International Journal on Emerging Technologies **10**(3): 268-275(2019)

ISSN No. (Print) : 0975-8364
ISSN No. (Online) : 2249-3255

Application of Machine Learning Models in Drug Discovery: A Review

N. Priya[1] and G. Shobana[2]
[1]Associate Professor, Department of Computer Science,
SDNB Vaishnav College for Women (Affiliated to University of Madras), Chennai, India.
[2]Research Scholar, Department of Computer Science,
SDNB Vaishnav College for Women (Affiliated to University of Madras), Chennai, India.

(Corresponding author: N. Priya)
(Received 31 May 2019, Revised 17 August 2019 Accepted 28 August 2019)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: Drug discovery involves identification of a target protein causing the disease and find the drug inhibitor molecule that restricts the growth of the target protein. Among several inhibitor molecules available, identification of the most appropriate one is crucial. Machine learning models can be applied to make accurate predictions when abundant data is available. In this paper, we explore various machine learning techniques that are applied to the bioinformatics and cheminformatics data to achieve accurate prediction for identifying active inhibitors of diseases in the process of drug discovery. We also investigate different model evaluation metrices. Various prediction analysis show that Support Vector Machine (SVM) and Random Forest (RF) produces best result.

Keywords: Machine learning, Drug discovery, Bioinformatics, Cheminformatics.

I. INTRODUCTION

Machine learning algorithms can be applied to various applications of bioinformatics like sequence homology analysis, drug design, predictive functions, genomics, proteomics and genome mapping. Cheminformatics extracts data from the chemical structures. In recent years, biological databases have increased profoundly. The lead features are collected and applied to the machine learning models for better inhibitor predictions. The need for accurate and efficient learning models for prediction has also increased considerably. Depending upon nature of the data set, the machine learning algorithms are applied to it**.** Bioinformatics-oriented approach provides an important advantage where, various biological problems such as sequence analysis, gene expression data analysis and genetic analysis, system biology and biomedical applications of the target protein are examined. In biomedical applications, biomedical texts and medical images can be manipulated for relevant data using machine learning techniques [1].

Some of the machine learning models used for prediction in bioinformatics are as follows: Decision Trees, Random Forest, Support Vector Machine (SVM), Linear Models (GLM), Neural Network(NN), M5P, Decision Stump, cubist, fobaetc [2]. The analysis of compound diversity, prediction of compound activity, molecular datamining and several numerical features are extracted to form Cheminformatic data. These are called Chemical descriptors. Chemical descriptors may vary from one dimensional (1D) to four dimensional (4D). Chemical fingerprints are vectors with high dimension. These are generally used in analysis of chemometric and virtual screening applications based on similarity. Chemical descriptors values are the elements obtained from these processes. Chemical similarity search is a fundamental technique for ligand-based drug discovery. Its objective is to identify and return data-based compounds with structures and bioactivities similar to query compounds [3].

Some of the supervised machine learning methods are: Multiple regression analysis, K nearest neighbor, Naïve Bayes, Random forest, Neural network and deep learning, Support vector machine [3]. Some machine-learning algorithms used in cheminformatics are: Ant Colony, Relevance Vector Machine(RVM), Parzen-Rosenblatt Window, Fuzzy Logic, Rough Sets, Support Vector Inductive Logic Programming(SVILP),Winnow, Decision Tree, Linear Discriminant Analysis(LDA), k-Score, Projection to Latent Structures(PLS) etc. [4].

II. MACHINE LEARNING IN DRUG DISCOVERY PROCESS

Molecular docking methods explore the ligand conformations adopted within the binding sites of macromolecular targets [5]. In computational docking, a large number of binding poses are evaluated and ranked using a scoring function. The scoring function is a mathematical predictive model that produces a score that represents the binding free energy and hence the stability of the resulting complex molecule. The key to computer-aided drug design is hence the design of an efficient, accurate and highly scoring function using machine learning technique [6]. Maciej Wojcikowski *et al* investigated the structure -based Virtual Screening that aims at identifying compounds with previously unknown affinity for a target from its three-dimensional (3D) structure. They used three machine learning scoring functions for building models [7].

Bioinformatics addresses genes, proteins and other larger chemical compounds, whereas cheminformatics has mainly dealt with small molecules (Fig. 1). Cheminformatics and bioinformatics complement each other for biomolecular processes, like structure and function of proteins, the binding of a ligand to its binding site, the conversion of a substrate within its enzyme receptor and the catalysis of a biochemical reaction by an enzyme [8].

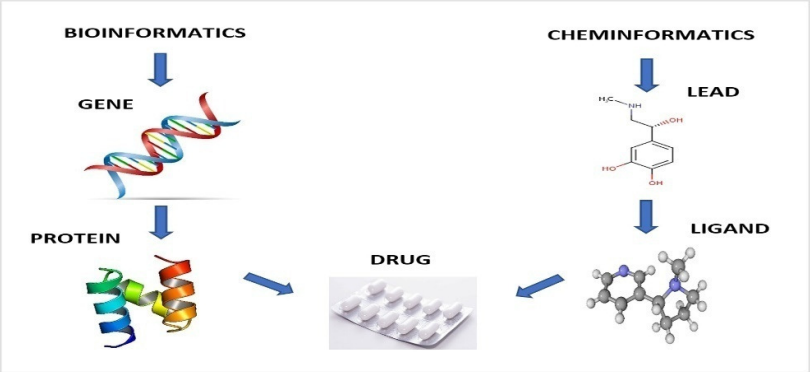


Fig. 1. The inter-related fields of bioinformatics and cheminformatics (Firdaus Begam *et al* 2012).

III. THE COMMON APPROACH

Redundant data might occur during data collection. To reduce unwanted 'noise' and redundant data, various pre-processing techniques are employed. The data is feature engineered and high ranked features are obtained. Feature engineering is the act of extracting features from raw data and transforming them in to formats that are suitable for the machine learning model [9]. Selected and effective features are provided as inputs to machine learning models for the most accurate prediction analysis as shown in Fig. 2.

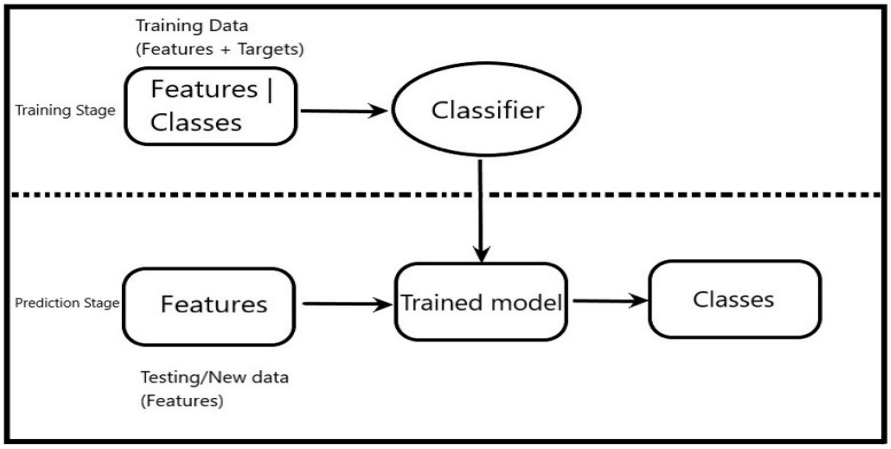


Fig. 2. The common approach for prediction analysis using machine learning (Yuxi, 2017).

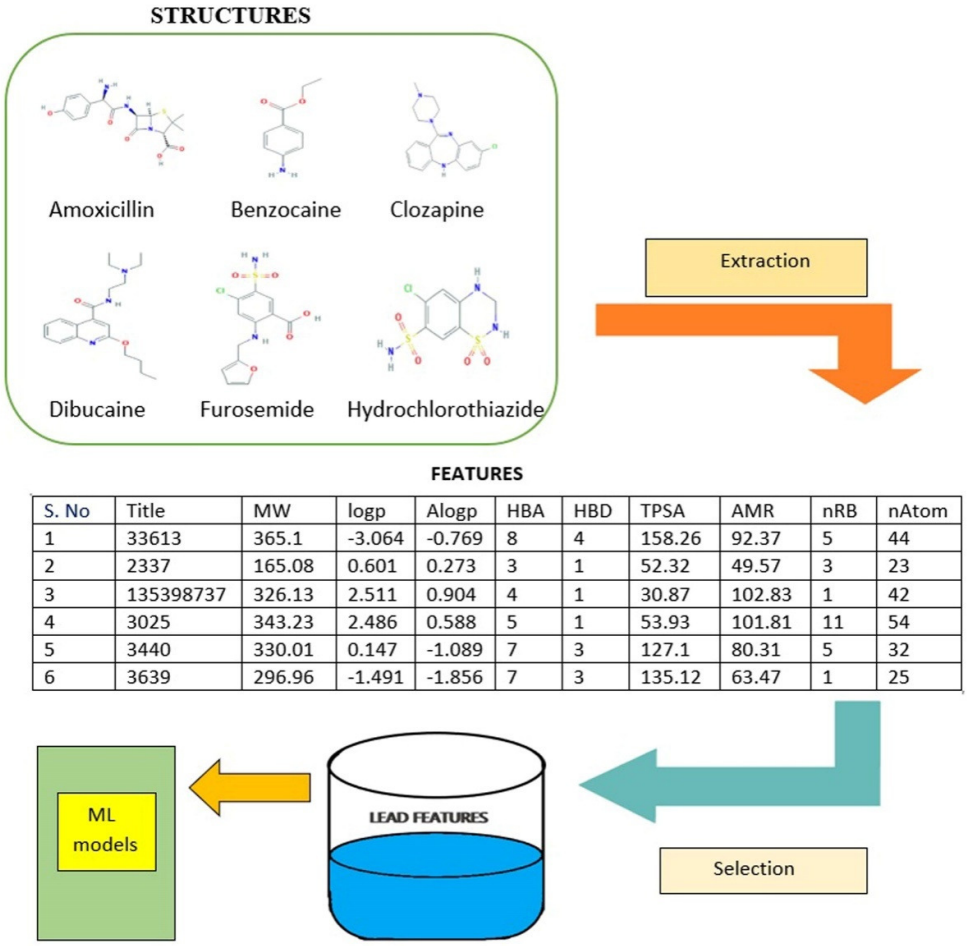
IV. FEATURE EXTRACTION

Feature selection techniques prune away non useful features. This greatly reduces the complexity of the resulting model. In general, feature selection falls in to three classes namely: filtering, wrapper methods and embedded methods [9]. Generally, the drug discovery process starts with a particular disease, identification of the target protein, identification of the drug molecule, which acts as an inhibitor. The pharmacophoric features are extracted and the lead features are selected. Using the machine learning models, best inhibitor is predicted as shown in Fig. 3.

Molecular features can be extracted from drugs using various tools available online. Swiss ADME is a free web-based tool used to evaluate physicochemical properties of drugs. DrugLiTo (Drug Likeness Tool) is simple and user-friendly application for determining pharmacokinetics. The ADMET (Absorption, Distribution, Metabolism, Excretion and Toxicity) properties can be evaluated using various free online tools like Pre ADMET (ADMET Prediction), Molinspiration, Pre ADMET(Toxicity Prediction) etc. Some of the commercial tools are MDL, ChemTree, Volsurf, GRID, Tsar 3.2, MetaSite, Shop, TOPKAT, Metabolism, ADMET, Metabase, ADME/Toxicity property calculator etc. Lipinski's rule of five or Pfizer's rule of five determines whether a drug can be orally taken by humans. Molecular weight (MW), Molecular refractivity (MR), Polar Surface Area (PSA), Topological Polar Surface Area (TPSA), logp (Lipophilicity), water solubility etc., are some of the features that can be extracted for the drugs using these tools.

1. Harish Bhaskar *et al* focussed primarily on general aspects of feature and model parameter selection. They also investigated issues affecting the application of machine learning tools [10].
2. Inza *et al* have explored the characteristics of main data pre-processing, Feature selection and classifier evaluation that have a deep impact current bioinformatics. Machine learning technique has become an essential tool in any biomarker discovery process [11].
3. Venkatesan *et al* have stated that mutual information and Chi-square test are the two most frequently used feature selection methods [12].

4. Nandhini *et al* have discussed various feature selection techniques like Greedy and Heuristic methods for the classification of heart disease [13].



S. No	Title	MW	logp	Alogp	HBA	HBD	TPSA	AMR	nRB	nAtom
1	33613	365.1	-3.064	-0.769	8	4	158.26	92.37	5	44
2	2337	165.08	0.601	0.273	3	1	52.32	49.57	3	23
3	135398737	326.13	2.511	0.904	4	1	30.87	102.83	1	42
4	3025	343.23	2.486	0.588	5	1	53.93	101.81	11	54
5	3440	330.01	0.147	-1.089	7	3	127.1	80.31	5	32
6	3639	296.96	-1.491	-1.856	7	3	135.12	63.47	1	25

Fig. 3. Feature extraction process (John, 2014).

V. APPLICATION OF MACHINE LEARNING MODELS

A. Naïve Bayes

Naïve Bayes performs well for a relatively small dataset, if its features are independent. It is a very simple algorithm and training of Naive Bayes is usually faster than any other algorithms due to its computational simplicity. However, this may sometimes lead to high bias condition. Coi *et al* reported a new strategy to predict DPP-IV inhibitors using machine learning techniques like Naïve Bayesian (NB) and Recursive Partitioning (RP). With 1307 known DPP-IV inhibitors, they used optimized molecular properties and topological fingerprints as descriptors. The accuracy achieved by these optimized models were greater than 80%[14]. Leena Sarvaiya *et al* explored various machine learning algorithms like Naïve Bayes, Decision Tree, K-nearest neighbour, SVM and Neural Networks used in diagnosing heart diseases [15].

B. Decision Trees (DT)

A decision tree, assigns a class label to each leaf node. The root and other internal nodes, that are non-terminal nodes contain attribute test conditions to separate The classification of HIV-1 protease Inhibitors was done by Li *et al* using machine learning methods like k-nearest neighbour (k-NN), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM) and Deep Neural Network (DNN). 98.37% was the best prediction accuracy, obtained by the model 3C, that was built by RF. The Random Forest was based on CORINA Symphony descriptor [16].

records that have different characteristics. Training sample data is divided in to successive subsets and the dividing process is further repeated in a recursive manner. Fig. 4 and Fig. 5 are examples of a simple decision tree and its graph representation.

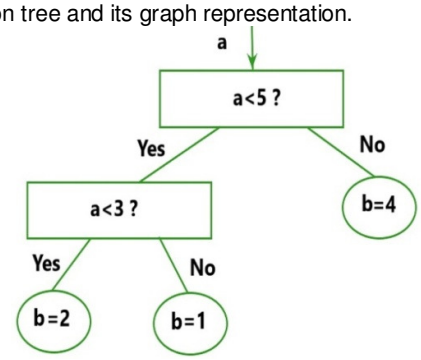


Fig. 4. Example of a simple decision tree (Michael Bowles 2015).

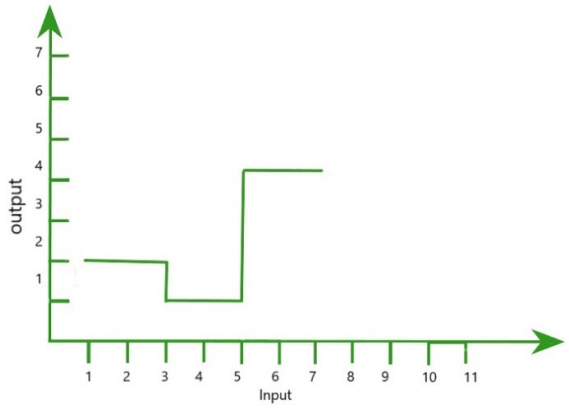


Fig. 5. Graphical representation of the tree(Michael Bowles 2015).

C. Random Forest (RF)

Random forest is considered as one of the most efficient ensemble techniques. It combines the predictions made by multiple decision trees, where each tree is generated based on the values of an independent set of random vectors. Multiple classification trees can be constructed from an input vector using the random forest method [18].

Matteo Lo Monte et al developed a computation tool for the prediction of ADP - ribosylated sites. The tool named AD Predict was developed using machine learning techniques and principal component analysis. To interpret the dataset they used the random forest (RF) method. To derive predictive models, they also applied the Support Vector Machine Model (SVM) [19]. Freya klepsch et al proposed Ligand and Structure Based Classification models for prediction of P-Glycoprotein inhibitors. They used machine learning models like k-NN, RF and SVM for the prediction of P-gp inhibitors and noninhibitors. Random Forest and SVM achieved the best results of 75% accuracy [20].

D. Support Vector Machines (SVM)

SVMs are most popular machine learning methods used in Bioinformatics and cheminformatics. The data is mapped in to high-dimensional space in SVMs. SVM is a statistical method that identifies a hyperplane that is low in dimension and using a non-linear kernel, it maximises the separation of data. This separating process is achieved by the margin maintenance between hyperplanes and is called support vectors. SVM is versatile to adapt to the linear separability of data. Fig. 6 and 7. Shows the linearly separable and non-separable classes by margins. Very high accuracy can be achieved by SVM with the right kernel parameters. It is used in bioactivity prediction that includes drug repurposing, kinase inhibition, estrogenic receptor agonists and opioid activity. The SVM is often used to predict toxicity-related properties such as HERG blockade, mutagenic toxicity, toxicity classification and phospholipidosis. Applications in physicochemical property prediction include solubility, pka, logp and melting point. Support Vectors decide the best margins possible, both in the case of linearly separable and non-separable classes.

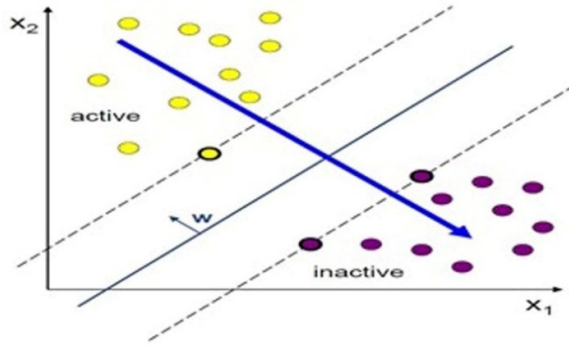


Fig. 6. Shows the active and inactive elements separated by margins (Yuxi, 2017).

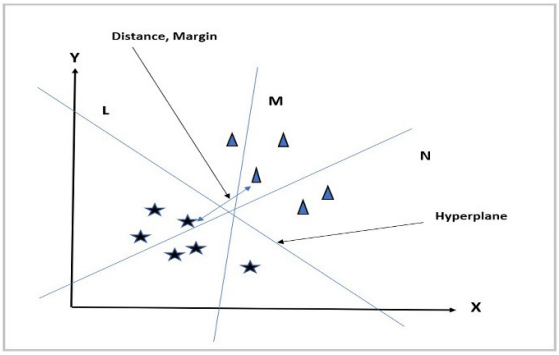


Fig. 7. Shows the linearly separable and non-separable classes by margins (Yuxi, 2017).

Kernel functions are convenient mapping functions that allow SVMs to obtain a transformed dataset of limited size which is equivalent to a more complicated and data-intensive non-linear transformation. Kernels can be implemented in both R and Python. SVM offers a large range of non-linear kernels. The most common and fastest kernel is the Radial Basis Function (RBF). It can almost map any nonlinear function, if its shape parameter, gamma is provided as shown in Table 1.

Table 1: Types of kernels and their parameters with mapping type.

Type of kernel	Parameter	Mapping type
Linear	No extra parameters	
Radial Basis Function	Shape parameters	Gamma
Polynomial	Shape parameters	Gamma, degree and CoefΘ
Sigmoid	Shape parameters	Gamma and CoefΘ
Custom – made kernels	Depends upon the kernel	

Liu *et al* investigated the derivatives of ethyl 2-[(3-methyl-2,5-dioxo(3-pyrrolinyl)) amino]-4-(trifluoromethyl) pyrimidine-5-carboxylate. SVM was used in the QSAR study of transcription factors activator protein (AP)-1and nuclear factor (NF)-kB [21]. Kinnings *et al* used SVM to predict the inhibitors that involve directly in M. tuberculosis (M. tb) Inh A. Molecular docking was performed for retrieving several associated energy terms. With adequately known binding affinity data of

individual compound, machine learning model was applied [22].

The inhibition activity of vascular endothelial growth factor (VEGF)-2 was explored by Nekoei, M. *et al*. SVM was used in combination with genetic variable selection approach and various structural features of aminopyrimidine 5-Carbaldehyde oxime derivatives were identified [23]. Vasanthanathan *et al* classified cytochrome P450 1A2 inhibitor and non-inhibitors by applying various machine learning techniques. Binary quantitative structure activity relationship, Support Vector Machine (SVM), Random Forest (RF) and Decision Tree (DT) methods were used. Among these, the best predictions were obtained using SVM, RF and K-NN in combination with the best first variable selection method [18].

Machine learning methods like Support Vector Machine (SVM), and C4.5 decision tree (C4.5 DT) and k-nearest neighbour(k-NN) were explored by Lv *&* Xue for predicting the inhibitors of Acetylcholinesterase (AchE). Molecular descriptors were used for improving the accuracy of prediction. The prediction accuracies were 76.3%~88.0% for AchEIs and 74.3%~79.6% for non-AchEIs [24]. Asma Aziz khan and Vipin Verma proposed an ensemble approach using SVM, KNN and GA models for the diagnosis of diabetes [25]. Dimitri SK. Lakovidis *et al* have developed a novel system to detect gastrointestinal adenomas. They combined the color-texture analysis methodologies and intelligent processing techniques in to a framework of sound pattern recognition [26]. Maurício Boff de Ávila etal proposed a computational model to predict inhibition of DHQD (3-dehydroquinate dehydratase). They built new machine learning models and used polynomial scoring function [27]. Jain, S. and Sood, M. have explored different kernel functions like linear, Gaussian, RBF etc to train the SVM and found that the results obtained with linear kernels are faster [28]. Ranilakshmi *et al* discussed various machine learning models like Support Vector Machine (SVM), GA (Genetic Algorithm), C4.5 decision tree and k-nearest neighbour(k-NN), Naïve Bayes etc to investigate the risk factor involved in heart disease [29].

VI. PERFORMANCE EVALUATION

Hassan *et.al* presented a cheminformatics model that was generated using LDA algorithm. The LDA algorithm had better accuracy when compared to random forest model shown in Fig. 8, that was proposed by Wahi *et al* in 2015 to predict the inhibitors of USP1/UAF1 activity of unknown compound. Cross-validation experiment was applied on the dataset, accuracy rate and area under curve [30]. Li *et al* obtained 4855 HIV-1 Protease inhibitors from ChEMBL. Machine Learning models like k-nearest neighbors (k-NN), decision tree(DT), random forest (RF), Support Vector Machine (SVM) and deep neural network (DNN) were applied on the data set for predicting the active inhibitors. The molecular structures were characterized by fingerprint descriptors and physicochemical descriptors.

The fingerprint descriptors included MACCS fingerprints and PubChem fingerprints. The physicochemical descriptors were characterized by CORINA Symphony. Models 1A, 2A, 3A, 4A and 5A were analysed in Fig. 9. Models 1B, 2B, 3B, 4B and 5B were analysed in Fig. 10. Models 1C, 2C, 3C, 4C and 5C were analysed in Fig. 11. The best accuracy was achieved by the model 4A with 83.07% [16].

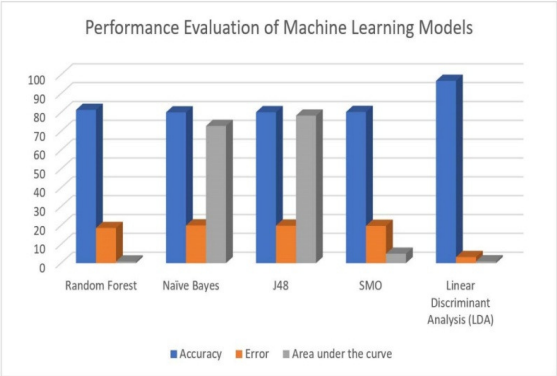


Fig. 8. Bar chart indicating the performance evaluation of Machine Learning Models (Hassan *et.al,* 2017).

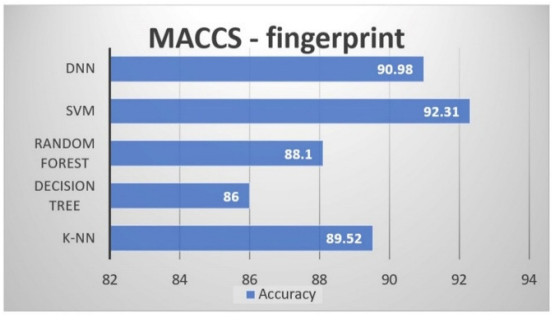


Fig. 9. Models 1A, 2A, 3A, 4A and 5A with MACCS – fingerprint (Li *et al* 2018)**.**

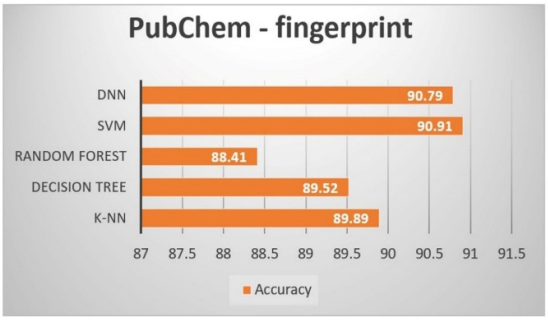


Fig. 10. Models 1B, 2B, 3B, 4B and 5B with PubChem – fingerprint (Li *et al.,* 2018)**.**

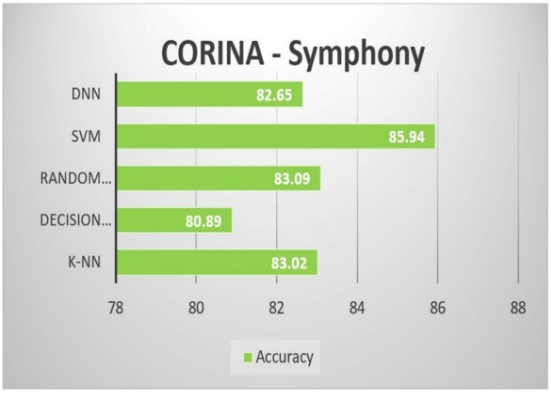


Fig. 11. Models 1C, 2C, 3C, 4C and 5C with CORINA – Symphony (Li *et al* 2018)**.**

VII. EVALUATION METRICS FOR MACHINE LEARNING MODELS

Machine Learning models can be evaluated using various metrics. Rana *et al,* evaluated the machine learning models on Correlation, R^2, RMSE and accuracy.

Correlation(r):

The statistical relationship between the predicted and actual values can be defined using Correlation as follows:

$$\text{Correlation(r)}=\frac{\sum_{k=1}^{m}(x_k-\bar{x})(y_k-\bar{y})}{\sqrt{\sum_{k=1}^{m}(x_k-\bar{x})^2(y_k-\bar{y})^2}}$$

Here, x is the actual value and y is the predicted value. Mean of all the actual values is $\bar{x}$ and the mean of all the predicted values is $\bar{y}$. m is the number of instances. The correlation value lies between 0 and 1. When the correlation value moves towards 1, it is considered to be good result.

Co-efficient of Determination(R^2)

The explanatory power of the regression model is defined by the Co-efficient of determination(R^2).

R^2 is computed as follows

$R^2 = r * r$

The proportion of variance of the dependent variable rendered by the regression model is defined by R^2. When the value R^2 is 1 the regression model is perfect. When the value of R^2 is 0, the regression model is zero.

Root Mean Squared Error (RMSE)

The error rate of a regression model is measures using RMSE.

$$\text{RMSE}=\sqrt{\frac{\sum_{k=1}^{m}(p_k-a_k)^2}{m}}$$

Here, p is the predicted target value. The actual target value is *a* and *m*, the total number of instances.

$$\text{Accuracy}=\frac{100}{m}\sum_{k=1}^{m}q_k$$

$$q_k=\begin{cases}1 & if\ abs(p_k-a_k)\le err\\ 0 & otherwise\end{cases}$$

Here, the predicted target is p. The actual target value is a. err is the accuracy error. The total number of instances is *m* [2].

When bio-molecules are used as target for the drug discovery process, various metrics are used to compare sets of molecules, observe their diversity and similarity using various statistical properties. Daniil Polykovskiy et al have utilized five metrics that can be used to compare a generated set G and the reference set of molecules R. The five metrics are Fragment similarity, Scaffold similarity, Nearest neighbor similarity, Internal diversity and $Fréchet$ Chem Net Distance as shown in Table. 2. They also presented a set of auxiliary metrics useful for the drug design process and that could be extended for other applications as well [31].

Table 2. Metrics for evaluating Generative models.

Name	Formula	Description
Fragment similarity (Frag)	$Frag(G,R)=1-\cos(f_G,f_R)$	Similarity of two sets of molecules at the level of chemical fragments
Scaffold similarity (Scaff)	$Scaff(G,R)=1-\cos(s_G,s_R)$	Similarity between scaffolds in generated and reference model
Nearest neighbour similarity (SNN)	$SNN(G,R)=\frac{1}{\lvert G\rvert}\sum_{m_G\in G}\max_{m_R\in R}T(m_G,m_R)$	Analysis of the chemotypes and the chemical space
Internal diversity ($IntDiv_p$)	$ntDiv_p(G)=1-\sqrt[p]{\frac{1}{\lvert G\rvert^2}\sum_{m_1,m_2\in G}T(m_1,m_2)^p}$	Detects common failure case of generative models
Frechet Chemnet Distance (FCD)	$FCD(G,R)=\lVert\mu_G-\mu_R\rVert^2+Tr(\sum_G+\sum_R-2(\sum_G\sum_R)^{1/2})$	Predicts biological activities of drugs

The Internal diversity and External diversity between molecules were evaluated using Tanimoto-similarity was presented by Mostapha Benhenda. Mostapha quantified the internal chemical diversity and considered Reinforcement Learning model and Objective-Reinforced Generative Adversarial Network (ORGAN) [32]. Alghamedy *et al* used Youden's index, AUC, Accuracy, F1, Precision and Recall as metrics for comparing binding predictions from the docking score. Three different data models were developed and the maximum Youden's index (or J value) is calculated for each model [33].

Miao & Niu have presented two common metrics to evaluate the performance of clustering as given in Table 3 [34].

Vukovic *et al* have explored about drug-target ligandability. A prerequisite for druggability is ligandability and therefore complex pharmacodynamics and pharmacokinetic properties of the ligand has to be investigated. The metrics for target ligandability is given as

$$LIG_{exp}=\frac{pK_i>7}{N}$$

The metric is formulated on the concept of effort and reward. To generate a high-affinity inhibitor, if less effort is required then a target is highly ligandable. The total number of K_i values in Binding DB (N) for the effort metric. The number of reported compounds in Binding DB with a $pK_i>7$ was used. The best possible discrimination between targets is provided when the threshold is 7.0 and that maximises the variance in LIG_{exp} [35].

Table 3. Metrics to evaluate the performance of clustering.

Name	Formula	Description
Clustering Accuracy (Acc)	$Acc = \frac{\sum_{i=l}^{n} \delta(map(l_i), y_i)}{n}$	Compare the label obtained from clustering with true label
Normalized Mutual Information (NMI)	$NMI = \frac{\sum_{i=l}^{c}\sum_{j=l}^{c} n_{ij} log\frac{n_{ij}}{n_i \hat{n}_j}}{\sqrt{(\sum_{i=l}^{c} n_i log\frac{n_i}{n})(\sum_{j=l}^{c} \hat{n}_j log\frac{\hat{n}_j}{n})}}$	Evaluate the quality of clusters

VIII. CONCLUSION

In traditional drug discovery process, several newly synthesized drugs are tested for wide range of biological reactions in "WET Labs". Each drug has to tested against many different target proteins. Some of the Cervical cancer cell lines are 5J6R,5Y9E,4J96,3J6R,2R5K etc. There may be numerous other cell lines to be investigated for the same drug. This procedure is time-consuming and involves huge cost. Predicting inhibitors and noninhibitors for any type of target protein using machine learning techniques has given best results. Random Forest (RF) and Support Vector Machine were found to give more accurate results. Using Machine learning techniques, the most potential drugs can be identified. In the process of drug discovery, feature extraction and pre-processing of raw data plays an important role in accurate prediction. Apart from data extracted from Protein-Ligand affinity, other factors like protein's biological homology, ligand's physicochemical properties like ADMET (Absorption, Distribution, Metabolism, Excretion and Toxicity) can also be drawn and the collected data can be feature engineered. With efficient pre-processing and effective machine learning models, the most relevant drug discovery is possible.

IX. FUTURE SCOPE

Random Forest and Support Vector machine learning models have proved to give best results. Whenever the data availability is comparatively enormous, deep learning, ensemble and hybrid machine learning techniques can be applied for more accurate predictions.

Conflict of Interest. The authors declare no conflict of interest.

REFERENCES

[1]. Qian Xu & Qiang Yang. (2011). A Survey of Transfer and Multitask Learning in Bioinformatics. *Journal of Computing Science and Engineering*. *5*(3), 257-268.

[2]. Prashant Singh Rana, Harish Sharma, Mahua Bhattacharya & Shukla, A. (2015). Quality assessment of modelled protein structure using physicochemical properties. *Journal of bioinformatics and computational biology*,*13*(02).

[3]. Lo, Y.C., Rensi, S.E., Torng, W., & Altman, R.B. (2018). Machine learning in chemoinformatics and drug discovery. *Drug discovery today*, *23*(8), 1538-1546.

[4]. John B.O. Mitchell. (2014). Machine learning methods in Chemoinformatics. *WIREs Computational Molecular Science.4* .468-481.

[5]. Leonardo G. Ferreira, Dos Santos R.N., Oliva G. & Andricopulo A.D. (2015). Molecular Docking and Structure-Based Drug Design Strategies. *Molecules. 20(7)*. 13384-13421.

[6]. Khamis, Mohamed, Gomaa, Walid & Ahmed, Walaa. (2015). Machine learning in computational docking. *Artificial Intelligence in Medicine*. 81. 10.1016/j.artmed. 2015.02.002.

[7]. Wójcikowski, Maciej, Ballester, Pedro & Siedlecki, Pawel. (2017). Performance of machine-learning scoring functions in structure-based virtual screening. Scientific Reports. 7. 46710. 10.1038/srep46710.

[8]. B. Firdaus Begam & J. Satheesh Kumar (2012). A Study on Cheminformatics and its Applications on Modern Drug Discovery. *Procedia Engineering. 8*, 1264 – 1275.

[9]. Alice Zheng & Amanda Casari. (2018). Feature Engineering for Machine Learning.First Edition O Reilly,SPD.

[10]. Harish Bhaskar, David C. Hoyle & Sameer Singh (2006). Machine learning in bioinformatics: A brief survey and recommendations for practitioners, *Computers in Biology and Medicine*.*36*(10) ,1104-1125.

[11]. Inza, I., Calvo, B., Armañanzas, R., Bengoetxea, E., Larrañaga, P., & Lozano, J. A. (2010). Machine learning: an indispensable tool in bioinformatics. In *Bioinformatics methods in clinical research* (pp. 25-48). Humana Press.

[12]. Venkatesan, D., Vithiya Ruba, K. & Sekar, K.R. (2019). Investigation of Various Techniques and Classification Methods on Cognitive Sentimental Learning. *International Journal on Emerging Technologies, 10*(2): 15-18.

[13]. C. Usha Nandhini & P.R. Tamilselvi (2018). A Review on Feature Selection Approaches for Heart Disease Classification. *International Journal of Theoretical & Applied Sciences, Special Issue 10* (1a): 63-67

[14]. Cai, J., Li, C., Liu, Z., Du, J., Ye, J., Gu, Q., & Xu, J. (2017). Predicting DPP-IV inhibitors with machine learning approaches. *Journal of computer-aided molecular design*, *31*(4), 393-402.

[15]. Sarvaiya, L., Yadav, H. & Agrawal C. (2019). A Literature review of Diagnosis of Heart Disease using Data Mining Techniques. *International Journal of Electrical, Electronics and Computer Engineering, 8*(1): 40-45.

[16]. Li, Y., Tian, Y., Qin, Z., & Yan, A. (2018). Classification of HIV-1 Protease Inhibitors by Machine Learning Methods. *ACS omega*, *3*(11), 15837-15849.

[17]. Michael Bowles. (2015). *Machine Learning in Python: Essential Techniques for Predictive Analysis*. United States of America. WILEY.

[18]. Vasanthanathan, P., Taboureau, O., Oostenbrink, C., Vermeulen, N. P., Olsen, L., & Jørgensen, F. S. (2009). Classification of cytochrome P450 1A2 inhibitors and noninhibitors by machine learning techniques. *Drug Metabolism and Disposition*, *37*(3), 658-664.

[19]. Matteo Lo Monte, Candida Manelfi, Marica Gemei, Daniela Corda & Andrea Rosario Beccari (2018). AD Predict: ADP – ribosylation site prediction based on physicochemical and structural descriptors. *Bioinformatics, 34*(15), 2566-2574.

[20]. Freya Klepsch, Poongavanam Vasanthanathan, & Gerhard F. Ecker. (2014). Ligand and Structure-Based

Classification Models for Prediction of P-Glycoprotein Inhibitors. *J. chem. Inf. Model*, *54*(1), 218-229.
[21]. H. X. Liu, R. S. Zhang, X. J. Yao, M. C. Liu, Z. D. Hu & B. T. Fan. (2003). QSAR study of ethyl 2- [(3-methyl-2, 5-dioxo (3-pyrrolinyl)) amino]-4-(trifluoromethyl) pyrimidine-5-carboxylate:an inhibitor of AP-1 andNF-kB mediated gene expression based on support vector machines.*Journal of Chemical Information and Computer Sciences. 43*(4), 1288-1296.
[22]. Kinnings, S. L., Liu, N., Tonge, P. J., Jackson, R. M., Xie, L., & Bourne, P. E. (2011). A machine learning-based method to improve docking scoring functions and its application to drug repurposing. *Journal of chemical information and modeling*, *51*(2), 408-419.
[23]. Nekoei, Mehdi & Mohammad Hosseini, Majid & Pourbasheer, Eslam. (2015). QSAR study of VEGFR-2 inhibitors by genetic algorithm-multiple linear regressions (GA-MLR) and genetic algorithm support vector machine (GA-SVM): a comparative approach. *Medicinal Chemistry Research, 24*. 3037-3046.
[24]. Lv, W., & Xue, Y. (2010). Prediction of acetylcholinesterase inhibitors and characterization of correlative molecular descriptors by machine learning methods. *European Journal of medicinal chemistry*, *45*(3), 1167-1172.
[25]. Khan, Asma Aziz & Verma, V. (2017). Prediction of Diabetes Disease Using Entropy and Gain based Data Mining Approach. *International Journal of Electrical, Electronics and Computer Engineering 6*(1): 164-172.
[26]. Iakovidis, D. K., Maroulis, D. E., & Karkanis, S. A. (2006). An intelligent system for automatic detection of gastrointestinal adenomas in video endoscopy. *Computers in Biology and Medicine*, *36*(10), 1084-1103.
[27]. Maurício Boff de Ávila, Walter Filgueira de Azevedo Jr. (2018). Development of machine learning models to predict inhibition of 3-dehyroquinate dehydratase. *Chemical Biology and Drug Design. 92.1468–1474.*
[28]. Jain, S. and Sood, M. (2019). SVM Classification of Cell Survival/Apoptotic Death for Color Texture Images of Survival Receptor Proteins. *International Journal on Emerging Technologies*, *10*(2): 23-28.
[29]. S. Ranilakshmi & R. Mallika (2018). Survey on Heart Disease: Characteristics, Symptom and Prediction Method. *International Journal of Theoretical & Applied Sciences, Special Issue 10*(1a): 08-12.
[30]. Hassan, S.A., & Osman, A. H. (2017). An Improved Machine Learning Approach to Enhance the Predictive Accuracy for Screening Potential Active USP1/UAF1 Inhibitors. *International Journal of Advanced Computer Science and Applications*, *8*(4), 144-148.
[31]. Polykovskiy, D., Zhebrak, A., Sanchez-Lengeling, B., Golovanov, S., Tatanov, O., Belyaev, S., Kurbanov, R., Artamonov, A., Aladinskiy, V., Veselov, M., Kadurin, A., Nikolenko, S.I., Aspuru-Guzik, A., & Zhavoronkov, A. (2018). Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *ArXiv, abs/1811.12823*.
[32]. Mostapha Benhenda (2017). ChemGAN challenge for drug discovery: can AI reproduce natural chemical diversity? arXiv:1708.08227v3[stat.ML].
[33]. Alghamedy, F., Bopaiah, J., Jones, D., Zhang, X., Weiss, H.L., & Ellingson, S.R. (2018). Incorporating protein dynamics through ensemble docking in machine learning models to predict drug binding. *AMIA Summits on Translational Science Proceedings*, *2018*, 26.
[34]. Miao, J., & Niu, L. (2016). A survey on feature selection. *Procedia Computer Science*, *91*, 919-926.
[35]. Vukovic, S., & Huggins, D.J. (2018). Quantitative metrics for drug–target ligandability. *Drug discovery today*, *23*(6), 1258-1266.
[36]. Yuxi Hayden Liu. (2017). *Python Machine Learning By Example*. Birmingham UK. Packt Publishing Ltd.

How to cite this article: 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.