A Machine Learning-Based Method to Predict of Drug–Target Association Based On Multiple Feature Information Selection and Reduction Techniques

[1]Deepak Srivastava, [2] Dr.Pramod Kumar,[3] Dr. Sunil Ghildiyal

[1]Assistant Professor, Department of Computer & Information Sciences, Himalayan School of Science & Technology, Swami Rama Himalayan University, Dehradun, India,

[2]Professor, Krishna Engineering College, Ghaziabad, U. P., India

[3]Associate Professor, UIT, Uttaranchal University, Dehradun, India

ABSTRACT

Identifying drug–target (protein) interactions is critical for research and development of innovative drugs, providing a significant benefit to pharmaceutical businesses and patients. However, predicting Drug Target Indications by clinical trial procedures is typically costly and time consuming. As a result, many machine learning-based algorithms have been created for this goal, yet significant unknown interactions remain. Additionally, feature selection and reduction concerns are a key barrier in drug-target datasets, since they might affect classifier performance if not handled well. This study offered a unique approach for predicting drug–target interactions. To begin, the amino acid composition (AAC), dipeptide composition (DC), and tripeptide composition (TC); and drug SMILES substructure fingerings are used to extract the protein sequence's feature vectors. PCA (Principal Component Analysis) is used to eliminate superfluous and redundant characteristics in order to get the most optimum features. Finally, balanced and optimum features are supplied to SVM with RBF kernel function in order to detect Drug Target indications, and the proposed approach's prediction capacity is evaluated using the 10-fold CV validation test method. The prediction findings suggest that the proposed model outperforms other current approaches in predicting Drug Target interaction.

KEYWORDS: Drug Repurposing, Support Vector Machine (Kernel), Principal Component Analysis, Feature selection and reduction, Classification

I. INTRODUCTION

Predicting novel drug–target association is a critical stage in the pipelines of drug discovery and design [1–3]. Drug repurposing is a rising area in pharmaceutical science, with an emphasis on uncovering previously undiscovered interactions between current drugs and novel target proteins. The advancement of the entire genes and the expansion of the molecular biology project provide valuable information for predicting novel therapeutic targets. Numerous attempts have been made in recent years to identify new treatments, but relatively few have been approved by the Food and Drug Administration (FDA) and reached patients, while a large number of pharmaceuticals have been rejected in clinical trials due to unacceptable toxicity. DTI wet-lab investigations are often time-consuming, labor-intensive, and expensive; as a result, such failures are difficult to accept and result in significant financial loss. As a result, researchers are particularly motivated to develop machine learning (ML)-based algorithms for detecting Drug Target Association [3], which may successfully narrow the search space of drug–target possibilities to be evaluated in wet-lab trials, therefore reducing work and expense. Recently, machine learning-based computational approaches have grown increasingly advantageous due to the vast amount of heterogeneous pharmacological and protein data.

chemogenomic approaches often make advantage of the genomic and chemical information associated with target proteins and medicines. As a result, chemogenomics techniques are becoming increasingly common for identifying Drug Target associations. The chemogenomic model's prediction challenge might be addressed utilising powerful machine learning methods [12].

Numerous machine learning classifiers, including deep learning, SVM, fuzzy logic, and closest neighbour, have been successfully used to various sorts of prediction tasks. Whereas feature-based algorithms use input vectors of drug chemistry and protein sequence features and display the class label as a binary value (1 or 0).

1.1 Feature Selection and Reduction

Feature Selection in Computer Aided Diagnosis (CAD) is a difficult module to learn when it comes to classifying Drug Target indications. This is mainly due to the increased number of features to be analyzed with high desirable accuracy. When the feature sets are huge, or the input dataset is voluminous, the classification becomes a highly time-consuming task. Feature selection generally deals with selecting the most appropriate useful features and minimizes the redundancy in improving the performance of classification subsystem. Thus, the optimal feature selection will effectively increase the accuracy, reduce the time complexity, and improve the performance of any CAD system. [11].

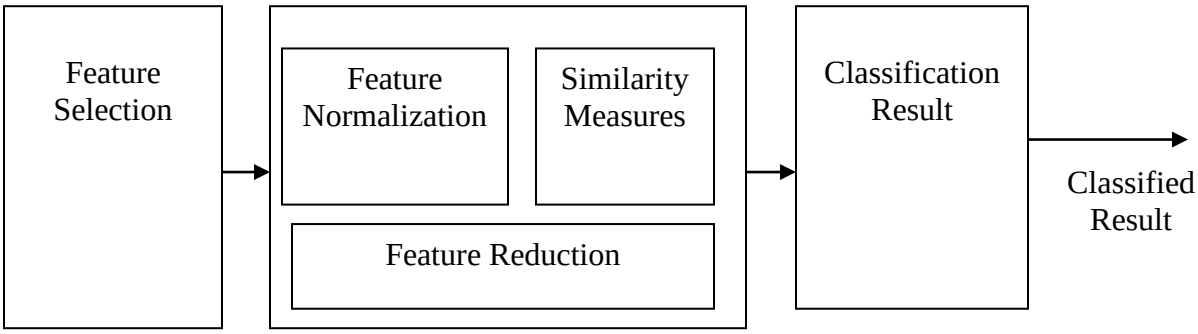

Figure 1: Depict Feature Selection and reduction

In this paper, a combined feature selection algorithm that uses multi Zero Mean and Unit Variance Normalization with Correlation Distance Method is presented and analyzed. The single-objective feature selection algorithms provided only a single bag of optimal solution. This approach circumvented the limitations of typical single goal algorithms by generating a collection of optimum solutions that trade off distinct objectives. The multi objective approach ensured that the minimal features with high impact on classification were selected and it achieved improved accuracy with lesser.

The multi-objective feature selection algorithm's system architecture is divided into three phases: feature selection, feature reduction, and classification. The selection phase's input was produced by features collected from drugs and targets. Feature representation and neighborhood formation were carried out during the selection phase. [20] During the feature reduction step, redundant features and those with a negligible effect on categorization were filtered. The classification step was used to classify the massive amount of input data into distinct categories. Cross validation, data trimming, nearest-neighbor computation, and normal distribution model were all performed throughout the classification phase. To classify input data into training and testing sets, the cross validation approach was applied. To compute the class probability, data reduction was performed. To compare the real input features to the training set of features, the nearest class algorithm was utilized. Finally, the categorized results were determined using a normal distribution model.

II. BACKGROUND/ LITERATURE SURVEY

The PCA-based technique outlined in [11] was used to perform a survey on drug repurposing, which has gotten a lot of interest in recent years. It is capable of providing effective solution for applications with certain limits after several years of study.

Orawan et al.[17] built the system by collecting Fuzzy Co-occurrence Matrix and fractal dimension characteristics, and then used PCA to reduce the system's dimensionality. Multi-class SVM is used to

classify cancer patients with a 91.7 percent correct classification rate, 93 percent sensitivity, and 91 percent specificity.

Zhu et al. [14] used principle component analysis (PCA) and the K-means clustering technique to construct an improved logistic regression model for diabetes prediction. PCA is used in this proposal to translate diabetic data to a lower dimension. Integration of PCA enhanced the accuracy of K-means clustering and logistic regression, as demonstrated by simulation results.

III. DATA COLLECTION AND METHODOLOGY

We created a prediction model that employs Support Vector Machine with RBF kernel nonlinear classification techniques to determine the potential of particular drug-target associations. However, prior research indicates that the dataset is classifiable using linear models. Feature selection and reduction is done into two parts that is feature normalization and feature similarity. External validation and tenfold cross-validation were performed to determine the accuracy of each prediction model. Below steps are involved to determine Drug Target association. Methodology is explained in below steps: [4]

Step 1: Data Collection

I acquired data from Drug Bank in order to utilize it for Drug Repurposing. A Drug Bank that has a huge number of drugs and target information. It is composed of a diverse set of licenced small molecule drugs, biotech pharmaceuticals, and experimental drugs that are linked to non-redundant protein sequences.

Step 2: Data Pre-processing:

Pre-processing is used to eliminate undesired noise and increase contrast between regions of varying brightness. To eliminate undesirable items, pre-processing is used to separate the arithmetic data from the non-numerical data.

Step 3: Compute Descriptors

To begin, the amino acid composition (AAC), dipeptide composition (DC), and tripeptide composition (TC) of the protein sequence are retrieved, as well as the drug SMILES substructure fingerings. The word "descriptors" refers to the terms used to describe the chemical, topological, and geometrical properties of drugs and targets. We gathered 591 medication descriptions by binding Drugs and targets.

Step 4: Check Similarity

In this work, immuno-oncology proteins (UniProt ids - Q6UWE3, P42677, P63173, and Q9Y243) were compared to known protein structures from a drug repository to uncover new biomarkers. Generally, the immune system kills cancer cells. We compared the immune-oncology compound's similarity scores.[15]

Step 5: Feature selection and representation

The retrieved collection of characteristics served as the initialization phase's input. The characteristics were represented as a two-dimensional array, each of which was seeded with a random particle. The grid was constructed using the neighbours of the first random particle that generated the leader. A random array with a size equal to the number of features was formed and randomly filled with values ranging from 0 to 1 using a uniform random function defined by the position of each particle. If the associated index I in the array was greater than the threshold value of 0.45, a feature was picked. [17]

Step 6: Feature Reduction and Extraction

We focus our efforts on the two phases of feature normalization and similarity in order to produce dependable and flexible recognition of features. Accuracy, sensitivity, specificity, precision and F1-Score were utilized as performance indicators for the evaluation of three distinct feature descriptor databases used for PCA-based recognition. Zero Mean and Unit Variance Normalizing feature normalization and similarity Correlation Distance approaches are used in this study to evaluate feature descriptor performance based on principal component analysis (PCA).

- **Zero Mean and Unit Variance Normalization**

Zero Mean and Unit Variance Normalization [9] [10] normalize all of the elements ai (i=1, 2,...,d) of a, it translate and scale the axes so that all the feature vector have zero mean and unit variance. Following expression will produce the normalized feature vector a’.

$$a'_i = \frac{a_i - \mu}{\gamma} \quad (1)$$

Where, μ and γ are the vector mean and the vector standard deviation of that feature respectively.

- **Correlation Distance methods**

Similarity measure used to match the similar subjects (persons) as well as being able to discriminate dissimilar one. Let x, y be the feature vectors of length n. then we can calculate the following distances between these feature vectors

$$\partial(a, b) = 1 - \sum_{i=1}^{n} \frac{(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \bar{a})^2 \sum_{i=1}^{n}(b_i - \bar{b})^2}} \quad (2)$$

Where,

$$\bar{a} = \sum_{i=1}^{n} a_i \quad \text{And } \bar{b} = \sum_{i=1}^{n} b_i \quad (3)$$

Apply PCA based algorithm on the dataset to select the best feature set PCA-200, PCA-500 and PCA1000.

Step 7: Classification Phase:

Performs training and testing only on the best features set selected by PCA based Algorithm and create a model using support vector machine with RBF kernel classifier. [19]

Step 8: Performance

Evaluate the performance of this model based on some parameters like accuracy, sensitivity, specificity, precision, F1-score and analyse the prediction using mean AUROC.

IV. RESULT AND DISCUSSION

This section contains result and discussion about prediction of Drug target association for breast cancer. For implementing the proposed technique, we have used python. The proposed system has been tested on the data sets. These three dataset repository uses 591 associations for the purpose of classification. Confusion matrix for the model is represented in Table 1.

Table 1: Confusion matrix for Model

Confusion Matrix		200 Descriptor		500 Descriptor		1000 Descriptor	
		Predicted					
		P	N	P	N	P	N
Actual	P	443	17	447	15	453	11
	N	19	112	13	116	9	118

Simulation results comparing expected interactions between a drug and its target. The suggested system's accuracy may be considerably enhanced by employing the rule base. These metrics also validate the specificity and sensitivity of the proposed system. The suggested system is assessed using the following metrics. Table 2 displays the result for model.

Table 2: Table 5.2 Comparative analysis of proposed system using nonlinear SVM with RBF function

Parameter (%)	**200**	**500**	**1000**
Accuracy	93.9	95.3	96.6
Sensitivity	95.8	97.2	98.1
Specificity	86.8	88.5	91.5
Precision	96.3	96.7	97.6
F1 – Score	96.0	96.9	97.8
AUROC	96.9	97.9	98.3

Evaluation findings for classification-based prediction models are depicted in Table 2. In this part, we represented the results of our calculations.

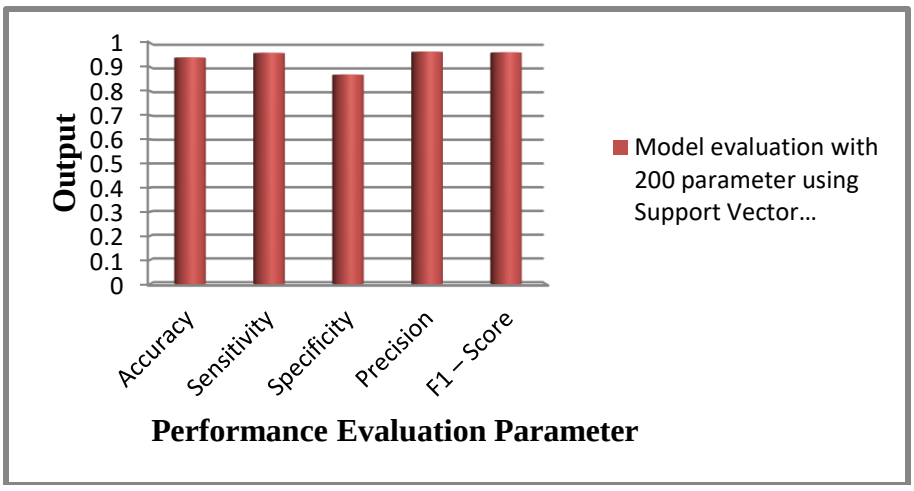

Figure 2: Comparative representation for proposed system with 200 feature attribute

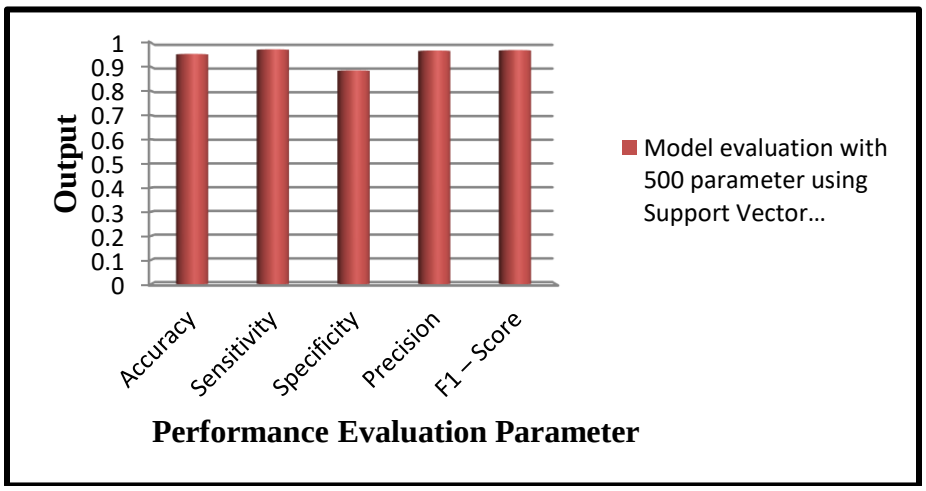

Figure 3: Comparative representation for proposed system with 500 feature attribute

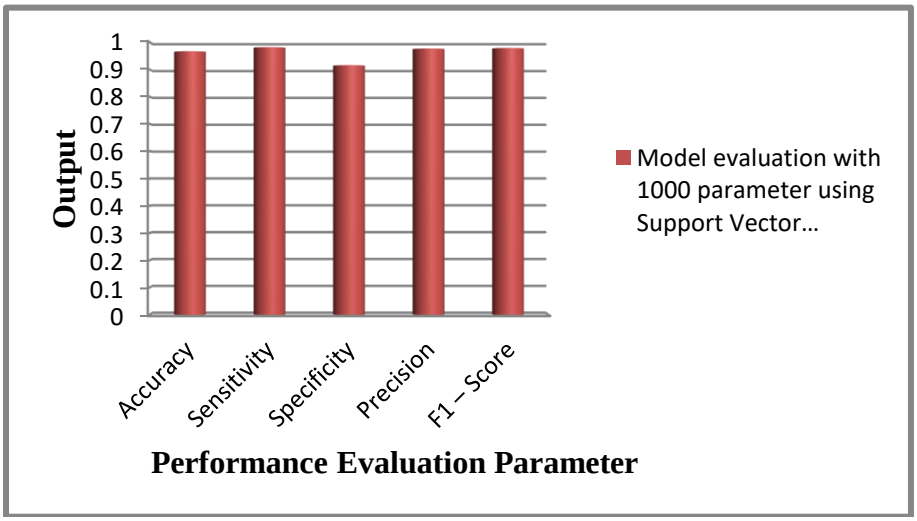

Figure 4: Comparative representation for proposed system with 1000 feature attribute

AUROC measures the model's capacity to distinguish between "cases" (positive instances) and "non-cases" in terms of performance (negative examples.) Assuming that 90 percent of the time, a model accurately assigns a higher absolute risk to a randomly picked patient with an incident than to another randomly selected patient without an event, this indicates that the model has strong discriminating capacity. Comparative analysis of proposed system with different number of feature selection using Support Vector Machine using RBF kernel function below:

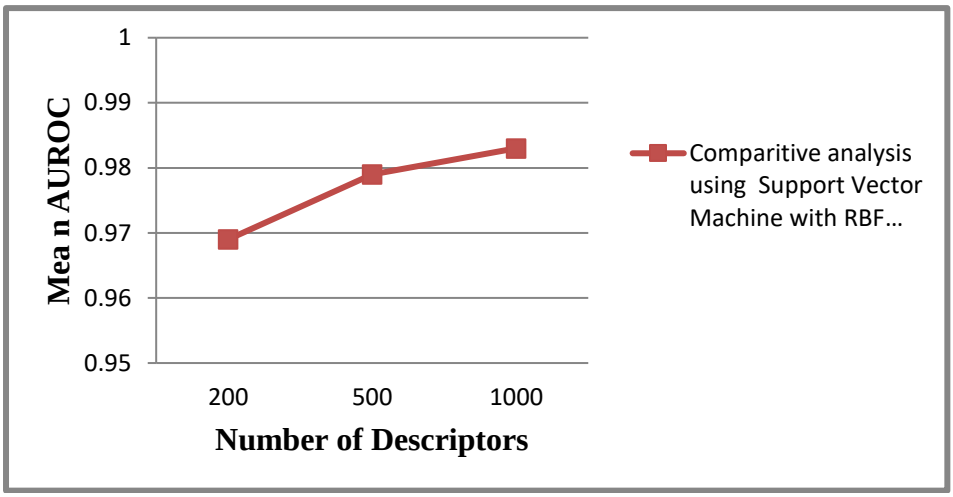

Figure 5: Average Mean of AUROC values for different number of feature descriptors using SVM with RBF kernel function

V. SUMMARY AND CONCLUSION

In this paper, a combined feature selection algorithm was presented to select the best subset of features from bag of features. Best features were chosen using feature selection with feature similarity. The model was used to forecast connections between immuno-oncology agents and disease. The data set consisted of 591 Drug target associations. Zero Mean and Unit Variance Normalization and Correlation Distance methods extracted the features from the dataset. The performance analysis of the present feature selection model was compared with Support Vector Machine with RBF kernel function. The present model performed better in terms of performance parameters against all the aforementioned algorithms with 200, 500 and 1000 features. For the dataset considered, the present algorithm was effective due to selection of lesser number of features in sequential processing of data. We performed cross validation on each model to determine its performance The model's accuracy and mean AUROC were greater than 95%, while increasing the number of descriptor features. The findings indicated that immuno-oncology compounds may be useful as therapeutic candidates for a variety of disorders cancer treatment. The proposed prediction models can aid in drug development by identifying the potential for immuno-oncology compounds to be repurposed for cancer treatment.

VI. REFERENCES

1. Ali Ezzat, Peilin Zhao, Min Wu, Xiao-Li Li, and Chee-KeongKwoh, "Drug-Target Interaction Prediction with Graph Regularized Matrix Factorization", in IEEE/ACM Transactions on Computational Biology and Bioinformatics (2015).
2. Antonio Lavecchia, “Machine-learning approaches in drug discovery: methods and applications,” in Elsevier, Drug Discovery Today Volume 00, Number 00 November 2014.

3. Ashis Kumer Biswas, Nasimul Noman and Abdur Rahman Sikder, “Research article Machine learning approach to predict protein phosphorylation sites by incorporating evolutionary information,” Eleventh BMC Bioinformatics 2010, 11:273.
4. Brian Delavan , Ruth Roberts, Ruili Huang, WenjunBao, Weida Tong and Zhichao Liu, "Computational drug repositioning for rare diseases in the era of precision medicine," in Elsevier, Drug Discovery Today, Volume 00, Number 00, October 2017.
5. Christopher C. Yang, Mengnan Zhao, " Mining heterogeneous network for drug repositioning using phenotypic information extracted from social media and pharmaceutical databases ",in Elsevier, Artificial Intelligence In Medicine 96 (2019) 80–92 .
6. Fabian Pedregosa , Gael Varoquaux , AlexandreGramfort, Vincent Michel and Bertrand Thirion, "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research 12 (2011) 2825-2830.
7. HuiminLuo, Jianxin Wang, Min Li, Junwei Luo, Peng Ni, Kaijie Zhao, Fang-Xiang Wu, and Yi Pan, "Computational drug repositioning with random walk on a heterogeneous network," IEEE/ACM Transactions On Computational Biologyand Bioinformatics, 2017.
8. Jianlin Cheng, Allison N. Tegge and Pierre Baldi, “Machine Learning Methods for Protein Structure Prediction”, IEEE Reviews In Biomedical Engineering, Vol. 1, 2008.
9. Jiaying You, Robert D. McLeod, Pingzhao Hu, "Predicting drug-target interaction network using deep learning model", in Elsevier, Computational Biology and Chemistry 80 (2019) 90–101.
10. Liang Yu, Ruidan Su, Bingbo Wang, Long Zhang, YapengZou, Jing Zhang, Lin Gao, "Prediction of novel drugs for hepatocellular carcinoma based on multi-source random walk”, IEEE/ACM Transactions on Computational Biology and Bioinformatics,2015.
11. Zhu, J., Xie, Q., & Zheng, K. (2015), “ An improved early detection method of type-2 diabetes mellitus using multiple classifier system”. Information Sciences, 292, 1-14.
12. Lu Zhang, Jianjun Tan, Dan Han and Hao Zhu, "From machine learning to deep learning: progress in machine intelligence for rational drug discovery," in Elsevier, Drug Discovery Today, Volume 22, Number 11, November 2017.
13. Majid Rastegar-Mojarad, Ravi kuma,r Komandur Elayavilli, Liwei Wang, Liwei Wang, Rashmi Prasad, Hongfang Liu, "Prioritizing Adverse Drug Reaction and Drug Repositioning Candidates Generated by Literature- Based Discovery", Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics October 2016 Pages 289–296.
14. Fangjun Kuang, Weihong Xu Siyang Zhang (2014). “A novel hybrid KPCA and SVM with GA model for intrusion detection”. Applied Soft Computing, Volume 18, May 2014, Pages 178-184.
15. Ahmadi, N., Nilashi, M., Samad, S., Rashid, T. A., & Ahmadi, H. (2019). An intelligent method for iris recognition using supervised machine learning techniques. *Optics & Laser Technology*, *120*, 105701.
16. Pemovska T, Johnson E, Kontro M, et al. Axitinib effectively inhibits BCR-ABL1 (T315I) with a distinct bre conformation. Nature. 2015;519:102–105.
17. Orawan, C., Panwadee, S., & Bandit, S. (2016). “Application of artificial neural networks on growth prediction of Staphylococcus aureus in milk”. International Food Research Journal, 23(1), 415.
18. R. Burbidge, M. Trotter, B. Buxton, S. Holden, “Drug design by machine learning: support vector machines for pharmaceutical data analysis”, in Elsevier, Computers and Chemistry 26 (2001) 5–14.
19. Thein, H. & Tun, K. M. M. (2015). “An approach for breast cancer diagnosis classification using neural network”. *Advanced Computing*, *6*(1), 1.