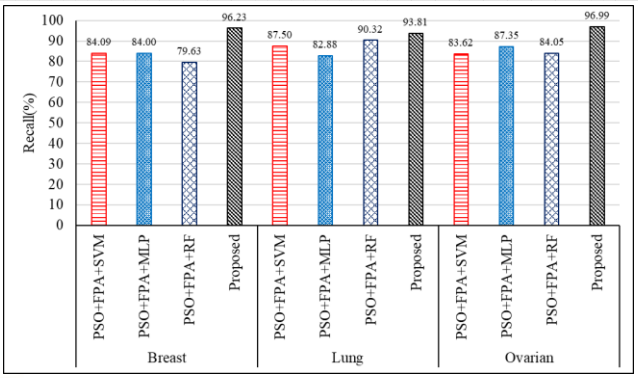

Figure 8. Recall Comparison of the proposed system in contrast to different hybrid models.

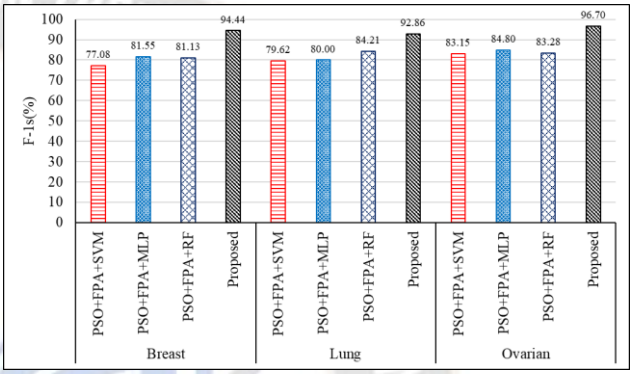

Figure 9. F-1 Score Comparison of the proposed system in contrast to different hybrid models.

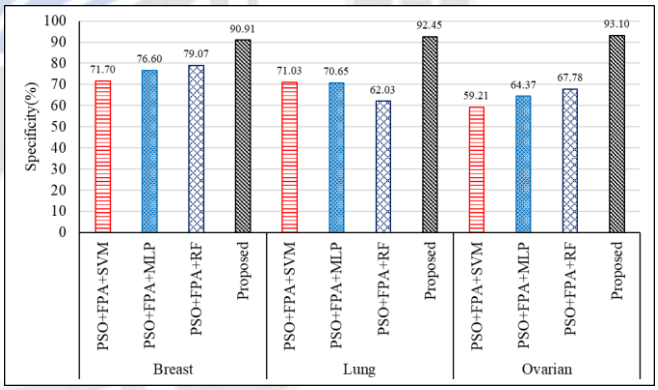

Figure 10. Specificity Comparison of the proposed system in contrast to different hybrid models

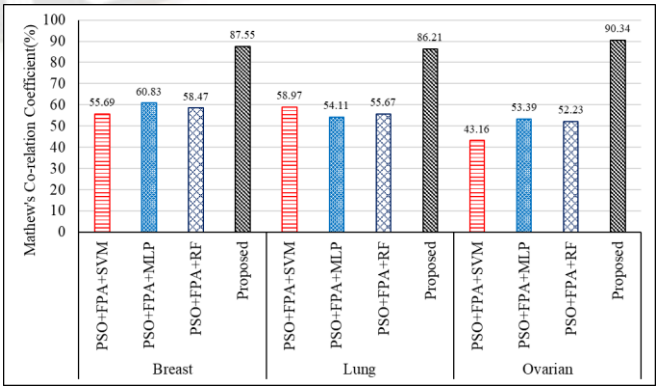

Figure 11. Mathew's Co-Relation Coefficient Comparison of the proposed system in contrast to different hybrid models

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7419
***Article Received**: 05 May 2023 **Revised:** 28 June 2023 **Accepted:** 23 July 2023*

V. CONCLUSION

The current research work aims to develop an effective ML-based ensemble model for cancer diagnosis. The PSO is employed as the feature selection algorithm to deal with the high dimensionality issue of the microarray data. Whereas the FPA is used as the optimization algorithm for selecting the most relevant feature set from the featured dataset. Finally, SVM, MLP, and RF classification algorithms are applied as the base learners. To the initial prediction the majority voting ensemble technique is used to develop the ensemble En-PaFlower. The proposed En-PaFlower is evaluated over 3 different cancer microarray dataset. The empirical analysis shows that the proposed En-PaFlower outperforms all other proposed hybrid methodologies PSO+FPA+SVM, PSO+FPA+MLP, and PSO+FPA+RF. For Breast Cancer dataset the En-PaFlower outperforms the above PSO+FPA+SVM, PSO+FPA+MLP, and PSO+FPA+RF by ~21.33%, ~16.67% and ~18.18% respectively in terms of accuracy. In case of Lung Cancer the proposed En-PaFlower outperforms the PSO+FPA+SVM, PSO+FPA+MLP, and PSO+FPA+RF by ~18.18%, ~20.38%, and ~17.39% respectively in terms of accuracy. Similarly, for Ovarian Cancer the En-PaFlower outperforms the PSO+FPA+SVM, PSO+FPA+MLP, and PSO+FPA+RF by ~25.4%, 20.39%, and ~22.22% respectively in terms of accuracy.

The future scope of this work is to test the robustness of the proposed ensemble model with different microarray datasets. The highest accuracy level obtained using En-PaFlower is 95.65% for Ovarian Cancer dataset. The performance level may be increased by implementing other metaheuristic algorithms as the optimizer.

REFERENCES

[1] M. Khalsan et al., "A Survey of Machine Learning Approaches Applied to Gene Expression Analysis for Cancer Prediction," in IEEE Access, vol. 10, pp. 27522-27534, 2022, doi: 10.1109/ACCESS.2022.3146312.

[2] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, “Cancer statistics, 2023.” CA: A Cancer Journal for Clinicians, vol. 73, no. 1, pp. 17-48, 2023, doi: 10.3322/caac.21763.

[3] W. Astuti and Adiwijaya, “Support vector machine and principal component analysis for microarray data classification.” Journal of Physics: Conference Series, vol. 971, p. 012003, 2018, doi: 10.1088/1742-6596/971/1/012003.

[4] I. M. Zubair and B. Kim, "A Group Feature Ranking and Selection Method Based on Dimension Reduction Technique in High-Dimensional Data," in IEEE Access, vol. 10, pp. 125136-125147, 2022, doi: 10.1109/ACCESS.2022.3225685.

[5] S. K. Senapati, M. Shrivastava and S. Mahapatra, "PCSVM: A Hybrid Approach using Particle Swarm Optimization and Cuckoo Search for Effective Cancer Diagnosis," 2023 2nd Edition of IEEE Delhi Section Flagship Conference (DELCON), Rajpura, India, 2023, pp. 1-5, doi: 10.1109/DELCON57910.2023.10127354.

[6] K. R. Kavitha, A. V. Ram, S. Anandu, S. Karthik, S. Kailas and N. M. Arjun, "PCA-based gene selection for cancer classification," 2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Madurai, India, 2018, pp. 1-4, doi: 10.1109/ICCIC.2018.8782337.

[7] A. Pati, M. Parhi, B. K. Pattanayak, B. Sahu, and S. Khasim, “CanDiag: Fog Empowered Transfer Deep Learning Based Approach for Cancer Diagnosis,” Designs, vol. 7, no. 3, p. 57, Apr. 2023, doi: 10.3390/designs7030057.

[8] A. Pati et al., “Breast Cancer Diagnosis Based on IoT and Deep Transfer Learning Enabled by Fog Computing,” Diagnostics, vol. 13, no. 13, p. 2191, Jun. 2023, doi: 10.3390/diagnostics13132191.

[9] A. Pati et al., “Diagnose Diabetic Mellitus Illness Based on IoT Smart Architecture,” Wireless Communications and Mobile Computing, vol. 2022, p. e7268571, Aug. 2022, doi: https://doi.org/10.1155/2022/7268571.

[10] Gawande, G. S. ., Kanwade, A. B. ., Deshmukh, A. B. ., Bhandari, S. ., Bendre, V. ., & Dakre, A. G. . (2023). IoT-based Weather Information Using WeMos. International Journal of Intelligent Systems and Applications in Engineering, 11(3s), 85–92. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2534

[11] A. Pati, M. Parhi, and B. K. Pattanayak, “A review on prediction of diabetes using machine learning and data mining classification techniques,” International Journal of Biomedical Engineering and Technology, vol. 41, no. 1, p. 83, 2023, doi: https://doi.org/10.1504/ijbet.2023.128514.

[12] 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, 2023, doi: https://doi.org/10.1016/j.procs.2023.01.095.

[13] W. Jia, M. Sun, J. Lian, and S. Hou, “Feature dimensionality reduction: a review,” Complex & Intelligent Systems, Jan. 2022, doi: https://doi.org/10.1007/s40747-021-00637-x.

[14] J. Zhu, “Classification of gene microarrays by penalized logistic regression,” Biostatistics, vol. 5, no. 3, pp. 427–443, Jul. 2004, doi: https://doi.org/10.1093/biostatistics/kxg046

[15] X. Zhou, K.-Y. Liu, and S. T. C. Wong, “Cancer classification and prediction using logistic regression with Bayesian gene selection,” Journal of Biomedical Informatics, vol. 37, no. 4, pp. 249–259, Aug. 2004, doi: https://doi.org/10.1016/j.jbi.2004.07.009.

[16] H. Zhao, S. Qi and Q. Dong, "Predicting prostate cancer progression with penalized logistic regression model based on co-expressed genes," 2012 5th International Conference on BioMedical Engineering and Informatics, Chongqing, China, 2012, pp. 976-980, doi: 10.1109/BMEI.2012.6512948.

[17] F. Morais-Rodrigues et al., “Analysis of the microarray gene expression for breast cancer progression after the application modified logistic regression,” Gene, vol. 726, p. 144168, Feb. 2020, doi: https://doi.org/10.1016/j.gene.2019.144168.

[18] [1]L. Fan, K.-L. Poh, and P. Zhou, “A sequential feature extraction approach for naïve bayes classification of microarray data,” Expert Systems with Applications, vol. 36, no. 6, pp. 9919–

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7419
***Article Received**: 05 May 2023 **Revised:** 28 June 2023 **Accepted:** 23 July 2023*

9923, Aug. 2009, doi: https://doi.org/10.1016/j.eswa.2009.01.075.

[19] M. -Y. Wu, D. -Q. Dai, Y. Shi, H. Yan and X. -F. Zhang, "Biomarker Identification and Cancer Classification Based on Microarray Data Using Laplace Naive Bayes Model with Mean Shrinkage," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 9, no. 6, pp. 1649-1662, Nov.-Dec. 2012, doi: 10.1109/TCBB.2012.105.

[20] S. Nagi and D. Kr. Bhattacharyya, "Classification of microarray cancer data using ensemble approach," Network Modeling Analysis in Health Informatics and Bioinformatics, vol. 2, no. 3, pp. 159–173, Jun. 2013, doi: https://doi.org/10.1007/s13721-013-0034-x.

[21] M. A. Mahfouz, A. Shoukry, and M. A. Ismail, "EKNN: Ensemble classifier incorporating connectivity and density into kNN with application to cancer diagnosis," Artificial Intelligence in Medicine, vol. 111, p. 101985, Jan. 2021, doi: https://doi.org/10.1016/j.artmed.2020.101985.

[22] U. Maulik and D. Chakraborty, "Fuzzy Preference Based Feature Selection and Semisupervised SVM for Cancer Classification," in IEEE Transactions on NanoBioscience, vol. 13, no. 2, pp. 152-160, June 2014, doi: 10.1109/TNB.2014.2312132.

[23] Y. Huo, L. Xin, C. Kang, M. Wang, Q. Ma, and B. Yu, "SGL-SVM: A novel method for tumor classification via support vector machine with sparse group Lasso," Journal of Theoretical Biology, vol. 486, p. 110098, Feb. 2020, doi: https://doi.org/10.1016/j.jtbi.2019.110098.

[24] N. Almugren and H. M. Alshamlan, "New Bio-Marker Gene Discovery Algorithms for Cancer Gene Expression Profile," in IEEE Access, vol. 7, pp. 136907-136913, 2019, doi: 10.1109/ACCESS.2019.2942413.

[25] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," Neurocomputing, vol. 408, pp. 189–215, Sep. 2020, doi: https://doi.org/10.1016/j.neucom.2019.10.118.

[26] Maria Gonzalez, Machine Learning for Anomaly Detection in Network Security , Machine Learning Applications Conference Proceedings, Vol 1 2021.

[27] B. Dai, R. -C. Chen, S. -Z. Zhu and W. -W. Zhang, "Using Random Forest Algorithm for Breast Cancer Diagnosis," 2018 International Symposium on Computer, Consumer and Control (IS3C), Taichung, Taiwan, 2018, pp. 449-452, doi: 10.1109/IS3C.2018.00119.

[28] H. -J. Chiu, T. -H. S. Li and P. -H. Kuo, "Breast Cancer–Detection System Using PCA, Multilayer Perceptron, Transfer Learning, and Support Vector Machine," in IEEE Access, vol. 8, pp. 204309-204324, 2020, doi: 10.1109/ACCESS.2020.3036912.

[29] Z. Guo, L. Xu, and N. Ali Asgharzadeholiaee, "A Homogeneous Ensemble Classifier for Breast Cancer Detection Using Parameters Tuning of MLP Neural Network," Applied Artificial Intelligence, pp. 1–21, Jan. 2022, doi: https://doi.org/10.1080/08839514.2022.2031820.

[30] A. Panigrahi, S. Bhutia, B. Sahu, M.G.Galety, and S.N. Mohanty, "BPSO-PSO-SVM: An Integrated Approach for Cancer Diagnosis," pp. 571–579, Jan. 2022, doi: https://doi.org/10.1007/978-981-19-2177-3_53.

[31] B. Sahu, A. Panigrahi, S. Pani, S. Swagatika, D. Singh and S. Kumar, "A Crow Particle Swarm Optimization Algorithm with Deep Neural Network (CPSO-DNN) for High Dimensional Data Analysis," 2020 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2020, pp. 0357-0362, doi: 10.1109/ICCSP48568.2020.9182181.

[32] P. E. Mergos and X.-S. Yang, "Flower pollination algorithm with pollinator attraction," Evolutionary Intelligence, Jan. 2022, doi: https://doi.org/10.1007/s12065-022-00700-7.

[33] Ólafur, S., Nieminen, J., Bakker, J., Mayer, M., & Schmid, P. Enhancing Engineering Project Management through Machine Learning Techniques. Kuwait Journal of Machine Learning, 1(1). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/112

[34] S. Lalljith, I. Fleming, U. Pillay, K. Naicker, Z. J. Naidoo and A. K. Saha, "Applications of Flower Pollination Algorithm in Electrical Power Systems: A Review," in IEEE Access, vol. 10, pp. 8924-8947, 2022, doi: 10.1109/ACCESS.2021.3138518.

[35] O. A. Alomari, A. T. Khader, M. A. Al-Betar and Z. A. Alkareem Alyasseri, "A Hybrid Filter-Wrapper Gene Selection Method for Cancer Classification," 2018 2nd International Conference on BioSignal Analysis, Processing and Systems (ICBAPS), Kuching, Malaysia, 2018, pp. 113-118, doi: 10.1109/ICBAPS.2018.8527392.

[36] B. Sahu and A. Panigrahi, "Efficient Role of Machine Learning Classifiers in the Prediction and Detection of Breast Cancer," SSRN Electronic Journal, 2020, doi: https://doi.org/10.2139/ssrn.3545096.

[37] B. Sahu, A. Panigrahi, S. K. Rout and A. Pati, "Hybrid Multiple Filter Embedded Political Optimizer for Feature Selection," 2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP), Hyderabad, India, 2022, pp. 1-6, doi: 10.1109/ICICCSP53532.2022.9862419.