Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(4s), 542-561 | ISSN:2251-6727

Breast Cancer Prediction Using Hybrid Machine Learning and Nature-Inspired Adaptive Optimization Algorithms

Tamilselvi Madeswaran[1*], Aruna Kumar Kavuru[2], Padma Theagarajan[3], Nasser Al Hadrami[4], Ohm Rambabu[5], Maya Al Foori[6]

[1*]Lecturer, IT department, University of Technology and Applied Sciences, Nizwa, Sultanate of Oman
Email:- tamilselvi.madeswaran@nct.edu.om
[2]Lecturer, IT department,University of Technology and Applied Sciences, Nizwa, Sultanate of Oman
Email:- arun.kavuru@nct.edu.om
[3]Professor, Department of Computer Applications, Sona College of Technology Salem, Tamilnadu, India
Email:- padmat@sonatech.ac.in
[4]Senior Lecturer, IT department,University of Technology and Applied Sciences, Nizwa, Sultanate of Oman
Email:- nasser.almur@nct.edu.om
[5]Technical Support, ETC department,University of Technology and Applied Sciences, Nizwa, Sultanate of Oman
Email:- ram.babu@nct.edu.om
[6]Oncology Nursing and Clinical Educator,National Oncology Center, Royal Hospital, Sultanate of Oman
Email:- lianobaby@gmail.com

***Corresponding Author:-** Tamilselvi Madeswaran
*Lecturer, IT department, University of Technology and Applied Sciences, Nizwa, Sultanate of Oman
Email:- tamilselvi.madeswaran@nct.edu.om

KEYWORDS:

Breast cancer prediction, Intelligent predictors, Hybrid modelling strategies, Adaptive optimization approaches, Nature-inspired optimization, UCI data repository.

ABSTRACT

An early stage of breast cancer occurrence does not have pain as a symptom. This asymptomatic nature of cancer necessitates the need of timely and accurate prediction using other potential indicators. Breast cancers are highly curable if predicted and diagnosed at the earliest. This research explores the capacity of hybridizing intelligent learning models with nature-inspired optimization algorithms for breast cancer prediction. This integration becomes pivotal to optimize internal parameters of the dataset while aiming for augmented accuracy of the proposed predictive models. Each of the four intelligent machine learning models including K-Nearest Neighbor (KNN), Naïve Bayes (NB), Support Vector Machine (SVM), and Artificial Neural Network (ANN) are hybridized with all the three adaptive optimization approaches including Adaptive Particle Swarm Optimization (APSO), Adaptive Genetic Algorithm (AGA) and Adaptive Venus Flytrap Optimization (AVFO). Thus twelve hybridized predictive models were derived namely APSO-KNN, AGA-KNN, AVFO-KNN, APSO-NB, AGA-NB, AVFO-NB, APSO-SVM, AGA-SVM, AVFO-SVM, APSO-ANN, AGA-ANN and AVFO-ANN. These hybridized models were investigated through UCI data repository's breast cancer dataset namely WBC, WDBC and WPBC. The experimental results were validated against the respective learning models with and without feature selection. Based on the performance measures such as accuracy, F-measure and G-mean, the outperformed learning models with WBC, WPBC and WDBC datasets are AVFO-SVM, AVFO-ANN and APSO-ANN respectively. Conclusively ANN and SVM machine learning algorithms fused with AVFO for feature selection are robust for all the three datasets. The derived hybrid intelligent models trained with tuned datasets optimize the prediction ability of existing breast cancer prediction models.

1. INTRODUCTION

Breast cancer (BC) is a most common lethal cancer worldwide and is the primary cause of cancer deaths among women, strangely affecting low- and middle-income countries. There are more than 2.3 million cases of breast cancer that occur each year and in 95% of countries breast cancer is the leading cause of female cancer deaths. According to world health organization

(WHO), data estimates indicates that a substantial upsurge in cancer mortality to nearly one million deaths per year by 2030, without crucial and daring interventions and estimated number of new cases in females of age group 0-59 from 2020 to 2030 is 1.36 million [1]. By 2040, breast cancers diagnosed anew are projected to develop over 40% for about 3 million cases each year. Also, deaths due to breast cancer are expected to exceed 50%, from 685,000 in 2020 to 1 million in 2040 [2]. The Breast Cancer is a development of cancerous cells in the lining part of glandular breast tissues normally produces no chronic pain and do not develop symptoms. The painless thickening or lumps, redness, inflations in the breast region are some common symptoms to consult health practitioners. The BC has effective treatments and higher probability of curing when it is diagnosed in earlier stages. The cancer death happens when the cancer spreads out to other organs in the metastasis stages, thus required to be treated promptly before spreading. There are no invasive clinical tests for BC diagnosis. The medical imaging, textural and shape measures are used for diagnosing the BC. These test reports are the basic sources arrived at based on the biologists and supports medical practitioners in diagnoses. Simultaneously technology based analysis of the associated parameters of medical imaging and textural measures provide early detection of the diseases. Therefore, by way of analysing the clinical data and to predict the presence of BC an intelligent system is desirable which will positively reduce the false positives and false negatives. The primary purpose of Artificial Intelligence analysis tools for medical imaging is to aid clinicians in their decision-making by combining multiple factors into a model that returns an actionable output.

In literature, there are various intelligent machine learning models for detecting and predicting the BC. But these intelligent models are trained for specific limited set of data and they may not be adaptable for other datasets. The intelligent models have to be trained with large number of appropriate patterns for each class to attain better performance. The model behaviour depends on the quality of knowledge inputs given to the models as well as the relevancy of the data to the responses. Most of the related research works are not focused on selecting the relevant predictors. In such scenario, the presence of irrelevant predictors may let-down the system. The selection of feature or predictors plays vital role in developing an intelligent model. To overcome these limitations this research work proposes hybrid intelligent learning models with adaptive feature selection approaches and the models are trained with three different breast cancer datasets to find out the optimal model for BC prediction. Here twelve hybridized learning models are proposed which are the combinations of four learning models with the three optimization algorithms for adaptive feature selection. Generally in feature selection approaches, the number of required features to be selected should be mentioned priorly. At the same time deciding the number of features required itself is a complicated task. In order to handle this checkpoint, nature-inspired adaptive optimization approaches are proposed in this research in order to augment the internal parameters of the intelligent learning models. Adaptive feature selection is the process of automatically choosing the number and set of features based on the model performance and correlation of predictors with response variable. So, the system automatically decides relevant predictors in each of the iterations based on the defined cost function. The pertinent predictors may be either the entire feature set or the feature subset. The learning algorithms based cost functions are also proposed in this research. Further the proposed research work is organized as: discussions about research background followed by the implementation and experimentation of proposed hybrid intelligent prediction models for breast cancer prediction and finally summary and future enhancements are discussed.

2. RESEARCH BACKGROUND

In recent times, breast cancer detection and prediction becomes a greatly spotlighted issue in the literature. Machine learning models play a vital role in early detection and prediction of breast cancer. The Machine Learning (ML) approaches such as Decision Trees (DT), Naïve Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbor (KNN), Linear Regression (LR), Multi-Layer Perceptron (MLP) and Artificial Neural Network (ANN) are generally used for breast cancer prediction. The research study for BC prediction summarized in Table 1, shows that DT, KNN, SVM, LR are the benchmark classification approaches among the said machine learning methods. It was also observed that these approaches provide more than ninety percentage accuracy in BC classification and prediction. Majority of the researchers studied in the literature tested these approaches using the Wisconsin Breast Cancer (WBC) dataset. It is one of the popular and benchmark dataset for breast cancer prediction obtained from the UCI repository. Various related research works regarding with BC prediction found in the literature during the study are listed in Table 1. Out of fifty research articles studied, twenty six most appropriate researches are highlighted in this table. The significance of these research articles are emphasized as table columns and are detailed as follows:

Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(4s), 542-561 | ISSN:2251-6727

a) **Author and Year**: The list of authors of the proposed research and the year of findings.
b) **Dataset**: The dataset focussed in the particular research.
c) **Objective**: Main goal of the proposed research work such as breast cancer disease detection, prediction or diagnosis.
d) **Models Explored**: The specific model proposed or experimented in the particular research for instance decision trees, k-nearest neighbors, etc.
e) **Measures**: The qualitative and quantitative measures used for the performance analysis of the proposed model for example accuracy, sensitivity, specificity etc.
f) **Significance**: It highlights the importance of the respective research work proposed.
g) **Limitations**: Emphasized all the restrictions of the models proposed in the corresponding research article.

Table 1. Highlights of breast cancer prediction related research works

S.No	Author and Year	Dataset	Objective	Model(s) Explored	Measures	Significance	Limitations
1.	Borges & Rodrigues, 2015 [1]	WBC	Detection	K-Means	Positive Predictive Value	– Effects of K-means analysed effectively	– Experimented on random samples – Missing values are not handled
2.	Agarap, 2018 [2]	WDBC	Detection	GRU-SVM, Linear Regression, MLP, Nearest Neighbor search, Softmax Regression, SVM	Accuracy, Sensitivity, specificity	– Explored six ML models – Obtained higher test accuracy of 99%	– Regression approaches are used for classification problem – Parameters are fixed manually
3.	H, Alshayeji, Ellethy, Abed, & Gupta, 2021 [3]	WBC and WDBC	Detection	Shallow ANN	Accuracy, Sensitivity, Specificity, AUC	– Efficiently classifies BC tumours with higher accuracy without feature selection	– Network used is dataset dependant – Trained Network may misbehave on strange dataset
4.	Mushtaq, Yaqub, Sani, & Khalid, 2019 [4]	WBC and WDBC	BC Classification	KNN	Accuracy, Sensitivity, Specificity, ROC, AUC	– Classifies BC effectively with Feature Selection – Comparative analysis is carried out with existing approaches	– Selection of K is limited – Suits well for WBC dataset alone – Benchmark Feature selection approaches are not considered
5.	Mohammed, Darrab, Noaman, & G, 2020 [5]	WBC and BC	Detection	Decision Tree (J48), Naïve Bayes (NB), and Sequential Minimal Optimization	True positive, False positive, ROC curve, Standard deviation, and Accuracy	– Detects BC effectively with SMO and J48 – Data Resampling is used to handle imbalanced dataset – 10 fold cross validation used	– Suitable measures for imbalanced dataset is not considered – No discussions were seen about early stage BC prediction
6.	Shawarib, Latif, Al-Zatmah, & Abu-Naser, 2020 [6]	WBC and BC survival Dataset	Diagnosis and Survival Prediction	ANN	Accuracy	– ANN trained to achieve higher classification accuracy	– Limited data is used for experimentation – More quality measures can be used to highlight model behaviours
7.	Ahmed, Imtiaz, & Karmakar, 2020 [7]	WBC	Prediction	NB, SVM, MLP, J48, Random Forest	Accuracy, Kappa statistic, Precision, Recall, F-measure, MCC, ROC area, PRC area	– All the models are trained for WBC dataset	– Applied existing methods for prediction – Features selection done manually
8.	Assegie, 2020 [8]	WBC	Detection	Optimized KNN	Accuracy	– Hyper-parameter is tuned for KNN	– Optimal parameter values are given manually – Limited comparative analysis

Journal of Chemical Health Risks
www.jchr.org
JCHR (2023) 13(4s), 542-561 | ISSN:2251-6727

S.No	Author and Year	Dataset	Objective	Model(s) Explored	Measures	Significance	Limitations
9.	Showrov, Islam, Hossain, & Ahmed, 2019 [9]	WBC	Detection	SVM, ANN, NB	Accuracy	– Performance analysis of classifiers are done	– Limited measures are used for analysis
10.	Benbrahim, Hachimi, & Amine, 2020 [10]	WBC	Detection	ML algorithms	Accuracy	– More than 11 Machine learning algorithms are studied	– Limited measures are used for analysis – Feature selection is not focused
11.	Naji, Filali, Aarika, Benlahmar, Abdelouhahid, & Debauche, 2021 [11]	WDBC	Diagnosis and prediction	SVM, RF, DT, KNN	Accuracy	– Model trained effectively to predict BC	– Restricted with only one dataset – Limited quality measures
12.	Islam, Haque, Iqbal, Hasan, Hasan, & Kabir, 2020 [12]	WBC	Prediction	SVM, KNN, ANN, LR	Accuracy, Sensitivity, Specificity, Precision, F1 Score	– Prediction performance is studied among these approaches	– Feature selection is not considered
13.	Assegie, Tulasi, & Kumar, 2021 [13]	WBC	Prediction	Decision Tree, Adaptive Boosting	Accuracy	– Prediction of BC with tree based approaches	– Limited comparative analysis – Lack of benchmark approaches
14.	Ghosh, Azam, Hasib, Karim, Jonkman, & Anwar, 2021 [14]	WDBC	Prediction	Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU)	Accuracy, Sensitivity, Specificity, Precision, F1 Score	– Deep learning approaches are used for BC prediction	– Lack of benchmark approaches – Features subset discussions are not seen
15.	Sahu, Mohanty, & Rout, 2019 [15]	WBC	Prediction	PCA –ANN	Accuracy, Sensitivity, F-Score	– Feature reduction approach PCA is hybridized for effective prediction	– Focused PCA method alone – Comparative analysis for PCA is missing
16.	Khan, et al., 2022 [16]	WDBC	Prediction	random forest, logistic regression, DT, K-NN	Accuracy	– Prediction of BC is performed	– Feature selection not considered – No significant methods are proposed
17.	Kumar, Mishra, Mazzara, Thanh, & Verma, 2020) [17]	WBC	Detection	DT, KNN, J48, J-RiP, Logistic Regression, SVM, MLP, RF, etc	Accuracy	– Twelve different ML approaches are compared for effective BC diagnosis	– Feature selection not considered – Focused only classification approaches
18.	Ara, Das, & Dey, 2021 [18]	WBC	Detection	SVM, NB, RF, DT, LR, KNN	Accuracy	– Automatic diagnostic system is built	– Lack of quality measures – Features subset selection is not discussed

Journal of Chemical Health Risks
www.jchr.org
JCHR (2023) 13(4s), 542-561 | ISSN:2251-6727

S.No	Author and Year	Dataset	Objective	Model(s) Explored	Measures	Significance	Limitations
19.	Dhanya, Paul, Akula, Sivakumar, & Nair, 2019 [19]	WDBC	Detection	LR, NB, RF	Accuracy	– Feature selection based BC detection is proposed	– Trained for one dataset alone – Lack of Feature selection benchmark approaches
20.	Mridha, 2021 [20]	WDBC	Prediction	KNN, NB, LR, SVM, Gradient Booster, ANN	Accuracy, Cross-validation, Sensitivity, Specificity	– ANN is trained to predict BC	– Limited handling of Irrelevant Features
21.	Tiwari, Bharuka, Shah, & Lokare, 2020 [21]	WDBC	Prediction	LR, SVM, MLP, ANN, CNN	Accuracy	– DL methods are trained for BC prediction	– DL may behave abruptly for strange datasets – Require more data for training
22.	Chaurasia & Pal, 2020 [22]	WDBC	Prediction	CART, SVM, LR, NB, MLP, KNN	Accuracy, Run time analysis	– Statistical Feature selection approach is proposed	– More qualitative analysis is required to justify the performance of the approach
23.	Shaikh & Ali, 2018 [23]	WBC and WDBC	Prediction	NB, J48, K-NN, SVM	Accuracy	– Feature selection is discussed	– Lack of benchmark feature selection approaches – No novel approaches are proposed
24.	Gupta & Garg, 2020 [24]	WBC	Prediction	K-NN, LR, DT, NB, SVM, Adam Gradient Learning	Accuracy	– Deep Learning approach is trained to explore non-linear relation in data	– Model trained is dataset dependant – Scalability is a major issue
25.	Thomas, Pradhan, & Dhaka, 2020 [25]	WBC	Prediction	KNN, SVM, NB, DT, K-means, ANN	Accuracy	– Comparative study on prediction approaches are carried out	– No novel approach is proposed – Feature reduction is not focused
26.	Afolayan, Adebiyi, Arowolo, Chakraborty, & Adebiyi, 2022 [26]	WBC	Prediction	PSO_DT	Accuracy	– Efficiency of prediction is increased using PSO	– Feature Selection must be adaptive – Lack of comparative analysis on Feature selection

This existing literature study notifies the following important aspects that are to be considered while developing intelligent predictive models.

a) Model selection: The selected model should be robust enough to fit data and its characteristics in order to explore hidden knowledge and provide good predictive accuracy.

b) Data pre-processing: The raw data must be prepared using various transformations such as data normalization, discretization, scaling, handling and imputing missing values to match up and stand firm the selected model for better results.

c) Data splitting: Separating data is another major aspect of the learning process that splits the dataset as training data and testing data for effective learning.

d) Defining objective function: This function mimics the research objectives; thus choosing the existing mathematical functions or defining a new function is a crucial part of model development. Besides various distance measures, similarity measures, activation functions, and training algorithms are also chosen for better performance.

e) Determining performance quantifiers: The selection of qualitative and quantitative measures to analyse the model performance and emphasize the model's significance to the current research problem scenario.

The overall limitations observed in the related research works are:

a) The models are trained for one particular dataset, which will make the model rigid on other datasets. The most common approaches and their behaviour on different datasets have to be identified.

b) Generally, the irrelevant predictors plays major role in lowering the model performance. In most of the research works studied, the feature subset selection or irrelevant feature reduction are extremely lacking.

3. EXISTING MACHINE LEARNING AND OPTIMIZATION APPROACHES

The learning models and optimization approaches applied in this research work are briefly narrated in this section. For learning purpose four machine learning models such as K-Nearest Neighbor, Naïve Bayes, Support Vector Machine and Artificial Neural Network are applied. For optimal feature subset selection the optimization approaches including Particle Swarm Optimization, Genetic Algorithm and Venus Flytrap Optimization methods are used.

3.1 MACHINE LEARNING MODELS

The machine learning models are used to explore the key insights of data. These models are usually built to understand and to learn the patterns of each group/class in the dataset [27, 28, 29]. The data is normally split into training data and testing data to train, test and tuning the model. This process will be continued till the better results are acquired. The trained model can be used for predicting the similar new dataset. The significance of all the machine learning models used in this research is explained as below:

3.1.1 K-Nearest Neighbor

It is one of the supervised learning algorithms used to solve classification problem. It is used as a benchmark classification model which suits most of the real time problems and produce accurate results. It classifies the new data points based on the similarity between the existing data points [30, 31, 32].

3.1.2 Support Vector Machine

This method creates the best decision boundary to separate the different classes in n-dimensional space. It uses the extreme data points to the decision line as support vectors to create a hyper plane as decision boundary. It exactly defines the boundary for each class and produce accurate class approximation [33].

3.1.3 Naïve Bayes

This probabilistic supervised classification model uses Bayes theorem for classification. It suits for the high dimensional data. It is a scalable model with predictors as well as with data points. It is a fastest model and suits for real-time predictions. It is insensitive to irrelevant features [34].

3.1.4 Artificial Neural Network

It is a sub-category of machine learning which is designed based on the biological neurons to make intelligent decisions with limited assistance. This is because the method can learn and model the relationship between predictors and responses that are non-linear and compound in nature. There are different varieties of neural networks available. These method ensemble problems of various domains and produces better results [35].

3.2 OPTIMIZATION APPROACHES

Optimization is the process which trains a model iteratively that results in a maximum and minimum function evaluation. This approach is a most important phenomenon in machine learning to get improved results. Iteration results at the end of each loop are compared by altering the hyper-parameters in each step until the optimum results are reached. Thus such methods create an accurate model with less error rate [36].

3.2.1 Particle Swarm Optimization

This meta-heuristic swarm intelligence algorithm is devised based on the flocking behaviour of the birds. It is simple and easy to implement; less number of parameters are required to tune and highly robust. It uses fixed inertia weights; decline too slow and extremely easy to fall into local optimum solution. This algorithm supports high-dimensional data [36].

3.2.2 Genetic Algorithm

This algorithm is devised based on the natural selection and biological evolution. It is a search based optimization algorithm. This simple and efficient algorithm, suits well to any optimization problem, continuously changes the individuals and converges to optimal solution [37].

3.2.3 Venus flytrap Optimization

Venus Flytrap Optimization is one of the meta-heuristic non-swarm intelligence algorithm devised based on foraging mechanism of Venus flytrap plant. It is free from local stagnation and easy to implement. The main-

Journal of Chemical Health Risks
www.jchr.org
JCHR (2023) 13(4s), 542-561 | ISSN:2251-6727

support criteria induce more exploration of search space; suits for multi-dimensional data; accelerates the individuals periodically; has autonomous individuals and achieves better results [38, 39].

4. PROPOSED HYBRID INTELLIGENT MODELS (HIM)

In order to overcome the limitations identified based on the literature review, this research work proposes development of hybrid intelligent learning models as a solution by means of fusing the machine learning algorithms with adaptive feature subset selection approaches. The objectives of the proposed hybrid model are:

a) To propose adaptive feature subset selection using various optimization approaches including Particle Swarm Optimization, Genetic Algorithm and Venus Flytrap Optimization.
b) To propose machine learning based cost functions for these suggested adaptive feature selection approaches to build the hybrid intelligent learning models.
c) To analyse the behaviour of these hybrid intelligent models on three different breast cancer datasets.

4.1 PREDICTION USING PROPOSED HYBRID INTELLIGENT MODEL

The hybrid intelligent learning approaches proposed for breast cancer prediction are detailed in this section. The basic architecture of HIM is as shown in the Figure 1. The hybrid intelligent model is the fusion of adaptive feature selection approach and the learning model. The adaptive feature selection is the process of selecting the best predictors based on the prediction error found by the learning model. In each epoch, the learning model trains the model with selected predictors and feds the prediction error to the adaptive feature selection approach for further selection of the tuned feature subset subsequently. This process will continue until either the convergence takes place or the maximum iterations reached. Convergence of the learning model helps in defining how many iterations of training a learning model will require producing minimum errors. The learning rate and the number of epochs are made proportional while modelling the hybrid intelligent model in order to avoid convergence failure.

The required processed dataset is the input to this hybrid intelligent model to train the intelligent model and finally the expert hybrid intelligent model is achieved with optimal predictors. The adaptive feature selection approaches are achieved through applying the optimization algorithms including PSO, GA and VFO whose performances are enhanced using the novel cost functions. The learning models are fixed as the cost function which will determine the prediction error. While executing an optimization algorithm, at each iteration the learning model is trained and predicts the output. This proposed hybrid intelligent model is implemented with three optimization algorithms and four cost functions that are discussed further in this section.

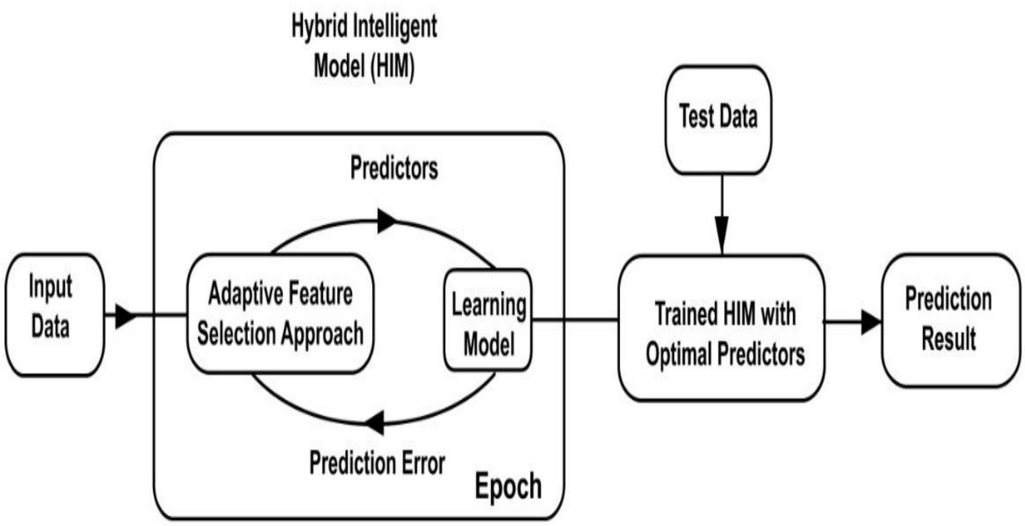

Figure 1. Architecture of proposed Hybrid Intelligent Models (HIM)

4.1.1 PROPOSED ADAPTIVE PARTICLE SWARM OPTIMIZATION WITH ML (APSO_ML)

The adaptive particle swarm optimization is proposed for optimal selection of relevant predictors to train the intelligent models. The particles initiated with different feature subset and start searching for the best predictors at each iteration. The prediction error of the ML algorithm is defined as a cost function. The mapping between the algorithmic terminologies and the problem domain is given as in Table 2. In this table the terms like swarm, swarm size, particle, particle type and length, cost function and other parameters are mapped with the problem domain.

Algorithm_1 represents the pseudo code for APSO_ML approach. Initially, the particles are generated with different predictor subset. Then other parameters are also initiated and start searching for predictor subset. Here the ML is hybridized on the cost function $cost_{ML}(\ )$ described in Algorithm_4.

Table 2. Mapping the APSO terminologies with the problem domain

APSO Algorithmic Terms	**Problem Description**
Swarm	Set of random predictor subsets
Particle	Predictor set
Swarm size	Number of predictor subsets chosen for searching
Particle type	Binary vector
Particle length	Total number of predictors
Cost Function	Cost function is used to evaluate the particles' current search using $Cost_{ML}(\ )$ as in Algorithm_4
Initial Position	Random positions chosen for starting the search in the search region
Initial Velocity	Initially the velocity is zero for all particles
Particle position	Current position of the particle updated by global best particle in each iteration
Particle velocity	Updated by best positions to proceed further
Pbest position	Best position of the particle which achieved the best cost.
Gbest particle	Global best particle which has best position in the swarm

Algorithm_1: APSO_ML

Input: *Processed Data*
Output: *Optimal Predictors*
Step 1 Define the swarm parameters*: swarm size (n), cost function $cost_{ML}(\ )$ as in Algorithm_4, stopping criteria, weight values.*
Step 2 Generate the swarm *with 'n' particles: position, velocity, pbest, gbest*
Step 3 Repeat until met the stopping criteria
3a. *For each particle P*
i. *Calculate cost of $cost_{ML}(P_i)$*
ii. *Evaluate the Pbest of P_i*
End for
3b. *Update gbest*
3c. *Update particle position and velocity*
End
Step 4 Return Optimal particle

4.1.2 PROPOSED ADAPTIVE GENETIC ALGORITHM WITH ML (AGA_ML)

The adaptive genetic algorithm is proposed for optimal selection of relevant predictors using evolutionary approach to train the intelligent models. Table 3 shows the mapping between the GA terms and problem domain. The terms like chromosome, gene type and length, cost function and other parameters are described in this table.

Table 3. Mapping the AGA terminologies with the problem domain

AGA Algorithmic Terms	**Problem Description**
Chromosome	Set of random predictor subsets
Gene	Predictor set
Chromosome size	Number of predictor subsets chosen for searching
Gene Type	Binary vector
Gene Length	Total number of predictors
Cost Function	Cost function $cost_{ML}(\ )$ is used to evaluate the particles' current search using Algorithm_4
Gbest_pop	Best individual which has best fitness in the chromosome

Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(4s), 542-561 | ISSN:2251-6727

Algorithm_2: AGA_ML

Input: *Processed Data*
Output: *Optimal Predictors*
Step 1 ***Define the Chromosome (P)****: chromosome size (n), number of generations, gene_length, cost function* $cost_{ML}(\ )$ *as in Algorithm_4, stopping criteria, weight values.*
Step 2 ***Generate chromosome*** *with 'n' genes* (P_n)
Step 3 ***Repeat until stopping criteria***
i. *Apply crossover and mutation operator to P*
ii. *Calculate cost of all individuals in P using* $cost_{ML}(P_i)$
iii. *Select the best individuals and generate new generation*
iv. *Replace the worst individuals with new individuals*
End
Step 4 ***Return Optimal Gene***

Algorithm_2 depicts the pseudo code of AGA_ML. In this approach the population is a set of individuals. Each individual stands as a solution to the problem need to be solved. Here an individual is characterized by a set of parameters (variables) known as Genes and the genes are linked into a chain to form a Chromosome (solution). These genes are initialized with different predictors set. The optimal search is based on the genetic evolution of generations. The fittest individual, the predictors with the minimal prediction error of the ML algorithm is defined as a cost function. The major operations of GA are selection, crossover and mutation. The selection of individuals for new generation is done with Roulette Wheel Selection (RWS) technique. The single point crossover operator is used for creating new offspring with probability of 0.9. The higher crossover rate supports to explore the solution space wider and provides the algorithm a global search capability. The mutation operator presents diversity into the sampled population and is used in an attempt to avoid local minima by stopping the population of chromosomes from becoming too similar to each other, thus slowing or even preventing convergence to the global optimum. Mutation probability is kept far smaller so that the new offspring is imputing the new characteristics different from parents. The low value ensures only a tiny fraction of the population is mutated at each generation.

4.1.3 PROPOSED ADAPTIVE VENUS FLYTRAP OPTIMIZATION WITH ML (VFO–ML)

The adaptive Venus flytrap optimization is proposed for optimal selection of relevant predictors using non-swarm intelligence for training the intelligent models. The flytraps are used to hunt optimal prey for their nutritional intake. The flytraps are initiated with different feature subsets and start hunting the best predictors in each iteration. The predictors are the main criterion whereas the support criterion is the correlation of the predictors to response. The prediction error of the ML algorithm is defined as a cost function. The mapping between the algorithmic terminologies with the problem domain is given in the Table 4. In this table the terms like flytrap plant, plant size, flytrap, flytrap type and length, prey, cost function and other parameters are mapped with the problem domain.

Table 4. Mapping the AVFO terminologies with the problem domain

AVFO Algorithmic Terms	**Problem Description**
Flytap Plant	Set of random predictor subsets
Flytrap	Predictor set
Prey	Predictors
Flytap Plant size	Number of predictor subsets chosen for searching
Flytrap type	Binary vector
Flytrap size	Total number of predictors
Cost Function	Cost function is used to evaluate the flytrap current nutrition intake using *$cost_{ML}(\ )$ given* in Algorithm_4
Main Criterion	Prediction error
Support Criterion	Predictor correlation
Initial Charge	Random charge chosen for starting the prey hunt
Initial Potential	Initially the potential is zero for all flytraps
Flytrap Charge	Current charge of the flytrap updated by best flytrap in each iteration
Flytrap Potential	Updated by best charge to proceed further
Fbest	Best potential of each flytrap which yield maximal nutrition
Gbest	Global best flytrap which has best potential in the plant

The Algorithm_3 represents the pseudo code of AVFO_ML. The processed dataset is the input to this algorithm. Initially the flytrap plant is initiated with 'n' flytraps and their parameters including action potential, charge, max_iter are initialized. At first stimulation the action potential and charge accumulation are evaluated, further at second stimulation the action potential and charge are updated and the prey trapping will occur. The fitness is evaluated using cost function $cost_{ML}(\ )$ to ensure the presence of prey in the trap. The best flytrap is then sealed for further hunting.

Algorithm_3: AVFO_ML
Input: *Processed Data*
Output: *Optimal Predictors*
Step 1 : ***Define the plant parameters:*** *cost function $cost_{ML}(\ )$ as in Algorithm_4, Max_iter*
Step 2 : ***Generate flytrap plant*** *with n flytraps ft_i (i = 1, 2, ..., n)*
Step 3 : ***Repeat until stopping criteria***
3a. *For each flytrap ft*
i.*At $t_1=0$,*
– *Evaluate Action Potential (u_t) and Charge accumulation (C) using $cost_{ML}(\ )$*
ii.*At t_2, if $t_2<=T$ then*
– *update action potential and charge*
– *evaluate the fitness and object status*
3b. *Find the current best flytrap*
3c. *Seal the best flytrap until another best flytrap arrives*
Step 4 : ***Return optimal flytrap***

4.1.4 PROPOSED ML COST FUNCTION

The cost function is a function that evaluates the rate of expenditure of a particular job or task. The major role of using cost function in most of the meta-heuristic algorithms is to reduce the rate of expenditure of the problem. In this research, the cost represents the prediction error of a machine learning approach.

The pseudo code of the cost function is given in the Algorithm_4. Here the input is the predictors selected by the optimization approach. The ML approach is then trained and tuned using the selected data subset. Finally, the random samples from the data subset are used to test the model and prediction error is computed. This error value is treated as cost in the optimization approaches. In this research work, the learning models namely K-Nearest Neighbor, Naïve Bayes, Support Vector Machine and Artificial Neural Network are used for breast cancer prediction. The ML model in Step 3 of the Algorithm_4 holds any of these four learning models. This cost function contributes in terms of deciding the credibility and reliability of the HIM models for the BC prediction.

Algorithm_4: function $cost_{ML}(\ )$
Input: *Predictors, dataset*
Output: *Prediction_error*
Step 1 : *Fetch the columns of selected predictors from the dataset*
Step 2 : *Split the selected data subset into train_data and test_data*
Step 3 : *Continuously train and tune the ML model with train_data and test_data till stopping criteria*
Step 4 : *Test the trained model with random samples*
Step 5 : *Evaluate the prediction error using the following equation*

$$pred_{error} = \frac{number\ of\ samples\ wrongly\ predicted}{sample\ size}$$

Step 6 : *Return prediction_error*

4.1.5 DATASET DESCRIPTION

The hybrid intelligent models are experimented with the three benchmark Breast cancer datasets that are obtained from the UCI repository. These datasets are Wisconsin Breast Cancer Dataset (WBC), Wisconsin Diagnostic Breast Cancer Dataset (WDBC), and Wisconsin Prognostic Breast Cancer Dataset (WPBC). Table 5 describes these datasets.

Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(4s), 542-561 | ISSN:2251-6727

Table 5. Dataset Description

	WBC	WDBC	WPBC
# Samples	699	569	198
# Predictors	10	31	32
# Responses	1	1	2
# Classes	2	2	2

4.1.6 EXPERIMENTAL ANALYSIS OF HIM

The HIM prediction models are implemented on 4GHz, 8GB RAM, Intel Core-7 machine in Microsoft Windows 10 platform. The code is developed and experimented in MATLAB 2019b environment. In this research, three adaptive optimization techniques: APSO_ML, AGA_ML, AVFO_ML are incorporated with four learning models KNN, NB, SVM and ANN. Totally twelve hybrid intelligent models are proposed for BC prediction. These models are compared with the same four learning models without feature selection and with feature selection using Principal Component Analysis (PCA). Overall, twenty models are experimented with three different Breast Cancer datasets. Performances of these models are evaluated using various qualitative and quantitative measures [40, 41] including Accuracy, Sensitivity, Specificity, Precision, Recall, F-Measure and G-Mean. These measures are computed using the equations as given in Table 6. The accuracy measure is used to analyse the prediction correctness of the model. Though accuracy measure is commonly used in most of the research works, the accuracy may vary for different test data. So, other measures are also included in this research for analysing the performance of HIM. The sensitivity of the learning models quantifies how far the learning models identify the positive objects. The specificity of the model is used to learn the models' behaviour on true negative prediction results. Similarly, the precision and recall measures are used to quantify how far the model can identify the positive instance over all predicted positive instances and exact positive instances.

Table 6. Mathematical formulae for quality measures

Quality Measures	**Mathematical formulae**
Accuracy	$Accuracy = \dfrac{(TP+TN)}{(TP+FP+FN+TN)}$
Sensitivity	$Sensitivity = \dfrac{TP}{(TP+TN)}$
Specificity	$Specificity = \dfrac{TN}{(TN+FP)}$
Precision	$Precision = \dfrac{TP}{(TP+FP)}$
Recall	$Recall = \dfrac{TP}{(TP+TN)}$
F-Measure	$FMeasure = \dfrac{(2 \times precision\ \times recall)}{(precision + recall)}$
G-Mean	$Gmean = \sqrt{TP \times TN}$

4.1.7 EXPERIMENTAL RESULTS

The experimental results of all the twenty models with three different datasets are given in Tables 7 to 9. All the performance measures are highlighted for each of the proposed and existing models in these tables. Columns of these tables represent the four various machine learning models applied and its proposed enhancements. Each machine learning model comprises of four models with feature selection using PCA, PSO, GA and VFO, and one model without feature selection. The benchmark PCA approach is used for comparative analysis since it is one of the adaptive models; it does not demand for the number of required features for processing. Hence this approach is chosen for comparative analysis. The Table 7 shows the performance of the proposed hybrid models against existing models on WBC dataset.

Table 7. Comparative analysis of proposed Hybrid Intelligent Models on WBC Dataset

	K-Nearest Neighbor					Naive Bayes				
Measures	**KNN**	**KNN with PCA**	**APSO-KNN**	**AGA-KNN**	**AVFO-KNN**	**NB**	**NB with PCA**	**APSO-NB**	**AGA-NB**	**AVFO-NB**
Accuracy	95.65	94.71	95.69	95.65	95.70	95.85	94.56	96.13	96.71	96.63
Sensitivity	92.53	88.38	92.53	92.11	92.94	97.09	96.68	97.92	98.75	97.90
Specificity	97.59	98.03	97.37	97.59	97.16	95.19	93.45	95.19	95.63	95.94
Precision	95.29	95.95	94.89	95.27	94.51	91.40	88.59	91.47	92.24	92.85
Recall	92.53	88.38	92.53	92.11	92.94	97.09	96.68	97.92	98.75	97.90
F-Measure	93.49	92.01	93.69	93.67	93.72	94.16	92.46	94.52	95.39	95.31
G-Mean	95.03	93.08	94.92	94.81	95.53	96.14	95.05	96.55	97.18	96.92
	Support Vector Machine					**Artificial Neural Network**				
	SVM	**SVM with PCA**	**APSO-SVM**	**AGA-SVM**	**AVFO-SVM**	**ANN**	**ANN with PCA**	**APSO-ANN**	**AGA-ANN**	**AVFO-ANN**
Accuracy	96.42	96.42	96.56	96.56	96.92	95.99	95.85	96.57	96.14	96.85
Sensitivity	95.43	94.19	95.43	95.02	96.65	97.10	97.51	97.51	97.10	97.93
Specificity	96.94	97.60	97.16	97.37	97.07	95.41	94.98	96.07	95.63	96.29
Precision	94.26	95.38	94.65	95.02	94.67	91.76	91.09	92.89	92.13	93.28
Recall	95.43	94.19	95.43	95.14	96.65	97.10	97.51	97.51	97.10	97.93
F-Measure	94.84	94.78	95.04	95.02	95.65	94.35	94.19	95.14	94.55	95.55
G-Mean	96.18	95.88	96.29	96.19	96.86	96.25	96.24	96.79	96.36	97.10

Observing the overall performance of all the models on WBC dataset, the AVFO_SVM provides the highest accuracy of 96.92 percent, highest precision, specificity and F-Measure of 94.67, 97.07 and 95.65 percent respectively. The AVFO_ANN model has highest G-Mean of 97.10 percent, which indicates the performance on the imbalanced datasets whereas the G-Mean quantifies the class-wise accuracy. The response of AVFO_ANN model over the positive instances is high when compared to the AVFO_SVM, which is evaluated using sensitivity of the prediction models. The F-Measure of these two models is quite similar, so the behaviour of AVFO_SVM and AVFO_ANN on WBC dataset is alike. The performance of proposed adaptive feature selection models APSO and AGA provided improved results than the ML models with the benchmark PCA. Conclusively the Adaptive Venus Flytrap Algorithm enhances results and best suits for the KNN, SVM and ANN machine learning algorithms and adaptive genetic algorithm provides improved results for NB machine learning model in BC prediction on WBC dataset.

Table 8. Comparative analysis of proposed Hybrid Intelligent Models on WDBC Dataset

	K-Nearest Neighbor					Naïve Bayes				
Measures	**KNN**	**KNN with PCA**	**APSO-KNN**	**AGA-KNN**	**AVFO-KNN**	**NB**	**NB with PCA**	**APSO-NB**	**AGA-NB**	**AVFO-NB**
Accuracy	94.55	87.34	93.49	94.72	93.84	92.61	86.46	91.38	92.97	92.79
Sensitivity	91.50	75.47	91.50	92.45	89.62	89.15	70.28	86.79	88.67	88.67
Specificity	96.35	94.39	94.67	96.07	96.35	94.67	96.07	94.11	95.51	95.23
Precision	93.71	88.88	91.07	93.33	93.59	90.86	91.41	89.75	92.15	91.70
Recall	91.50	75.47	91.50	92.45	89.62	89.15	70.28	86.79	88.67	88.67
F-Measure	92.60	81.63	91.29	92.89	91.56	90.05	79.46	88.24	90.38	90.16
G-Mean	93.90	84.40	93.08	94.24	92.92	91.87	82.17	90.38	92.03	91.90
	Support Vector Machine					**Artificial Neural Network**				
	SVM	**SVM with PCA**	**APSO-SVM**	**AGA-SVM**	**AVFO-SVM**	**ANN**	**ANN with PCA**	**APSO-ANN**	**AGA-ANN**	**AVFO-ANN**
Accuracy	97.36	95.95	97.01	97.18	96.48	99.47	98.95	99.47	98.95	99.12
Sensitivity	94.33	91.03	93.86	93.86	91.98	98.58	97.17	99.53	97.17	98.58
Specificity	99.15	98.87	98.87	99.15	99.15	100	100	99.44	100	99.44
Precision	98.52	97.96	98.02	98.51	98.48	100	100	99.06	100	99.05
Recall	94.33	91.03	93.86	93.86	91.98	98.58	97.17	99.53	97.17	98.58
F-Measure	96.38	94.37	95.90	96.13	95.12	99.29	98.56	99.29	98.56	98.82
G-Mean	96.71	94.8777	96.34	96.47	95.50	99.29	98.57	99.48	98.57	99.01

The experimental results of the proposed hybrid intelligent models for WDBC dataset are shown in Table 8. The overall performance of these learning models shows that, the hybrid ANN models yield the accuracy, f-measure and g-mean above 98%. Out of them, the ANN and APSO- ANN obtained high accuracy of 99.47%, f-measure and g-mean of 99.29% and 99.48% respectively, which is better than other models. The AVFO-ANN also earned 99.12% accuracy as well as 99.01% g-mean. Thus out of the four learning models

Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(4s), 542-561 | ISSN:2251-6727

ANN models are performing well for the WDBC dataset. Followed by ANN models, the SVM models yield better results for this WDBC dataset. The APSO and AVFO feature selection algorithms are effective in learning this dataset through ANN. The ANN and SVM models are good in learning this dataset without feature selection. It infers that all the features in WDBC dataset are important in BC prediction. There is no significant outcome based on the PCA approach in all these twenty learning models on WDBC dataset. The entire machine learning models without feature selection outperforms all the proposed feature selection approaches. It shows that there is significant correlation found among the predictors and response of this dataset, and there is no irrelevant predictor in WDBC dataset.

Table 9. Comparative analysis of proposed Hybrid Intelligent Models on WPBC Dataset

	K-Nearest Neighbor					**Naïve Bayes**				
Measures	**KNN**	**KNN with PCA**	**APSO-KNN**	**AGA-KNN**	**AVFO-KNN**	**NB**	**NB with PCA**	**APSO-NB**	**AGA-NB**	**AVFO-NB**
Accuracy	69.59	67.53	70.62	72.16	69.59	64.43	73.71	69.07	67.53	65.46
Sensitivity	43.48	28.26	39.13	39.13	34.78	43.48	26.09	47.83	41.30	43.48
Specificity	77.70	79.73	80.41	82.43	80.41	70.95	88.51	75.68	75.68	72.30
Precision	37.74	30.23	38.30	40.91	35.56	31.75	41.38	37.93	34.55	32.79
Recall	43.48	28.26	39.13	39.13	34.78	43.48	26.09	47.83	41.30	43.48
F-Measure	40.40	29.21	38.71	40.00	35.16	36.70	32.00	42.31	37.62	37.38
G-Mean	58.12	47.47	56.09	56.79	52.88	55.54	48.05	60.16	55.91	56.07
	Support Vector Machine					**Artificial Neural Network**				
	SVM	**SVM with PCA**	**APSO-SVM**	**AGA-SVM**	**AVFO-SVM**	**ANN**	**ANN with PCA**	**APSO-ANN**	**AGA-ANN**	**AVFO-ANN**
Accuracy	76.88	79.90	78.35	78.87	79.90	88.66	90.21	90.72	92.78	93.30
Sensitivity	26.3	26.09	19.57	19.57	17.39	60.87	73.91	82.61	76.09	89.13
Specificity	95.95	96.62	96.62	97.30	99.32	97.30	95.27	93.24	97.97	94.59
Precision	14.29	70.59	64.29	69.23	88.89	87.50	82.93	79.17	92.11	83.67
Recall	26.3	26.09	19.57	19.57	17.39	60.87	73.91	82.61	76.09	89.13
F-Measure	44.4	38.10	30.00	30.51	29.09	71.79	78.16	80.85	83.33	86.32
G-Mean	15.89	50.21	43.48	43.63	41.56	76.96	83.91	87.77	86.34	91.82

The experimental outcomes of the WPBC dataset with twenty intelligent models are given in the Table 9. The AVFO-ANN model yield highest value for the measures accuracy, f-measure and g-mean with values of 93.30%, 86.32% and 91.82% respectively. ANN models perform exceptionally well compared with all other learning models with 93.30%, 92.78% and 90.72% accuracy for the hybrid AVFO-ANN, AGA-ANN and APSO-ANN models respectively. All the remaining models hold accuracy value relatively very low. This shows that the hybrid ANN learning models explore well the hidden patterns of the WPBC dataset.

The proposed feature selection approaches advance the results in all ML models, which evidence that, the presence of irrelevant predictors in the WPBC dataset. The benchmark model KNN shows 69.59% accuracy, which indicates that the dataset has no sufficient samples for each class. Similarly the results of Naïve Bayes model are improved by PCA. This indicates that there exist highly irrelevant features and insufficient samples to learn about each class. Amongst all the ML models, AVFO feature selection algorithm enhanced the results significantly. Thus, proposed AVFO with all learning algorithms better identified the relevant predictors from the WPBC dataset.

4.1.8 DISCUSSIONS BASED ON HIM MODELS

The KNN models produce maximal results of 96%, 94.55% and 72.16% accuracy; G-Mean values as 95.53%, 94.24% and 58%; F-Measure scores of 94%, 92.8% and 40.40% for WBC, WDBC and WPBC datasets respectively. The AVFO_KNN and AGA_KNN algorithms identify the relevant predictors from WBC and WDBC datasets. For WPBC dataset the AVFO-ANN, AGA-ANN and APSO-ANN models are efficient in identifying significant predictors. The performance plots of KNN models on different datasets are shown in the Figure 2, where these three plots highlight the Accuracy, F-Measure and G-Mean of these KNN models and it was observed that KNN model is less appropriate for WPBC dataset.

Journal of Chemical Health Risks
www.jchr.org
JCHR (2023) 13(4s), 542-561 | ISSN:2251-6727

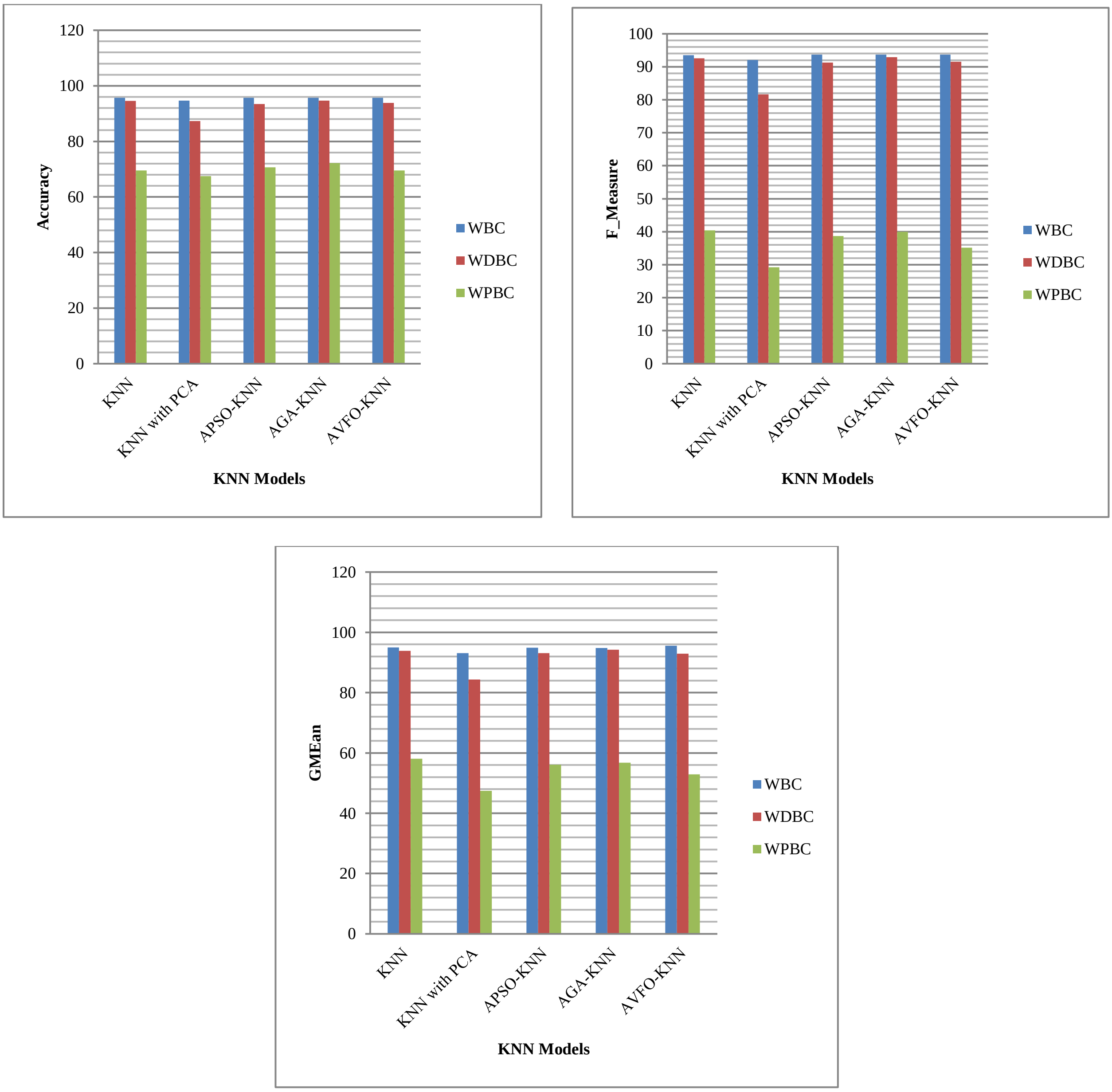

Figure 2. The overall performance of KNN models on different datasets

The comparative analysis of NB models for all the three datasets are shown in Figure 3. The AGA-NB model generates the highest scores for the measures accuracy of 96.71%, 92.79% and 73.71%; F-Measure of 95.39%, 90.38% and 42% and G-Mean of 97.18%, 92.03% and 60.16% for WBC, WDBC and WPBC datasets respectively. The AGA algorithm is suited for exploring the relevant features in all the datasets for Naïve Bayes approach. The F-Measure of NB models for WPBC very low, which is below 40%. As the NB is a probabilistic model, it easily learns the patterns in the model. Thus, it was inferred that the dataset has insufficient instances for each class.

Journal of Chemical Health Risks
www.jchr.org
JCHR (2023) 13(4s), 542-561 | ISSN:2251-6727

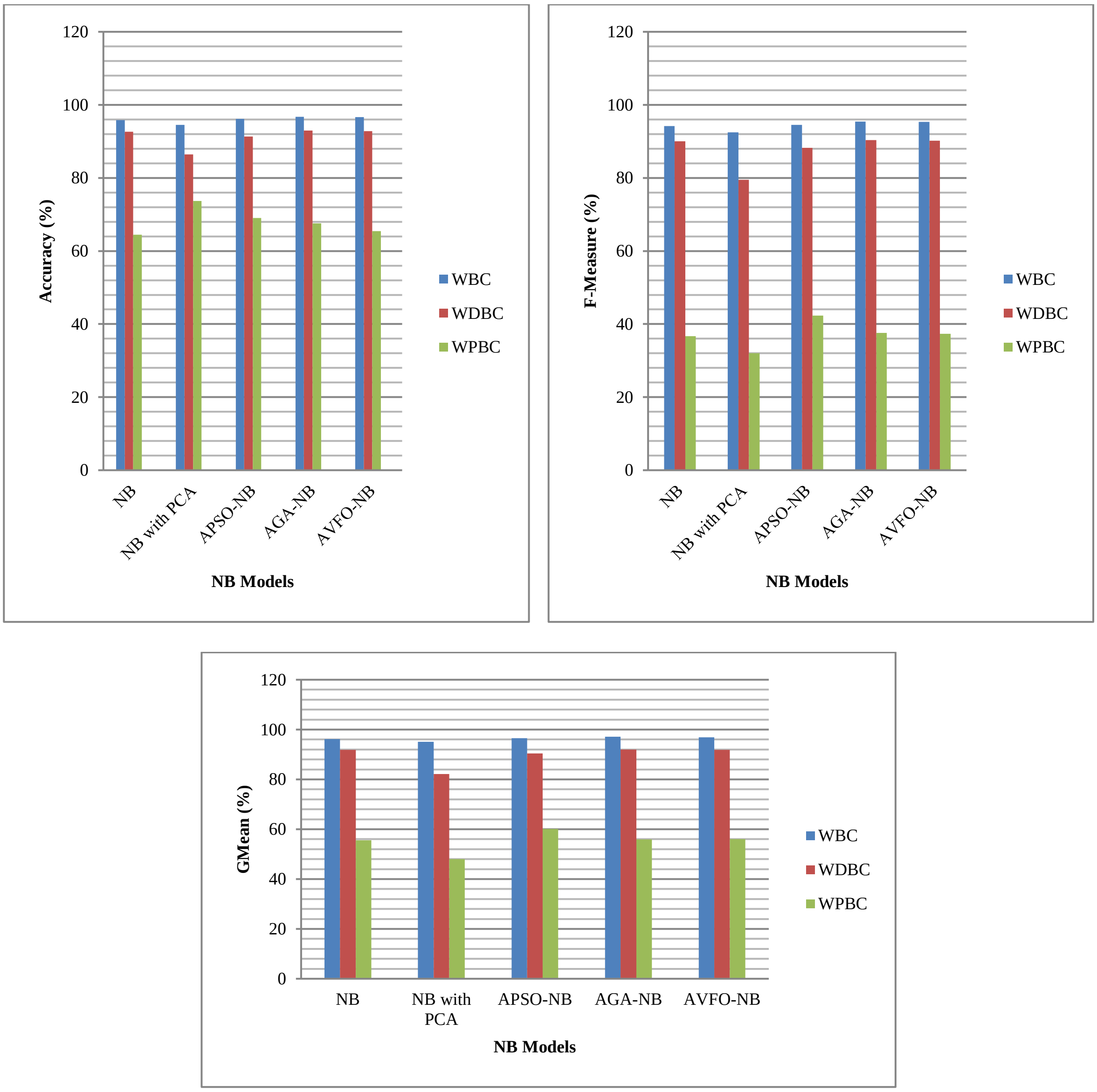

Figure 3. The overall performance of NB models on different datasets

The comparative performance analysis of SVM hybrid models is given in Figure 4. The SVM models are well suited for all the datasets. It is most appropriate for the WBC and WDBC datasets. It yields better results of 96.92%, 97.36% and 79.9% accuracy as well as 96.86%, 96.71%, and 50.21% G-Mean for these three datasets respectively. The AVFO algorithm improvises the prediction accuracy of SVM on WBC and WPBC datasets. For WDBC dataset, the feature selection approaches have not shown any significant progress in the performance, which shows that there may not be any irrelevant predictors in this dataset. The performance of SVM models on WPBC dataset indicates that there are inadequate samples for each class.

Journal of Chemical Health Risks
www.jchr.org
JCHR (2023) 13(4s), 542-561 | ISSN:2251-6727

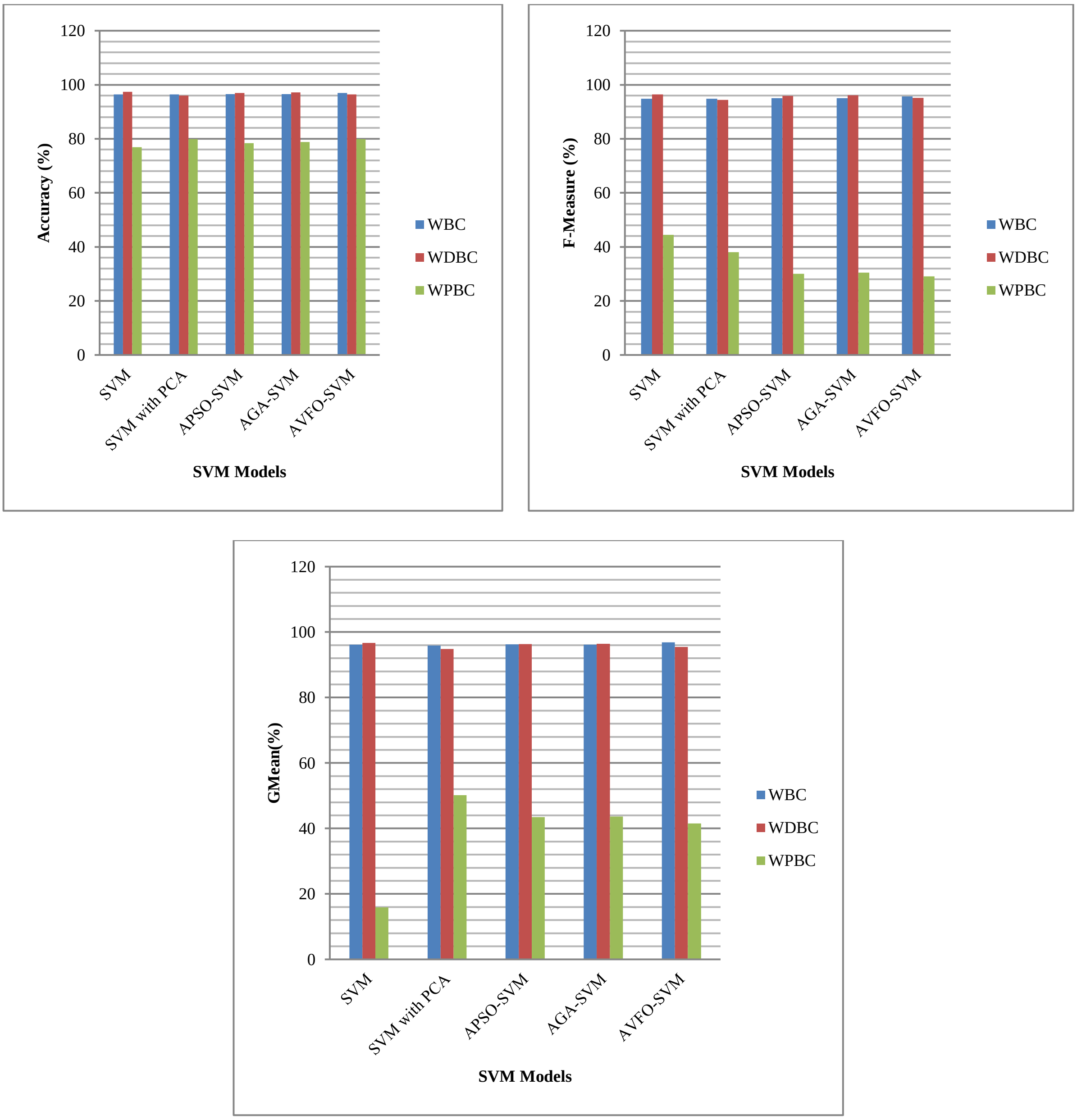

Figure 4. The overall performance of SVM models on different datasets

The overall performance of ANN models are shown in Figure 5. The ANN model is well suited for all the datasets. The AVFO-ANN hybrid models yield highest scores of 96.85%, 99.47% and 93.3% accuracy; 99.5%, 99.29% and 86.32% of F-Measure and 97.1%, 99.48% and 91.82% of G-Mean. The AVFO algorithm best fits for hybridized with ANN model for learning all the datasets. The ANN model can understand the complex relationship between predictors and responses in the WPBC dataset and yield above 90% result. The AVFO can explore the optimal predictors for almost all the ML models more specifically to the ANN model.

Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(4s), 542-561 | ISSN:2251-6727

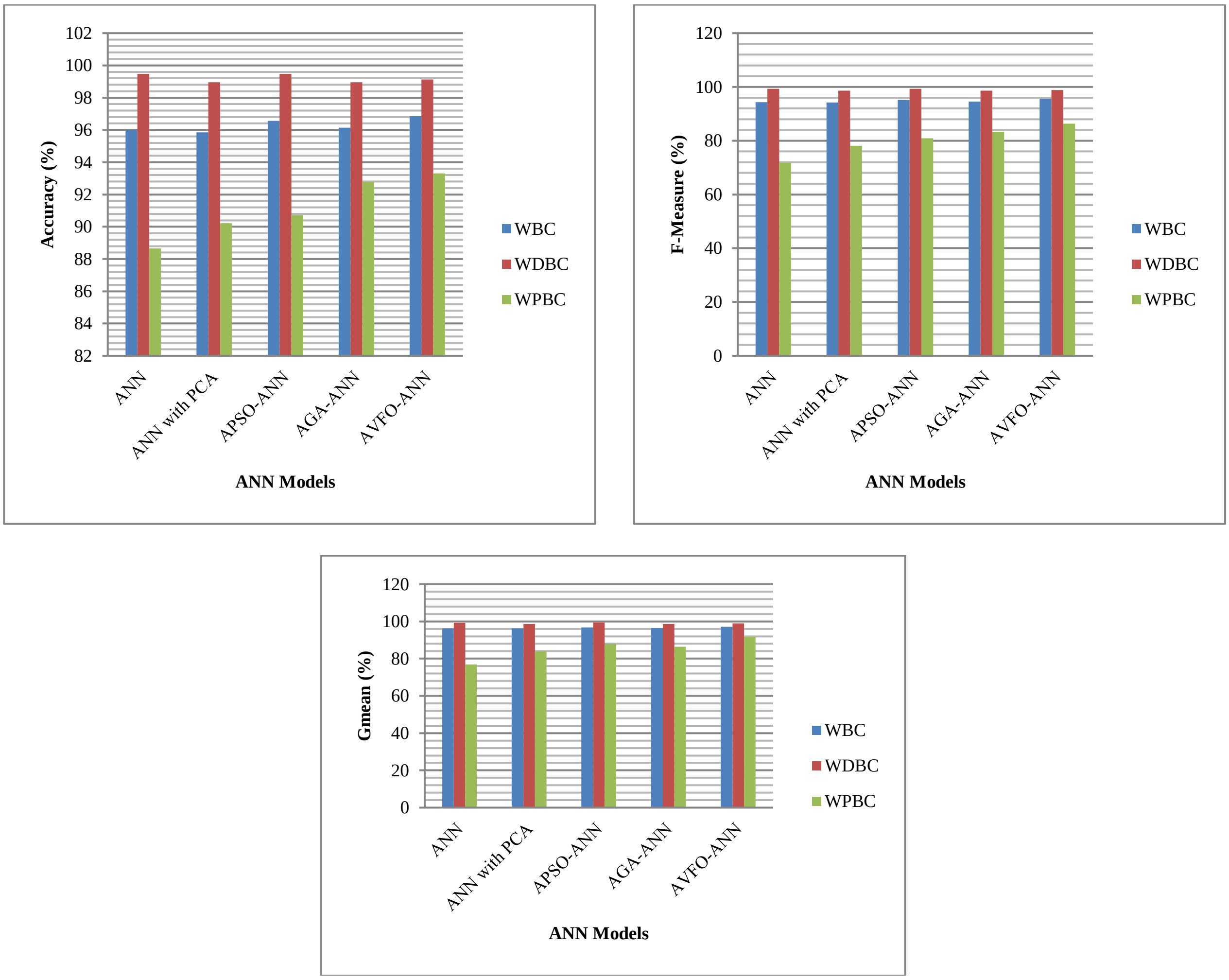

Figure 5. The overall Performance of ANN models on different datasets

Figure 6 shows the suitability quotient of each ML model for all the three datasets based on the accuracy measure. From this analysis, it was observed that the ANN and SVM models are well suitable for WBC and WDBC datasets. For WPBC dataset, only the ANN model attained maximal results. On the whole, ANN model and SVM model are efficient for Breast Cancer Prediction on three benchmark datasets WBC, WDBC and WPBC.

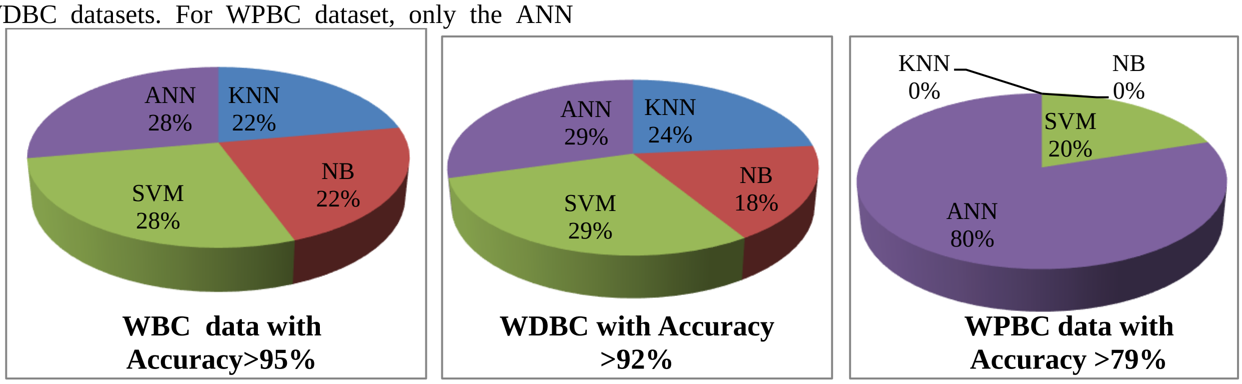

Figure 6. ML models suitability over the datasets based on accuracy measure

Journal of Chemical Health Risks
www.jchr.org
JCHR (2023) 13(4s), 542-561 | ISSN:2251-6727

Figure 7 shows the suitability proportion of the derived adaptive optimization algorithms over the three datasets. It was observed that proposed AVFO algorithm perform well for WBC and WPBC datasets. As observed in Figure 4, WDBS dataset may not have any irrelevant predictors. Further analysis on observing the maximal quotient of 'No FS' (No Feature Selection) model in Figure 7 shows that the WDBS dataset probably has all relevant predictors Also the AGA algorithm performs well on WDBC dataset.

In conclusion AVFO algorithm performs well for all three datasets. Similarly, ANN model and SVM model are efficient for breast cancer prediction on three benchmark datasets WBC, WDBC and WPBC.

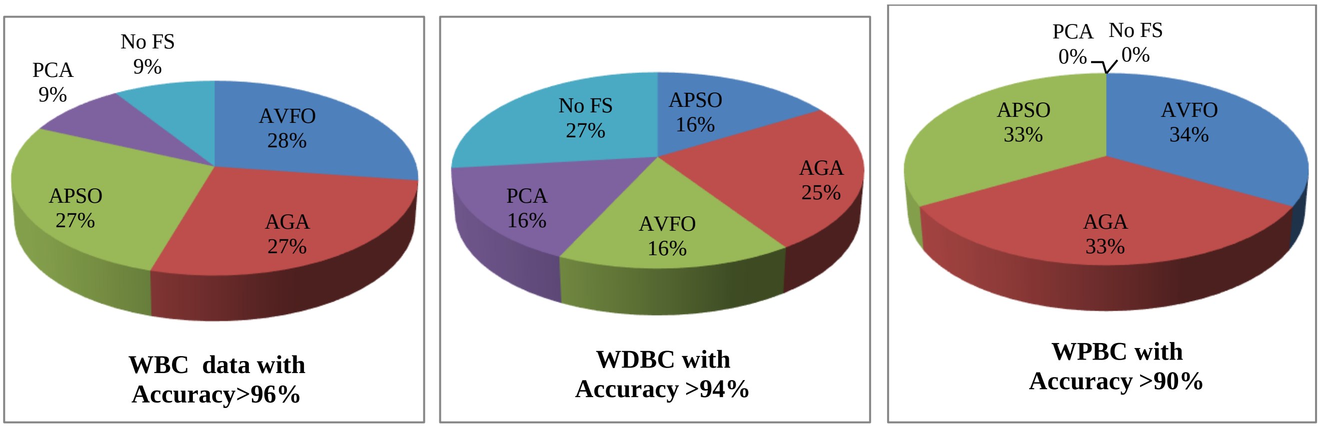

Figure 7. Adaptive optimization algorithms' suitability over the datasets based on accuracy measure

5. CONCLUSIONS

The current research work proposed a Hybrid Intelligent Prediction Model for breast cancer prediction. This work overcomes the existing limitations by means of specifically considering the relevant predictors and training various models with different datasets. This research work proposed twelve HIM models which are attained through hybridizing the four machine learning models (KNN, NB, SVM and ANN) and three optimization algorithms (APSO, AGA and AVFO) for adaptive feature selection. These algorithms are compared with the same four machine learning models with and without feature selection. Totally twenty intelligent models including proposed models are experimented on three UCI repository datasets namely WBC, WDBC and WPBC. The experimental analysis depicted that the ANN and SVM are compatible for all the datasets in combination with AVFO for feature selection. These improved models can be trained and tuned with various parameter set up and new datasets for optimal real-time predictions.

DECLARATIONS

Competing interests

No conflict of interest was declared by the authors.

Funding

Financial support from The Research Council (TRC), Oman under Research Grant Funding (Grant ID.: BFP/RGP/ICT/20/142) is gratefully acknowledged.

REFERENCES

1. WHO: Breast Cancer. [accessed 2022 Mar 23]. https://www.who.int/news-room/fact-sheets/detail/breast-cancer.
2. Melina A, Eileen M, Harriet R, Allini M, Deependra S, Mathieu L, Jerome V, Julie R.G, Fatima C, Sabine S, Isabelle S. 2022. Current and future burden of breast cancer: Global statistics for 2020 and 2040. The Breast. 66:15-23.
3. Borges LR. 2015. Analysis of the Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection. Proceedings of XI Workshop on Computer Vision; Oct 5-7; New York, USA.
4. Agarap AFM. 2018. On breast cancer detection: an application of machine learning algorithms on the Wisconsin diagnostic dataset. Paper presented at: ICMLSC 2018. Proceedings of the 2nd International Conference on Machine Learning and Soft Computing; Feb 2; New York, USA.
5. Alshayeji MH, Ellethy H, Abed S, Gupta R. 2022. Computer-aided detection of breast cancer on the Wisconsin dataset: An artificial neural networks approach. Biomed Signal Process Control. 71(Part A):103141.
6. Mushtaq Z, Yaqub A, Sani S, Khalid A. 2020. Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets. J Chin Inst Eng. 43(1):80-92.
7. Mohammed SA, Darrab S, Noaman SA, Saake G. 2020. Analysis of Breast Cancer Detection Using

Different Machine Learning Techniques. Paper presented at: DMBD 2020. Proceedings of the 5th International Conference on Data Mining and Big Data; Jul 14–20; Belgrade, Serbia.

8. Shawarib MZA, Latif AEA, Al-Zatmah BEEl-D, Abu-Naser SS. 2020. Breast Cancer Diagnosis and Survival Prediction Using JNN. Int J Eng Inf Syst. 4(10):23-30.
9. Ahmed MT, Imtiaz MN, Karmakar A. 2020. Analysis of Wisconsin Breast cancer original dataset using data mining and machine learning algorithms for breast cancer prediction. J Sci Technol Environ Inform. 9(2):665-672.
10. Assegie TA. 2020. An optimized K-Nearest Neighbor based breast cancer detection. JRC. 2(3):115-118.
11. Showrov MIH, Islam MT, Hossain MD, Ahmed MS. 2019. Performance Comparison of Three Classifiers for the Classification of Breast Cancer Dataset. Paper presented at: EICT 2019. Proceedings of the 4th International Conference on Electrical Information and Communication Technology; Dec 20-22; Khulna, Bangladesh.
12. Benbrahim, H, Hachimi, H, Amine A. 2020. Comparative Study of Machine Learning Algorithms Using the Breast Cancer Dataset. In: Ezziyyani, M. editors. Advanced Intelligent Systems for Sustainable Development AI2SD 2019. Advances in Intelligent Systems and Computing; Jul 8-11; Marrakech. Morocco:Springer. p. 83-91.
13. Naji MA, Filali S El, Aarika K, ELH Benlahmar, Abdelouhahid RA, Debauche O. 2021. Machine Learning Algorithms for Breast Cancer Prediction and Diagnosis. Paper presented at: SA2IOT 2021. Proceedings of the first International workshop on Edge IA-IoT for Smart Agriculture; Procedia Computer Science; Aug 9-12; Leuven, Belgium.
14. Islam MM, Haque MR, Iqbal H, Hasan MM, Hasan M, Kabir MN. 2020. Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques. SN Comput. Sci. 1(290):1-14.
15. Assegie TA, Tulasi RL, Kumar NK. 2021. Breast cancer prediction model with decision tree and adaptive boosting. IAES Int. J. Artif. Intell. 10(1):184-190.
16. Ghosh P, Azam S, Hasib KM, Karim A, Jonkman M, Anwar A. 2021. A Performance Based Study on Deep Learning Algorithms in the Effective Prediction of Breast Cancer. International Joint Conference on Neural Networks (IJCNN), pp. 1-8, 2021.
17. Sahu B, Mohanty SN, SK Rout. 2019. A Hybrid Approach for Breast Cancer Classification and Diagnosis. EAI Endorsed Transactions on Scalable Information Systems. 6(20):1-8.
18. Khan MM, Islam S, Sarkar S, Ayaz FI, Kabir MM, Tazin T, Albraikan AA, Almalki FA. 2022. Machine Learning Based Comparative Analysis for Breast cancer Prediction. J Healthc Eng. 2022:1-15,.
19. Kumar V, Mishra B, Mazzara M, Thanh D, Verma A. 2020. Prediction of Malignant and Benign Breast Cancer: A Data Mining Approach in Healthcare Applications. Paper presented at: ICDSM 2019. Proceedings of the 2nd International Conference on Decision Science and Management; Sep 20-21; Changsha, Hunan, China
20. Ara S, Das A, Dey A. 2021. Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms. Paper presented at: ICAI 2021. Proceedings of the first International Conference on Artificial Intelligence; Apr 5-7; Islamabad, Pakistan.
21. Dhanya R, Paul IR, Akula SS, Sivakumar M, Nair JJ. 2019. A Comparative Study for Breast Cancer Prediction using Machine Learning and Feature Selection. Paper presented at: ICCS 2019. Proceedings of the International Conference on Intelligent Computing and Control Systems; May 15-17; Madurai, India.
22. Mridha K, 2021. Early Prediction of Breast Cancer by using Artificial Neural Network and Machine Learning Techniques. Paper presented at: CSNT 2021. Proceedings of the tenth IEEE International Conference on Communication Systems and Network Technologies; Jun 18-19; Bhopal, India.
23. Tiwari M, Bharuka R, Shah P, Lokare R. 2020. Breast Cancer Prediction Using Deep Learning and Machine Learning Techniques. SSRN. 1-5.
24. Chaurasia V, Pal S. 2020. Applications of Machine Learning Techniques to Predict Diagnostic Breast Cancer. SN Comput Sci. 1(270):1-11.
25. Shaikh TA, Ali R. 2018. Applying Machine Learning Algorithms for Early Diagnosis and Prediction of Breast Cancer Risk, Paper presented at: ICCCN 2018. Proceedings of 2nd International Conference on Communication, Computing and Networking; Mar 29-30; NITTTR Chandigarh, India.
26. Gupta P, Garg S. 2020. Breast Cancer Prediction using varying Parameters of Machine Learning Models. Procedia Comput. Sci. 171:593-601.
27. Thomas T, Pradhan N, Dhaka VS. 2020. Comparative Analysis to Predict Breast Cancer using Machine Learning Algorithms: A Survey. Paper presented at: ICICT 2020. Proceedings of the International Conference on Inventive

Computation Technologies; Feb 26-28; Coimbatore, India.
28. Afolayan JO, Adebiyi MO, Arowolo MO, Chakraborty C, Adebiyi AA. 2022. Breast Cancer Detection Using Particle Swarm Optimization and Decision Tree Machine Learning Technique. Intelligent Healthcare. 286:61-83.
29. Janiesch C, Zschech P, Heinrich K. 2021. Machine learning and deep learning. Electron Markets. 31:685–695.
30. Shalev-Shwartz S, Ben-David S. 2014. Understanding Machine Learning: From Theory to Algorithms. New York (NY): Cambridge University Press.
31. Rokach L, Maimon O. 2005. Decision Trees. In: Maimon O, Rokach L, editors. Data Mining and Knowledge Discovery Handbook. Springer; Boston; p. 165–192.
32. Zhang Z. 2016. Introduction to machine learning: k-nearest neighbors. Ann Transl Med. 4(11):218.
33. Shmilovici A. 2009. Support Vector Machines in Data Mining and Knowledge Discovery Handbook. Boston: Springer.
34. Rooney N, Patterson D, Galushka M. 2004. A comprehensive review of recursive Naïve Bayes Classifiers. Intell Data Anal. 8(601): 615–628.
35. Grosan C, Abraham A. 2011. Artificial Neural Networks. In: Intelligent Systems. Intelligent Systems Reference Library. Springer; Berlin; Heidelberg; p. 281–323.
36. Gad AG. 2022. Particle Swarm Optimization Algorithm and Its Applications: A Systematic Review. Arch. Comput. Methods Eng. 29:2531–2561.
37. Sastry K, Goldberg DE, Kendall G. 2014. Genetic Algorithms. In: Burke EK, Kendall G, editors. Search Methodologies. Springer; Berlin; p. 97-125.
38. Gowri R, Rathipriya R. 2015. Biclustering using Venus Flytrap Optimization. Paper presented at: CIDM 2015. Proceedings of the 1st International Conference on Computational Intelligence in Data Mining, Dec 5-6; Odisha, India.
39. Gowri R, Rathipriya R. 2021. Non- Swarm Intelligence Algorithms: A Case Study. Computing. 103(9):1-43.
40. Gowri R, Rathipriya R. 2015. Venus Flytrap Optimization. In: Muthukrishnan S, Vijayalakshmi R, Shina S, Veeramani C, Anthony B, Lynn B, editors. Computational Intelligence, Cyber Security and Computational Models. Springer; p.519-531.
41. Gowri R, Rathipriya R. 2023. A novel cancer drug target module mining approach using nonswarm intelligence. In: Janmenjoy N, Danilo P, Bighnaraj N, Manohar M, Muhammad K, Al-Dabass D, editors. Computational Intelligence in Cancer Diagnosis Progress and Challenges. 1st ed. Academic Press; p. 359-387.