Research Article

International Journal of Advanced Computer Research, Vol 10(47)
ISSN (Print): 2249-7277   ISSN (Online): 2277-7970
http://dx.doi.org/10.19101/IJACR.2019.940150

Traditional machine learning and big data analytics in virtual screening: a comparative study

Sahar K. Hussin[1*], Yasser M. Omar[2], Salah M. Abdelmageid[3] and Mahmoud I. Marie[4]
Assistant Lecturer, Department of Communication and Computers Engineering, Alshrouck Academy, Cairo, Egypt[1]
Assistant Professor, Department of Computer Science, Arab Academy for Science Technology and Maritime Transport, Cairo, Egypt[2]
Professor, Department of Computer Science, Collage of Computer Science and Engineering, Taibah University, Saudi Arabia[3]
Professor, Department of Systems and Computers, Al-Azhar University, Cairo, Egypt[4]

Received: 08-December-2019; Revised: 24-February-2020; Accepted: 02-March-2020

Abstract

Nowadays, the massive amount of data that needs to be processed is increased. High-performance computing (HPC) and big data analytics are required. In the identical context, research on drug discovery has reached an area where it has no preference, but the use of HPC and huge data processing systems to perform its targets at a reasonable time. Virtual screen (VS) is one of the costliest tasks in terms of computation requirements. It is considered as an intensive and heavy task. At the same time, it plays an essential role in new drug design. This research investigates machine learning and big data analytics in VS. It tries to use a ligand base and a structural base and rank molecular databases as active against a specific target protein. The machine learning algorithms, including random forests, naive Bayesian classifiers, nerve networks, decision trees, support vector machines, and deep-learning strategies have been developed for both Ligand-based and structure-based docking. Also, this paper introduces a review of previous research conducted on the utilization of machine learning as well as big data analytics framework in VS. The paper outlines the current progress in the use of traditional methods for machine learning and massive data analytic applications in a multi-node dataset. This article compares the estimation of machine learning approaches and broad ligand-base theoretical system. It also explores how machine learning approaches can improve the performance of various problems of virtual screening classification in broad repositories. Finally, various challenges and solutions of the virtual screening dataset in the machine learning and big data analytics are discussed.

Keywords

Drug discovery, Virtual screening, Descriptors, Machine learning and Big data analytics frameworks.

1.Introduction

An unprecedented development in biomedical data has been observed in latest years. The capability to analyze a large portion of this data will offer many opportunities that will in turn affect the future of health care [1]. In this age, traditional storage and processing techniques are not sufficient to meet the demand and hence, computing techniques must scale to handle the huge volume of data. The main difficulty in managing these data is the speed at which they are generated, that is, data generation is much faster than the available computer resources for data analysis.

The acquisition and processing of big data (BD) are useful for researchers in various fields [2], such as drug discovery, which involves the searching and identifying of drugs. The process of drug discovery is an extremely long, complicated and expensive process. It may take 12 to 15 years and cost more than $1 billion, with the risk of failure [3]. In order to limit candidates, large sets of molecules, could be thousands, must be processed and selected.

Decision processes are, however restrained due to the growth in data generation, which poses a challenge to the development of data-based solutions that can effectively and accurately improves the decision-making in the drug discovery. High throughput screening (HTS) is tools that are widely utilized in

[*] Author for correspondence

the drug discovery and screening process, where large molecular libraries are screened in fully automated environments [3]. Nevertheless, libraries screening size limitations as well as their cost lead to few numbers of success stories with high error rate, including high false (positive and negative) rate [4]. As an alternative, VS is a pre-screening technique that is cheaper and faster than HTS. It has successfully been applied to reduce the number of compounds to be screened by generating new drug leads [5].

There are two VS approaches: Ligand-based-VS (LBVS) and structure-based -VS (SBVS) [6]. LBVS depends on the existing information about the ligands. It utilizes knowledge about a set of ligands that are known to be active for the given drug target. This type of virtual screening uses mining of BD analytics. Train binary classifier by a small part of ligand can be employed and very large sets of ligands can be easily classified into two classes: dockable ligands and non-dockable ones. SBVS, on the other hand, is used to dock experimentally. However, 3D protein structural information is required [7].

Machine learning (ML) plays a vital role in VS for drug discovery. It is a branch of artificial intelligence. Nowadays, adaptive ML algorithms can be utilized to redact Quantitative-Structure-Activity Relationship (QSAR) and illustrate, with high accuracy, how biochemical modifications could influence biological behaviour. In chemo-informatics, machine learning is usually utilized to classify molecules as active or in-active against a particular target or against multiple targets [8]. In addition, ML algorithms can be utilized in docking molecular methods. Traditional ML learning methods can be used in datasets of small molecules and still give the best result. Nevertheless, due to the increasing number of molecules in the library and the unstructured data format, traditional methods cannot achieve all the set objectives. Therefore, as a promising solution, BD analytics techniques can be utilized in VS [9]. The number of compounds in the chemical libraries has increased significantly. Libraries of molecules contain 1010 records (refers to the “volume” of data) and this value continues to rise. These libraries can be stored in different formats such as SDF or smiles files (refers to the “variety” of data) and the high rate of data generation refers to the “velocity”, “Volume”, “variety” and “velocity” are data characteristics that signify BD [10] The application of ML techniques to large libraries (big data) in the VS process is computationally costly [11]. There is a growing need to develop sophisticated frameworks of efficient BD analytics [12]. Some of the most commonly deployed BD analytics systems are Apache Hadoop and Apache Spark [13]. Recently BD analytics methods have been utilized by research in drug discovery to achieve high-performance and short-term objectives [14].

This study contains a summary of drug discovery ML and BD analysis. Next, a detailed study of chemical descriptors and properties which aid to the analysis is performed using the correct function output. Moreover, the paper compares data mining techniques that manage biological activities of compounds on an exhaustive screening basis. Second, the theoretical observations of recent papers on the conventional machine learning, in-depth learning and large data systems are discussed. Ultimately, this paper is a step forward towards the problem of virtual screening and classification algorithms to draw the future research directions, especially in large datasets. It also advises that the issues found in this area to be solved.

This paper is structured as follows: Section 2 explains the VS process in drug discovery. In Section 3, literature review on ML and BD in VS is discussed. In Section 4, machine learning algorithms in VS are studied as a based solution. Section 5 explains the study outcomes and performance evaluations while Section 6 addresses unanswered issues and potential areas of research. The article concludes, finally, in section 7.

2.Preliminary

VS strategies can be categorized into SBVS and LBVS, as in [15]. SBVS strategies are physical communications among compounds and a protein target. In recent times, SBVS uses ML algorithms and docking software to calculate the activity of molecules to a certain target. The issue with these methods is that they need a 3D structure of the protein that is not applicable to all proteins. If the target's 3D configuration is unattainable, LBVS is used [16]. This is usually denoted as similarity tests. Few of ligand data sets proven to be active or inactive for a particular protein are used as a preparation and an estimation of what is unknown as *Figure 1*. A large number of classification methods are used.

Usually, libraries of virtual screening datasets, in their chemical form, are saved as SDF or smiles files. Training dataset shows a vital role in the classification of molecules as active or inactive. In many public libraries, label and non-label data are accessible to model such molecules at a high-performance scale. Until date, numerous databanks have been established to focus on drug-like or non-drug-like ligand and also on molecular docking to predict high and low scoring molecules [17]. There are many libraries of ligands and proteins that are used in virtual screening [18], as shown in *Table 1*.

A preprocessing step involving descriptors is used to convert the chemical library format to .csv file. There are two types of descriptors that describe features of molecules, namely, numeric descriptors and chemical fingerprints. The descriptors as presented in *Table 2* are described in the following section.

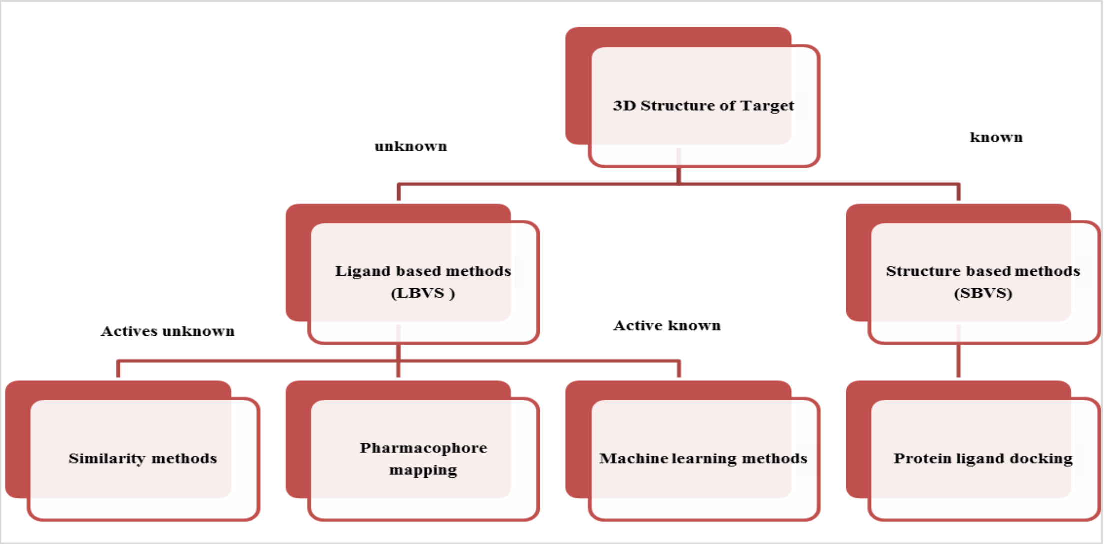

Figure 1 Taxonomy for 3D structure of virtual screening methods

Table 1 Compound library dataset

Compound library	No of compound	Link
ChemSpider	~62 million	http://www.chemspider.com
ChEMBL	~2 million	https://www.ebi.ac.uk/chembldb/index.php
PubChem	~92million	http://pubchem.ncbi.nlm.nih.gov
Enamine	~1.7 million	http://www.enamine.net
Chembridge	~1 million	http://www.chembridge.com
Drug Bank	~9591 (D)	http://www.drugbank.
STITCH	~500 000	http://stitch-beta.embl.de
Binding DB	~ 635 301	http://www.bindingdb.org .
BindingMoad	~12 440	http://bindingmoad.org
KEGG	~18 211	http://www.kegg.jp
ZINC	~35 million	http://zinc.docking.org
eMolecules	~7 million	https://www.emolecules.com
ChemDv	~1.6 million	http://www.chemdiv.com

Table 2 Descriptor software

Software	Description	Web Site
PaDEL	Calculates 1876 molecular descriptors and 12 type fingerprints	http://www.yapcwsoft.com/dd/padeldescriptor/
RDKit	RDKit is open source software and a collection of cheminformatics and machine learning software written in C++ and Python.	http://www.rdkit.org/
Dragon	Computes 5,270 atom descriptors by	http://www.talete.mi.it/

Software	Description	Web Site
	using various theoretical approaches	
Open Babel	Computes 4 types of fingerprints - FP2, FP3, FP4 and MACCS	http://openbabel.org/
Kit (CDK)	Open source tools that uses Java library to give descriptor and fingerprint of molecules	https://cdk.github.io/

2.1Chemical descriptors

They are numerical features derived from ligand data processing chemical structures, hybrid complexity analysis and compound behaviour predictive. One, two, three or four-dimensional descriptors (1D, 2D, 3D or 4D) are conceivable. Looking at single-dimensional descriptors, they use different measuring devices that are able to display different information such as 1) the number of atoms, 2) the bond count, 3) the molecular weight, 4) the sum of atomic properties, and 5) the number of parts. On the other hand, the 2D descriptors are topological descriptors, which illustrate a compound with bonds between atoms and features like number of atomic bonds, substructure information and molecular connectivity index. Though, the geometrical descriptors are (3D descriptors) for 3D auto-correlation and surface properties, and 4D descriptors are 3D coordinates and conformations. In literature, there are many of the software descriptors that are used for dataset feeding to Machin learning algorithms [19]. Some of them are presented in *Table 2* such as PaDEL, Dragon, RDKit and CDK toolkit. They generate high dimension features according to the type of descriptors.

2.2Chemical fingerprints

A fingerprint is a specific case of the dual vector, where each place contains 1 or 0 to show the absence or presence of a descriptor. The size of this type of fingerprint grows with the number of representative descriptors but, in order to use space efficiently, it can be compressed to the shortest ones [20].

In the next section, we describe the taxonomy of virtual screening-based solutions. The taxonomy refers to two different solutions including machine learning-based solutions and big data analytics frameworks-based solutions. Current machine learning methods include 1) decision tree, 2) naive Bayesian, 3) natural, 4) k-nearest neighbours, 5) artificial neural network, 6) support vector and 7) deep-learning algorithms. Certain methods are used in LBVS and in SBVS. Spark, Hadoop or MapReduce are used as for Big Data Analytics Frameworks, see *Figure 2*.

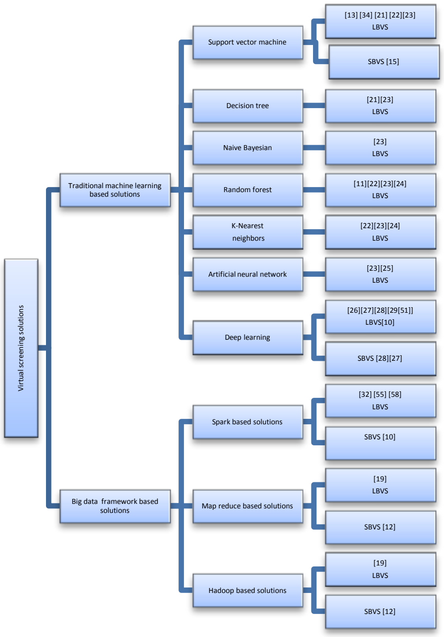

Figure 2 Taxonomy of virtual screening solutions

3.Literature review

This section is dedicated to the research that explore machine learning performance used in VS. it reviews traditional machine learning methods and machine learning along with big data frameworks used for VS.

3.1Traditional machine learning based solution

Virtual Screening is typically employed to remove unwanted molecules (i.e. in-active or toxic) from a compound library. ML algorithms can be used for VS by analyzing the important feature of molecules with well-known activity or in-activity. Support vector machine (SVM), wrapper method (WM) and subset selection subset (SS) have been used to classify ligand as drug-like and non-drug-like [21]. PaDEL was used as software descriptor to calculate attributes of ligand. The model's accuracy levels were 88% for SVM, 90% for WM and 91% for SS. Authors use SVM, random forest and deep learning algorithms in another study [22] to identify this combination as being active and/or inactive. Fingerprints provides 94 percent accuracy as descriptors. Other researchers introduced a new paradigm in [23], which included three ensemble classifiers. Such models are used to determine the resemblance between drugs with the Tree Ensemble Classifier. In [24], the author used ML algorithms including DTs, SVM, Random Forest (RF), Naive bays, Rotation forest and k-Nearest Neighbour (KNN); in addition, the classification accuracy of all these algorithms were compared. These models were utilized to predict drug-likeness using tree-based ensemble classifier. In [25], data mining procedure was derived using a workflow consisting of two main stages - data visualization using the t-SNE method and six different algorithms, namely, DT, RF, SVM, KNN, linear regression (LG), artificial neural network (ANN) and ensemble-algorithm. These were used to build the classification models, distinguish drugs and non-drugs and generate three major classes of drug compounds. In [26] deep learning was used to predict the potential targets of new drugs and drug indicators using extensive chemical genome data while improving the performance of predicting the target reaction of the drug. This model generated an accuracy rate of 98%. In the same context, in [27], author introduced a deep learning-based approach that can identify ligands of targets. The performed experiments showed that deep learning outperforms the two widely used methods, which are Auto Dock Vina and Smina. The developed method accomplished higher values of area under curve (AUC). In [28], authors developed a model based on deep learning and deep synergy and predicted, with high accuracy, dozens of synergies of drug combinations for cancer cell lines. Scientists have proven that deep synergy is able to provide the best predictions in the preparation of mutual verification with external test groups, outperforming other methods by a wide margin. Preparation of drug combinations based on deep synergy predictions at AUC of 0.90 can already reduce the time and costs spent on experimental verification. In [29], deep learning and machine learning methods were introduced to determine the impact of these modern methods in predicting new compounds against specific targets. Prediction and similarity of targets help to examine potential compounds based on already approved drugs.

3.2Big data analytics framework-based solutions

Models for big data analytics are used in many fields such as social media, education, wireless network and drug discovery (especially in virtual screening). Big data means massive amount of data which is difficult to handle throughout conventional ways. There are four VS which define the features of large data: quantity, size, variety and truthfulness. In big data situation, many research works have been concerned with VS including Apache Spark and Apache Hadoop. In [30], deep learning algorithms utilized in Apache SparkH2o on big data set to classify a compound as a drug-like and non-drug-like reached one million ligands at a high accuracy rate. Dataset of drug discovery is highly dimensional. Deep learning is one of the algorithms that can handle large number of dimensions without the need for feature selection; however, it needs huge training datasets. In [31], the authors recalled five models where the suggested algorithms were designed to be used on different big data platforms including Hadoop/MapReduce. Then, the authors selected three algorithms random forest, multilayer Perceptron, and naive based to develop the ensemble classifier and calculate the activity of ligand. In [32], authors offered a pretty new method that is based on Apache Spark and the ensemble learning model to upgrade the performance of large-scale VS processes. Three classifiers are used in combination, which include SVM, multi-layer perceptron, and DT, to create the ensemble learning approach, in which the method of aggregation had the common vote.

4.Machine learning algorithms (ML)

4.1Support vector machine-based solution

SVM are algorithms subject to machine learning supervision to facilitate compound classification, forecasting and regression-based property values [33]. For example, to distinguish drugs from non-drug [34] or between compounds that have or do not have specific activity [35, 36], in artificial accessibility [37], or water solutions [38], SVMs are usually used for predicting property or binaries. First, complex libraries are projected onto a high-dimensional space where it is anticipated that the particles, represented as descriptive vectors, will become linearly separated, as shown in *Figure 1*. SVMs is used to solve the classification problem using nonlinear kernel functions that maps high-dimensional space data by finding a hyper-plane. The hyperplane is suitable for maximizing the margin between SVM, and points closest to the accuracy limits expressed as a linear set of data points [39]. Linear, polynomial, and radial basis function (RBF) are often the preferred grain. If there is no kernel that can convert data points into hyper-space that is completely linear separation, then you should choose a basic and hyper-plane optimization that reduces the number of misclassified training point. SVMs are generally among the best performance in ML and VS comparison studies (*Figure 3*).

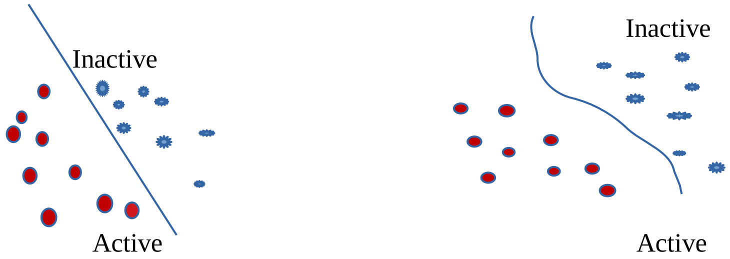

Figure 3 Support vector machine

4.2Naive Bayesian classifier-based solution

Naïve Bayesian (NB) classifier is one of the simpler ML techniques. NB classifiers are utilized in cheminformatics and analyze against different classifiers in order to predict organic as opposed to physicochemical properties. Practical applications of this technique have been seen in the virtual screening field as well as other fields, for example, the expectation of toxicity of the compound [40], and protein target and bioactivity grouping for medication-like particles. It is on a fundamental level to utilize NB for regression; however, this is rarely observed in cheminformatics. Bayesian techniques are based on the Bayes' hypothesis, which provides a mathematical equation to describe the possibility of an event that may be the result of any two or more causes, as in Equation 1 [41].

$$p\left({}^{A}\!/_{B}\right) = \frac{p\left({}^{B}\!/_{A}\right)(A)}{p(B)} \qquad (1)$$

The term P((B/A) P(A)) is progressively being utilized given its flexibility, strength and usability. The most important weakness is that even though the chance of estimation is inaccurate, the number of cases of NB during which it will act well, partly as a result of the classifications created, will still be optimum. Inaccuracy in estimation stems from feature dependence, as in [42], where combined NB models together with SVM were used to create an enhanced ensemble model.

4.3Decision tree-based solution

DT is usually depicted as an actual tree with its root at the upper and the leaves at the bottom. Starting from the root, the tree is divided from one square into two or more branches. Each branch can be divided into two or more branches. This continues until the leaf is reached, which means there is no more split knot. The division of the branch is referred to as the internal node of the tree. It is also referred to as root and nodes. Each leaf node is set with a target property while a non-yellow node (root or inner node) is set with a molecular descriptor that becomes a test case. Outer branches are assigned to groups with different properties [43]. It is used as an LBVS to differentiate between active and in-active ligand.

4.4Random forest-based solution

RF is an algorithm for supervised learning. RF has many resolution trees and fuses them to make a better and more stable forecast. The main advantage of random forests is that it can be used for both classification and regression problems, which make up the majority of current machine learning systems, as in [44]. Random forests are suitable when the target protein is not selected or unknown because random forests can find good combinations of features from the many features available. It is also called the band method. The band method is an effective technique for predicting situations and providing better performance. RF contains a set of machine learning algorithms called bootstrap aggregation or packing. It is known that packing reduces the variance in the algorithm. RF takes a subset of observations and factors in order to build a decision tree. It then creates a different decision tree and combines the new tree with the old in order to get a more accurate, stable and similar prediction. RF is also able to handle missing values. Random forest classifier does not over-fit the model when there are more trees in the forest. One of the main advantages of random forest is the ability to handle a large number of datasets with higher dimensions. It possesses a powerful strategy for assessing missing information and maintaining precision when a majority of the information is missing, as in [45]. RF is considered as an important algorithm that recently used in VS.

4.5K-Nearest neighbours-based solution

The KNN is an algorithmic rule that could be an easy and intuitive methodology to forecast the category, rank [46] of a ligand that is supported by the nearest training examples of the feature area. KNN could be a reasonable instance-based learning or lazy learning. In either case, the performance of KNN is simply approximated regionally and every calculation is postponed until classification. KNN could also be utilized for regression. KNN is one of the best ML algorithms. A ligand e is assessed by a common vote of its neighbours, with the ligand being allocated to the most typical category among its K nearest neighbours. In binary classification issues, it is useful to select K to be an odd range to avoid tied votes [47]. Typically, Euclidean distance is approved, though alternative distance measures like the Manhattan or Mahalanobis distance may, in theory, be used instead. A poor selection of distance metrics may lead to mindless classifications.

4.6Artificial neural networks-based solution

The two major forms of artificial neutral networks (ANN) are: feed-forward networks (which are supervised) and back propagation. Each network is made up of a series of connected neurons. A neuron takes multiple numerical inputs and outputs are reworked based on the weighted sum of the inputs, as shown in *Figure 4*. Common transformation functions embody the tanh and sigmoid functions. Neurons are organized into layers. ANN can contain a range of hidden layers [48] and the neurons will be connected only to those in the next layers, known as feed forward networks, multiplayer perceptions (MLP), or the functional radial basis network (RBN). ANN is commonly used as a classification tool and it examines sets of molecules that match neurons that are occupied by known actives and are additionally deemed to be active. The neural counter propagation network [49] is a closely connected architecture. The most popular technique used in ANN is a back-propagation mechanism in which the weights are set arbitrarily to reduce the output error from the closest layer to the input layer.

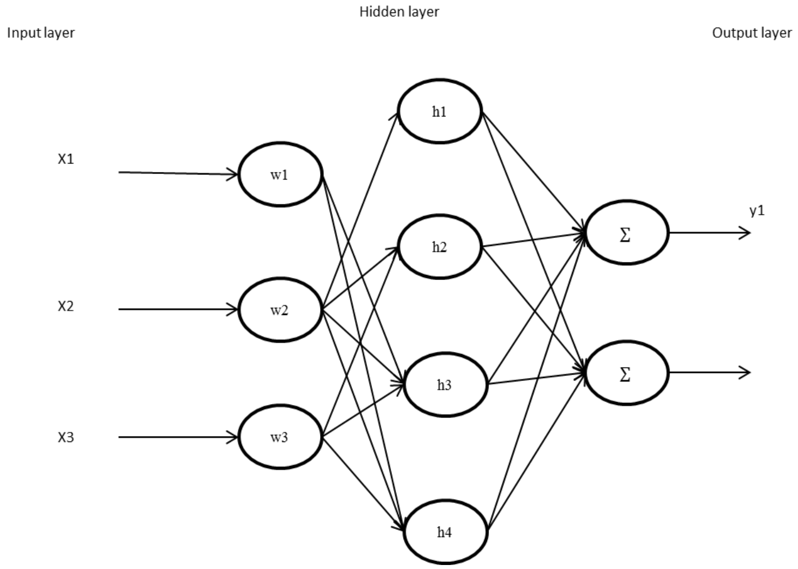

Figure 4 Artificial neural networks

4.7Deep learning-based solution

It converts the low-level features obtained from the input into more complex features of each subsequent layer [50]. The relationship between output and input values of a hidden unit is demonstrated for output value Yj of the node I and is calculated as follows in Equation 2.

$$\text{Yj} = g\left(\sum_j W_{ij} * a_j\right)g\left(\sum_j W_{ij} * a_j\right) \qquad (2)$$

DL is a new research algorithm in the field of drug discovery. DL is a great technology that can handle large datasets without feature selection requirements [51].

5.Performance comparisons

5.1Evaluation metrics and performance comparison of VS methods

There are many standard metrics for evaluating the working performance in a particular test group. These are based on the calculated amounts of the confusion matrix. Here, we tend to the widely used analysis scales employed in literature: Precision in Equation 3, Recall in Equation 4, F1-score in Equation 5, Accuracy in Equation 6, Matthews Parametric Statistic (MCC) in Equation 7 (formulations are given below in conjunction with quantitative ranges), and false positive rate (FPR) in Equation 8 [52].

$$\text{Precision} = \frac{TP}{TP+Fp} \qquad (3)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \qquad (4)$$

$$\text{F1score} = \frac{2*precision*Recall}{precision+Recall} \qquad (5)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (6)$$

$$\text{MCC} = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}} \qquad (7)$$

$$\text{False positive rate (FPR)} = \frac{FP}{FP+TN} \qquad (8)$$

FP, TP, TN and FN are the percentage of false positive, true positive, real and false negatives. Each of these metrics has completely different properties. For instance, precision denotes fractions of the right predicted samples (TP) for all positively predicted targets, although recall (i.e. TP rate) denotes the fraction of acceptably predicted samples of all truly positive samples. Assessing the performance of methods using only precision or only recall may yield unrealistic results.

5.2Datasets

Experimentally determined calculations of the performance evaluations for virtual screening data set on two cases were done. First, models were implemented that used machine learning algorithms on a Python platform. Second, machine learning algorithms in big data platform such as Apache Spark/Hadoop were used. In case one, the datasets of compounds that are used in this paper were retrieved from [21]. The first dataset contains 762 compounds, which classified into two classes: active (366 compounds) and inactive (396 compounds) as shown in *Table 3*. These samples are further split into 80% training and 20% test samples. Thirty-five molecular properties were chosen and calculated for each compound using ChemAxon1 and XLogP software. Second, datasets from PubChem were used [53], that is, AID 651820 (qHTS Assay for Inhibitors of Hepatitis C Virus HCV) 11,664 ligands as drug-like and 271,341 non-drug-like was collected for this protein. AID 651820 is an unbalanced dataset that has been analyzed in several articles [31, 32, 27, 28]. The ligands that are used in this work are represented by sets of descriptors (i.e., feature vectors). The molecular descriptors of all ligands were calculated using the cheminformatics software, PADEL, as a numeric and fingerprint descriptor. Numeric PADEL generates 1444 attributes showing information such as fragment counts, atom counts, molecular weight, bond counts and sums of atomic properties. PADEL is also used as a fingerprint descriptor in PubChem and it gives 881 attributes. Machine learning algorithms are implemented using Pyspark 2.4.3, Hadoop 2.7 version [54] and Python as program writing languages with the Jupyter 3.7 version notebook. The computer configuration used for experiments is a local machine - Intel core i7 with 2.8 GHz speed and 8 GB of RAM. Pyspark standalone cluster was created locally and then one spark master was first launched using the command line./sbin/start-master.sh. Next, three workers were launched using the command./sbin/start-slave.sh, where the master URL takes the following spark master format at spark: //sahar-Precision-M4800:707, as shown in *Table 3.*

Table 3 Dataset description

Dataset	Total data	Active compound	Non active compound	No.attribute
Breast cancer	762	366	396	35
AID 651820	283,005	11664	271341	1444

5.3Results and discussion

This paper experimentally analyzes some of articles [21, 23‒25] that use machine learning algorithm to predict ligand activity to specific proteins. Some papers [27, 28] used deep learning to predict drug activity. This paper also investigates other works that used big data platforms to predict activity in virtual screening [31, 32].

The experimental results for [21] are presented in *Table 4*, with support vector machine as classification method. Our analysis generated different results with 81% accuracy, specificity of 0.90 and sensitivity is 0.88 using the sub selection feature method (SS).

In addition, two different feature selection methods - wrapper (WM) and recursive feature elimination (RFE) method – are implemented. WM gives 89% accuracy; sensitivity is 0.89 and specificity is 0.72. RFE gives 77% accuracy, sensitivity is 0.87 and specificity is 0.70. Using only personal correlation gives the least accuracy of 76%, sensitivity of 0.89 and specificity of 0.64. The results indicate that SVM with sub-selection method (accuracy of 81% and specificity of 0.90) gives the best outcome in contrast to the other methods. In [23], three selection methods for features were used: genetic algorithm (GA), RF and minimum Redundancy Maximum Relevance (mRMR) on SVM classifier. GA gives 94 percent of the experimental results, 1 sensitivity and 0.93 specificity. In [24], authors used six different classification algorithms – RF, DT, SVM, NB, KNN and rotation forest (RT) – with the same feature selection method, that is, linear discriminant analysis (LDA) and quadratic discriminant analysis transformations. Our experimental results give the best accuracy of 95% in RT, sensitivity of 0.91 and specificity of 0.91. It is followed by RF with 94% accuracy, sensitivity of 0.9 and specificity is 0.9. In [25], authors used PCA and t-SNE to visualize datasets and reduce dimensions and then different six machines learning algorithms – RF, DT, KNN, SVM, ANN and AL Boosting – were used to evaluate results. Performance evaluations of AL Boosting and logistic regression (LG) give the best accuracy (96%), sensitivity (1) and specificity (0.96). The experimental results for four articles indicate that 96% was the highest accuracy for t-SNE/AL Boosting [25] followed by LDA/RT in model [24] by accuracy 95%. A summary to the results is given in *Figures 5* and *6*.

Table 4 Performance evaluations for traditional machine learning in four models

Ref	Method	Accuracy %	Sensitivity	Specificity	No. of features
[21]	Personal _Corr/SVM	76	0.89	0.64	11
	REF/SVM	77	0.87	0.70	7
	WM/SVM	80	0.89	0.72	7
	SS/SVM	81	0.88	0.90	6
[23]	mRMR/SVM	80	0.33	0.92	10
	GA/SVM	93	1	0.92	16
	RF/SVM	87	1	0.83	7
[24]	LDA/RT	95	0.91	0.91	12
	LDA /RF	94	0.90	0.90	
	LDA /DT	91	0.89	0.89	
	LDA /SVM	90	0.87	0.86	
	LDA /KNN	91	0.90	0.90	
	LDA /NB	89	0.90	0.87	
[25]	t-SNE/J48	92	0.67	1	14
	t-SNE/RF	95	1	0.96	
	t-SNE/KNN	92	0.75	0.95	
	t-SNE/SVM	92	0.75	0.95	
	t-SNE/ANN	96	0.75	0.96	
	t-SNE/LG	96	1	0.96	
	t-SNE/Al boosting	96	1	0.96	

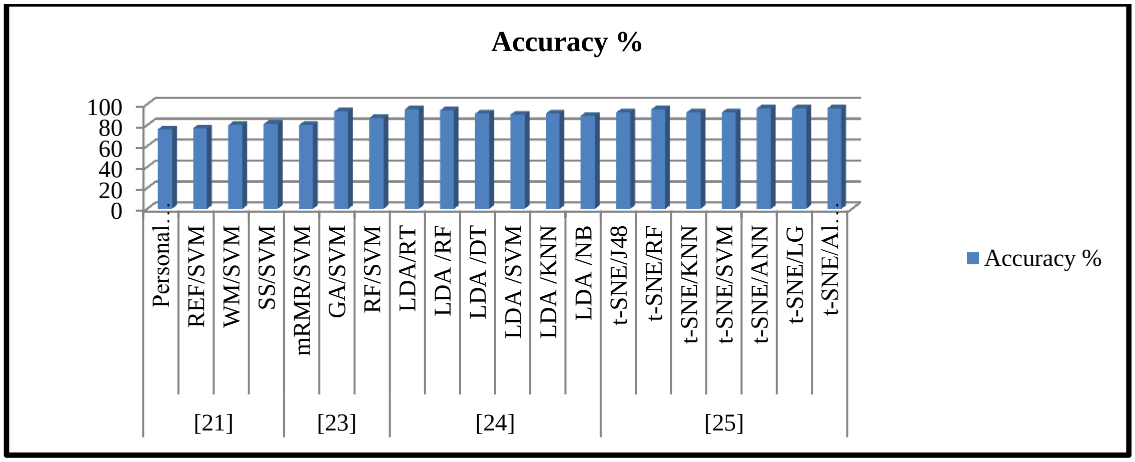

Figure 5 Accuracy of various models of different selection processes and common algorithms for machine learning

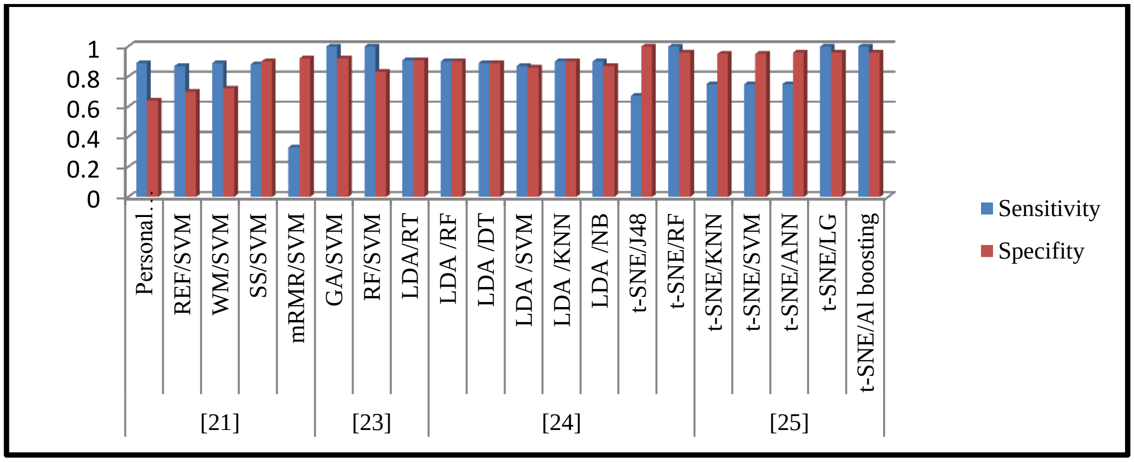

Figure 6 Sensitivity and specificity of various simulation approaches and of multiple algorithms for machine learning

High accuracy of results comes from the balance between active and inactive and also the selection of the best features of attributes. All these papers applied different features selection methods to reduce dimensions and then used machine learning algorithms to increase accuracy. In order to compare the methods, the difference between specificity and sensitivity was also calculated. The difference represents the typical balance between the two categories or the bias of the model towards one. Low difference values represent balanced models while higher difference values represent high bias and imbalanced models. By comparing all four models, the best model is t-SNE. The results of the test set for the t-SNE/AL Boost method show very similar results of specificity and sensitivity (0.96 and 1, respectively). Also, t-SNE /LG and t-SNE/RF also give the same results (0.92 and 1 for the specificity and sensitivity, respectively) followed by the genetic algorithm, which used support vector machine as a base classifier.

On the other hand, the papers that used DL, as in [27, 28], have summarized their experimental results in *Table 5*. The authors introduced an effective computational method that is semi-supervised DL based, a combination of stacked auto encoding and a supervised DL network. The aim was to forecast probable drug targets and new drug indications by using a large-scale chemo-genomics data while increasing performance. Authors used a numeric descriptor as a feature generation. Results for DL give the best results –accuracy (98%), sensitivity (0.94) and Specificity (0.95). DL gives the best performance and does not need feature selection methods. DL needs however a large number of training datasets. Authors in [28] also used DL with fingerprint descriptor as a feature generation. The experimental results give an accuracy of 90%, sensitivity of 0.85 and specificity of 0.95. Better performance was seen for DL in [27] because the authors used two sequential processes: first pre-training layer process is unsupervised that is used stacked auto encoders, the second is supervised layer, fine-tuning process of DL Stacked is one of the popular DL models, built with multiple layers of auto coding, in which each layer output is connected to the next layer input. It tries to reconstruct the same features at the output layer using its hidden activation. Accuracy in [27] is higher than in [28], as also shown in *Figure 7*. Sensitivity and specificity are depicted in *Figure 8*.

Table 5 Performance evaluations for deep learning in two models

Ref	**Method**	**Accuracy %**	**Sensitivity**	**Specificity**	**No.features**
[27]	DL	98	0.94	0.95	1444/numeric descriptor
	SVM	81.08	0.76	0.858	
	K-NN	82.26	0.785	0.853	
	EL	89.65	0.888	0.885	
	RF	81.84	0.795	0.84	
	DT	75.05	0.516	0.668	
[28]	DL	90	0.85	0.95	881/fingerprint
	GB	89	0.71	0.89	
	RF	87	0.60	0.94	
	SVM	81	0.69	0.77	
	Elastic Nets	77	0.62	0.78	

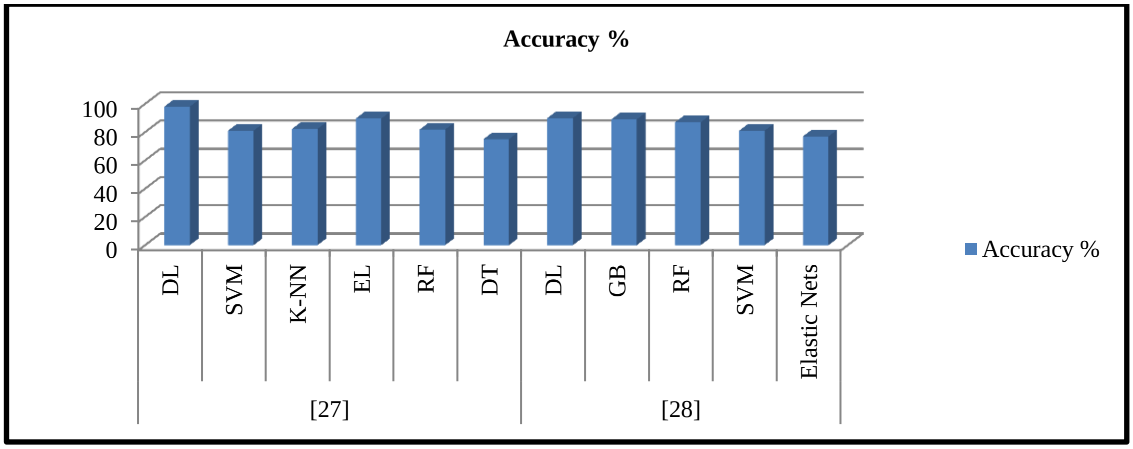

Figure 7 Comparisons between accuracies of deep learning

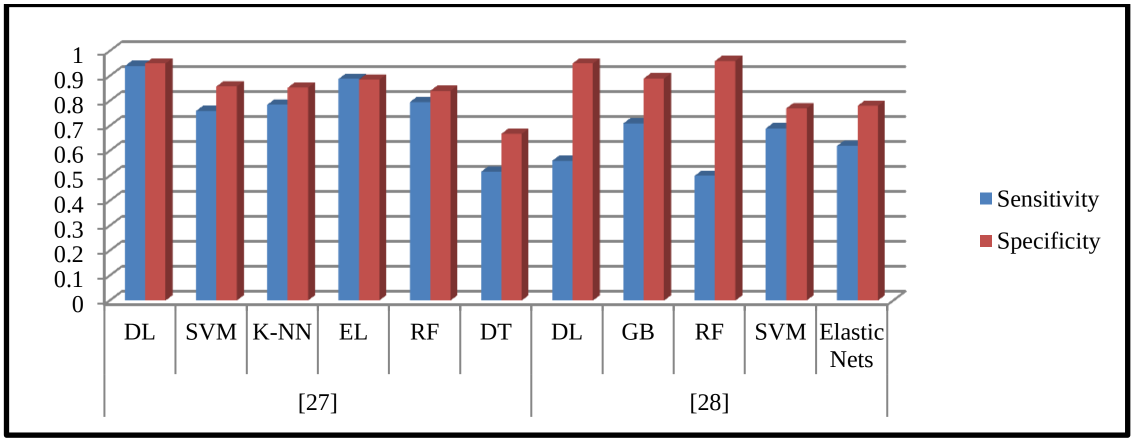

Figure 8 Sensitivity and specifity of different models in deep learning

AID 651820 dataset is implemented on Apache Hadoop, as in [31]. Big data analytics algorithms are used (RF, NB, SVM, MLP and ensemble learning technique (ET)). ET algorithm gives best accuracy results of 90% and precision of 0.86, but it takes 75 seconds. Apache Spark is used, as in [32], for different big data analytics methods (MLP, DT, NB, SVM and ET). ET gives the highest accuracy (94%) and precision (0.93), but it takes longer time than DT (*Table 6*). Based on the results, SVM algorithm takes the longest time (175 seconds) while DT algorithm take the shortest time (24 seconds). Comparing results of the two models in the same dataset, as shown in *Figure 9*, indicates that the accuracy of ET algorithm on Spark gives higher accuracy of 94% than ET on Apache Hadoop (90%) because of the differing descriptors.

Time of machine learning algorithms in Spark takes the shortest time than in Hadoop for all machine learning algorithms, as shown in *Figure 10.* From experiential observation, Apache Spark as big data analytics tool is much faster and gives accurate results than Hadoop. Also, using numerical descriptors gives better results than using fingerprint. The use of different descriptors, numeric and fingerprint, has an effect on the accuracy of classifiers and their speed. Ensemble learning in Spark gives best accuracy and precision of 94% and 0.93, respectively. However, it takes 48 seconds to be executed. Although DT gives the best execution time (24 seconds), it is still lower than ET in accuracy and precision. MLP follows ET, with 88% and 0.85 as accuracy and precision, respectively. Imbalance between active and inactive in big data affects the accuracy and precision of all classifiers.

Molecular descriptors act a vital role, particularly in drug activity researches. The selection of important descriptors is also essential and affects the performance and overfitting of the system. There are two types of descriptors – numeric and fingerprint. It can be complicated to understand descriptors and pattern of drug activity relationship. Nevertheless, with a few numbers of attributes are generated from descriptors, the basic mechanism of drug activity can be easily detected. Moreover, the selection of important attributes is useful for increasing both the speed and accuracy of ML methods, especially in BD set.

Table 6 Performance evaluation of the experimental results of machine learning big data frame work for two models

Ref	Method/framework	Accuracy	Precision	Time/s	Descriptors
[19]	RF/hadoop	83	0.78	49	fingerprint /OpenBabel
	SVM/hadoop	86.5	0.83	500	
	NB/hadoop	82	0.77	**38**	
	MLP/hadoop	88	0.85	67	
	ET/hadoop	**90**	0.86	75	
[32]	MLP/spark	88	0.85	37	Numeric descriptors/CDK
	SVM/spark	87	0.83	175	
	NB/spark	82	0.78	28	
	DT/spark	83	0.80	**24**	
	ET/spark	**94**	0.93	48	

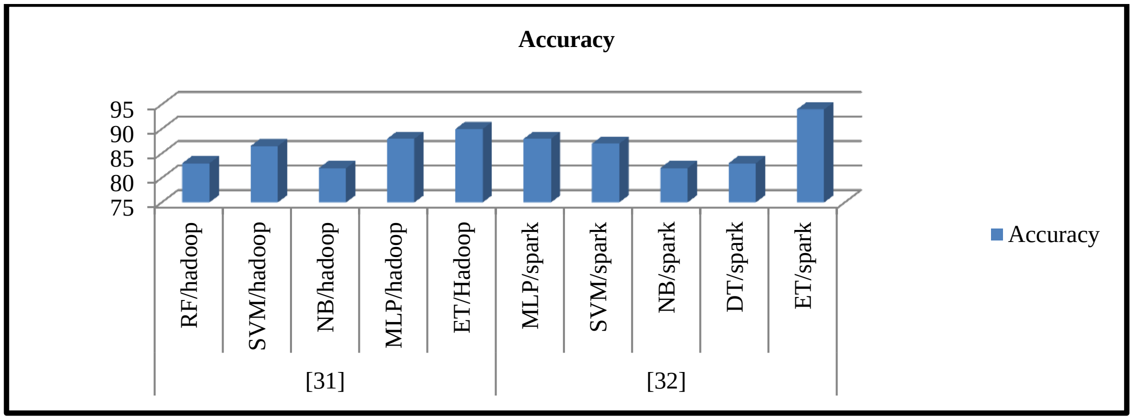

Figure 9 Accuracy comparison

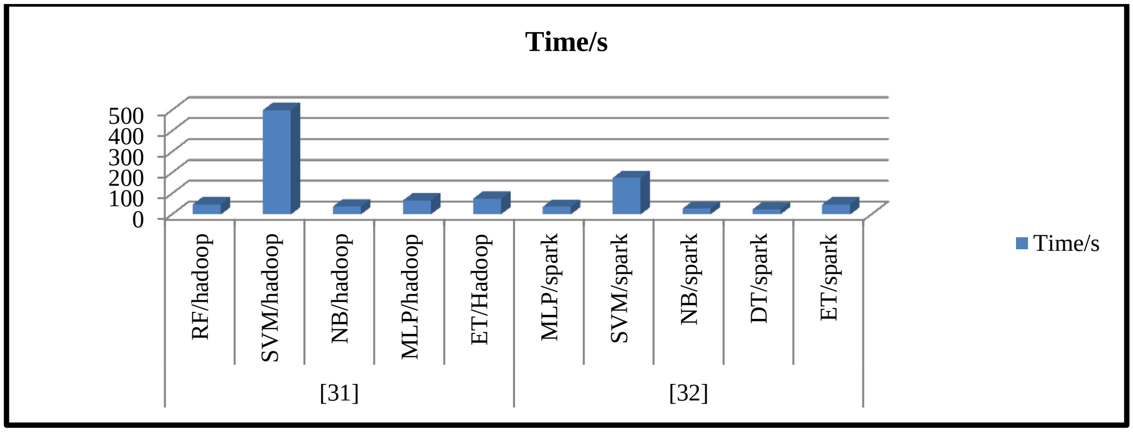

Figure 10 Execution times in two models

6.Open problems and future direction

As discussed above, many solutions for virtual screening have been introduced. However, there are still many problems:

1) **Large volume of data:** Chemical libraries contain up to 10^{10} records and this value continues to rise. A large volume of data renders traditional machine learning algorithms insufficient [55], but these algorithms can be used appropriately for these datasets in default scan. Apache Hadoop, MapReduce and Apache Spark are used as tools for big data. Apache Spark, for example, has ML library, uses all machine learning algorithms and can handle big data faster than Hadoop and MapReduce by segmenting datasets to tasks. It also has more workers. To produce faster results, each worker has been allocated to some of these tasks

2) **High dimension:** Virtual screening dataset has been applied in many articles and the authors suggested different solutions to this problem of high dimension such as filtering, wrapped, embedded and genetic algorithms [56]. However, it is challenging for traditional methods to reduce high dimension in big data.

3) **Imbalanced dataset: R**esearchers has used a library of chemical datasets to search for active compounds based on a specific protein such as breast cancer, colon cancer, Alzheimer and human immunodeficiency viruses. The majority of these compounds are inactive to these proteins and the minority is active as datasets in PubChem database. Classification of imbalanced dataset is inaccurate due to minority data given low accuracy. Many research articles addressed this problem and have obtained the best results using different analytical framework. Balance methods involve data sampling or data level, cost-sensitive or algorithm level, and hybrid methods [57]. Nevertheless, it is still challenging to find solutions for dataset imbalances.

4) **Unsupervised dataset:** More than 90% of the dataset used in virtual screening is unsupervised data. A small number is tested according to some chemical labels such as PubChem database and drug bank. Labeling large unsupervised datasets is still the main challenge for big data analytics and ML algorithms. Clustering and self-organized map techniques are two of the used unsupervised methods. This leads to labeling only few dataset records leaving huge records unlabeled. A semi-supervised learning could be another solution such as active learning by self-training or co-training [58]. It can manage the imbalance of virtual screening dataset.

ML methods have been widely used in bioinformatics to identify and design new drugs with greater biological activities. One of the key areas of innovation in this area is the combination of big data and machine learning to predict a broader range of biological phenomena.

The general protocol for the construction of screening contains of numerous modular stages including chemical informatics and ML techniques, as discussed earlier. The first step is "molecular coding" or feature generation where chemical features and properties are generated from chemical structures. Feature generation can be done by different software descriptors such as RDKit, PADEL and Open Babel. These descriptors give two types of descriptors – numeric and fingerprint, as described in Section 2. The second step is the preprocessing step containing two-phase feature selection to reduce dimensions and data sampling when dataset suffers from imbalanced data. The feature selection step is implemented where learning techniques are used to identify the most relevant characteristics, decrease the feature vector dimensions and reduce over-fitting. Chemical dataset also suffers from imbalance so data sampling is used to overcome this problem. Finally, at the learning stage, a supervised ML model is applied to generate an experimental function (either explicitly or implicitly) that can optimize input feature vectors and biological responses. In fact, the creation of a particular model involves the analysis and collection of datasets used for model training and validation. The full framework of the VS classification process is used to classify and predict ligand activity to a specific protein target, as in *Figure 11*.

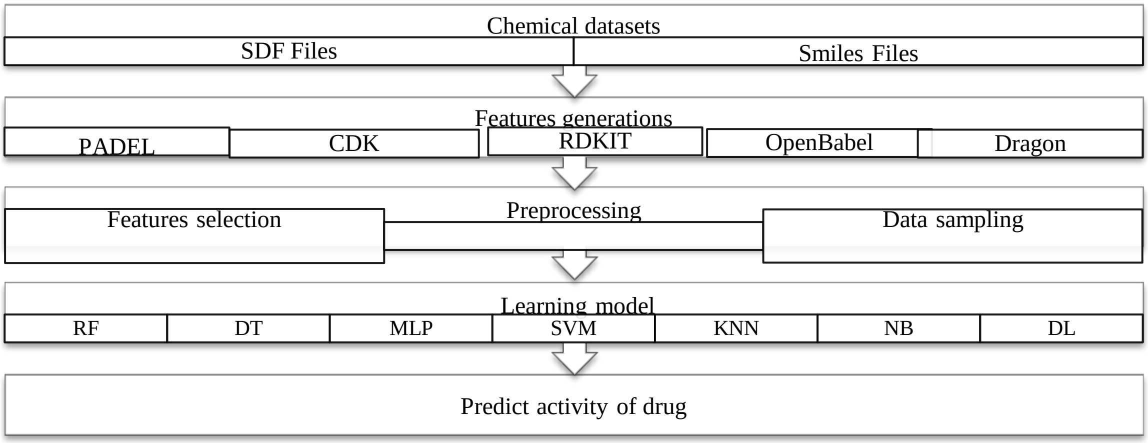

Figure 11 The classification block diagram of virtual screening

In what follows, the future directions in VS will be discussed. The importance of LBVS has grown in recent decades because of its ability to find lead compounds and scaffolding that limit the number of vehicles available for testing. However, in the age of big data, growth in chemical databases is expected, both in terms of the size and diversity of information. Future similarity algorithms may not be accurate enough in themselves to reach the accuracy of predictions. However, the combination of methods of a different nature, such as the above-mentioned numeric, fingerprint descriptors and SBVS/LBVS approaches, is expected to be important in the coming years because they have already been proven successful in previous works. New automated learning models of great versatility that are abled of processing BD in large quantities, velocity and veracity are also required. The recent development of DL networks has assured to promise well for effective learning from large data sets to modern drug discovery campaigns. The major challenge of virtual screening is the combined use of DL with BD analytics frameworks such as Apache Hadoop and Apache Spark. These frameworks have become general because they are easily reachable, both as open source and commercially. Traditional feature selection methods have become expensive for computation because of the slow learning process when dealing with a high dimension of large dataset. There are two methods that are used with deep learning (DL). Those methods are utilized to decrease the dimensions and make processing faster. One of these methods is relegated, stacked noising auto encoders, which scale effectively for high-dimensional data and are computationally faster than regularly stacked noising auto encoders. High dimensional data are also classified by another method which is convolution neural networks. In the virtual screening domain, the input dataset consists of a mix of both labelled and unlabelled data. In such cases, DL algorithms can include semi-supervised training methods to define standards for quality learning to represent data. In future work, we will introduce solutions to some of these problems in large datasets in big data platforms such as Apache Spark. Pharmaceutical companies can use big data analysis techniques effectively to find a lead molecule to develop into acceptable and active drugs. The article discusses specific methodological methods that can be used in various stages of virtual screening process (VSP). Some of the approaches are proposed for efficient purposes. The framework will provide guidelines for potential research in drug discovery big data analysis field. In the future, the implementation of these advanced analytical approaches that support technology can improve the success rate of drug discovery. As big data analytics technologies become more important, other issues including the privacy and security of biomedical data and generation of standard bioinformatics tools and databases will continue to be considered.

7.Conclusion

This article provides a comprehensive analysis of chemical descriptors and properties that contribute to the use of VS for appropriate feature generation. In this paper, seven advanced ML algorithms are discussed. These algorithms are commonly used in cheminformatics as supervised classification algorithms and they are compared in virtual screening for drug discovery. These algorithms are SVM, DT, RF, k-NN, NB, ANN and deep learning methods. The possibilities and weaknesses of these algorithms are systematically analyzed here, with particular emphasis on their realistic pertinence and importance. LBVS and SBVS are common classification methods and output comparisons are made. Both SVM and Bayesian methods currently dominate the LBVS field. LBVS ' success depends heavily on the size and diversity of the training. SBVS uses machine learning algorithms and docking software to calculate the activity of molecules to a specific target. This paper gives an overview of the big data platform that can be utilized in the virtual screening process. There is also a comparative study of latest research utilizing conventional computing machines and software algorithms in Big Data. Finally, this paper experimentally implements four articles that used traditional machine learning algorithms and two papers that used machine learning in big data platform (Apache Spark and Hadoop). The accuracy and precision of all methods were calculated. Finally, open problems were also discussed. In future work, we plan to develop a new model to solve big imbalanced dataset in virtual screening and to predict a large number of unsupervised libraries using a small number of labelled dataset.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Ross K. Protein bioinformatics: from protein modifications and networks to proteins. Humana Press. 2017.

[2] Chen B, Harrison RF, Papadatos G, Willett P, Wood DJ, Lewell XQ, et al. Evaluation of machine-learning methods for ligand-based virtual screening. Journal of Computer-Aided Molecular Design. 2007; 21(1-3):53-62.

[3] Yang H, Chen J, Tang S, Li Z, Zhen Y, Huang L, et al. New drug R&D of traditional Chinese medicine: role of data mining approaches. Journal of Biological Systems. 2009; 17(3):329-47.

[4] Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Research. 2012; 40(D1):D1100-7.

[5] Maltarollo VG, Kronenberger T, Espinoza GZ, Oliveira PR, Honorio KM. Advances with support vector machines for novel drug discovery. Expert Opinion on Drug Discovery. 2019; 14(1):23-33.

[6] Shoichet BK. Virtual screening of chemical libraries. Nature. 2004; 432:862-5.

[7] Afolabi LT, Saeed F, Hashim H, Petinrin OO. Ensemble learning method for the prediction of new bioactive molecules. PloS One. 2018; 13(1):1-14.

[8] Neves BJ, Braga RC, Melo-Filho CC, Moreira-Filho JT, Muratov EN, Andrade CH. QSAR-based virtual screening: advances and applications in drug discovery. Frontiers in Pharmacology. 2018; 9:1-7.

[9] Huang HJ, Yu HW, Chen CY, Hsu CH, Chen HY, Lee KJ, et al. Current developments of computer-aided drug design. Journal of the Taiwan Institute of Chemical Engineers. 2010; 41(6):623-35.

[10] Liu X, Xu Y, Li S, Wang Y, Peng J, Luo C, et al. In silicotarget fishing: addressing a “Big Data” problem by ligand-based similarity rankings with data fusion. Journal of Cheminformatics. 2014; 6:1-14.

[11] Lavecchia A. Machine-learning approaches in drug discovery: methods and applications. Drug Discovery Today. 2015; 20(3):318-31.

[12] Ahmed L, Edlund A, Laure E, Spjuth O. Using iterative MapReduce for parallel virtual screening. In 5th international conference on cloud computing technology and science 2013 (pp. 27-32). IEEE.

[13] Ballester PJ, Mitchell JB. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. Bioinformatics. 2010; 26(9):1169-75.

[14] Thai KM, Nguyen TQ, Ngo TD, Tran TD, Huynh TN. A support vector machine classification model for benzo [c] phenathridine analogues with topoisomerase-I inhibitory activity. Molecules. 2012; 17(4):4560-82.

[15] Lionta E, Spyrou G, K Vassilatis D, Cournia Z. Structure-based virtual screening for drug discovery: principles, applications and recent advances. Current Topics in Medicinal Chemistry. 2014; 14(16):1923-38.

[16] https://en.wikipedia.org/wiki/Virtual_screening. Accessed 21 November 2019.

[17] Banerjee P, Preissner R. BitterSweetForest: a random forest based binary classifier to predict bitterness and sweetness of chemical compounds. Frontiers in Chemistry. 2018; 6:1-10.

[18] Xiong Y, Qiao Y, Kihara D, Zhang HY, Zhu X, Wei DQ. Survey of machine learning techniques for prediction of the isoform specificity of cytochrome

P450 substrates. Current Drug Metabolism. 2019; 20(3):229-35.

[19] Ponzoni I, Sebastián-Pérez V, Martínez MJ, Roca C, De la Cruz Pérez C, Cravero F, et al. QSAR classification models for predicting the activity of inhibitors of beta-secretase (BACE1) associated with alzheimer's disease. Scientific Reports. 2019; 9:1-13.

[20] Muegge I, Mukherjee P. An overview of molecular fingerprint similarity search in virtual screening. Expert Opinion on Drug Discovery. 2016; 11(2):137-48.

[21] Korkmaz S, Zararsiz G, Goksuluk D. Drug/nondrug classification using support vector machines with various feature selection strategies. Computer Methods and Programs in Biomedicine. 2014; 117(2):51-60.

[22] Li Y, Kong Y, Zhang M, Yan A, Liu Z. Using support vector machine (SVM) for classification of selectivity of H1N1 neuraminidase inhibitors. Molecular Informatics. 2016; 35(3-4):116-24.

[23] Kumar A, Verma DK, Purohit R. Conceptual modelling of telapathic network. Metabolomics. 2012; 2(5).

[24] Ani R, Manohar R, Anil G, Deepa OS. Virtual screening of drug likeness using tree based ensemble classifier. Biomedical and Pharmacology Journal. 2018; 11(3):1513-9.

[25] Yosipof A, Guedes RC, García-Sosa AT. Data mining and machine learning models for predicting drug likeness and their disease or organ category. Frontiers in Chemistry. 2018; 6:1-11.

[26] Bahi M, Batouche M. Deep semi-supervised learning for virtual screening based on big data analytics. In international conference on big data, cloud and applications 2018 (pp. 173-84). Springer, Cham.

[27] Bahi M, Batouche M. Drug-target interaction prediction in drug repositioning based on deep semi-supervised learning. In international conference on computational intelligence and its applications 2018 (pp. 302-13). Springer, Cham.

[28] Khan A, Kaushik AC, Ali SS, Ahmad N, Wei DQ. Deep-learning-based target screening and similarity search for the predicted inhibitors of the pathways in Parkinson's disease. RSC Advances. 2019; 9:10326-39.

[29] Ramsundar B, Kearnes S, Riley P, Webster D, Konerding D, Pande V. Massively multitask networks for drug discovery. arXiv preprint arXiv:1502.02072. 2015.

[30] Inglese P, McKenzie JS, Mroz A, Kinross J, Veselkov K, Holmes E, et al. Deep learning and 3D-DESI imaging reveal the hidden metabolic heterogeneity of cancer. Chemical Science. 2017; 8:3500-11.

[31] Constantine RM, Batouche M. Drug discovery for breast cancer based on big data analytics techniques. In international conference on information & communication technology and accessibility 2015 (pp. 1-6). IEEE.

[32] Sid K, Batouche M. Ensemble learning for large scale virtual screening on apache spark. In IFIP international conference on computational intelligence and its applications 2018 (pp. 244-56). Springer, Cham.

[33] Byvatov E, Fechner U, Sadowski J, Schneider G. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. Journal of Chemical Information and Computer Sciences. 2003; 43(6):1882-9.

[34] Zernov VV, Balakin KV, Ivaschenko AA, Savchuk NP, Pletnev IV. Drug discovery using support vector machines, the case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. Journal of Chemical Information and Computer Sciences. 2003; 43(6):2048-56.

[35] Warmuth MK, Liao J, Rätsch G, Mathieson M, Putta S, Lemmen C. Active learning with support vector machines in the drug discovery process. Journal of Chemical Information and Computer Sciences. 2003; 43(2):667-73.

[36] Jorissen RN, Gilson MK. Virtual screening of molecular databases using a support vector machine. Journal of Chemical Information and Modeling. 2005; 45(3):549-61.

[37] Podolyan Y, Walters MA, Karypis G. Assessing synthetic accessibility of chemical compounds using machine learning methods. Journal of Chemical Information and Modeling. 2010; 50(6):979-91.

[38] Cheng T, Li Q, Wang Y, Bryant SH. Binary classification of aqueous solubility using support vector machines with reduction and recombination feature selection. Journal of Chemical Information and Modeling. 2011; 51(2):229-36.

[39] Camps-Valls G, Bruzzone L. Kernel-based methods for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing. 2005; 43(6):1351-62.

[40] Sato T, Honma T, Yokoyama S. Combining machine learning and pharmacophore-based interaction fingerprint for in silico screening. Journal of Chemical Information and Modeling. 2010; 50(1):170-85.

[41] Von Korff M, Sander T. Toxicity-indicating structural patterns. Journal of Chemical Information and Modeling. 2006; 46(2):536-44.

[42] Abdo A, Chen B, Mueller C, Salim N, Willett P. Ligand-based virtual screening using bayesian networks. Journal of Chemical Information and Modeling. 2010; 50(6):1012-20.

[43] Gleeson MP, Waters NJ, Paine SW, Davis AM. In silico human and rat V ss quantitative structure−activity relationship models. Journal of Medicinal Chemistry. 2006; 49(6):1953-63.

[44] Ai S, Bai Y, Liu X. Virtual screening for COX-2 inhibitors with random forest algorithm and feature selection. In proceedings of the international conference on bioinformatics research and applications 2017 (pp. 9-14).

[45] Lee K, Lee M, Kim D. Utilizing random forest QSAR models with optimized parameters for target identification and its application to target-fishing server. BMC Bioinformatics. 2017; 18(16):75-86.

[46] Kauffman GW, Jurs PC. QSAR and k-nearest neighbor classification analysis of selective cyclooxygenase-2 inhibitors using topologically-based numerical descriptors. Journal of Chemical Information and Computer Sciences. 2001; 41(6):1553-60.

[47] Itskowitz P, Tropsha A. K-nearest neighbors QSAR modeling as a variational problem: theory and applications. Journal of Chemical Information and Modeling. 2005; 45(3):777-85.

[48] Patel JL, Patel LD. Artificial neural networks and their applications in pharmaceutical research. Pharmabuzz. 2007; 2:8-17.

[49] Soyguder S. Intelligent control based on wavelet decomposition and neural network for predicting of human trajectories with a novel vision-based robotic. Expert Systems with Applications. 2011; 38(11):13994-4000.

[50] Behrmann J, Etmann C, Boskamp T, Casadonte R, Kriegsmann J, Maaβ P. Deep learning for tumor classification in imaging mass spectrometry. Bioinformatics. 2018; 34(7):1215-23.

[51] Pérez-Sianes J, Pérez-Sánchez H, Díaz F. Virtual screening meets deep learning. Current Computer-Aided Drug Design. 2019; 15(1):6-28.

[52] Koutsoukas A, Lowe R, KalantarMotamedi Y, Mussa HY, Klaffke W, Mitchell JB, et al. In silico target predictions: defining a benchmarking data set and comparison of performance of the multiclass Naïve Bayes and Parzen-Rosenblatt window. Journal of Chemical Information and Modeling. 2013; 53(8):1957-66.

[53] https://pubchem.ncbi.nlm.nih.gov. Accessed 14 November 2019.

[54] https://spark.apache.org/. Accessed 14 November 2019.

[55] Fathima AJ, Murugaboopathi G. A novel customized big data analytics framework for drug discovery. Journal of Cyber Security and Mobility. 2018; 7(1):145-60.

[56] García-Sosa AT, Oja M, Hetényi C, Maran U. DrugLogit: logistic discrimination between drugs and nondrugs including disease-specificity by assigning probabilities based on molecular properties. Journal of Chemical Information and Modeling. 2012; 52(8):2165-80.

[57] Khaldy MA, Kambhampati C. Resampling imbalanced class and the effectiveness of feature selection methods for heart failure dataset. International Robotics & Automation Journal. 2018; 4(1):37-45.

[58] Jahan S, Shatabda S, Farid DM. Active learning for mining big data. In international conference of computer and information technology (ICCIT) 2018 (pp. 1-6). IEEE.

Sahar k. Hussin received the B.Sc., M.Sc. degrees in Systems and Computers Engineering Department from Al-Azhar University in 2010 and 2016, respectively. She is currently a Ph.D. student and Assistant Lecturer at Communication and Computers Engineering Department Alshrouck academy, Cairo, Egypt. Her current research interests are Bioinformatics include Machine Learning, Data Mining and Big Data Analytics.
Email: saharkamal26@yahoo.com

Yasser M.K. Omar received a Ph.D. degree in Biomedical Engineering from Cairo University, Cairo, Egypt. He has been an Assistant Professor in the Department of Computer Science, Faculty of Computing and Information Technology, Arab Academy for Science Technology & Maritime Transport (AASTMT). His research interests are Bioinformatics, Medical Imaging, Data Visualization, Machine Learning, and Computing Algorithms.
Email: dr_yaser_omar@yahoo.com

Salah M. Abdel-Mageid received his M.S. and Ph.D. in Systems and Computers Engineering from Al-Azhar University in 2002 and 2005, respectively. He is a professor in Computer Engineering at College of Computer Science and Engineering, Taibah University, Saudi Arabia. From 2017 until 2019, he was the head of Systems and Computers Engineering Department at Al-Azhar University. He performed his postdoctoral research in 2007 and 2008 in the Computer Science and Engineering Department, School of Engineering, the Southern Methodist University in Dallas, TX, USA. He was a software development manager of TEMPO (Tool for Extensive Management and Performance Optimization) project in Cairo University and Vodafone Egypt as an industrial partner in 2014 and 2015. His research interests include Mobile Computing, Cellular Networks, Sensor Networks, Cognitive Radio Networks, Vehicular Ad-hoc Networks, Big Data and Data Analysis, Internet Services and Applications.
Email: sabdelmageid@taibahu.edu.sa

Mahmoud I. Marie received his B.Sc, M.Sc and PhD in Electronic and Communication engineering from Cairo University on 1972, 1981, 1985, respectively. Currently he is a professor of communications at Computer and System Engineering Department AlAzhar University, Cairo, Egypt. His fields of interest include Digital Communication, Computer Networks and Protocols Development, Face Recognition and Big Data Analytics.
Email: marie@azhar.edu.eg