RESEARCH ARTICLE

Role of Machine Learning in Computational Toxicity Prediction

Ankur Omer*
Government College Silodi, Katni, MPHED, Madhya Pradesh

Corresponding Author*

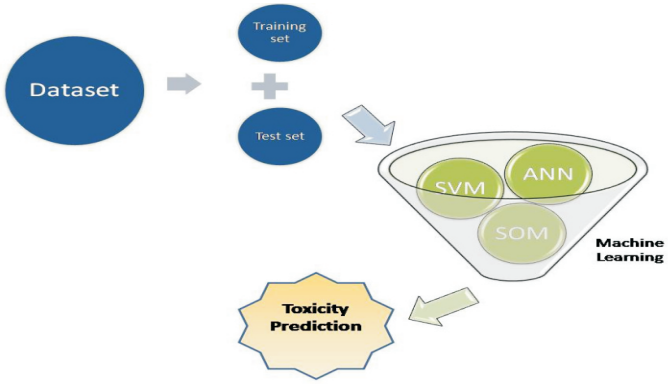
Ankur Omer

Email

ankuromer@gmail.com

MS No. 010123

Submitted: *18-10-2022,* ***Accepted:*** *21-12-2022,* ***Published:*** *14-03-2023*



KEYWORDS :. Machine learning, Predictive toxicology, SVM, ANN, SOM, Toxicity prediction, *in silico*

SUMMARY

It is necessary to do study on how to predict toxicity since actually conducting toxicity testing may be both time-consuming and expensive. Bioinformatics tools can save time and money. Ever since its start, it has consistently delivered results. The process of analysing and classifying data is an essential component of bioinformatics. Because of their speed and low cost, *in silico* approaches have gained popularity in recent years for evaluating the kinetic and toxic behaviour of drugs. Machine learning is a potent tool for exploring *in vitro* and *in vivo* data for previously undiscovered complicated combinatorial associations. It has found useful applications in areas as varied as predicting pharmacodynamic characteristics and protein activities, identifying spam, locating oil spills, and recognising human voices. Algorithms such as Support Vector Machines (SVMs), Artificial Neural Networks (ANNs), and Self Organizing Maps (SOMs), as well as the difficulties they present, the potential ties they may one day forge, and the web-based toxicity prediction tools have been discussed in this article.

INTRODUCTION

Principles of Predictive Toxicology

When chemicals used in industrial processes leak into the surrounding environment, they may cause harm. There is a need to determine the relative toxicity of each of these substances because of their widespread use. Human, mouse, and calf receptors and other biological materials have been used in a variety of experimental approaches for screening the activity of drugs (*in vitro* and *in vivo* tests). But, current experimental approaches may be expensive, time-consuming, and may even yield harmful by products. As a result, there has been a lot of focus on creating computational algorithms as an alternate tool for predicting chemical characteristics. Since we know that a chemical's qualities stem from its molecular structure, it stands to reason that there are connections between those properties which can be used for predicting toxicity[1].

The Learning Algorithms: Machine learning

The study of algorithms that can learn, improve, or change their performance on a given job based on previous runs is known as machine learning [2]. Machine learning, like many subfields in AI, has grown rather specialised. The purpose of machine learning is, in part, to bridge the gap between the rigidity and inflexibility of computers and the malleability and fortitude of human thought. Learning all the reasons why certain substances are toxic while others are non-toxic may be of tremendous relevance and scientific use for predicting toxicity.

(a) Types of Machine learning Algorithms

Depending on the intended outcome, machine learning may take many different forms. Some of the more common kinds are (Figure 1.)

(i) Supervised Learning

It is utilised in classification and regression systems on a fairly regular basis. The objective here is to teach a computer how to use a human-made categorization system to maximise accuracy while minimising input noise. Classification learning works well for issues when it is both simple and helpful to produce a classification. It is the primary method for training neural networks and decision trees[3-8].

(ii) Unsupervised Learning

Unsupervised learning is a method of machine learning in which models are not regulated by utilising training datasets. Instead, the models themselves uncover the previously unrecognized patterns and insights contained within the data [9-11].

There are various kinds of algorithms:

When it comes to the field of supervised learning, which focuses heavily on categorization. There are many types of algorithms, some of the most significant of which are [12]:

- Linear Classifiers
- Support Vector Machine
- Quadratic Classifiers
- K-Means Clustering
- Boosting
- Decision Tree
- Neural networks
- Bayesian Networks

In this articles, we will examine the performance of three major machine learning approaches in relation to the actual prediction of toxicological datasets [12].

a) ANN (Artificial Neural Network)

b) SVM (Support Vector Machine)

c) SOM (Self Organising Maps)

(a) Artificial Neural Network (ANN)

A neural network is a data simulation tool capable of capturing and illustrating complicated input/output interactions. Neural network technology arose from the ambition to create a "smart" system capable of performing tasks akin to those of the

Academic editor- Dr. Sandeep Singh, PhD, Kanpur, (208021) Uttar Pradesh.

human brain [13]. Human brain and neural network (ANN) system have many things in common:

- Gaining knowledge through learning
- Acquire knowledge and stores within neural connections called as synaptic weights [14,15].

The "Neuron" is the basic computational unit of a neural network. A neuron gathers data from multiple sources, processes it using a non-linear function, and then transmits the outcome. Network designers may continue to enhance their systems by learning from the biological brain [16,17]. A group of neurons is called a neural network. Neurons, their layering and grouping, and the connections between them, as well as other mathematical functions like summation and transfer, are all components of a fully working neural network. Although some networks have a single layer of neurons and others have several, the basic vocabulary needed to explain them is the same [17].

(i) ANN Toxicity Prediction

A dataset is compiled by mining many databases for information on harmful compounds used in manufacturing, pharmaceuticals, and other fields. Subsequently, the compounds are randomly split into a training set and a test set, with the former comprising 70% and the latter 30%.

The next step is to compute a large number of descriptors, which may be done using a variety of applications, including CODESSA, Hyperchem, and PaDEL-descriptor [18], Hyperchem [19], PaDEL-descriptor [20] Table 1.

The choice of descriptors needs to be governed by the factors that are most effective in describing the molecules. It's vital to keep in mind the possibility that certain descriptors don't provide information but make it more difficult to analyse the result by raising the noise level, while choosing an insufficient number of variables would reduce the effectiveness of the model [21].

(ii) Training Neural Networks

The optimal method of training or learning involves collecting a large number of instances with the objective of including all potential variations of the issues within the training set [22]. In order for the system to get used to the natural variability and probable noise in the actual data, it is often necessary to introduce some noise or other

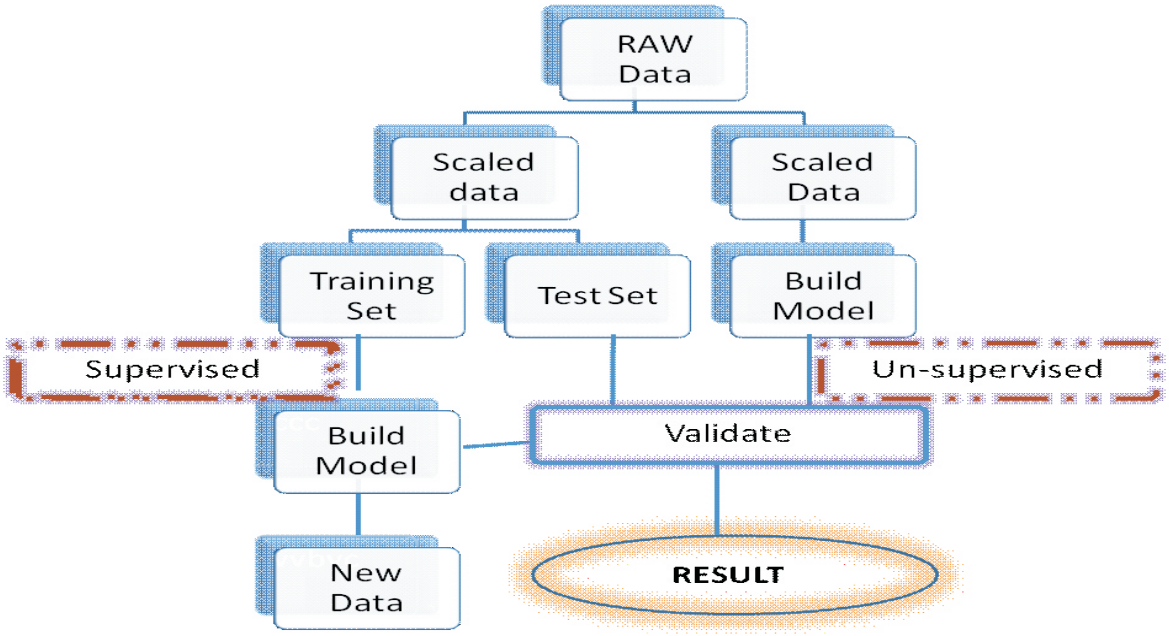


FIGURE 1. How data can be used in training and validating the model using supervised and unsupervised method

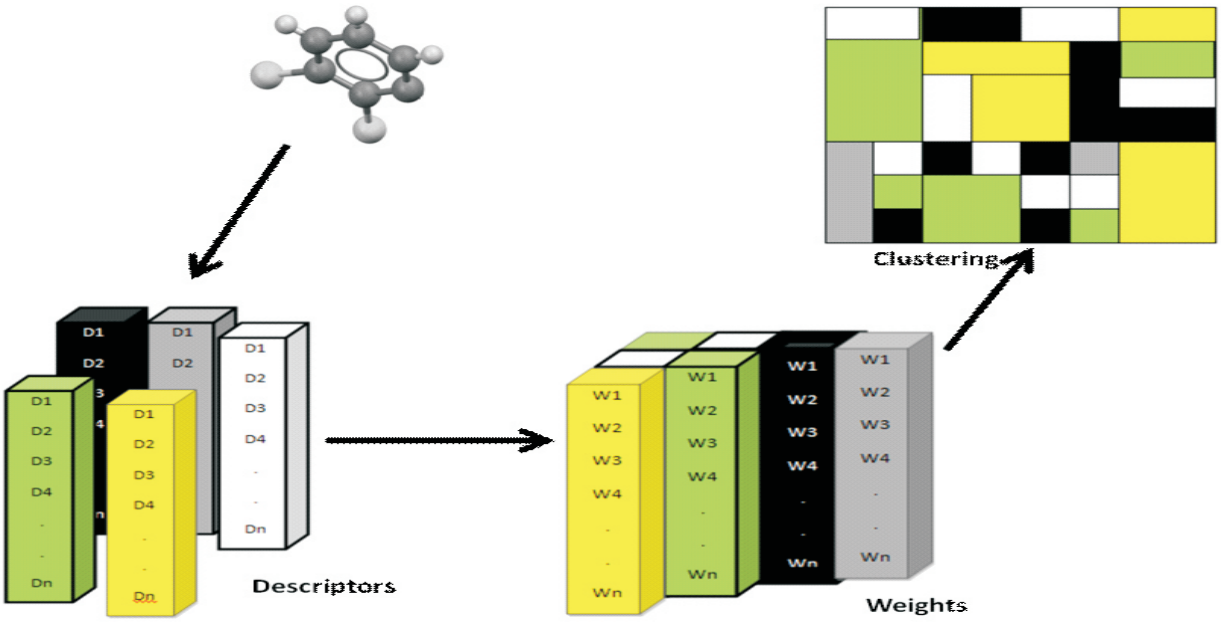


FIGURE 2. Representation of the process of Self Organising Maps.

TABLE 1. List of some *in silico* tools for toxicity prediction

Tools	Availability
ToxiM	http://metagenomics.iiserb.ac.in/ToxiM/
TEST	https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test
TOPKAT	www.omictools.com/toxicity-prediction-by-komputer-assisted-technology-tool
Toxtree	https://toxtree.sourceforge.net/
VEGA Caesar	www.caesar-project.eu
ADMET Predictor	https://www.simulations-plus.com/software/admetpredictor/toxicity/
HazardExpert	https://www.compudrug.com/hazardexpertpro
MultiCASE/MC4PC	www.multicase.com/case-ultra-models
Derek Nexus	www.lhasalimited.org
SVM Light	http://svmlight.joachims.org/
SOM_PAK	http://www.cis.hut.fi/research/som_lvq_pak.shtml
OpenTox	www.opentox.net/
SOM Toolbox	http://www.cis.hut.fi/projects/somtoolbox/
Protox-II	https://tox-new.charite.de/protox_II/
eTOX	www.etoxproject.eu/
ToxBoxes	www.acdlabs.com/
Mutagenicity (Ames test)	www.vegahub.eu
Toxit	https://www.toxit.it/en/services/software/topkat
KnowTox	https://volkamerlab.org/projects/knowtox/
DeepTox	http://www.bioinf.jku.at/research/DeepTox/

randomness inside the training set. The input and output layers are connected by a network of neurons. The higher the weight of a neuron the more effective it is at receiving signals from other layers. This process of adjusting weights is called training or learning [23]. As complexity rises, the number of instances should likewise increase to include all conceivable issue features in the training set [22].

(iii) Choosing the number of neurons

The more layers of hidden neurons a network has, the more efficient it will be. However, the network's performance degrades in terms of its ability to generalise to new data. On the other hand, if there aren't enough hidden layers, the network won't be capable of learning the various combinational connections between the data, and it won't be able to minimise the error to a satisfactory level [24].

(b) Support Vector Machine

It's a brand-new algorithm developed by experts in the field of machine learning. Over the last several years, it has risen to prominence as a central issue in the field of machine learning. In addition to a solid theoretical basis, it has shown superior practical performance and competitiveness when compared to established approaches like neural networks and decision trees. It's built in a way that mitigates the negative impact that higher input dimensions have on generalisation performance [1,25,26].

The SVM classifier has been designed to be able to take in all types of data and provide the best generalisation results on unknown data. It is able to find a unique hyperplane with maximum margin for separating the data in two classes measured along a line perpendicular to the hyper plane. There are a number of existing applications of SVM and they are dramatically increasing for example handwriting recognition, text categorization, object recognition, classification and database marketing [27-29].

To find the line with the highest margin, the gap between the two classes in the training set should be maximized. Keep in mind that maximising the distance is done so in order to reduce the structural risk, as described by Vapnik. SVM apart from performing linear classification can also be used when a linear classifier is not able to separate two classes. In such a scenario, the object's features are transformed into a feature space by use of a set of non-linear functions known as feature functions. High-dimensional feature space allows a linear classifier to distinguish between the two classes

(i) Training and Learning of SVM

Training and learning follow the ANN training technique. Two factors must be addressed to succeed. First, the training data set should be impartial and follow the same probability distribution as the test set. If the learning machine's capacity is inadequate, it cannot learn all conceivable scenarios. If the capacity is too great, too many functions congruent with the training examples will perfectly memorise those instances and be ineffective at discriminating unknown data [31,32].

(ii) SVM approach to toxicity Prediction

Because SVM is a supervised learning and classification approach, it requires training set to determine in advance how to group similar datasets. Informed by an exhaustive review of the relevant literature, a dataset of known hazardous compounds is compiled. A training set and a test set are generated from this dataset. Toxic compounds (positively labelled as belonging to the same functional class) and non-toxic compounds should be distinguishable after the SVM has been trained using the provided training data. Toxic chemicals, drugs, and compounds may have their pharmacophoric information extracted using a support vector machine [33,34].

(c) Self Organising Maps (SOM)

Self-Organizing Map (SOM), Self-Organizing Feature Map (SOFM), and Kohonen networks are used significantly differently than other networks since they are mainly intended for unsupervised learning [35,36]. As the SOM includes input that is not associated with any output tag, it is fascinating to consider how and from what a network might learn. The solution comes in the fact that it seeks to discover the data's structure [12].

The distance between an input pattern and each neuron in a SOM is computed during training or use. When the distance between two neurons is small enough, the neuron closest to the input pattern is chosen as the winner and given the task of representing the input pattern. The fundamental, infinitely repeating Kohonen SOM method simply cycles through a set number of epochs, whereby each epoch performs all training cases simultaneously using the following technique [37].

- Choose the winning node or neuron (the one with closest centre to the input case)
- The winning neuron's weights are changed such that they more closely resemble the input scenario [12].

To compute the weighted sum and ensure that the adjustments get more exact as the epochs pass, the method makes use of a time-decaying function termed learning rate (Figure 2) [12].

In the Kohonen SOM method, the neuron modification affects not just the victorious neuron but also the rest of the neighbours at the moment. By incorporating the idea of a neighbourhood into an algorithm, we are able to achieve the topological ordering property.

By stimulating groups of neurons that share topological features, the network establishes a rough topological ordering. Eventually, the fine-tuning of the individual neurons is achieved when the neighbourhood shrinks and the learning rate slows, allowing for more nuanced distinctions to be made within the map region [12,36].

SOMs' greatest strength is their intuitive nature. SOMs are quite straightforward; points are considered to be similar if they are near together or if there is a grey line joining them, and points are considered to be distinct if there is a black ravine separating them. This is a beautifully tuned machine. To determine the quality of a map (i.e., how excellent a map is) and analyse the similarities between objects, they perform extremely well when classifying data and then evaluating it for its own quality [38,39].

FUTURE DIRECTION AND CONCLUSION

The process of developing intelligent settings is fraught with difficulties. Systems that help things along are literally everywhere. These systems are always linked to their surrounding environment, and as a result, they both affect and are affected by that environment. Therefore, the development of multi-levelled, learnable, and adaptable systems is essential for ambient intelligence. Drug and environmental toxicity prediction is difficult. In vivo toxicity research in rats and other animals are costly and seldom reveal mechanism of action. Pharmaceutical companies utilise in vitro or cell-based tests to screen drugs and prioritise them for effectiveness and safety assessment. Environmental chemical safety assessment currently uses these in vitro screening methods. Machine learning is greatly improving toxicity prediction. Though supervised machine learning is better for classification tasks than unsupervised, this technique performed similarly well or better in certain circumstances. With technology developments lowering computing costs and new data sources, additional toxicity prediction models are projected in the future. The fundamental challenge for machine learning in predictive toxicology is data quality and quantity. Despite agreements with pharmaceutical firms and publicly available databases, certain gene or protein targets and toxicological end points cannot be accurately determined owing to data gaps. These data gaps must be filled to build a full computational model or a group of computer models covering all of human toxicity to address predictive toxicology's unbalanced data, gather and share compound-specific unfavourable experimental outcomes.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

REFERENCES

1. M. W. H. Wang, J. M. Goodman, T. E. H. Allen (2020). Machine Learning in Predictive Toxicology: Recent Applications and Future Directions for Classification Models. Chemical Research in Toxicology, 34(2), 217-39.
2. J. A. Nichols, H. W. Herbert Chan, M. A. B. Baker(2018). Machine Learning: Applications of Artificial Intelligence to Imaging and Diagnosis. Biophysical Reviews, 11(1), 111-118.
3. X. Wang and D. M. Wilkes (2020). Supervised Learning for Data Classification Based Object Recognition. 179-94.
4. M. I. Jordan and T. M. Mitchell (2015). Machine Learning: Trends, Perspectives, and Prospects. Science, 349(6245), 255-260.
5. D. R. Schrider and A. D. Kern (2018). Supervised Machine Learning for Population Genetics: A New Paradigm. Trends in Genetics, 34(4), 301-312.
6. M. W. Libbrecht and W. S. (2015). Noble Machine Learning Applications in Genetics and Genomics. Nature Reviews Genetics, 16(6), 321-332.
7. M. Längkvist, L. Karlsson and A. Loutfi (2014). A Review of Unsupervised Feature Learning and Deep Learning For Time-Series Modeling. Pattern Recognition Letters, 42, 11-24.
8. Q. V. Le (2013). Building High-Level Features Using Large Scale Unsupervised Learning. 85, 95-108. 11.
9. R. Raina, A. Madhavan, A. Y. Ng (2009). Large-Scale Deep Unsupervised Learning Using Graphics Processors. 1-8.
10. C. M. Signorell (2018). Can Computers Become Conscious and Overcome Humans? Frontiers in Robotics and AI, 5.
11. A. Apparaju and O. Arandjelović (2022). Towards New Generation, Biologically Plausible Deep Neural Network Learning. Sci, 4(4), 46.
12. A. R. Katritzky, S. Perumal, R. Petrukhin and E. Kleinpeter (2010). CODESSA-Based Theoretical QSPR Model for Hydantoin HPLC-RT Lipophilicities. Journal of Chemical Information and Computer Sciences. 41(3), 569-74.
13. C. W. Yap (2011). Padel-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. Journal of Computational Chemistry, 32(7), 1466-1474.
14. P. Mazzatorta, E. Benfenati, C. D. Neagu and G. Gini. Tuning Neural and Fuzzy-Neural Networks for Toxicity Modeling. Journal of Chemical Information and Computer Sciences, 43(2), 513-518.
15. D Decoste and B. Schölkopf (2002). Machine Learning. 46(1/3),161-190.
16. Z. R. Yang (2004). Biological Applications of Support Vector Machines. Briefings in Bioinformatics, 5(4), 328-338.
17. O. Ivanciuc (2007). Applications of Support Vector Machines in Chemistry. 2007:291-400.
18. J. B. O. Mitchell. Machine learning methods in chemoinformatics. WIREs Computational Molecular Science, 4(5), 468-481.
19. I. H. Sarker (2004). Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science, 2(3).
20. A. Omer, P. Singh, N. K. Yadav and R. K. Singh (2014). An Overview of Data Mining Algorithms in Drug Induced Toxicity Prediction. Medicinal Chemistry, 14(4), 345-54.
21. F. Baçço, V. Lobo and M. Painho (2005). The self-organizing map, the Geo-SOM, and relevant variants for geosciences. Computers & Geoscience, 31(2):155-63.

Citation: *Ankur Omer (2023),* Role of Machine Learning in Computational Toxicity Prediction. *Journal of Recent Advances in Applied Sciences*, 9(1), 1-4.