

A Novel Feature Selection Approach to Classify Breast Cancer Drug using Optimized Grey Wolf Algorithm

G. Shobana^{1†} and Dr. N. Priya^{2††},

Department of Computer Applications
Madras Christian College, INDIA¹,

PG Department of Computer Science
SDNB Vaishnav College for Women, INDIA²

Abstract

Cancer has become a common disease for the past two decades throughout the globe and there is significant increase of cancer among women. Breast cancer and ovarian cancers are more prevalent among women. Majority of the patients approach the physicians only during their final stage of the disease. Early diagnosis of cancer remains a great challenge for the researchers. Although several drugs are being synthesized very often, their multi-benefits are less investigated. With millions of drugs synthesized and their data are accessible through open repositories. Drug repurposing can be done using machine learning techniques. We propose a feature selection technique in this paper, which is novel that generates multiple populations for the grey wolf algorithm and classifies breast cancer drugs efficiently. Leukemia drug dataset is also investigated and Multilayer perceptron achieved 96% prediction accuracy. Three supervised machine learning algorithms namely Random Forest classifier, Multilayer Perceptron and Support Vector Machine models were applied and Multilayer perceptron had higher accuracy rate of 97.7% for breast cancer drug classification.

Keywords:

Supervised Machine learning, Grey Wolf Algorithm, Random Forest, Support Vector Machine, Multilayer Perceptron.

1. Introduction

Among diseases, cancer has one of the highest mortality rates among women across the globe.

Many types of medications are offered to the patients that include surgery and chemotherapy. Chemotherapy involves oral and parenteral drugs that resists the growth of cancer cells.

Drug discovery involves tedious procedures and is time consuming. The identification of the target molecule and possible inhibitor is the first step in the drug discovery process. The ligand or inhibitor's chemical characteristics are investigated. Chemical information is extracted from compound structures using several computational approaches, which are then applied to the machine learning process. Breast cancer medications include Gemcitabine hydrochloride, Capecitabine, Fluoxymesterone, Epirubicin, and others. Clofarabine, Nelarabine, Cytarabine, Dasatinib, Dexamethasone, and other Leukemia medicines are utilised. There are a variety of medications that are used to treat various disorders. KEGG is a database that contains a wealth of information on diseases and drugs [1]. Fig.1 shows the proposed novel Multi-Level Median Based Feature Ranking Method (MLMBFRM) workflow. Data is extracted via several techniques such as similarity analysis, retrieval of chemical graphs, development of descriptors, and fingerprint synthesis, among others. The subsequent procedures are built after the preceding computational method has been

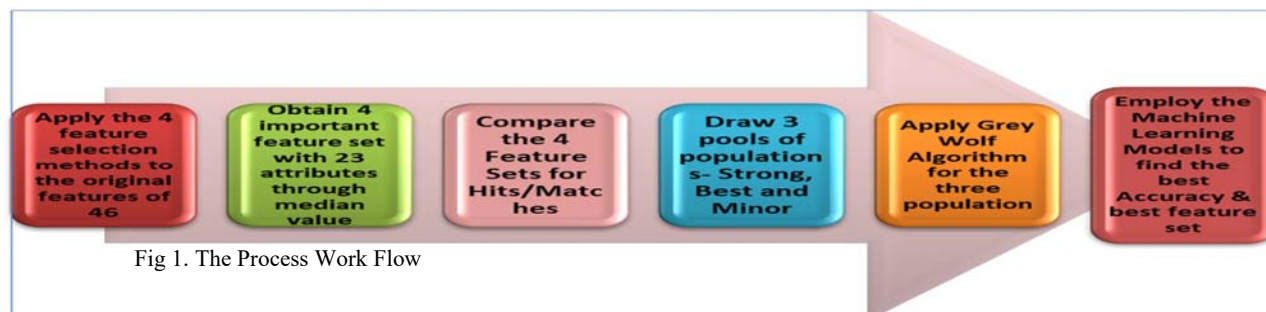


Fig 1. The Process Work Flow

Women from all sections of the society, urban or rural are affected due to this slowly progressing disease.

successfully developed and completed, and this has a significant impact on the quality of the chemical data

provided to the machine learning process. Both data and images are manipulated using machine learning algorithms to classify the relevant data [2]. Several machine learning models are used for exploring bioinformatic data. Support vector machine has proved as the best achieving model in the field of Cheminformatics and Bioinformatics [3].

In this paper, we obtain physicochemical and pharmaceutical properties of breast cancer drugs from reputed repositories. Feature engineering is applied to the dataset and the important attributes are selected for optimized performance by the machine learning algorithms. Three feature subsets or three types of population are generated and the Grey Wolf optimization technique is implemented. This technique improves the overall performance of the ML models. Three machine learning techniques were applied, Multilayer perceptron achieved an accuracy 97.7% while with the Leukemia dataset, it achieved 96%.

Cysts that can be malignant or benign can be diagnosed either through biopsy or mammogram. Depending on the stage of the disease treatments are given to the patients. The growth of the cancer cells can be prevented by administering oral or parenteral drugs to the affected patients in the procedure of systematic therapy. Some of the breast cancer drugs and their structure are shown in Fig 2.

In this study, we explore the breast cancer drug and Leukemia drug classification using the proposed methodology, where Multi-Layer Perceptron achieves higher prediction accuracy in both the cases.

2. Related Work

Many scientists have sought to categorise cancer kinds as benign or malignant. They looked at a dataset that was easily accessible. Primary data was obtained from databases for this study, and a new dataset was constructed. The Wisconsin Breast Cancer Database present in the UCI repository is a commonly used dataset. Breast cancer was classified by Tina et al as benign or malignant using the LR, which follows a Regression Classification approach. The WBCD was used to create the dataset. The experiment employed 699 samples with 11 different attributes. To delete irrelevant features, Recursive Feature Elimination (RFE) was utilised. With four key features, Logistic Regression improved prediction accuracy [4].

Breast Cancer was classified and compared by Ebru et al using several machine learning approaches. WBCD was used to extract data from 699 samples with 11 attributes. The dataset was used to train SVM and ANN. SVM correctly identified the cancer type 96.9% of the time [5]. RF and SVM were employed by Haruka et al to rank P53 Inhibitor candidates. The candidates were ranked using the Pareto approach. In the ranking, SVM worked efficiently [6].

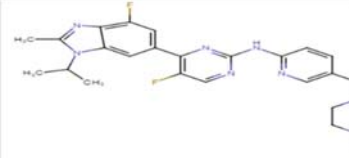
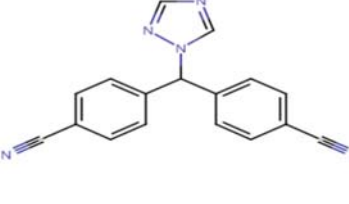
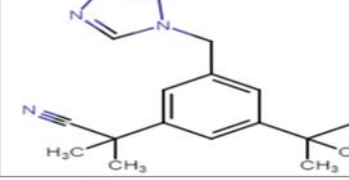
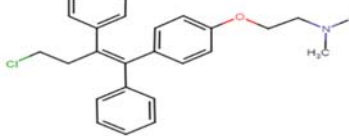
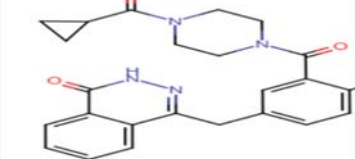
Breast Cancer Drug Name	Structure
Abemaciclib	
Letrozole	
Anastrozole	
Toremifene	
Olaparib	

Fig. 2. Breast Cancer Drugs and their Structure

To predict Chronic Leukemia illness, David Chen et al used ML classifiers such as Logistic Regression model (LR), Random Forest classifier (RF), and Gradient Boosting Machine (GBM) [7]. Dharani T et al gathered 200 photos from the open repository ALL_IDB. Blood smear photographs were taken from both leukaemia patients and healthy people. The dataset was pre-processed before being classified

using the Support Vector Machine (SVM) classifier, which achieved 97.8% accuracy [8].

For image processing, Preeti Jagadev et al gathered 220 pictures of leukaemia patients. For image pre-processing, they used the K-Means Clustering technique and the Marker Controlled Watershed Algorithm, and SVM was used to classify four forms of leukaemia [9]. Jakkrich Laosai et al proposed a Neural Network based Deep Learning technique for classifying leukaemia. Convolutional based Neural Network [CNN] was used to pre-process and classify the images, resulting in better accuracy [10].

Sachin Paswan et al classified leukaemia kinds using K-Nearest Neighbour, CNN, and SVM. CNN had the highest accuracy of the classifiers, at 98 percent [11]. To predict breast cancer inhibitor protein, Dejun Jiang et al used a variety of machine learning methods. Traditional machine learning algorithms such as NB, K-Nearest Neighbour, LR, and SVM were applied. They assessed the drug's features of absorption, distribution, metabolism, excretion, and toxicity (ADMET). The SA algorithm was paired with RF to implement feature selection. The prediction accuracy of XGBoost and Deep Neural Network was higher than the four classic ML algorithms [12].

Luz Adriana Borrero et al performed toxicity prediction utilizing Machine Learning procedures. ANN, NB, K-NN, RF, Decision Tree and SVM were utilized for toxicity classification. The Decision tree accomplished a prediction accuracy of 89%. Data was obtained from admetSAR web for the examination [13]. Delora Baptista et al talked about different Deep Learning structures utilized to foresee cancer drugs. They concluded that DNN had high prediction accuracy compared to other conventional models. DNN has demonstrated exceptionally high prediction accuracy in the domain of Drug Discovery [14]. Alex P. Lind et al explored the activity of 225 potent drugs against 990 different cancer cell lines by utilizing diverse machine learning methods. RF machine learning algorithm performed well. [15]. Alok Kumar Jha et al illustrated the utilization of Graph based Convolutional Neural Network to foresee cancer. GCNN demonstrates to perform way better than RF, SVM and basic Neural Network [16].

Chen Chen et al presented Particle Swarm based Optimization based Grey Wolf Optimizer to enhance the effectiveness of Intrusion Detection System by employing Support Vector Machine [17]. Jingyi Liu et al proposed and improvement in the Grey Wolf algorithm. They integrated Lion optimizer and dynamic weights to increase the efficiency of Grey Wolf Algorithm [18]. Ming-Zhen Tsai et al have proposed a strategy to enhance the efficiency of the Grey Wolf Algorithm through a linear type of descent in the presence of convergence factor [19]. Sridevi et al proposed an Intrusion detection Methodology for wireless with sensor networks using Boruta Feature Selection Algorithm with grid search based Random Forest machine learning model [20].

3. Datasets and Attributes

Two datasets are taken for the study. The names of breast cancer drugs were taken from the KEGG database. The names of FDI-approved drugs were also gathered from NCI. All of the medications used in the study were licensed and are presently being used to treat breast cancer. For augmentation, 85 percent comparable medicines are taken from the ChEMBL [21] and the dataset is prepared. For each medication, 46 characteristics are created and pre-processed. Cancer drugs account for 256 of the total, while non-cancer drugs account for 157.

The Leukemia Data samples of 513 was drawn from SWISS ADME database. 265 samples were Leukemia drugs and 248 were non- Leukemia drugs. Kyoto-Encyclopedia of Genes and Genomes (KEGG) is a Japanese website that provides details regarding approved drugs for several diseases. The SwissADME tool calculates these qualities or variables. SwissADME [22] is a technology that generates diverse medicinal chemical characteristics of medicines, making drug discovery more efficient. This user-friendly programme is maintained by the Swiss Institute of Bioinformatics and is frequently used by researchers. Several data sets generated by this programme are widely used. Several datasets generated by this programme are frequently utilised in computational chemistry, pharmaceutical research, bioinformatics, cheminformatics, and, most recently, drug toxicity prediction. Pharmacokinetics, Drug Likeness, Lipophilicity, Water Solubility, Medicinal Chemistry, and Physicochemical Properties of drug molecules are defined by the characteristics.

Molecular Refractivity, Molecular Weight, Presence of Heavy Atoms, Presence of Aromatic Heavy Atoms, Fraction Csp3 are all examples of physicochemical properties.

Presence of rotatable bonds, H-bond acceptors, H-bond donors, and TPSA (Topological Polar Surface Area). iLOGP, MLOGP, WLOGP, X LOGP, SILICOS-IT, and Consensus LOGP, which is the average of the other five defines lipophilicity. Water soluble properties include LogS (ESOL), LogS (Ali), and LogS (SILICOS-IT), with the Solubility class specified by the LogS Scale, which ranges from 0 to 1.

Lipinski (Pfizer Filter)	Ghose Filter
MW \leq 500	160 \leq MW \leq 480
MLOGP \leq 4.15	-0.4 \leq WLOGP \leq 5.6
N or O \leq 10	40 \leq MR \leq 130
NH or OH \leq 5	20 \leq atoms \leq 70

Fig. 3. Lipinski and Ghose Filter Parameters

for toxicity prediction or drug similarity properties are the Lipinski and Ghose filters, as shown in Figure 3. The drugs considered for the experiment are already approved drugs and come from a reputable database. Their biological activity against the indicated diseases has already been demonstrated. Experiments with 46 features make 413 observations. The first step in pre-processing is data cleaning. It removes noise and redundant data and fills in the missing data. All category data is converted to numbers to facilitate the machine learning process. Figure 4 shows a sample dataset.

4. Proposed Methodology

We propose a Multi-Level Median Based Feature Ranking Method with the following Four levels of feature ranking.

- Multiple Tree Based
- ANOVA F-Test
- RFE
- BORUTA

46 features were input to the four methods and median based 23 features were selected for further optimizing using multiple populated Grey Wolf algorithm. Fig.5 shows the proposed novel methodology. Every Feature Extraction method, Filter methods, Wrapper methods or Statistical methods has its own

MW	#Heavy atoms	#Aromatic heavy atoms	Fraction Csp3	#Rotatable bonds	#H-bond acceptors	MR	TPSA	iLOGP	XLOGP3	...	CYP1A2 inhibitor	CYP2C9 inhibitor	CYP2D6 inhibitor
336.44	24	0	0.85	0	4	91.45	57.53	2.71	2.13	...	0	0	0
334.42	24	0	0.80	0	4	90.49	54.37	2.67	2.11	...	0	0	0
364.45	26	0	0.81	1	5	96.46	74.60	2.44	1.77	...	0	0	0
378.48	27	0	0.82	1	5	101.27	74.60	2.74	2.34	...	0	0	0
348.45	25	0	0.81	1	4	95.26	54.37	2.70	2.45	...	0	0	0

Fig. 4 Sample Dataset

Drug toxicity is one of the major concerns during the synthesis of new compounds. The degree of toxicity varies from drug to drug and depends on the route of administration. Some commonly used filters

disadvantages. Moreover, when these methods are performed in sequence, insignificant features of one method might be considered significant by another method. The proposed methodology, considers the

significant features generated by all these techniques and obtains three pools of populations.

models to identify the suitable model for the Cheminformatic data.

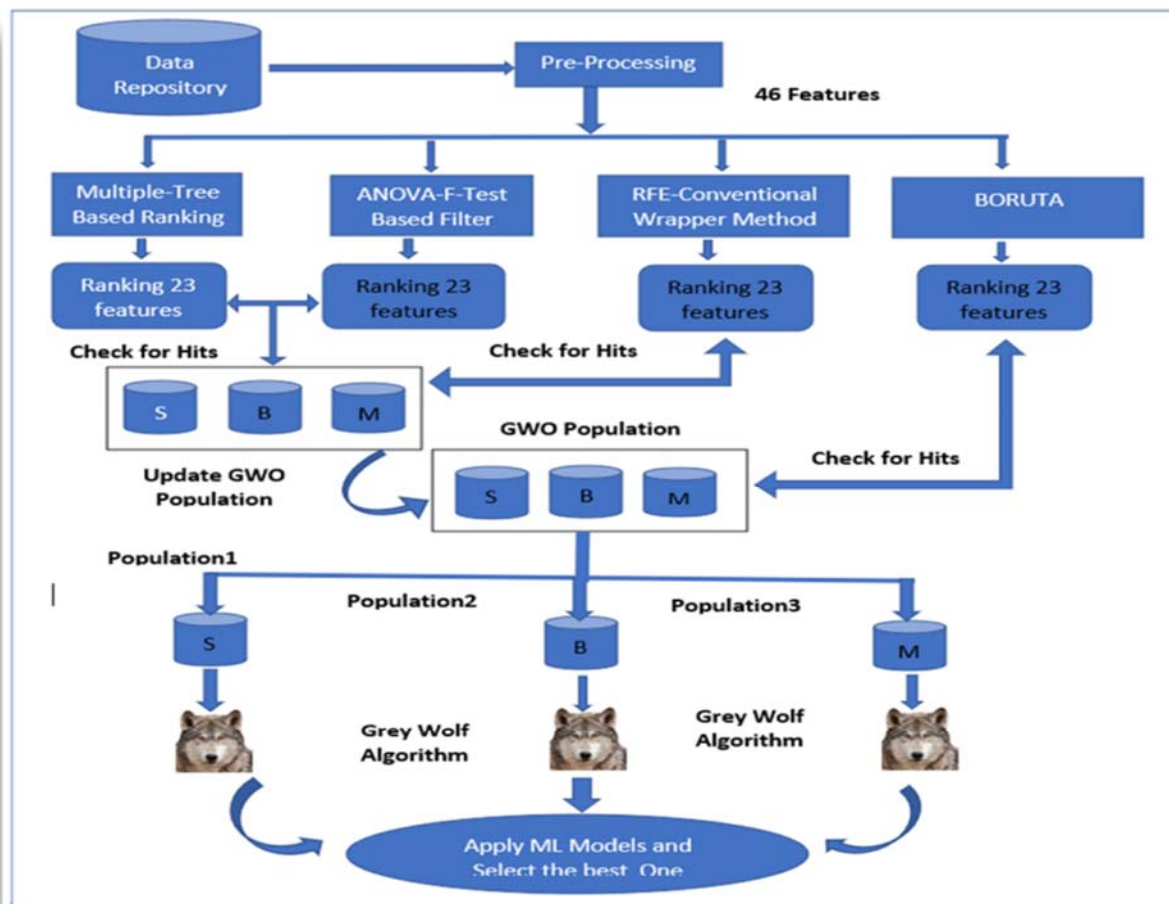


Fig.5 The Proposed Framework

Consider the original feature set of 46. Apply the dataset to the four feature determination models namely Tree Based, F-Test, RFE and BORUTA. Using Median, obtain 23 features from each process. Create three pools for Strong Features, Best Features and Minor features. Compare output of 1st process and 2nd process. Common hits are added to Strong pool. Compare the pools for hits from the 3rd process and update. Compare the updated pool with 4th process output and update the three population, Strong, Best and Minor. Finally apply the three populations to the grey wolf optimizer. Employ the machine learning

Multi-Level Median Based Feature Ranking Method (MLMBFRM).

Step1: Consider the original feature set $\{O_i\}$ of 46 features where i ranges from 0 to 45.

Tree Based Intrinsic Feature Selection:

Step 2: Let $\{RF_i\}$ be the feature importance computed by Random Forest Classifier for 46 attributes, where i ranges from 0 to 45.

Step 3: Let $\{ET_i\}$ be the feature importance computed by Extra Trees Classifier for 46 attributes, where i ranges from 0 to 45.

Step 4: Let $\{DT_i\}$ be the significant attributes determined by Decision Tree model for 46 attributes, where i ranges from 0 to 45.

Step 5: Let $\{GB_i\}$ be the feature importance computed by Gradient Boosting Classifier for 46 attributes, where i ranges from 0 to 45.

Step 6: Let $\{XG_i\}$ be the feature importance computed by Extreme Gradient Boost (XG Boost) Classifier for 46 attributes, where i ranges from 0 to 45.

Step 7: For $i = 0$ to 45

$$TOT_i = \{RF_i\} + \{ET_i\} + \{DT_i\} + \{GB_i\} + \{XG_i\}$$

Step 8: For $j = 0$ to 45

$$AVG_j = TOT_j / 5$$

Step 9: Compute the Median M of AVG_j , for $j = 0$ to 45.

Step 10: For the threshold = M , Obtain all the attributes with feature importance greater than the threshold M as a Sub-Feature Set F_1 .

Sub Feature set F_1 has 23 features, by the end of Tree Based Intrinsic Feature Selection.

ANNOVA -F Test Statistical Filter:

Spreading of datapoints around the mean is the variance measure. When individual points of data move apart from the mean, high variances occur.

Step 11: Using the test value, extract 23 Sub Feature Set F_2 .

Recursive Feature Elimination (RFE):

Recursive feature elimination is a conventional method that follows backward feature elimination. It first fits the model with all the features in a given existing set, then progressively one by one we remove the insignificant features. Every time it is re-fitted, until the desired number of features are left and the parameter `n_features_to_select` is set. Scikit-learn has this routine.

Step 12: By applying RFE, compute the 23 important features as shown in Table 1 and create a Sub Feature Set F_3 .

Table. 1 RFE Ranking

Column	Rank	Features
11	1	WLOGP
19	1	Ali Log S
26	1	Silicos-IT class
3	1	Fraction Csp3
42	1	PAINS #alerts
30	1	CYP1A2 inhibitor
36	1	Lipinski #violations
33	1	CYP2D6 inhibitor
31	1	CYP2C19 inhibitor
32	1	CYP2C9 inhibitor
27	2	GI absorption
45	3	Synthetic Accessibility
23	4	Silicos-IT LogSw
2	5	#Aromatic heavy atoms
15	6	ESOL Log S
9	7	iLOGP
44	8	Leadlikeness #violations
10	9	XLOGP3
14	10	Consensus Log P

BORUTA:

Boruta algorithm is an advanced wrapper algorithm that is built around the random forest technique. It efficiently captures all the significant attributes you might have in your dataset with respect to an outcome variable. In the first step, it duplicates the features of the dataset, creating a shadow like effect and then shuffles the values.

Boruta is a wrapper algorithm used for significant feature selection. Traditional feature selection methods rely on a sub-feature set of attributes and produces a minimal error on any selected classifier. In every iteration, the variables are eliminated. Whereas Boruta algorithm is an advanced feature selection method that is most suitable for Cheminformatic Data. Any suitable type of classifier can be used for ranking and in this methodology, XGBoost has been utilized and its performance is better than the regularly used Random Forest Classifier.

Boruta algorithm selects the important features and rejects the insignificant features using Z- Score. In Table. 2 the outcome of the Boruta algorithm for the dataset is shown where features ranked as 1 are important features.

Algorithm:

Step 1: Let the Original feature set O_i where $i = 0, 1, \dots, n$.
 Step 2: Create a shadow variable for each attribute O_i as Sh_i where $i = 0, 1, \dots, n$.
 Step 3: A classifier is fit to the data and the Z-Score is computed for all the original features and shadow features.

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad (1)$$

Where \bar{x} is the sample mean, n indicates the average sample size, μ represents the mean of population and

σ indicates the standard deviation of the population.

Step 4: Compute the Maximum Z-Score among the shadow features and assign that to Mx_S .

Step 5: For $j = 0, 1, \dots, m$, Check $R_j > Mx_S$

R_j indicates the Z-score obtained by the features.

If $R_j > Mx_S$, Choose the feature R_j as important and

Else if $R_j = Mx_S$, Assume it as tentative|

Else if $R_j < Mx_S$ then Reject the feature.

Step 6: End when all features are checked.

The procedure ends when all the attributes are either accepted or rejected.

Step 13: By applying BORUTA, compute the 23 important features as shown in Table 2 and create a Sub Feature Set F_4 .

Table 2. Boruta Feature Importance

Column	Rank	Features
45	Synthetic Accessibility	1
13	Silicos-IT Log P	1
32	CYP2C9 inhibitor	1
11	WLOGP	1
35	log Kp (cm/s)	1
8	TPSA	1
7	MR	1
9	iLOGP	1
42	PAINS #alerts	1
2	#Aromatic heavy atoms	1
36	Lipinski #violations	1
20	Ali Solubility (mg/ml)	1
23	Silicos-IT LogSw	2
25	Silicos-IT Solubility (mol/l)	2
1	#Heavy atoms	3
43	Brenk #alerts	4
3	Fraction Csp3	5
24	Silicos-IT Solubility (mg/ml)	5
30	CYP1A2 inhibitor	7
0	MW	7
4	#Rotatable bonds	9
5	#H-bond acceptors	10
41	Bioavailability Score	11

Step 14: Compare Sub Feature Sets F_1 and F_2

Three populations or Feature sets are obtained namely Strong (S), Best (B) and Minor (M).

Strong feature set is obtained in three consecutive steps of $S_1, S_2, \text{ and } S_3$.

Best feature set is obtained in three consecutive steps of $B_1, B_2, \text{ and } B_3$.

Minor feature set is obtained in three consecutive steps of $M_1, M_2, \text{ and } M_3$.

Phase 1:

Step 15: $S_1 = F_1 \cap F_2, B_1 = \{\}$

Step 16: $M_1 = F_1 \Delta F_2$

Phase 2:

Compare Sub Feature Sets S_1, B_1, M_1 and F_3

Step 17: $S_2 = S_1 \cap F_3$

Step 18: $B_a = S_2$

Step 19: $B_b = S_1 - S_2$

Step 20: $B_c = M_1 \cap F_3$

Step 21: $B_2 = B_a \cup B_b \cup B_c$

Step 22: $M_a = S_2 \Delta F_3$

Step 23: $M_b = M_1 \Delta F_3$

Step 24: $M_2 = M_a \cup M_b$

Phase 3:

Compare Sub Feature Sets S_2, B_2, M_2 and F_4

Step 25: $S_3 = S_2 \cap F_4$

Step 26: $B_a = S_3$

Step 27: $B_b = S_2 - S_3$

Step 28: $B_c = B_2 \cap F_4$

Step 29: $B_d = M_2 \cap F_4$

Step 30: $B_3 = B_a \cup B_b \cup B_c \cup B_d$

Step 31: $M_a = S_3 \Delta F_4$

Step 32: $M_b = M_2 \Delta F_4$

Step 33: $M_3 = M_a \cup M_b$

Step 34: $S = S_3, B = B_3, M = M_3$

After Phase 3, three population sets S, B and M are obtained.

Grey-Wolf Algorithm:

Step 35: For three populations $x_i = S, x_i = B$ and $x_i = M$, apply Grey Wolf Algorithm.

```

Initialize the grey wolf population  $X_i (i = 1, 2, \dots, n)$ 
Initialize d, A and C
Generate the Randomly Positions of Search Agent
Calculation the fitness of each search agent
 $X_a$  = the best search agent
 $X_b$  = the second-best search agent
 $X_c$  = the third best search agent
While (t < max number of iterations)
  for each search agent
    Update the position of the current search agent by  $\vec{x}(t+1) = \vec{x}_1 + \vec{x}_2 + \vec{x}_3 / 3$ 
  End for
  Update d, A and C
  Calculation the fitness of all search agents
  Update  $X_a, X_b$  and  $X_c$ 
End while
Return  $X_a$ 
    
```

Apply the three Population Sets to the Grey wolf Optimizer [23]. The algorithm further using the meta heuristic technique, selects important features as shown in Fig. 6.

```

RangeIndex: 413 entries, 0 to 412
Data columns (total 17 columns):
#   Column                               Non-Null Count  Dtype
---  ---                               ---
0   MW                                     413 non-null    float64
1   #Heavy atoms                          413 non-null    float64
2   #Aromatic heavy atoms                 413 non-null    float64
3   Fraction Csp3                         413 non-null    float64
4   #Rotatable bonds                      413 non-null    float64
5   MR                                     413 non-null    float64
6   TPSA                                   413 non-null    float64
7   WLOGP                                  413 non-null    float64
8   Consensus Log P                       413 non-null    float64
9   ESOL Log S                            413 non-null    float64
10  Ali Log S                              413 non-null    float64
11  Ali Solubility (mol/l)                 413 non-null    float64
12  Silicos-IT LogSw                      413 non-null    float64
13  CYP1A2 inhibitor                      413 non-null    float64
14  Ghose #violations                     413 non-null    float64
15  Bioavailability Score                 413 non-null    float64
16  Synthetic Accessibility                413 non-null    float64
dtypes: float64(17)
memory usage: 55.0 KB
    
```

Fig 6. Features Selected by Grey Wolf Algorithm

Three machine learning methods like neural network based Multilayer Perceptron, tree based Random Forest and kernel-based Support Vector Machine were applied to the data with 17 features. An efficient multilayer perceptron (MLP) is one type of

Neural Network. It is based on multilayers of perceptron with three layers of input, the middle hidden and the final output layer.

There are numerous hidden layers that serve as a black box, and the output of one layer becomes the input for the following layer. It is a supervised machine learning technique in which the input data is trained using the backpropagation technique. The total number of hidden layers in the Neural Network is represented by the length of the tuple. For multilayer perceptron, there are a variety of activation functions. There were 17 inputs to the network. The hidden layer sizes were set to (500,20), resulting in 20 layers with 500 nodes each. The option max_iter has been set to 1000, which determines the number of epochs or iterations that the MLP will do. One epoch is equal to one cycle of the feed forward and backpropagation phases. The Max_iter option has been set to 1000, which determines the number of epochs or iterations that the MLP will do. The RELU (rectified linear unit) activation function was utilised. As a solver, 'adam' was utilised.

5. Results and Discussions

The features were selected using four feature selection methods like Tree Based, F-score, Recursive Feature Elimination and BORUTA Feature selection method. The methodology is implemented using scikit-learn [24]. Various metrics were computed as shown in Table. 3.

Table 3. Evaluation Metrics

METRICS	DEFINITION
Precision	$\frac{True\ Positives}{True\ Positives + False\ Positives}$
Recall	$\frac{True\ Positives}{True\ Positives + False\ Negatives}$
F1 – Score	$\frac{2 * Precision * Recall}{Precision + Recall}$
Sensitivity	$\frac{True\ Positives}{Positives}$
Specificity	$\frac{True\ Negatives}{Negatives}$
Accuracy	$Sensitivity * \frac{Positives}{Positives + Negatives} + Specificity * \frac{Negatives}{Positives + Negatives}$

The models are evaluated using the confusion matrix, Sensitivity, Accuracy, Specificity, F1-Score, Precision and Recall as shown in TABLE III.

When the performance of the machine learning model is relatively high, the other important factors to be analyzed are overfitting and underfitting. Fig. 7 shows the performance of the three machine learning models Multi-Layer Perceptron, RF and kernel based SVM on the three pools of population after applying the grey wolf Optimizer. Random Forest and Multilayer Perceptron had higher prediction accuracies while comparing to the Support Vector Machines.

Sometimes, the data is over trained but testing performs less. The data should neither be over trained or less trained, which greatly has impact on the prediction accuracy



Fig.7. Performance of the ML classifiers

Traditional machine learning models and ensemble models were applied to the dataset. Multilayer perceptron achieved highest prediction accuracy with 97.7%. while Random Forest had a prediction accuracy of 92%. For cheminformatic data, Multilayer perceptron performs efficiently compared to other machine learning models. Table 4. Shows the performance of the Strong feature Set. Pipelined framework helps to overcome data leakage issues [25].

Table 4. Confusion Matrix- STRONG Population

MLP	0	1
0	45	8
1	2	76

Training	95.5
Testing	92.4

RF	0	1
0	43	10
1	1	77

Training	92.1
Testing	91.6

SVM	0	1
0	41	11
1	5	67

Training	87.7
Testing	87.1

Table 5. Confusion Matrix- BEST Population

MLP	0	1
0	52	1
1	2	76

Training	98.7
Testing	97.7

RF	0	1
0	47	4
1	6	67

Training	94.4
Testing	92.4

SVM	0	1
0	46	3
1	11	64

Training	89.2
Testing	88.7

Table 6. Confusion Matrix- MINOR/TOTAL population

MLP	0	1		
0	48	5	Training	94.7
1	3	75	Testing	93.9

RF	0	1		
0	46	5	Training	92.7
1	7	66	Testing	90.3

SVM	0	1		
0	47	5	Training	88.7
1	12	60	Testing	86.2

Table 4, 5 and 6 show the Confusion Matrices for the three populations Strong, Best and Minor/Total. The colored diagonal elements show the correctly classified elements.

Table 7. Performance of the ML models on the Leukemia Dataset

Classifiers	Training	Testing
MLP (Strong)	0.985	0.953
RF	0.975	0.917
SVM	0.897	0.869
MLP (Best)	0.994	0.962
RF	0.994	0.928
SVM	0.911	0.897
MLP (Total)	0.983	0.946
RF	0.953	0.906
SVM	0.877	0.854

Table 7 shows that for both the classifications, the Best pool of population had higher accuracy of 96%. MLP has performed better than the other 2 models.

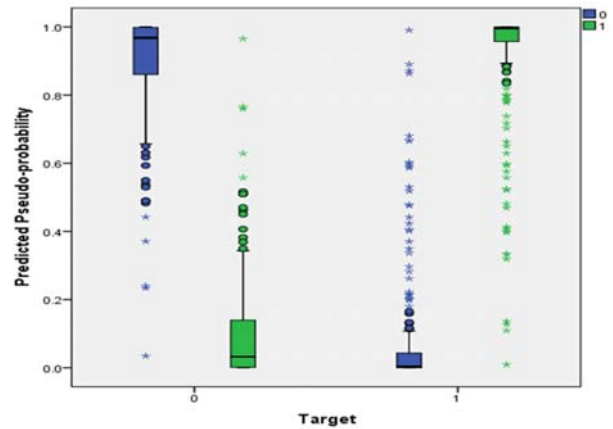


Fig 8. Box Plot

ROC

Area Under the Curve		Area
Target	0	.989
	1	.989

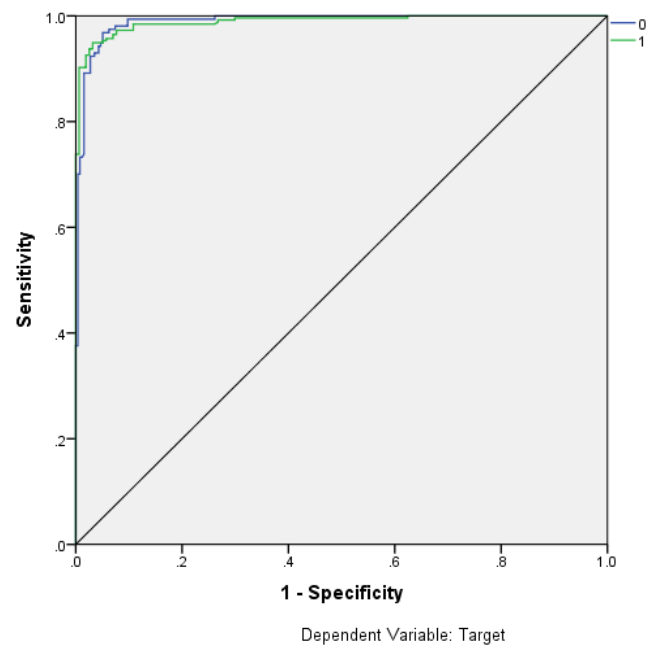


Fig 9. MLP generated ROC Curve

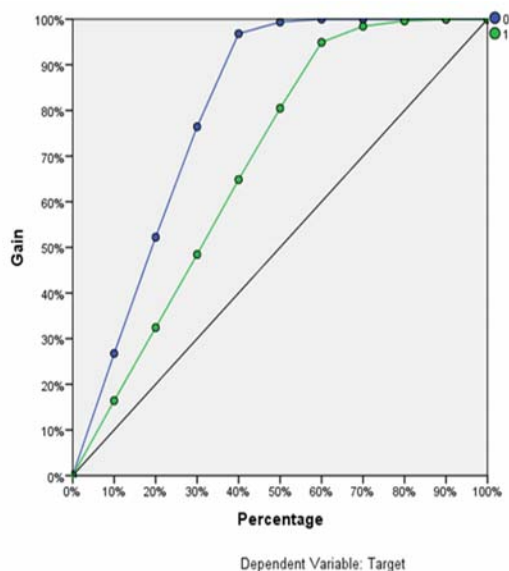


Fig 10. MLP generated Cumulative Gain Chart

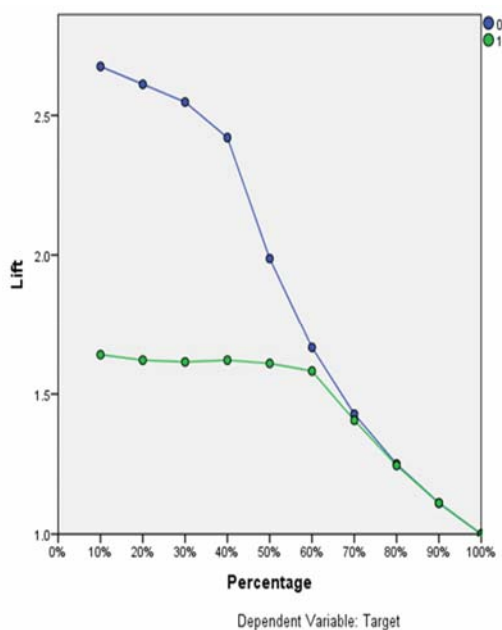


Fig 11. MLP generated Lift Chart

Fig 8. And Fig 9. Shows the Box Plot and ROC generated by the Multilayer Perceptron. Similarly, Fig. 10 and Fig. 11 shows the Cumulative Gain Chart and Lift Chart generated by the MLP [26]. Multilayer perceptron has proved to be the best performing ML models in the domain of drug classification and disease prediction [27][28].

6. Conclusion and Future Work

Huge investment is made by the pharmaceutical companies and research laboratories to produce a novel drug. There are millions of other compounds with medicinal properties. Repurposing of drugs is very crucial since there are already discovered enormous drugs. The proposed methodology has engineered the features and the novel procedure has increased the prediction accuracy of cancer drugs to 97.7%. With the Leukemia dataset, MLP achieved 96% accuracy. The most significant features can be obtained through this Multi-Level Median Based Feature Ranking Method (MLMBFRM). With increase in the prediction accuracy, the immediate challenge that arises is either the overfitting or underfitting. The other issue is the error function. This methodology has efficiently, overcome both the issues. The existing methodology can further be effectively extended by enhancing several feature engineering methods to optimize the classification accuracy of the ML models. Fine tuning of hyperparameters would further contribute to the increased prediction accuracy of the models. Augmentation of feature set and data can be done. Novel methodologies and frameworks can be designed to suit the cheminformatic data. Predictions can be optimized using hybrid optimization algorithms. Select the best feature set for optimal Performance by the models. Novel approaches may be implemented to increase high prediction accuracy for other diseases.

References

- [1] <https://www.genome.jp/kegg/drug>
- [2] Qian Xu & Qiang Yang. (2011). A Survey of Transfer and Multitask Learning in Bioinformatics. *Journal of Computing Science and Engineering*, 5(3), 257-268.
- [3] Prashant Singh Rana, Harish Sharma, Mahua Bhattacharya & Anupam Shukla. (2015). Quality assessment of modelled protein structure using physicochemical properties. *Journal of bioinformatics and computational biology*.13(02).
- [4] Mathew,T.E. (2019). A Logistic Regression with Recursive Feature Elimination Model for Breast Cancer Diagnosis. *International Journal on Emerging Technologies*, 10(3): 55–63.
- [5] E. A. Bayrak, P. Kırıcı and T. Ensari, "Comparison of Machine Learning Methods for Breast Cancer Diagnosis," 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT),

- Istanbul, Turkey, 2019, pp. 1-3, doi: 10.1109/EBBT.2019.8741990.
- [6] H. Motohashi, T. Teraoka, S. Aoki and H. Ohwada, "Regression Models and Ranking Method for p53 Inhibitor Candidates Using Machine Learning," 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, 2018, pp. 708-712, doi: 10.1109/BIBM.2018.8621142.
- [7] David Chen, Gaurav Goyal, Ronald Go, Sameer Parikh, Che Ngufor, "Predicting Time to First Treatment in Chronic Lymphocytic Leukemia using Machine Learning Survival and Classification Methods", IEEE International Conference on Healthcare Informatics, IEEE, 2018.
- [8] Dharani T and Hariprasath S, "Diagnosis of Leukemia and its types Using Digital Image Processing Tehniques", Proceedings of the International Conference on Communication and Electronics Systems (ICCES), IEEE, 2018.
- [9] Preeti Jagadev and H.G.Virani, "Detection of Leukemia and its Types using Image Processing and Machine Learning", International conference on Trends in Electronics and Informatics (ICEI), IEEE 2017.
- [10] Jakkrich Laosai and Kosin Chamnongthai, "Deep-learning-Based Acute Leukemia Classification Using Imaging Flow Cytometry and Morphology", International Workshop on Smart Info-Media Systems in Asia (SISA), IEEE, 2018.
- [11] Sachin Paswan and Yogesh Rathore, "Recognition and Arrangement of Blood Cancer from Microscopic Cell pictures Utilizing Support Vector Machine K-Nearest Neighbor and Deep Learning", International Conference on Communication, Computing and Internet of Things (IC3IOT), IEEE, 2018.
- [12] Jiang, D., Lei, T., Wang, Z. et al. ADMET evaluation in drug discovery. 20. Prediction of breast cancer resistance protein inhibition through machine learning. *J Cheminform* 12, 16 (2020). <https://doi.org/10.1186/s13321-020-00421-y>.
- [13] Borrero, Luz & Guette, Lilibeth & Lopez, Enrique & Pineda, Omar & Buelvas, Edgardo. (2020). Predicting Toxicity Properties through Machine Learning. *Procedia Computer Science*. 170. 1011-1016. [10.1016/j.procs.2020.03.093](https://doi.org/10.1016/j.procs.2020.03.093).
- [14] Baptista D, Ferreira PG, Rocha M. Deep learning for drug response prediction in cancer. *Brief Bioinform*. 2021 Jan 18;22(1):360-379. Doi: 10.1093/bib/bbz171. PMID: 31950132.
- [15] Lind AP, Anderson PC. Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties. *PLoS One*. 2019 Jul 11;14(7): e0219774. Doi: 10.1371/journal.pone.0219774. PMID: 31295321; PMCID: PMC6622537.
- [16] A. Jha, G. Verma, Y. Khan, Q. Mehmood, D. Rebholz-Schuhmann and R. Sahay, "Deep Convolution Neural Network Model to Predict Relapse in Breast Cancer," 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), 2018, pp. 351-358, Doi: 10.1109/ICMLA.2018.0005
- [17] C. Chen, L. Song, C. Bo and W. Shuo, "A Support Vector Machine with Particle Swarm Optimization Grey Wolf Optimizer for Network Intrusion Detection," 2021 International Conference on Big Data Analysis and Computer Science (BDACS), 2021, pp. 199-204, doi: 10.1109/BDACS53596.2021.00051.
- [18] J. Liu, X. Wei and H. Huang, "An Improved Grey Wolf Optimization Algorithm and its Application in Path Planning," in *IEEE Access*, vol. 9, pp. 121944-121956, 2021, doi: 10.1109/ACCESS.2021.3108973.
- [19] M. -Z. Tsai, P. -Y. Yang, F. -I. Chou and J. -H. Chou, "Parameters Optimization for Improved Grey Wolf Optimizer by Using Uniform Experimental Design," 2021 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), 2021, pp. 1-2, doi: 10.1109/ISPACS51563.2021.9651072.
- [20] S. Subbiah, K. S. M. Anbananthen, S. Thangaraj, S. Kannan and D. Chelliah, "Intrusion detection technique in wireless sensor network using grid search random forest with Boruta feature selection algorithm," in *Journal of Communications and Networks*, vol. 24, no. 2, pp. 264-273, April 2022, doi: 10.23919/JCN.2022.000002.
- [21] <https://www.ebi.ac.uk/chembl/>
- [22] <http://www.swissadme.ch/>
- [23] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," *Adv. Eng. Softw.*, vol. 69, pp. 46–61, Mar. 2014
- [24] <https://scikit-learn.org/stable/>
- [25] G Shobana, N Priya, A New Multi-Phase Feature Selection Framework for The Prediction of Breast Cancer Drug Using Machine Learning Techniques, *Journal of Algebraic Statistics* 13 (2), 300-312(2022).
- [26] <https://www.ibm.com/in-en/products/spss-statistics>
- [27] G. Shobana and S. N. Bushra, "Classification of Myopia in Children using Machine Learning Models with Tree Based Feature Selection," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1599-1605, doi: 10.1109/ICECA49313.2020.9297623.
- [28] G. Shobana and S. N. Bushra, "Prediction of Cardiovascular Disease using Multiple Machine Learning Platforms," 2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICES), 2021, pp. 1-7, doi: 10.1109/ICES52305.2021.9633797.



G. Shobana is currently working as Assistant Professor in the Department of Computer Applications, Madras Christian College, Chennai. She completed her M.E (Computer Science and Engineering) in 2006 from Sathyabama University, Chennai. She completed her Masters in

Bioinformatics from Bharathiar University with distinction. She has more than 15 years of teaching experience and has published more than 20 research articles. Her research areas include Machine Learning, Bioinformatics, Cheminformatics and Cloud Computing.



Dr. N. Priya has more than 17 years of teaching experience and has published more than 30 research articles. She completed both her UG and PG Degrees with distinction. She received her Ph. D from Bharathiar University Coimbatore. She has chaired many sessions at international

conferences and holds several patents. She is a recognized research supervisor under the University of Madras. Her research areas include datamining, image processing, Neural Networks, Network programming and Fuzzy Logic. Currently she is working as Associate Professor in the Research Department of Computer Science, SDNB Vaishnav College for Women.