JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 300-312
https://publishoa.com
ISSN: 1309-3452

A New Multi-Phase Feature Selection Framework for The Prediction of Breast Cancer Drug Using Machine Learning Techniques

G. Shobana,
Assistant Professor, Department of Computer Applications of Madras Christian College, Chennai, India. Affiliated to University of Madras.
e-mail: gmshobana@gmail.com.

Dr. N. Priya,
Associate Professor, PG Department of Computer Science of Shrimathi Devkunvar Nanalal Bhatt Vaishnav College for Women, Chennai, India, Affiliated to University of Madras

ABSTRACT

Cancer is one of the slowly progressing diseases that exhibit symptoms only at the later stage of the disease. Cancer that is common among women is breast cancer and in recent years, the total number of women affected has elevated to a higher number across the globe. It is more prevalent in the western world than on the other side of the world due to varied food habits and stressful lifestyles. To understand the various factors that contribute to the development of the disease, to classify the disease as benign or malignant and for predicting the disease several machine learning models were employed. In a similar perspective, machine learning models can be also be utilized to identify or predict potential breast cancer drugs and classify them. This computational approach helps in reducing experimental costs that incur during the pre-clinal trials and enables to filter few potential drugs among millions of compounds available. The result relies on the type of feature set or attributes considered for the study. Prediction of the drug is determined based on the feature set that defines the physicochemical, lipophilicity, water-solubility, pharmacokinetics, and drug-likeness properties of the compound. In this paper, a new multiphase feature selection with pipelined methodology is proposed that enhances the prediction accuracy of the breast cancer drug. This study further investigates the significance of feature selection and its impact on the predicted result. Multilayer perceptron model obtained high accuracy of 94.7% compared to the other supervised machine learning models.

Index Terms—Cancer, Breast Cancer Drug, Machine Learning Models, Multilayer Perceptron

I. INTRODUCTION

Cancer is a global disease that occurs due to the irregular proliferation of human cells. WHO states that the current world cancer scenario has changed drastically over the past decade. According to the reports of the International Agency for Research on Cancer, the number of breast cancer cases has increased compared to lung cancer cases across the globe.

Breast cancer and Cervical cancer are the cancer types that affect many women throughout the world and their prediction at an early stage is a challenging task. Due to the present systematic and mechanical lifestyle, many women procrastinate their health check-ups, leading to the progress of

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 300-312
https://publishoa.com
ISSN: 1309-3452

the disease [1]. Currently, there are more than 7.8 million women suffering from breast cancer around the globe. Cysts
that develop inside the breast can be divided into benign and malignant. Fibroadenomas are benign tumours that can be treated completely with appropriate medication. Malignant tumours are persistent lumps that are detectable only through
biopsy of the affected tissue. This cancer gradually spreads through the lymph node and over time damages other vital organs like the brain, liver, and lungs. Metastasis is a stage of breast cancer where the cancer cells proceed to invade other organs. Breast cancer can affect women of any age after puberty but the higher risk group is women above forty. The genes PALB-2, BRCA1, and BRCA2, when undergoes mutation, the risk of breast cancer is higher and such patients are likely to have a family history.

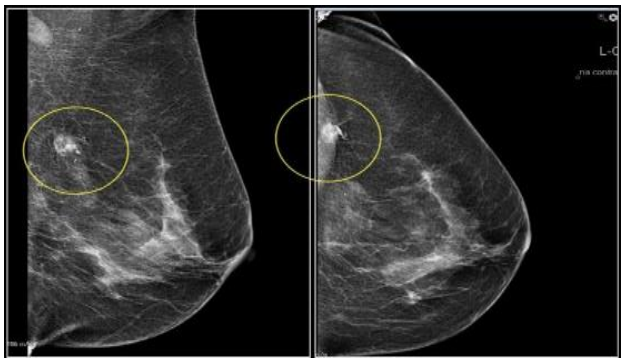

Fig. 1. Mammogram of Breast Cancer

The common symptoms are abnormal lump, discharge, or change in the skin. Depending on the stage or progress of the disease treatments are administered to the patients which include surgery, chemotherapy, hormonal therapy, biological therapy, and targeted therapy. Mastectomy is the surgical procedure that removes part or the entire breast to prevent the spread of the disease to other parts of the body [2]. Breast cancer tumours can be diagnosed through biopsy and a mammogram helps in identifying the location of the cyst as shown in Fig 1. Several treatments are given to the patients based on the level of the disease. In systematic therapy, the drugs are administered to the patients either orally or intravenously, which helps to prevent the growth of the cancer cells.

Early diagnosis of the disease greatly reduces the mortality rate. Researchers have employed machine learning methods to detect the disease using voluminous data which includes both images and quantitative data. They have classified the cancer disease as benign-type and malignant-type by applying statistical and advanced ML models. Accordingly, the same methodology can be applied in the classification of breast cancer drugs. In this paper, a novel methodology is proposed to overcome the challenges that occur during the feature selection process. Relevance of the feature plays a very crucial role in determining the prediction accuracy and the methodology also minimizes the overfitting of the models. Initially, the dataset was applied to the machine learning models without implementing the methodology and the results were observed. In the next step, the novel methodology was applied to the dataset and the results were recorded. The compared results indicate that the methodology implemented enhanced the overall prediction accuracy of all the models and Multilayered Perceptron achieved the highest prediction accuracy of 94.7%. The drug

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 300-312
https://publishoa.com
ISSN: 1309-3452

discovery domain has enormous data regarding compounds that have pharmaceutical properties and filtering the most potent drugs among millions of compounds is a tremendous, time consuming and challenging task. Proceeding the research only through chemical and biological lab experiments is time-consuming and incurs huge lab costs. To overcome these drawbacks machine learning algorithms with feature engineering can be applied and more prediction accuracy can be obtained.

II. Related work

Qian Li et al implemented the Deep Learning technique to predict cancer drugs. They proposed a multi-fusion procedure that involves CNN and LSTM (Long and Short-Term Memory). The dataset had gene expression data from GDSC and COSMIC databases for cell lines. The proposed integrated model of neural network achieved the prediction accuracy of 84% [3].

Yanpeng Qu proposed a novel approach for detecting breast cancer using mammographic images. They implemented feature selection using a Rough-fuzzy algorithm. They incorporated the Fuzzy-Rough algorithm with four machine learning models like NB, LR, K-Nearest Neighbour and RF. The proposed Multi-Label Fuzzy Rough Feature Selection had increased the overall performance of the machine learning models and the T-test revealed better classification results [4].

Luca Parca et al investigated pharmacogenomics data using different machine learning techniques. They applied Elastic Net-based Regression Model to overcome the limitations of LASSO and Ridge. They also utilized RF and Support Vector Regression for gene-based drug prediction [5].

Dejun Jiang et al used various machine learning models to predict breast cancer inhibitor protein. They used traditional machine learning algorithms like NB, K-NN, LR and SVM. They evaluated the Absorption, Distribution, Metabolism, Excretion and Toxicity (ADMET) properties of the drug. Feature selection was implemented using the SA algorithm combined with RF. XGBoost and Deep Neural Network had higher prediction accuracy than the four traditional ML algorithms [6].

Luz Adriana Borrero et al predicted toxicity using Machine Learning techniques. ANN, Decision Tree, NB, K-NN, RF and SVM were used for toxicity classification. The Decision tree achieved an accuracy of 89%. Data for the study was taken from admetSAR web [7].

Delora Baptista et al discussed various Deep Learning architectures used to predict cancer drugs. They concluded that DNN had high prediction accuracy compared to other traditional models. DNN has proved very efficient in the field of Drug Discovery [8]. Alex P. Lind et al investigated the action of 225 drugs against 990 cancer cell lines by using different machine learning techniques. Random Forest had high prediction accuracy [9].

Alok Kumar Jha et al demonstrated the use of Graph Convolution Neural Network to predict cancer. GCNN prove to perform better than RF, SVM and simple Neural Network [10].

III. Dataset and attributes

Breast Cancer drug names were drawn from the KEGG database [11]. FDI approved drug names from NCI were also obtained [12]. All the drugs taken for the experiment are approved and are used currently for breast cancer treatment. 85% of similar drugs are taken from the CHEMBL and the dataset for augmentation [13]. 46 features are generated for each drug and are pre-processed. 256 drugs are classified as cancer drugs and 157 drugs are classified as non-cancer drugs. These attributes or variables are computed by the SwissADME tool. SwissADME is a tool that generates

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 300-312
https://publishoa.com
ISSN: 1309-3452

various medicinal chemical properties of drugs and hence facilitates drug discovery procedures. Swiss Institute of Bioinformatics maintains this user-friendly tool which is widely used among researchers. Several data generated through this tool are widely used in various domains of computational chemistry, pharmaceutical field, bioinformatics, cheminformatics and most recently in the prediction of drug toxicity. The attributes define the Pharmacokinetics, Drug likeness, Lipophilicity, Water Solubility, Medicinal Chemistry and Physicochemical Properties of the drug molecules. Physicochemical Properties include the attributes like Molecular Refractivity, Molecular weight, presence of Heavy atoms, presence of Aromatic Heavy atoms, Fraction Csp3, Number of rotatable bonds, Number of H-bond acceptors, Number of H-bond donors and TPSA (Topological Polar Surface Area). Lipophilicity includes iLOGP, MLOGP, WLOGP, X LOGP, SILICOS-IT and Consensus LOGP which is the average of the other five. Water Solubility consists of LogS (ESOL), LogS (Ali) and LogS (SILICOS-IT) properties where Solubility class is defined by LogS Scale which range like Insoluble < -10 < poorly < -6 < Moderately < -4 < Soluble < -2 < very Soluble < 0 < Highly Soluble. Pharmacokinetics includes properties like Gastro Intestinal absorption, BBB permeation and SVM model based trained and tested p-gp Substrate, CYP1A2 inhibitor, CYP2C19 inhibitor, CYP2C9 inhibitor and CYP3A4 inhibitor. Skin Permeation (LogKp) is also included in the pharmacokinetics features. Medicinal Chemistry property is defined by Pan Assay Interference Structures, Brenk and Leadlikeness properties. The features also include

Bioavailability Score. Druglikeness is one of the crucial properties of the drug that helps in determining its toxicity.

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

Fig. 2. Lipinski and Ghose Filter Parameters

The Toxicity of the drug is one of the major concerns during the synthesis of a new compound [14]. The toxicity level varies from drug to drug and depending on their administration route, it differs. Some commonly used filters for toxicity prediction or drug-likeness properties are Lipinski and Ghose filters as shown in Fig.2. The drugs considered for the experiment are already approved drugs and is drawn from the reputed database. Their biological activity for the specified disease has already been proved. 413 observations are taken for the experiment with 46 features. The first step in pre-processing is the data cleaning where the noise, redundant data are removed and missing data are filled. All the categorical data are changed to numerical data to employ the machine learning process.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 300-312
https://publishoa.com
ISSN: 1309-3452

IV. PROPOSED METHODOLOGY

In Cheminformatic data, the feature set plays a crucial part in determining the prediction accuracy. When the attributes are too many, there is the possibility of features being correlated. During the feature selection procedures, highly collinear attributes or variables are eliminated before the commencing of the Selection Process. The existing challenge is that the feature selection procedure is given limited importance while applying machine learning models to cheminformatic data, especially during drug repurposing. The proposed model is divided into three major phases. The first two phases deal with selecting the most important features with irrelevant features removed. The third phase is a pipelined procedure where various classifiers are implemented after the traditional Recursive Feature Elimination. The data is normalized before applying the classifiers as shown in Fig.3. The need for the pipeline is to avoid data leak and at the same time parallel implementation of various machine learning classifiers is achieved.

The Feature selection process is a very crucial aspect that has to be performed before implementing the machine learning classifiers. The main objective of feature reduction is to prune highly correlated attributes which play an insignificant role in determining the outcome.

A. Mutual Information

The proposed Framework is divided into three major phases. The first phase deals with the basic filtering procedure. Based on Mutual Information, three insignificant features are eliminated from 46 features to reduce highly correlated attributes. This is mandatory because highly correlated variables might result in the declined performance of Boruta algorithm. Three insignificant features are eliminated. With fewer collinear features, the refined data is ready for the application of Boruta algorithm.

The data passes the filter section i.e., the dataset with 43 features is passed to the Second phase of the Boruta algorithm. Three features were found to be insignificant.

B. Boruta Algorithm

Boruta is a wrapper algorithm used for feature selection. Traditional feature selection methods rely on a sub-feature set of attributes and produce a minimal error on any selected classifier. In every iteration, the variables are eliminated. Whereas the 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.

Algorithm:

Step 1: Let each attribute in the data set be A_i where $i = 0,1, \dots n$.

Step 2: Create a shadow variable for each attribute A_i as S_i where $i = 0,1, \dots n$.

Step 3: Fit a classifier and compute the Z-Score for all the original features and shadow features.

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

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

σ is the standard deviation of the population.

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

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 300-312
https://publishoa.com
ISSN: 1309-3452

Step 5: For $j = 0,1, \ldots m$, Check O_j > Max_S

If O_j> Max_S, Select the feature O_j as important and
Else if O_j = = Max_S, Assume it as tentative
Else if O_j < Max_S then Reject the feature.

Step 6: End when all features are checked.

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

Boruta algorithm selects the important features. It rejects the insignificant features using Z-Score. In Fig. 4 the outcome of the Boruta algorithm for the dataset is shown where features ranked as 1 are important features.

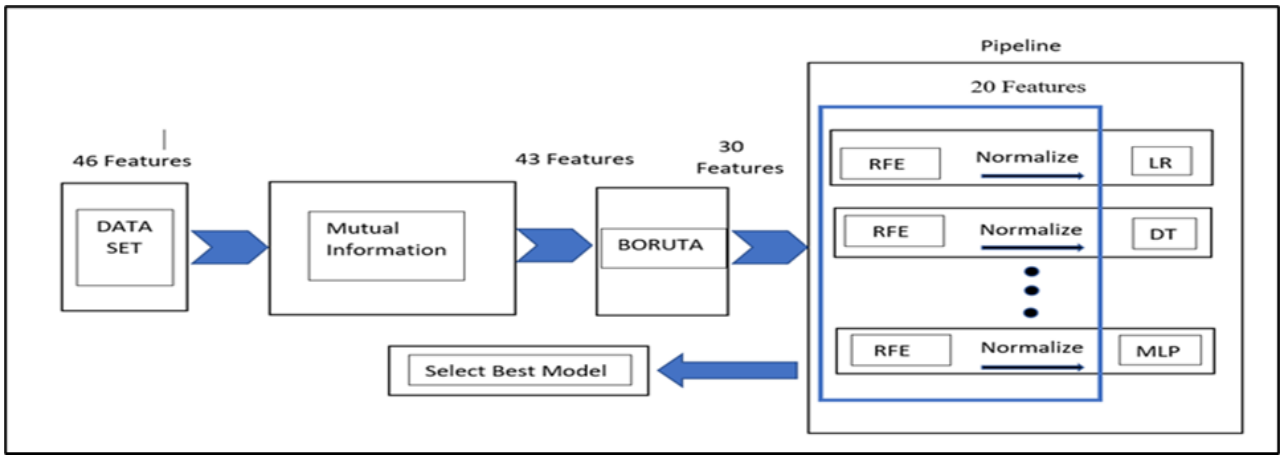

Fig. 3. Proposed Framework

Attribute Column	Feature	Ranking
0	MW	1
34	Lipinski#Violations	1
33	Log Kp (cm/s)	1
24	Silicos-IT Solubility (mol/l)	1
23	Silicos-IT Solubility (mg/ml)	1
22	Silicos-IT LogSw	1
17	ESOL Solubility (mol/l)	1

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 300-312
https://publishoa.com
ISSN: 1309-3452

16	ESOL Solubility (mg/ml)	1
15	ESOL LogS	1
14	Consensus Log P	1
13	Silicos-IT Log P	1
12	MLogP	1

Fig.4 Boruta Ranking of Features

C. Pipelined Classifiers

After the Boruta technique was applied to the data, the features were reduced to 30. The reduced features were passed to pipelined multiple classifiers. A pipeline is used when pre-processing ends with a model. Any type of transformation steps can be implemented and the classifiers can be added to the pipeline. The greatest advantage of the pipeline is that it reduces data leakage and simplifies the coding. When the same feature selection method is applicable for all the classifiers then pipeline would be the best technique.

- Apply Recursive Feature Elimination.
- Reduce the important features to 20.
- Normalize the data.
- Employ the Classifiers

After the third phase, it is observed that among all the classifiers, Multi-layer Perceptron performed better than the other traditional machine learning methods. A multilayer perceptron is one type of ANN which is a feedforward technique. It is based on a multilayer of perceptron with three layers of input, hidden and the final output layer. The output of one layer becomes the input for the next layer and there are multiple hidden layers that act as a black box. It is a supervised machine learning technique that used the backpropagation technique for training the input data. There are many activation functions available for multilayer perceptron. The inputs given to the network was 20. The input data were scaled using the standardized method and one hidden layer with 8 nodes was used. Hyperbolic tangent was the activation function used in the hidden layer. The output layer has 2 units and the activation function used was SoftMax, while the error function investigated was Cross-Entropy.

Original Features			**Reduced Features**		**New Methodology**	
Classifiers	Train	Test	Train	Test	Train	Test

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 300-312
https://publishoa.com
ISSN: 1309-3452

LR	0.913	0.790	0.921	0.839	0.879	0.855
DT	0.914	0.790	0.927	0.831	0.892	0.843
SVM	0.960	0.718	0.971	0.863	0.994	0.880

TABLE. I. Performance of new methodology

V. RESULTS AND DISCUSSION

The features were selected using RFE during initial research and the result was observed. This new methodology increased the performance of the three machine learning models LR, DT and SVM. Table 1 show the training and testing details of the dataset with original features, with only RFE reduction and with new sequential methodology [16]. The methodology is implemented using scikit-learn [15].

Fig.5 shows the comparison of the performance of the three ML models with RFE reduced Feature Set1(FS-1) and the Feature Set2 (FS-2) selected by adapting the new methodology [17]. Among the three models SVM performed better and in all the three models the new methodology has yielded better results [18]. Table 1 is a comparative result of research articles where the data was processed with only

traditional feature selection technique [18] [19].

TABLE. II. Metrics

Metrics	**Definition**
Precision	True Positives/(True Positives+False Positives)
Recall	True Positives/ (True Positives + False Negatives)
F1-Score	(2*Precision*Recall)/(Precision+Recall)
Sensitivity	True Positives/Positives
Specificity	True Negatives/Negatives
Accuracy	Sensitivity* ((Positives/(Positives+ Negatives)) + Specificity*((Negatives/(Positives+ Negatives))

The ML models were evaluated using the Sensitivity, Accuracy, Specificity, F1-Score, confusion matrix, Precision and Recall as shown in Table. II. Performance of all the models are given in

Table. III.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 300-312
https://publishoa.com
ISSN: 1309-3452

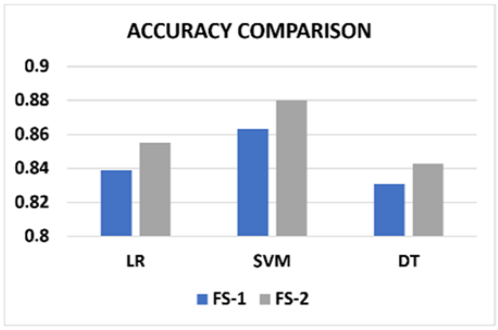

Fig.5 Compared Performance of the ML Classifiers

Table. III. ML Models and their relative performance

Weighted Average

Classifiers	Training	Testing	Precision	Recall	F1	TN	FP	FN	TP
LR	0.879	0.855	0.87	0.86	0.86	32	2	10	39
NB	0.776	0.687	0.70	0.69	0.65	31	3	4	45
KNN	0.952	0.916	0.92	0.92	0.92	12	22	4	45
DT	0.892	0.843	0.84	0.84	0.83	28	6	7	42
SVM	0.994	0.88	0.88	0.88	0.88	18	7	3	55
RF	0.938	0.916	0.93	0.92	0.91	29	7	0	47

When the performance of the machine learning model is relatively high, the other important factors to be analyzed are overfitting and underfitting. Fig. 6 clearly show the closer curved lines of training and testing. This proves that the result of the methodology adapted had overcome this problem and both the training and testing procedure has performed efficiently with less difference. 1, 2…7 in the X-axis indicates the machine learning models and their training and testing values have been plotted [19]. Sometimes, the data is over-trained but testing performs less. The data should neither be over-trained or less trained, which greatly has an impact on the prediction accuracy [20][22].

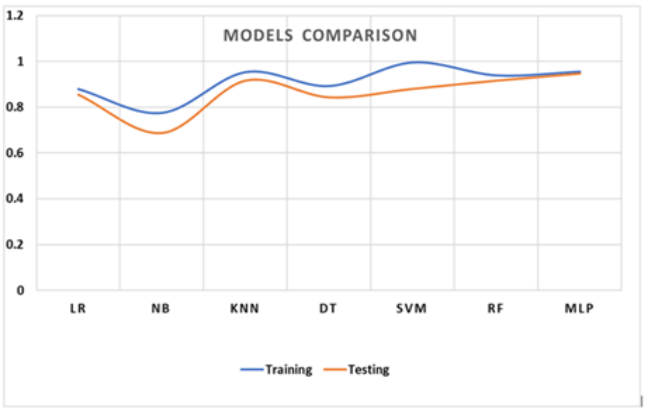

Fig. 6. Training and Testing Comparison

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 300-312
https://publishoa.com
ISSN: 1309-3452

Traditional machine learning models and ensemble methods were fitted to the dataset. Multilayer perceptron had the best prediction accuracy of 94.7%. K-NN and Random Forest had a prediction accuracy of 91%. Logistic Regression, Decision tree and SVM had a prediction accuracy of 85%, 84% and 88% respectively. Naive Bayes had a prediction accuracy of 68%.

Table. IV. Boosting Models

Boosting ML Models	Training	Testing	Precision	Recall	F1	TN	FP	FN	TP
GB	0.994	0.928	0.93	0.93	0.93	30	5	1	47
XGBOOST	0.979	0.940	0.94	0.94	0.94	31	4	1	47
LGB	0.934	0.916	0.93	0.92	0.91	28	7	0	48

The Boosting algorithms are also employed and all the three algorithms performed well with more than 90% accuracy as shown in Table IV.

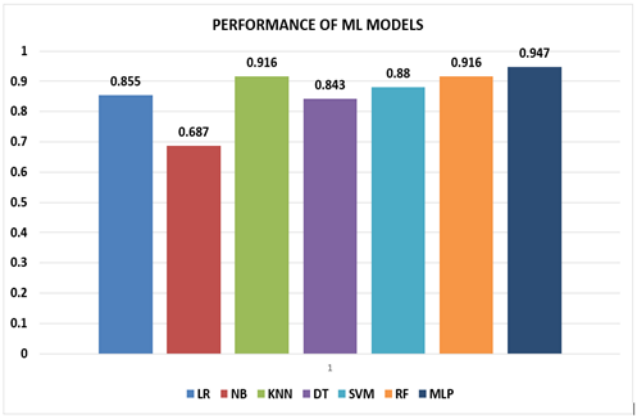

Fig. 7. Performance of traditional classifiers & MLP

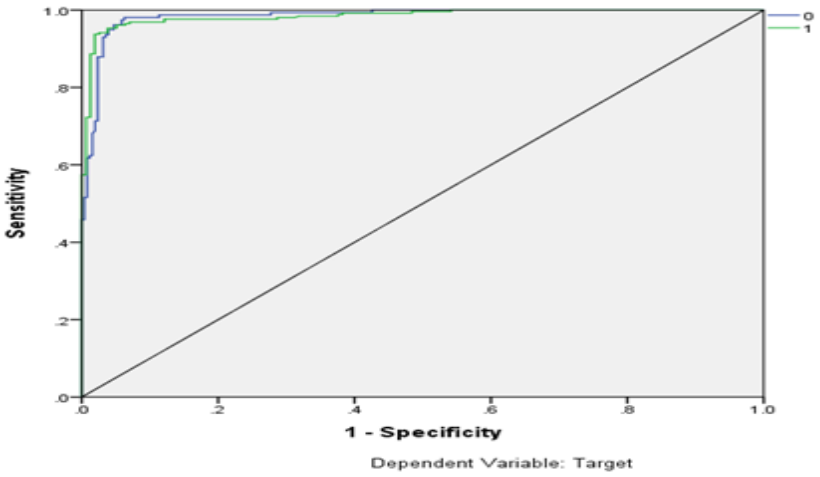

Fig.8. ROC generated by MLP

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 300-312
https://publishoa.com
ISSN: 1309-3452

Fig. 7 show the performance of both traditional and MLP classifiers. For cheminformatic data, Multilayer perceptron performs efficiently compared to other machine learning models. Fig. 8 show the ROC generated by Multi-Layer Perceptron.

Table. V. Performance of MLP

Multi-Layer Perceptron	**Without Feature Selection**	**Proposed Methodology**
Accuracy-Training	96.4%	95.4%
Accuracy-Testing	87.6%	94.7%
Cross Entropy Error	40.28%	24.62%
Incorrect Predictions	12.4%	5.3%

Table. V. shows the efficiency of the proposed methodology. The data was trained and tested in the ratio of 70:30. The model has achieved an accuracy of 94.7%. From the table, it is also known that the Cross-Entropy Error and incorrect prediction rate has drastically reduced, which proves the efficiency of the proposed methodology [23].

VI. Conclusion

Cancer is a disease that affects almost any part or organ of the body and relatively few drugs are available for breast cancer compared to other diseases. A potential drug undergoes various pre-clinical trials for several years before reaching the market. 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. Screening of the existing drug repositories would reveal more potent drug compounds. The proposed methodology has engineered the features and the pipelined procedure has increased the prediction accuracy of cancer drugs to 94.7%. With the increase in the prediction accuracy, the immediate challenge that arises is either overfitting or underfitting. The other issue is the error function. This methodology has efficiently, overcome both issues. Hence when a new set of features of any compound is given as input, the methodology would be able to classify the drug as a cancer drug or not. This procedure has proved efficient for breast cancer drug repurposing with supervised machine learning models. Ensemble methods and the traditional methods had relatively less accuracy compared to Multilayered perceptron. Further, a comparison between the existing and the proposed methodology has been demonstrated using the same dataset. Future enhancement of the research can incorporate toxicology tests as an integral module and application of novel advanced quantum machine learning models with feature augmentation and engineering. Implementation methodologies vary according to the dataset and features in particular. Hybrid Deep Neural Network would prove very efficient in drug classification with a large dataset.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 300-312
https://publishoa.com
ISSN: 1309-3452

REFERENCES

[1] https://www.cancer.gov/about-cancer/treatment/drugs/breast#2

[2] https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[3] Q. Li, J. Huang, H. Zhu and Q. Liu, "Prediction of Cancer Drug Effectiveness Based on Multi-Fusion Deep Learning Model," 2020 *10th Annual Computing and Communication Workshop and Conference (CCWC)*, 2020, pp. 0634-0639, doi: 10.1109/CCWC47524.2020.9031163.

[4] Qu Y, Yue G, Shang C, Yang L, Zwiggelaar R, Shen Q. Multi-criterion mammographic risk analysis supported with multi-label fuzzy-rough feature selection. *Artif Intell Med.* 2019 Sep; 100:101722. doi: 10.1016/j.artmed.2019.101722. Epub 2019 Sep 25. PMID: 31607343.

[5] Parca, Luca & Pepe, Gerardo & Pietrosanto, Marco & Galvan, Giulio & Galli, Leonardo & Palmeri, Antonio & Ferrè, Fabrizio & Ausiello, Gabriele & Helmer Citterich, Manuela. (2019). Modeling cancer drug response through drug-specific informative genes. *Scientific Reports.* 9. 10.1038/s41598-019-50720-0.

[6] 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

[7] [7] 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.

[8] 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.

[9] 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.

[10] 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.2018.00059.

[11] https://www.genome.jp/kegg/drug/

[12] https://www.cancer.gov/about-cancer/treatment/drugs/breast

[13] https://www.ebi.ac.uk/chembl/

[14] http://www.swissadme.ch/

[15] https://scikit-learn.org/stable/

[16] Priya, N. and Shobana, G. (2019). Application of Machine Learning Models in Drug Discovery: A Review. International Journal of Emerging Technologies, 10(3): 268–275.

[17] N. Priya and G. Shobana, "Multivariate Classification of Drugs using Parametric and Non parametric Machine Learning Models" (2020) "International Journal of Innovative Technology and Exploring Engineering", Volume-9(3).

[18] N. Priya and G. Shobana. "Leukemia Drug Prediction Using Machine Learning Techniques with Feature Engineering." 2020 Journal of Advanced Research in Dynamical and Control Systems – JARDCS, Volume 12, 04 Special Issue, Pages 141-146.

[19] N. Priya and G. Shobana, "Potential Breast Cancer Drug Prediction using Machine Learning Models," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India, 2020, pp. 1-6, Doi: 10.1109/ic-ETITE47903.2020.288.

[20] G. Shobana and DR. N. Priya, "Cancer Drug Classification using Artificial Neural Network with Feature Selection," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India, 2021, pp. 1250-1255, Doi: 10.1109/ICICV50876.2021.9388542.

[21]https://www.ibm.com/analytics/spss-statistics-software.

[22]G. Shobana and K. Umamaheswari, "Forecasting by Machine Learning Techniques and Econometrics: A Review," 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 1010-1016, Doi: 10.1109/ICICT50816.2021.9358514.

[23]D. Priya and G. Shobana, "Leukemia Drug Prediction Using Machine Learning Techniques with Feature Engineering", 2020 Journal of Advanced Research in Dynamical and Control Systems - JARDCS, vol. 12, no. 04, pp. 141-146.