JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 790 - 797
https://publishoa.com
ISSN: 1309-3452

An Artificial Intelligence Based Recommender System to analyze Drug Target Indication for Drug Repurposing using Linear Machine Learning Algorithm

[1]Deepak Srivastava, [2]Dr. Dheresh Soni , [3] Dr.Vibhor Sharma, [4] Dr. Pramod Kumar, [5]Dr. Anuj Kumar Singh

[1]Department of Computer & Information Sciences, Himalayan School of Science & Technology, Swami Rama Himalayan University, Dehradun, India,
[2] School of Computing Science and Engineering (SCSE), VIT Bhopal University, Bhopal, India
[3]Department of Computer Sciences and Engineering, Roorkee Institute of Technology, Roorkee, India
[4]Krishna Engineering College, Ghaziabad, U.P., India
[5]Krishna Engineering College, Ghaziabad, U.P., India

ABSTRACT
Drug Discovery and Development process refers to the process through which a new chemical compound is discovered, produced, and brought to market to treat a specific disease or medical condition. Today's datasets are often big in size and contain hundreds or thousands of characteristics so we need to extract meaningful information from them via automated content analysis. Machine Learning (ML) approaches encompass a diverse set of statistical algorithms for evaluating data, identifying shared patterns, deriving user models, and generating predictions. In this paper we have established biological interaction by analysing the drug molecule structure and protein structure for getting the relationship of Drug and Target for breast cancer that after we performed the optimization on prepared biological dataset that is Standard Gold Dataset (SGD). For optimization, we built machine learning model using linear machine learning algorithm such that Logistic Regression, Linear Discriminant Analysis (LDA) and Support Vector Machine and classify our Standard Gold Dataset (SGD). Logistic Regression is performing better above-mentioned linear machine Learning algorithms.

Keywords – Drug Discovery, Drug Target Interaction, Machine Learning, Feature selection, Classification

1. Introduction

The healthcare sector is an area in which the governments of all emerging and developed countries demonstrate their particular interest by dedicating special funds to this sector in order to give patients with smoother and less expensive treatments. However, many underdeveloped countries continue to struggle with the adaptation of computer support in their healthcare systems [1]. As chronic diseases such as cancer, nephrotic syndrome, and heart disease are rapidly spreading around the globe and resulting in a large number of fatalities each year, early identification and diagnosis of such diseases is a difficult undertaking in order to limit the number of deaths. Information technology can assist medical practitioners in making appropriate medical decisions based on data collected during the early stages of disease and also provides patients with a cost-effective method of therapy[1] [2].

Breast cancer is one of the diseases that claim a large number of lives each year around the world. Breast cancer is the second leading cause of mortality after lung cancer, accounting for around 30% of all cancer cases identified in women, accounting for 15% of all cancer-related fatalities [4 -5]. The properties of the cells found in the human body tend to alter and begin to behave improperly in this sort of cancer. As this disease is rapidly spreading, early identification of breast cancer is critical and vital. Two modelling methodologies are more frequently used to construct models for various chronic illnesses in order to ensure proper and early detection [3][6].

1.1 Machine Learning for Medical Data

Machine learning combines artificial intelligence and computational intelligence. Their role model is the human mind, on which they hope to build intelligent machines that can solve real-world problems. Probabilistic reasoning (including genetic algorithms), belief networks and learning theory are all part of it. [4] all of which provide the basis for the design, development, and deployment of intelligent discovery.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 790 - 797
https://publishoa.com
ISSN: 1309-3452

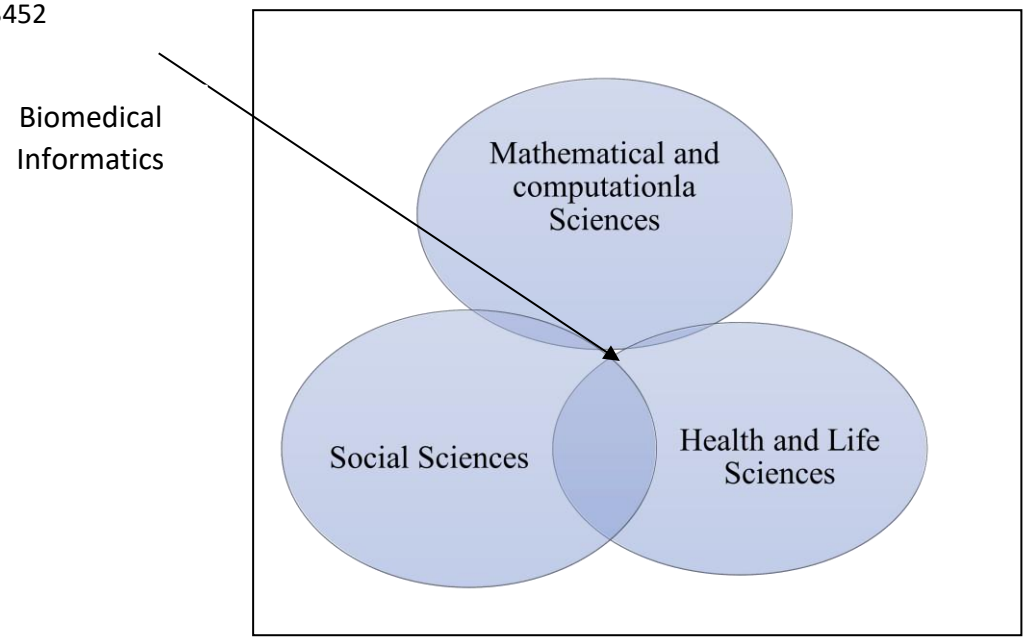

Figure 1: The relationship between mathematics and computational science, social life sciences, sciences and health.

1.2 Linear Machine Learning Algorithm

In this thesis, we use two types of machine learning algorithms that is linear machine learning algorithm and nonlinear machine learning algorithm. In this chapter, we discussed about linear algorithm and see the result that defines the type of cancer that is close to immuno-oncology therapeutic agents proteins or not. The reason behind choosing these classifiers is that these are the most influencing machine learning algorithms and ranked in the top 10 algorithms of machine learning [5][6]. Standard Gold Dataset(SGD) for Breast Cancer was used to build the prediction model using 10-fold cross-validation technique This dataset contains a total of 591 in which are categorized into two stages i.e., most closely linked cancer immuno-oncology therapeutic agents proteins and rest are other type of cancer with more than 8000 feature.
In this paper, we used logistic regression, linear discriminant models and SVM (Linear) models to optimize our Standard Gold Dataset (SGD).
The paper is divided into six section. In the first and second section, we described the theoretical context, explain why the work being done is significant, and the review of literature. In the third section I described the detailed methodology by explaining about the collected data and working mechanism. In fourth section, I described the output tables that hold the data related classification. Important information displayed in the form of tables and graph. In the next section, we concluded and summarized work.

2. Literature Review

Chemoinformatics is a vast field that combines computer science with chemistry to address a variety of chemistry-related challenges, including molecular graph mining, compound database searching, and chemical information retrieval and extraction (Varnek et al., 2011).
In view of the aforementioned, the present section discussed about the foundation base paper for this research below.
(Mohamed Hosni et al., 2019) examined a variety of ensemble approaches that are often used to perform prediction tasks in a variety of fields, including bioinformatics. The purpose of this study is to examine recent advances in ensemble classification techniques when applied to breast tumours in terms of nine characteristics, including publication domains, medical activities involved, experimental and research categories agreed upon, recommended ensembles, sole methodologies used to build the ensembles, validation structure used to examine the recommended ensembles, tools used to construct the ensembles, and optimization.
(Monteiro et al., 2018) used a variety of machine learning algorithms to cure schemic stroke patients. This study shown that when compared to the ASTRAL, DRAGON, and THRIVE tools, the machine learning technique produces only minor accuracy. However, it is observed that increasing the number of features in the prediction job considerably increases the effectiveness of the machine learning technique. This study focuses on the addition of additional characteristics throughout patient therapy. Actual patient data was combined with previously collected data to arrive at a diagnosis. As a result, the author focuses on the quality of health care application's services, as well.
(M. Daoud et al., 2016) recommend that programmed 3-D breast ultrasound be used as a balancing modality for mammography in order to aid in the early detection of breast cancer. To assist in the deciphering of such pictures, computer-aided identification methods are developed in which masses are segmented and handled as meaningful objects for feature extraction and temporal comparisons. Additionally, it is recognised how difficult automated mass segmentation is, given the enormous variation in volume, form, and quality of such 3-D objects.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 790 - 797
https://publishoa.com
ISSN: 1309-3452

(Kumar et al., 2015) suggested a system for classifying cancer using biologically interpretable morphology-based 115 features, including gray-level texture features, colour gray-level texture features, color-based features, Tamuras features, TEM features, and wavelet features. Prior to feature extraction, a contrast constrained adaptive histogram equalisation approach is used to improve the image's contrast and staining distribution. It is subsequently extracted from the image using the K-means clustering technique. The model's performance is evaluated using four classifiers: RF, Fuzzy K-NN, SVM, and K-NN. According to the experimental results, K-NN surpassed all other investigated classifiers in terms of accuracy, sensitivity, and specificity
(Qin et al. 2014), to employ an in silico technique for target identification, it is critical to catalogue the quantity, features, and biological diversity of authorised pharmacological targets.
In this chapter Literature survey of existing bioinformatics tools and supervised, unsupervised, neural network and hybrid classifiers, and usage and application of advanced Drug Repurposing techniques are discussed in detail.
In this section, we discussed literature survey of existing bioinformatics tools, machine learning algorithm classifier and their usage. In the next section we discussed the methodology in detail.

3. Methodology

3.1.1 Dataset Collection

I have prepared standard gold datasets with the help of publicly available medical datasets.[12][13][14] The usefulness of the proposed standard gold dataset is tested using data gathered using 200 descriptor features in the first experiment, 500 descriptor features in the second experiment, and 1000 descriptor features in the third experiment. The typical Standard Gold Dataset collects 591 values encompassing both negative and positive drug target sets. For the feature selection, I used PCA (Principle Component Analysis) with Correlation-matrix based feature selection is used in our study to analyses cancer data.[16][17] The primary benefit of feature selection is modelled by using these models, which describe the interdependencies between the various characteristics.

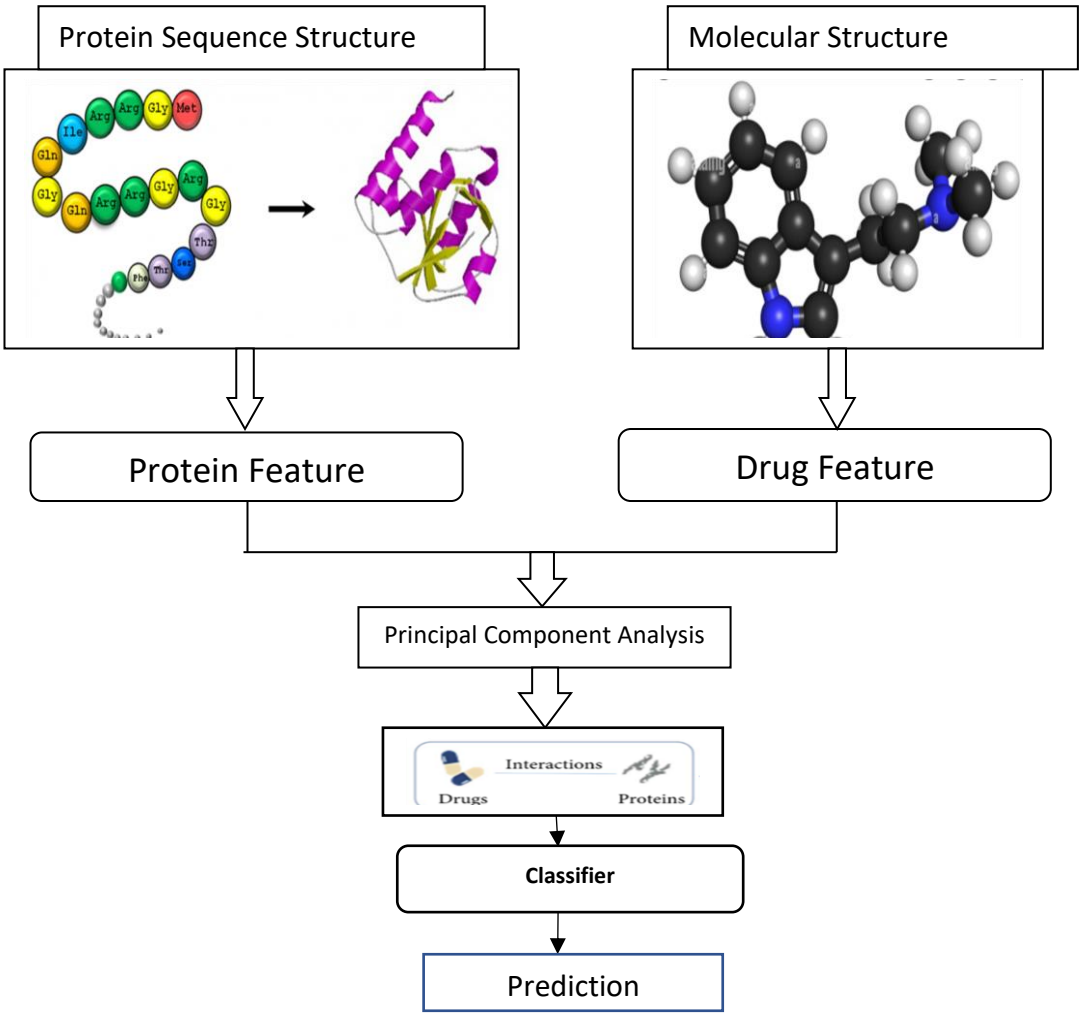

Figure 2: Drug repurposing for breast cancer using ML algorithm

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 790 - 797
https://publishoa.com
ISSN: 1309-3452

3.3.2 Working Mechanism

A drug is described as a chemical or a substance that can be used to diagnose, treat, cure, prevent, or relieve a disease. Drug repurposing is the process of discovering an agent/drug that can diagnose, cure, attenuate, prevent, or treat an ailment or disease. The Working mechanism of work is given below-

Target Similarity Based Algorithm

Input: Take Drug Molecule data (SMILES) and Protein Sequence (FASTA file) data for process

Output: Get output score and prediction for Interacting and non-interacting Drug-target indication.

Step-1: The software reads SMILLES and FASTA sequences and saves them in SequenceSet objects.

Step-2: Convert data into descriptor.

Step-3: Define matrix m_{ij} to calculate the pairwise score of descriptors. No score is computed less than 62% for a size by calculate BLOSUM.

Step-4: Use gradient boosting regression tree to get binding affinity score of Drug and Target.

Step-5: The beginning of the algorithm for the number of features from 1 to the total number of features .Check each feature one by one (until the end of the feature)

Step-6: Use PCA for feature class objects that encompass conversion mappings and quantification measures, making chaining simple.

Step 7: create Standard Gold Dataset (SGD) of three dataset such that PCA-Best 200, PCA-Best 500, and PCA-Best 1000.

Step-8: Build the model using above discussed machine learning.

Step -9: Perform classification and prediction.

4 Result

The usefulness of the proposed standard gold dataset is tested using data gathered using 200 descriptor features in the first experiment, 500 descriptor features in the second experiment, and 1000 descriptor features in the third experiment. The typical Standard Gold Dataset collects 591 values encompassing both negative and positive drug target sets. The properties of these datasets are explored above, and performance criteria like as accuracy, sensitivity, specificity, precision, F-1 score, and mean AUROC values are chosen for assessing the simulation outcomes. Initially, the confusion matrix is used to represent the simulation results of the aforementioned methodologies and the suggested Standard Gold Dataset (SGD). Confusion matrix is used to calculate performance characteristics.

Table 1: Confusion matrix of (a) Logistic Regression (b) LDA and (c) SVM technique using SGD

Confusion Matrix		200 Descriptor		500 Descriptor		1000 Descriptor	
		Predicted					
		P	N	P	N	P	N
Actual	P	286	45	311	38	331	34
	N	88	172	65	177	45	181

(a)

Confusion Matrix		200 Descriptor		500 Descriptor		1000 Descriptor	
		Predicted					
		P	N	P	N	P	N
Actual	P	265	55	273	47	297	38
	N	92	179	85	186	65	191

(b)

Confusion Matrix		200 Descriptor		500 Descriptor		1000 Descriptor	
		Predicted					
		P	N	P	N	P	N
Actual	P	284	48	305	38	327	38
	N	90	169	77	171	57	174

(c)

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 790 - 797
https://publishoa.com
ISSN: 1309-3452

The comparative analysis of simulation results of proposed drug target interaction prediction. The rule base can also enhance the accuracy of proposed system in significant manner. The specificity and sensitivity parameters also confirm the effectiveness of the proposed system. The proposed system measures based on the below performance parameters. Table 2 shows the performances of proposed Standard Gold dataset.

Table 4.2: Comparative analysis of proposed system for (a) Logistic Regression, (b) LDA and (c) SVM

Model	**Parameter (%)**	**200**	**500**	**1000**
Logistic Regression	Accuracy	77.4	82.5	86.6
	Sensitivity	76.4	82.7	88.0
	Specificity	79.2	82.3	84.1
	Precision	86.4	89.1	90.6
	F1 – Score	81.1	85.7	89.3
	AUROC	80.9	86.1	89.4
LDA	Accuracy	75.1	77.6	82.5
	Sensitivity	74.2	76.2	82.0
	Specificity	76.4	79.8	83.4
	Precision	82.8	85.3	88.6
	F1 – Score	78.2	80.5	85.2
	AUROC	78.4	80.3	85.1
SVM	Accuracy	76.6	80.5	84.7
	Sensitivity	75.9	79.8	85.1
	Specificity	77.8	81.8	82.0
	Precision	85.5	88.9	89.5
	F1 – Score	80.4	84.1	87.3
	AUROC	79.1	84.3	87.4

The Standard Gold Dataset (SGD) is considered for evaluating the efficiency of proposed drug target interaction prediction system for Drug Repurposing. The simulation results of proposed Drug Repurposing and Linear Machine Learning approach reported in above sections. Simulation results showed that proposed Drug Repuposing system achieves a prediction system that significantly improves the accuracy, sensitivity and specificity rates when we increasing number of features. Hence, it is stated that proposed Drug Repurposing and prediction system can determine the drug target interaction more efficiently.

Figure 3, 4 and 5 demonstrates the performance of proposed Standard Gold Dataset and all Linear Machine Learning techniques in graphical manner and it is observed that proposed system obtains good classification results. As see we got optimal result for 1000 descriptor features in all proposed system with different Machine Learning techniques. In below section we compared the performance of the entire linear machine learning algorithms that discussed in this paper above with 200, 500, and 1000 descriptor features.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 790 - 797
https://publishoa.com
ISSN: 1309-3452

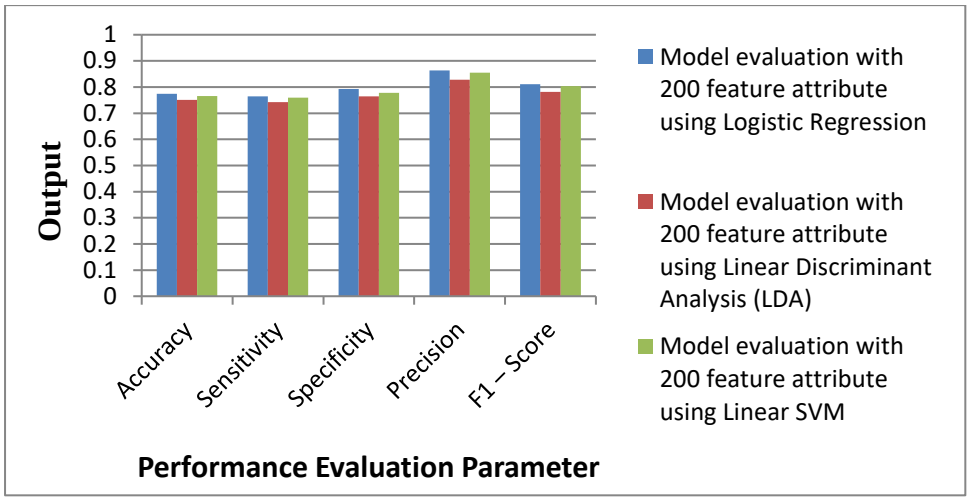

Figure 3: Comparative analysis of different linear Machine Learning Algorithm with 200 feature selection in cross validation

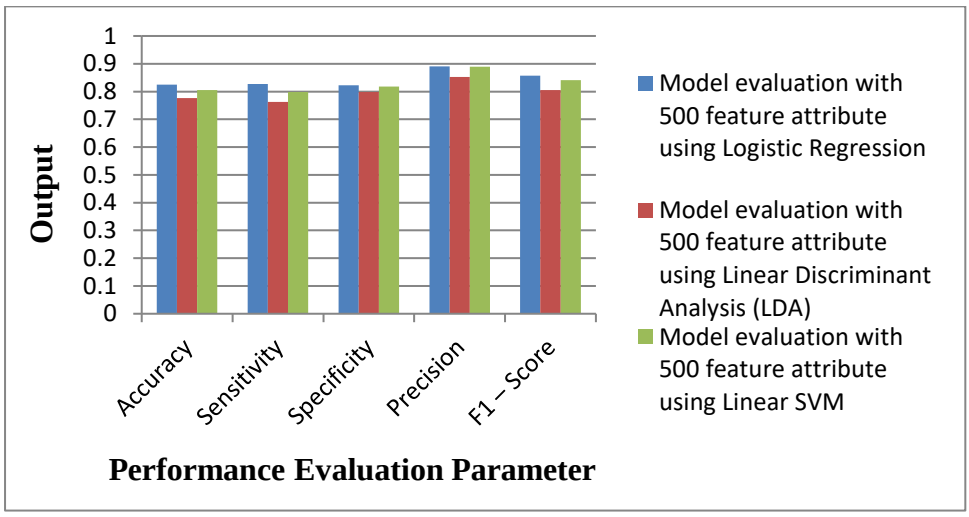

Figure 4: comparative analysis of different Linear Machine Learning Algorithm with 500 features

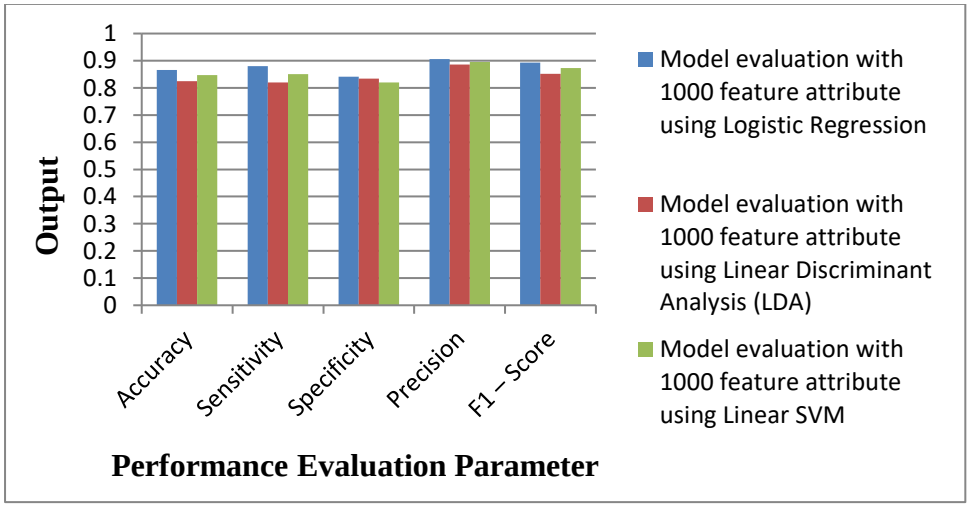

Figure 5: comparative analysis of different Linear Machine Learning Algorithm with 1000 features.
In figure 6 depicts the comparative analysis of different Linear Machine Learning Algorithm with 200, 500 and 1000 features selection below-

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 790 - 797
https://publishoa.com
ISSN: 1309-3452

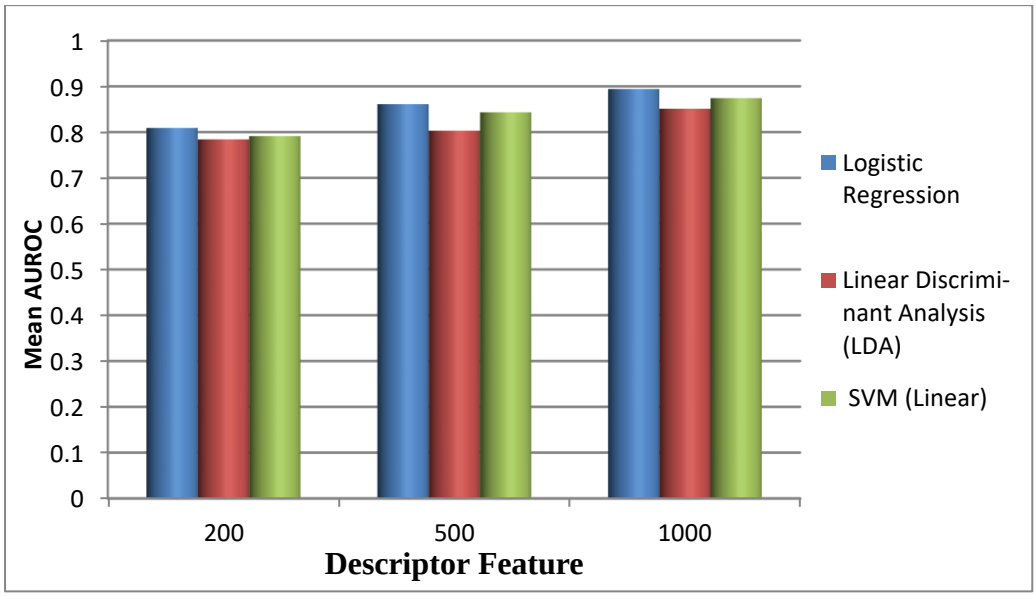

Figure 6: comparative analysis of different Linear Machine Learning Algorithm with 200, 500 and 1000 features

5. Conclusion and future Scope

An accurate breast cancer diagnosis can be made with the use of this chapter's prediction algorithm. Three datasets of 200, 500, and 1000 features are used in the suggested system. Performance measurements for the proposed Standard Gold Dataset are based on accuracy, sensitivity, specificity, precision, and F1-Score (SGD). This Standard Gold Dataset and prediction system shows good performance with all linear machine learning approaches but logistic Regression gives better result among all discussed above linear machine learning algorithms. Datasets are also evaluated in order to examine the effectiveness of the proposed approach and to obtain improved predictions of Drug Target interactions. It was found the best result of the classifier is 89.4% for a Logistic Regression classifier is pretty good among the linear classifier. In future we will see the use of nonlinear machine learning algorithms and see the performance of nonlinear machine learning algorithms of Standard Gold Dataset (SGD).

REFERENCES

1. A. Chan and J. A. Tuszynski, "Automatic prediction of tumour malignancy in breast cancer with fractal dimension," Royal Society Open Science, vol. 3, pp. 1-10, 2016.
2. Abdar, M., Książek, W., Acharya, U. R., Tan, R. S., Makarenkov, V., & Pławiak, P. (2019). A new machine learning technique for an accurate diagnosis of coronary artery disease. Computer methods and programs in biomedicine, 179, 104992.
3. Ali, L., Rahman, A., Khan, A., Zhou, M., Javeed, A., & Khan, J. A. (2019). An Automated Diagnostic System for Heart Disease Prediction Based on Statistical Model and Optimally Configured Deep Neural Network. IEEE Access, 7, 34938-34945.
4. Anitha K and Ventatesan P "Feature Selection by Rough Quick Reduction Algorithm", International Journal of Innovative Research in science, Engineering and Technology, Volume -2, Issue-8, August 2013.
5. Araújo, et al., "Classification of breast cancer histology images using convolutional neural networks," PloS One, vol. 12, pp. 1-28, 2017.
6. Arslan, A. K., Colak, C., & Sarihan, M. E. (2016). Different medical data mining approaches based prediction of ischemic stroke. Computer methods and programs in biomedicine, 130, 87-92.
7. Varnek A, Baskin II. Chemoinformatics as a Theoretical Chemistry Discipline. Molecular Informatics. 2011; 30:20-32.
8. Hosni M., Abnane, I., Idri, A., de Gea, J. M. C., & Alemán, J. L. F. (2019). Reviewing ensemble classification methods in breast cancer. Computer methods and programs in biomedicine, 177, 89-112.
9. Monteiro, M., Fonseca, A. C., Freitas, A. T., e Melo, T. P., Francisco, A. P., Ferro, J. M., & Oliveira, A. L. (2018). Using machine learning to improve the prediction of functional outcome in ischemic stroke patients. IEEE/ACM transactions on computational biology and bioinformatics, 15(6), 1953-1959.
10. Daoud, M. I., Alshalalfah, A. L., & Al-Najar, M. (2016, December). GPU accelerated implementation of kernel regression for freehand 3D ultrasound volume reconstruction. In 2016 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES) (pp. 586-589). IEEE.

11. Kumar, et al., "Detection and classification of cancer from microscopic biopsy images using clinically significant and biologically interpretable features," Journal of Medical Engineering, vol. 2015, pp. 1-15, 2015.
12. Qin, C., Zhang, C., Zhu, F., Xu, F., Chen, S.Y., Zhang, P., Li, Y.H., Yang, S.Y., Wei, Y.Q., Tao, L., Chen, Y.Z., (2014). 'Therapeutic target database update 2014: a resource for targeted therapeutics'. Nucleic acids research. Vol. 42, pp D1118-1123.
13. S. B Kotsiantis, Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. Artificial Intelligence Review, 26(3), 159-190
14. S. Cascianelli, et al., "Dimensionality reduction strategies for cnn-based classification of histopathological images," in International Conference on Intelligent Interactive Multimedia Systems and Services, 2018, pp. 21-30.
15. S. K. Ahammad Fahad & Yahya, A. E. (2018, July). Big Data Visualization: Allotting by R and Python with GUI Tools. In 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE) (pp. 1-8). IEEE.
16. S. Kothari, et al., "Histological image classification using biologically interpretable shape-based features," BMC Medical Imaging, vol. 13, p. 1-17, 2013.
17. Saha, S., Heber, S., (2006). 'In silico prediction of yeast deletion phenotypes'. Genetics and molecular research. Vol. 5, pp 224-232.