Research Article
Journal of Intelligent Systems: Theory and Applications 7(2) (2024) 116-128
DOI: 10.38016/jista.1468153

A Hybrid Metaheuristic based Feature Selection Framework for In-silico Mutagenicity Prediction

Özlem Yılmaz[1], Enis Gümüştaş[2], Ayça Çakmak Pehlivanlı[3*]

[1] Faculty of Science and Letters, Mathematics Department, Mimar Sinan Fine Arts University, İstanbul, Türkiye

[2, 3] Faculty of Science and Letters, Statistics Department, Mimar Sinan Fine Arts University, İstanbul, Türkiye

ozlem.yilmaz@msgsu.edu.tr, 20203107001@ogr.msgsu.edu.tr, ayca.pehlivanli@msgsu.edu.tr

Abstract

Mutagenicity is both a toxic risk to humans and an indicator of carcinogenicity. Hence, estimating mutagenicity in the early stages of drug design is crucial to minimize last-stage failures and withdrawals in drug discovery. Recently, in-silico methods have started to play critical and essential roles in the drug development process because they are low cost and low effort procedures. This study aims to predict mutagenicity of chemicals using in-silico methods. To achieve this goal, a two-phased flexible framework was proposed: 1) searching the effective and representative descriptors subset with Butterfly Optimization Algorithm (BOA) and Particle Swarm Optimization and 2) predicting mutagenicity of chemicals by the selected descriptor using gradient boosted tree-based ensemble methods. The study used two datasets: one including 8167 compounds for descriptor selection and modelling, and another containing 716 external compounds to validate the efficacy of our models. The datasets comprise 162 descriptors calculated using PaDEL. The results of both the cross-validation and the external data showed that descriptors reduced by nearly one-third by BOA (51 descriptors) yielded similar or slightly better predictive results than results obtained with the entire data set. The accuracy range attained by the proposed approach using BOA is approximately 91.9% to 97.91% for the external set and 83.35% to 86.47% for the test set. This research contributes that using optimization techniques for improving early drug design and minimizing risks in drug discovery can be considered as a valuable insights and advances in the field of drug toxicity prediction, based on the findings.

Keywords: Machine Learning, Feature Selection, Metaheuristics, Gradient Boosting Algorithms, Mutagenicity Prediction, In-Silico Modelling

In-silico Mutajenite Tahmini için Hibrit Metasezgisel Tabanlı Özellik Seçimi Çerçevesi

Öz

Mutajenite hem insanlar için toksik bir risk hem de kanserojenitenin bir göstergesidir. Bu nedenle, ilaç tasarımının erken aşamalarında mutajenitenin tahmin edilmesi, ilaç keşfinde son aşama başarısızlıklarını ve geri çekilmeleri en aza indirmek için çok önemlidir. Son zamanlarda, in-silico yöntemler, düşük maliyetli ve az çaba gerektiren prosedürler olmaları nedeniyle ilaç geliştirme sürecinde kritik ve önemli roller oynamaya başlamıştır. Bu çalışma, in-silico yöntemler kullanarak kimyasalların mutajenitesini tahmin etmeyi amaçlamaktadır. Bu amaca ulaşmak için iki aşamalı esnek bir çerçeve önerilmiştir: 1) Kelebek Optimizasyon Algoritması (BOA) ve Parçacık Sürü Optimizasyonu ile etkili ve temsili değişken alt kümesinin aranması ve 2) gradyan destekli ağaç tabanlı topluluk yöntemleri kullanılarak seçilen değişkenlere göre kimyasalların mutajenitesinin tahmin edilmesi. Çalışmada iki veri kümesi kullanılmıştır: biri değişken seçimi ve modelleme için 8167 bileşik, diğeri ise modellerimizin etkinliğini doğrulamak için 716 harici bileşik içermektedir. Veri kümeleri PaDEL kullanılarak hesaplanan 162 değişkeni içermektedir. Hem çapraz doğrulama hem de harici verilerin sonuçları, BOA ile neredeyse üçte bir oranında azaltılan değişkenlerin (51 adet), tüm veri setiyle elde edilen sonuçlara benzer veya biraz daha iyi tahmin sonuçları verdiğini göstermiştir. BOA kullanılarak önerilen yaklaşımla elde edilen doğruluk aralığı harici set için yaklaşık %91,9 ila %97,91 ve test seti için %83,35 ila %86,47'dir. Bu araştırma, bulgulara dayanarak, erken ilaç tasarımını iyileştirmek ve ilaç keşfindeki riskleri en aza indirmek için optimizasyon tekniklerinin kullanılmasının, ilaç toksisitesi tahmini alanında değerli bir içgörü ve ilerleme olarak kabul edilebileceğine katkıda bulunmaktadır.

Anahtar Kelimeler: Makine Öğrenmesi, Özellik Seçimi, Metasezgisel, Gradyan Boosting Algoritmaları, Mutajenite Tahmini, In-Silico Modelleme

* Corresponding Author
E-Mail: ayca.pehlivanli@msgsu.edu.tr

Recieved : 14 Apr 2024
Revision : 16 Aug 2024
Accepted : 20 Aug 2024

1. Introduction

Mutagenicity can be defined as the capacity of a compound to cause permanent mutations in the DNA sequence (Bakhtyari et al. 2013). It could lead to a toxic risk to humans. Moreover, it is the indicator of carcinogenicity which means healthy cell transforms themselves into cancer cells. Assessment of mutagenicity at the early stages of the drug approval process is crucial to swift eliminate such compounds from the drug development pipeline (Raghavan 2005). Among the toxicity tests, the in-vitro Ames test has become a standard for assessing mutagenicity (Zhang 2017). However, these in-vitro experiments are highly expensive, laborious, and time-consuming. On the other hand, both in-vivo experiments and in-vitro experiments have begun to give up their places to statistical and computational methods and tools developed in the computer environment without requiring laboratory experiments. Therefore, to prevent drug failure and withdrawal due to late-stage mutagenicity, it is necessary to predict mutagenicity by developing in-silico methods. Computational methods used for in-silico approaches can be grouped basically as expert rule-based systems, also referred to as structural alerts, and statistics-based models, known as quantitative structure-activity relationship tools (Bakhtyari et al. 2013; Honma 2019; Hansch 1980). Difficulties in explainability and interpretability, which are the main drawbacks of computational approaches, are almost non-existent in expert systems. Therefore, expert systems are widely used because they provide comprehensive outputs that can be understood, questioned and judged by the user. Despite this transparency, their prediction success is lower than that of statistical approaches (Wichard 2017). It is worth noting that expert systems for in-silico research are available both commercially and open access (Honma 2019; Çakmak Pehlivanlı and Çakmak 2022).

A well-designed in silico approach can yield several benefits, including the ability to plan studies with fewer animals, identify the concentration that will be used in advance, save time and money, and guide whether the information obtained about the molecule should proceed to laboratory experiments (Toropov et al. 2014).

Metaheuristic optimization algorithms are ideally suited for efficiently exploring the complex and high-dimensional feature spaces encountered in feature selection problems due to their stochastic, adaptable, and global search characteristics.

To the best of our knowledge, no studies address the estimation of drug toxicity, particularly mutagenicity, except for the limited number of studies in which metaheuristic optimization algorithms have been applied to drug discovery (Houssein et al. 2020; Algamal et al. 2020; Subaş and Çakmak Pehlivanlı 2020). The main contributions and scope of this paper are summarized as follows in order to fill this gap;

- We propose a flexible approach that hybridized metaheuristic algorithms with several machine learning algorithms to select descriptors and compare the classification models that promise the best prediction results of mutagenicity,
- We show that metaheuristic algorithms and machine learning algorithms can work together in in-silico studies such as drug toxicity prediction,
- We conclude whether metaheuristic approaches are suitable for searching the descriptor space and enhancing mutagenicity classification based on the chosen descriptors.

It should be noted that our motivation is not only the success of the optimization part but also mostly obtaining higher accuracy with fewer descriptors. To address these aforementioned aims, we introduce a flexible approach by hybridizing metaheuristic and machine learning algorithms on mutagenicity.

The rest of this paper is organized as follows. After introducing the related work of this paper in Section 2, dataset and the proposed model are described in Section 3. Experimental results, discussions and conclusions are presented in Sections 4, 5 and 6 respectively.

2. Related works

Early studies on systems based on rules and expert knowledge gained speed, especially at the end of the ’90s. The relationships between chemical structures and observed toxic effects and outcomes were examined, and various software was presented comparatively (Greene et al. 1999). This study was followed by several in-silico studies (White et al. 2003; Cariello et al. 2002).

In the early 2000s, modelling based on statistical learning algorithms was commonly used to predict mutagenicity in in-silico studies. Zheng et al. developed a mutagenic probability model with support vector machines (SVM) for the mutagenicity prediction and achieved better performance than the TOPKAT, a tool based on rules and expert knowledge (Zheng 2006). Liao et al. applied a combination of recursive partitioning (RP) and SVM on different data sets to predict mutagenic toxicity and achieved between 80.2% and 87.3% performances with two models (Liao et al. 2007). In order to improve in-silico methodologies used to predict mutagenicity in the first decade of the 2000s, Mazzatorta et al. proposed a novel system named robust hybrid classifier (RHC) by combining a fragment-based structure activity relationship (SAR) model and AI-based approaches on Bursi mutagenicity data set. The performance of the proposed methods was tested with external test data and obtained 85% both in sensitivity and specificity (Mazzatorta et al. 2007; Kazius et al. 2005). In order to build a public Ames mutagenicity data set, Hansen et al. constructed a data set that comprised about 6500 compounds, in the format of SMILES (simplified molecular input line entry specification) and

SDF (structure data format), with biological activity (Hansen 2009). In the same study, this benchmark data set was used to compare commercial tools (DEREK, Pipeline Pilot, and MultiCase) based on expert knowledge with machine learning algorithms (SVM, random forest (RF), k-nearest neighbour (KNN), and Gaussian process (GP)). As a result of this study, while the best performance was obtained by SVM with 0.86 AUC, DEREK yielded the lowest sensitivity and specificity (Hansen 2009).

Since toxicity is one of the most critical issues that cause late-stage drug failure or withdrawal, the in-silico studies in the prediction of mutagenicity gained speed in the last decade. In most of these studies, machine learning and statistical learning-based methods such as SVM, RF, artificial neural networks (ANN), KNN, genetic algorithms, radial basis function (RBF), partial least squares (PLS), naïve Bayes methods were preferred (Sharma et al. 2011; Webb et al. 2014b; Xu et al. 2012). Since the experimental screening of chemical compounds for biological activity is time consuming and expensive, Seal et al. applied supervised learning approaches on two different data sets to generate an alternative predictive model. As a result of the study, the RF algorithm achieved the best performance with 89.27% with a new mutagenicity data set comprising two well-known data (Seal et al. 2012). Webb et al. published a study emphasizing that interpretability of Ames mutagenicity prediction is more important than successful performance to interpret the model. They tried to extract the pattern of biological activity through the descriptors importance (Webb et al. 2014a). Zhang aimed to investigate the prediction of agents as mutagens and non-mutagens using a naive Bayes classifier in several studies. In addition to this purpose, they focused on identifying the most informative molecular descriptors related to mutagenicity. Although the prediction performance was similar to previous studies, their model identified four simple molecular descriptors (apol, number of H donors, number of rings, and Wiener) related to mutagenicity (Zhang 2017, 2015, 2016). Another research group has provided machine learning-based models for toxicity prediction of approximately 1500 diverse chemical compounds in various species. It has been reported that 70% of compounds were classified correctly based on the random forest algorithm and listed the physicochemical descriptors based on their importance (Moorthy et al. 2017).

Several kinds of research have recently been published on the prediction of toxicity and in-silico drug discovery with new approaches such as deep learning and ensemble methods such as XGBoost (Fan 2018; Rifaioğlu et al. 2019; Ji et al. 2019). Most recent studies in this area generally examine the current impact of AI studies on drug toxicity, potential challenges and future perspectives and potential (Tran et al. 2023; Zhang et al. 2019; Chu et al. 2021). In their 2023 review, Tran et al. provide an overview of recent AI driven advances in drug toxicity prediction, including machine learning and deep learning techniques on various toxicity traits (Tran et al. 2023). Zhang et al. conducted a study on chemical toxicity prediction with LightGBM, a machine learning algorithm, using Tox21 and Mutagenicity databases (Ji et al. 2019). Similarly, Chu et al. tried to present robust in silico models accurately estimate a compound's mutagenicity before synthesis to get around the limitations (costly, time consuming) of the Ames test (Chu et al. 2021).

Providing the most related descriptors that effectively classify mutagenic and non-mutagenic compounds also emerges as another important research area. Feature selection which can be conducted either based on the wrapper or filter approach, is considered a preprocessing for machine learning algorithms. It is generally hard to obtain the best feature subset set using traditional approaches. Therefore, metaheuristic approaches can be another alternative in order to select optimal subsets. In 2020, Houssein et al. built a novel hybrid Harris Hawks optimization (HHO) and SVM in drug discovery. As they reported, this was the first time HHO had been applied in the field of drug design (Houssein et al. 2020). Similarly, Algamal et al. developed the pigeon optimization algorithm with a new time varying transfer function to select the features most relevant to high dimensional QSAR / QSPR classification modeling (Algamal et al. 2020).

3. Material and Methods

3.1. Data preparation

This study used a combination of two popular data sets, namely The Benchmark and Bursi Mutagenicity data sets. The Benchmark data set consists of 6512 compounds, and the Bursi data set has 4337 compounds. These data sets were collected by Hansen et al. and Kazius et al., respectively (Kazius et al. 2005; Hansen 2009). According to Ames results, each compound in the data sets was given its canonical SMILES format and corresponding label, indicating whether it was mutagen or non-mutagen. In addition to this training data set, an external validation set has been included to study to make a fair measurement of the proposed approach. The external data set consisted of canonical SMILES format of 731 compounds was collected by Xu et al. (Xu et al. 2012). After removing duplicate compounds based on their SMILES format, 8167 unique compounds in the training set and 716 unique compounds in the external validation set were left. All the descriptors of molecules were calculated by PaDEL-Descriptor software (Yap 2010). Among 1444 1D and 2D physicochemical descriptors, i.e., properties, 225 descriptors were chosen based on several studies (Xu et al. 2012; Fan 2018; Gupta and Rana 2019; Guan et al. 2018). During the preprocessing and selection process, only the training data set was used. The limited number of missing values was imputed by using mean. Correlated descriptors

given in the correlation matrix across all pairs of descriptors with 0.95 or higher correlations were assumed to be redundant and removed from the data set. Finally, the entire data set used in this study consisted of 162 descriptors, and the details were presented in Table 1.

3.2. Butterfly optimization algorithm

BOA, proposed by Arora et al., is a metaheuristic algorithm that models the food foraging behavior of butterflies in nature (Arora and Singh 2019). Through chemoreceptors scattered on their bodies, butterflies can separate different fragrances of food (flowers), sense (smell) their intensities, and perform foraging movements (Tubishat et al. 2020). During their movements, butterflies can produce fragrance with an intensity that is directly proportional to their fitness. Butterflies communicate with each other by the fragrance they emit. BOA is a global optimization method based on the communication behaviors of butterflies. The intensity of the fragrance a butterfly emits is as much as other butterflies can feel it. The most crucial feature of BOA that differs from different metaheuristic algorithms is that the intensity of fragrance felt by the butterfly is unique. The most critical part of BOA algorithm is how the fragrance is calculated based on concepts of sensing and processing the modality like the smell, sound, temperature, etc. As reported in the original study of Arora et al., modality is fragrance in BOA Arora and Singh (2019). Three terms should be clearly explained for this; sensory modality (c), stimulus intensity (I), and power exponent (a). The formulation of the perceived fragrance intensity for each butterfly is given in Eq.(1) based on Steven's Law of Power (Arora and Singh 2019; Stevens 1986).

$$f = cI^a \tag{1}$$

where *f* is the emitted magnitude of the fragrance, i.e., how intensively other butterflies emit the fragrances within the search space, *c* is a proportionality constant taken as the sensory modality taken in the range [0, 1], *I* is the stimulus magnitude of the perceived fragrance by butterfly, and *a* is the power exponent characterizing the degree of absorption of sensory modality with its values over the range [0, 1]. Since *a* and *c* directly affect the convergence speed of the BOA algorithm, it is a crucial point to choose suitable values for both *c* and *a*. This can be expressed as, at the extreme points of the range, the fragrance emitted by the butterfly, if $a = 0$ it is not perceived by other butterflies if $a = 1$, it is perceived by other butterflies at the same intensity.

Butterflies share information with each other about their positions according to the fragrance intensity they produce. Thus, the butterflies change their positions towards the best butterfly closest to the food with the optimum fragrance intensity in the search space. This movement of butterflies is called global search and determined as in Eq. (2).

$$x_i^{t+1} = x_i^t + (r^2 \times g^* - x_i^t) \times f_i \tag{2}$$

where the solution vector x_i^t is the position of ith butterfly in movement t, g^* is the current fittest position decided among the available positions at the current movement of all butterflies, sensed fragrance magnitude of the butterfly is symbolized by f_i and r is a uniform random number in the range of [0, 1].

On the other hand, when butterflies cannot detect the fragrance of different butterflies in the search space, they move randomly. This movement of butterflies is called local search and formulated as in Eq. (3).

$$x_i^{t+1} = x_i^t + \left(r^2 \times x_j^t - x_k^t\right) \times f_i \tag{3}$$

where the solution vector x_j^t and x_k^t are the position of jth and kth butterfly in movement t. If x_j^t and x_k^t are located in the same neighborhood, and r is a random number in the range of [0, 1], then Eq. (3) turns out to be a local random stride. In order to control switching between global search and local search space in BOA, the switching probability (p) parameter is utilized.

3.3. Particle swarm optimization algorithm

First introduced by Kennedy et al., PSO is one of the metaheuristic search algorithms inspired by the bird's swarm's social behaviour (Kennedy and Eberhart 1995). PSO is a population-based algorithm that consists of the particles, i.e., a possible set of solutions. These particles move through in the multidimensional search space in order to find the best solution. While their movement, they have a memory in keeping track of their previous best position, namely best solution. Besides concerning their own best solutions, they considered the best solution of the swarm as well (Mirjalili and Lewis 2013). There are two types of particle positions, namely local (personal) best and global best.

Table 1. Distribution of the mutagens and non-mutagens in training and external validation set

Data Sets	Mutagens	Non-Mutagens	Total
Training data set (Kazius et al. 2005; Hansen 2009)	4524	3643	8167
External validation set (Xu et al. 2012)	599	117	716
Total	5123	3760	8883

Each particle owns certain information in order to update its position; the current position, the current velocity, distance to the local best solution (p), and distance to the global best solution (g^*). The mathematical definition of the PSO model consists of both the velocity of the ith particle at iteration $t + 1$ and also the new position of the ith particle given in Eq. (4) and Eq. (5), respectively.

$$v_i^{t+1} = wv_i^t + c_1.r(p_i - x_i^t) + c_2.r(g^* - x_i^t) \quad (4)$$

$$x_i^{t+1} = x_i^t + v_i^{t+1} \quad (5)$$

where v_i^t is the velocity of the ith particle at iteration t, c_1 and c_2 are acceleration constants, r is a uniform random number in the range of [0, 1], w is the inertia weighting function predefined by the user, x_i^t represents the current position of ith particle at iteration t, p_i is the best solution that obtained previously by ith particle, and global best solution $g*$ is the best position of all particles, i.e., swarm. Once v_i^{t+1} (the velocity of the ith particle at iteration $t + 1$) is obtained by Eq. (4), the position of the ith particle is updated by Eq. (5).

The general idea of Eq. (4) can be explained as the combination of exploration ability wv_i^t, individual thinking ($c_1 \times r \times (p_i - x_i^t)$) and collaboration of particles ($c_2 \times r \times (g^* - x_i^t)$). Initially, each particle placed in the search space has a random position, velocity, and fitness value calculated by the fitness function. At each iteration, the velocity and the position of particles are updated until the stopping criterion is met (Mirjalili and Lewis 2013).

3.4. Machine learning based models

Machine learning as a branch of artificial intelligence seeks to build analytical computational models by learning automatically from data and improving with experience (Mitchell 1997). In this work, several machine learning methods have been used both for feature selection and prediction phases. Support vector machine (SVM) as a binary learning machine is based on statistical learning theory introduced by Vapnik (Vapnik 1995). SVM aims to construct a decision hyperplane that the margin of separation between a set of objects of different classes is maximized (Haykin 2011). K-nearest neighbour (KNN), the most basic instance-based method, was designed to approximate real-valued or discrete valued target functions with instances consisting of the k closest training examples in the training data (Mitchell 1997; Cover and Hart 1967). Logistic regression (LR), known as the discriminative classifier, is one of the baseline machine learning algorithms used widely for binary and multinomial classifications. Random forest (RF), proposed by Breiman, is one of the commonly used ensemble algorithms. Basically, it combines the results of the tree predictors applying on the random subsamples from a standard data set. The strength of the individual trees in the forest and their correlation affects the generalization error of a forest (Breiman 2001). Randomness, the most crucial property of RF, can be defined as a combination of bagging and random subspace methods (Ho 1998). Extremely randomized trees (ExTrees) and RF can be similar ensemble algorithms that follow almost identical procedures. The main differences are in the subsampling approach and the split points selections. While RF obtains subsamples with replacement, ExTrees uses the original sample. The cut points are decided randomly in ExTrees, whereas RF selects the optimum cut point (Geurts et al. 2006). Extreme gradient boosting (XGBoost) is an ensemble of decision tree models utilized based on the principle of gradient boosting machines (Chen and Guestrin 2016). Although training based on the gradient boosting principle can be diffcult, it can achieve a lower model bias than the RF. XGBoost follows the idea to correct the previous mistakes done by the model and propagate the experience to the next step for improving the performance. Light gradient boosting (LightGBM) like XGBoost is also a supervised learning method based on the gradient boosting framework. The main difference between them is faster training speed, especially on a large data set. LightGBM is a histogram-based algorithm with low memory usage since it transforms numerical values to discrete bins (Ke et al. 2017).

3.5. Model evaluation

The data sets used in this study have binary classes as mutagen and non-mutagen. Since our main focus is to identify mutagen compounds, mutagen labelled class is assumed as 1 which indicates positive class, and non-mutagen labelled class is accepted as 0, i.e. negative class. True positive (TP) and true negative (TN) results show that the compound is correctly predicted to be mutagenic and non-mutagenic, respectively. On the other hand, false positive (FP) and false negative (FN) results indicate that the compound has been incorrectly predicted to be positive and negative, respectively (John et al., 2023). In this study, several statistics had been calculated based on the confusion matrix to measure the models' performance. All the experiments were evaluated in terms of sensitivity (recall), specificity, F-measure, and accuracy. In addition to these measurements, AUC and probability excess had been preferred to compare models. The reason to choose the probability excess is that relative class frequencies, i.e., imbalanced class distribution, do not affect the probability excess, whereas accuracy (success rate) is affected by imbalance class distribution (Yang et al. 2005).

3.6. Proposed approach

Since each metaheuristic optimization algorithm follows different search strategies, each of them may propose different subsets of features for a given dataset. The proposed approach consists of feature selection and

Algorithm 1. Hybrid Feature Selection and Prediction

Inputs

n: number of features
OP_i : Optimization algoritm, *i=1,…,nbOP, nbOP*: number of optimization algorithms
CL_j : Classification algorithm, j=*1,…,nbCL, nbCL*: number of classification algorithms
FS_j^i: Feature Subset selected with with OP_i and CL_j
error : error obtained by classifier
d: number of selected features
FFS^i : Final Feature Subset selected with with OP_i and CL_j
w_1 = importance of the classification error (w_1 = 1- w_2)
w_2 = importance of the number of selected features

Feature Selection Phase

for each OP_i; *i=1 to nbOP*
  for each CL_j ; *j=1 to nbCL*
    perform optimization algorithm OP_i to get the best subset of features
      calculate fitness value by objective function
        calculate classification error by CL_j
        evaluate the candidate subset and get *d*
        fitness ← w_1 x *error* + w_2 x (*d/n)*
    save the best subset of features found during the OP_i process, as FS_j^i
  end for
  *d*etermine the final feature set by *majority voting* strategy by selecting the most seen features among FS_j^i and save as FFS^i
end for

Prediction Phase

Evaluate the performance of the classification algorithms on dataset with FFS^i (Final Feature Subset obtained from Feature Selection Phase)

mutagenicity prediction phases described in the pseudo code seen in Algorithm 1. The feature selection phase provided in Algorithm 1 basically selects the final significant and representative feature subsets by consensus of several hybridization of optimization algorithms with classifiers. To obtain this outcome, different machine learning algorithms had been chosen as part of the fitness function, which is essential for the optimization algorithm. The fitness function allows to determine the distance of each unit (particle, butterfly, etc) to the best solution based on the nature of the chosen optimization algorithm. The units receive a fitness value by sending their position values to the fitness function. In this frame, the optimization had a two-fold aim; obtaining the lowest error with the minimum number of features (Arora and Singh 2019). The solution for this multi-objective problem had been given as a fitness function in Eq. (6)

$$fistness = w_1 \times error(classifier) + w_2 \times \frac{d}{n} \quad (6)$$

where $error(classifier)$ is the classification error rate of the $classifier$, d is the number of selected descriptors, and n represents the number of descriptors in the original data set. The importance of accuracy and the number of selected features were weighted by w_1 and w_2, respectively and chosen as $w_2 = 1 - w_1$. Thus, a balance was provided between classification accuracy and subset length utilizing the fitness function.

Once relevant and informative descriptors from each hybrid of optimization and classifier, the final subset is determined based on the majority voting strategy by selecting the most seen features obtained from each hybrid. Several machine learning algorithms were applied on data sets consisting of descriptor subsets obtained from the feature selection phase in the mutagenicity prediction phase. In order to validate the proposed model, besides cross-validation, an external data set was also used. It should be noted that the most promising property of the proposed approach is the flexibility. It can be used either the same optimization algorithms with different classifiers or different optimization algorithms with other classifiers.

4. Experimental Results

As stated earlier, the main purpose of this study is to select the most informative descriptor subset for the early prediction of the mutagenicity by following the flow given in Algorithm 1. In order to explore the performance and effectiveness of the proposed approach, BOA and PSO were hybridized with KNN, SVM, and LR respectively for the feature selection

phase, and DT, ExTree, LightGBM, RF, and XGBoost were involved for the early prediction of mutagenicity with the features obtained from feature selection phase. All the experiments were conducted by using both PaDEL data set and the external data set explained before.

Normalization was applied to the data sets before the feature selection process and the models fitted with the selected variables. In order to avoid the effect of the parameters, no data set-specific parameter optimization was performed. The parameters of the classification algorithms preferred in this study were predetermined and preferred as the same for all data sets.

The optimization part of the proposed hybrid framework was operated with classifiers using the fitness function. The hybrid framework was run with different parameter combinations using KNN, SVM, and LR for BOA and PSO, respectively. To ensure both reproducibility and diversity, combinations were run using different seeds randomly generated by a seed function for each combination. Accordingly, feature selection was made for BOA and PSO using KNN with different parameter sets including different K values, different distance metrics – Manhattan and Euclidean; SVM with different parameter sets including a different number of C coefficients, different maximum iteration numbers, and LR with different parameter sets including C, different iteration numbers.

Table 2 outlines the parameter setting for BSO and PSO. The parameter values were decided based on the outcomes of the preliminary runs. Besides these parameters, different number of population size and number of iterations were obtained applied in trial-and error manner. Considering huge number of computational time and the obtained results, population size and number of iterations were chosen as 15 and 50, respectively.

All simulations were carried out in a cloud environment on Intel(R) Xeon(R) CPU @ 2.00GHz, Linux operating system, and 16 GB RAM.

Table 2 Parameter setting for optimization algorithms.

Methods	Parameters	Values
PSO	Search domain	[0, 1]
	Interia w	0.9
	Acceleration constants [c_1, c_2]	[2, 2]
BOA	Search domain	[0, 1]
	Sensory modality (c)	0.01
	Power exponent (a)	Increased from 0.1 to 0.3 with iterations
	Switching probability (p)	0.8

The reason for preferring the cloud environment is the high resource requirement and time cost. On the other hand, mutagenicity prediction part was implemented using Python 3.8. All predictive models were performed on an AMD Ryzen 5 2600X @ 3.6GHz, Windows 10, and 16GB RAM computer.

Before feature selection, the PaDEL data set was partitioned into 80% training and 20% test set, and experiments for feature selection were conducted using only the training set. In the modelling phase, 5-fold cross validation was preferred. Different iteration numbers were tried with varying numbers of particles for BOA and PSO. As a result of this selection, about 54 trials of 3 models were conducted for both methods, and feature selection outputs were obtained. With the intention of evaluating if optimization algorithms work with mutagenetic datasets, a comprehensive statistical analysis of the best, worst, mean and standard deviations of the fitness scores, average number of features and computational time were provided in Table 3. It can be observed based on the statistical fitness measurements given in Table 3, while BOA-SVM has better fitness measures than PSO-SVM, PSO-KNN and PSO-LR are slightly better than the results of BOA-KNN and BOA-LR. Conversely, the models constructed with BOA outperformed PSO in terms of the average number of selected features and the average computational time.

Table 3 Statistical analysis obtained by the hybridized algorithms based on mean, standard deviation, best and worst of the fitness scores, average number of features and computational times

Algorithm	Mean ± SD	Best	Worst	Avg. Number of Features	Avg. Computational Time ± SD
BOA-KNN	0.2366± 0.0129	0.22616	0.26694	41.17	225.8189±97.04
PSO-KNN	0.2330± 0.0086	0.22463	0.26236	53.58	357.7957±76.22
BOA-LR	0.2755± 0.0088	0.26663	0.29611	40.83	907.8523±895.52
PSO-LR	0.2671± 0.0085	0.25900	0.29444	57.08	1490.8023±1102.32
BOA-SVM	0.2882± 0.0132	0.27664	0.31983	40.28	368.8617±279.68
PSO-SVM	0.2965± 0.0071	0.29209	0.32129	68.00	627.1484±356.65

SD: standard deviation, Avg: average

Table 4 Comparison of different classification methods with all features sets and feature subsets obtained by PSO and BOA for PaDEL Test Data

Classification Method	data sets (#of Features)	F1-score	Acc	Precision	Recall	Spec	ProbEx	AUC
DT	Baseline (162)	85.25	83.41	86.52	84.01	79.56	63.57	83.04
	PSO (87)	86.00	84.46	86.19	85.81	82.30	68.11	84.25
	BOA (51)	85.57	83.78	86.85	84.33	79.97	64.31	83.41
ExTrees	Baseline (162)	87.97	86.47	89.28	86.70	82.99	69.69	86.14
	PSO (87)	87.90	86.41	89.06	86.76	83.13	69.89	86.09
	BOA (51)	87.32	85.68	89.06	85.65	81.48	67.14	85.27
LightGBM	Baseline (162)	86.76	84.88	89.39	84.27	79.29	63.56	84.34
	PSO (87)	86.45	84.64	88.51	84.49	79.84	64.33	84.17
	BOA (51)	85.34	83.35	87.51	83.28	78.19	61.47	82.85
RF	Baseline (162)	87.71	86.23	88.73	86.72	83.13	69.84	85.93
	PSO (87)	87.40	85.92	88.18	86.64	83.13	69.77	85.65
	BOA (51)	87.89	86.47	88.62	87.17	83.81	70.99	86.22
XGBoost	Baseline (162)	87.78	86.17	89.72	85.93	81.76	67.68	85.74
	PSO (87)	87.86	86.41	88.73	87.00	83.54	70.54	86.13
	BOA (51)	87.79	86.29	88.95	86.65	82.99	69.64	85.97
Overall	Baseline (162)	87.09	85.43	88.728	85.53	81.35	66.87	85.04
	PSO (87)	87.12	85.57	88.134	86.14	82.39	68.53	85.26
	BOA (51)	86.78	85.11	88.198	85.42	81.29	66.71	84.75

DT: Decision Tree, ExTrees: Extra Trees, RF: Random Forest Acc: Accuracy, Spec: Specificity, ProbEx: Probability Excess, AUC: Area Under Curve

The variables selected from each experiment were combined, and the most repetitive (majority voting) features were chosen uniquely. By following these approaches for BOA and PSO separately, the 87 most repetitive variables among the variables selected for PSO and the first 51 most repetitive variables for BOA were chosen for the final feature subset due to the feature selection phase.

In the second phase of the study, models were fitted with PaDEL data set using treebased methods (DT, RF, XGBoost, ExTree, and LightGBM) with selected features, and predictions were obtained for both test set and External data set. In Tables 4-5, one can compare the results obtained by using a data set with full features named Baseline with 162 features and the data sets which were reduced by BOA and PSO involved in the proposed feature selection approach with 51 features and 87 features respectively in terms of F1-score, Accuracy (Acc), Precision, Recall, Specificity (Spec), Probability Excess (ProbEx) and Area Under Curve (AUC). All experiments were conducted by 5-fold cross validation.

The results presented in Table 4 were obtained using PaDEL test data reserved for testing at the beginning of the experiments. It can be analysed that although ExTrees got the highest F1 score with 162 features, there is no significant difference between the reduced data sets by using the proposed feature selection scheme. It is worth noting that results obtained with almost a third of the data set yielded similar or slightly better prediction results than the results obtained with the entire data set. It can be observed by analysing the results of BOA with RF in Table 4. The highest ProbEx, the unbiased measurement for evaluating prediction performance, was obtained with 51 features. On the other hand, the results of feature selection phased conducted by PSO were yielded by XGBoost based on the results given in Table 4. According to the overall results reported in Table 4, the proposed feature selection approach used with BOA provided almost the best results with the smallest number of features.

The proposed approach was applied to a completely unseen external data set explained in the Data Preparation section to meet the fair comparison. As given in Table 1, External data set can be assumed as an imbalanced data set. Since the relative class frequency does not influence ProbEx, most of the analyses and explanations for External data set given in Table 5 were done by ProbEx.

Table 5 Comparison of different classification methods with all features sets and feature subsets obtained by PSO and BOA for External Data

Classification Method	data sets (#of Features)	F1-score	Acc	Precision	Recall	Spec	ProbEx	AUC
DT	Baseline (162)	97.22	95.25	99.33	95.20	74.36	69.56	86.85
	PSO (87)	96.47	93.99	98.16	94.84	72.65	67.49	85.41
	BOA (51)	97.44	95.67	98.50	96.41	81.20	77.60	89.85
ExTrees	Baseline (162)	98.76	97.91	99.83	97.71	88.03	85.75	93.93
	PSO (87)	98.60	97.63	99.83	97.39	86.32	83.72	93.08
	BOA (51)	98.76	97.91	100.00	97.56	87.18	84.74	93.59
LightGBM	Baseline (162)	96.37	93.72	99.67	93.28	63.25	56.53	81.46
	PSO (87)	95.37	91.90	99.83	91.30	51.28	42.58	75.56
	BOA (51)	95.46	92.04	100.00	91.31	51.28	42.59	75.64
RF	Baseline (162)	98.60	97.63	100.00	97.24	85.47	82.71	92.74
	PSO (87)	97.88	96.37	100.00	95.84	77.78	73.62	88.89
	BOA (51)	98.27	97.07	99.83	96.76	82.91	79.67	91.37
XGBoost	Baseline (162)	97.87	96.37	99.83	95.99	78.63	74.62	89.23
	PSO (87)	96.69	94.27	99.83	93.73	65.81	59.54	82.82
	BOA (51)	97.56	95.81	100.00	95.23	74.36	69.59	87.18
Overall	Baseline (162)	97.76	96.18	99.73	95.89	77.95	73.83	88.84
	PSO (87)	97.00	94.83	99.53	94.62	70.77	65.39	85.15
	BOA (51)	97.50	95.70	99.67	95.45	75.39	70.84	87.53

DT: Decision Tree, ExTrees: Extra Trees, RF: Random Forest Acc: Accuracy, Spec: Specificity, ProbEx: Probability Excess, AUC: Area Under Curve

In the model-based analysis, a comparison of the overall metrics in both Table 4 and Table 5 reveals that ExTrees and RF consistently outperform other methods. They achieve high accuracy, precision, and AUC while exhibiting minimal declines in recall and specificity. This suggests that these methods are suitable for handling both the Baseline feature sets and the reduced feature sets without significant performance degradation. In contrast, methods such as LightGBM, although effective with the full set of features, demonstrate greater sensitivity to feature reduction, particularly impacting recall and ProbEx.

The Baseline feature set (162 features) consistently yields slightly better results across all metrics compared to the reduced feature sets (PSO and BOA). However, the differences are generally minimal. This observation indicates that while feature selection may lead to some loss in precision, recall, and other metrics, the trade-off is justified by the reduction in computational complexity and the potential for avoiding overfitting. By eliminating redundant or irrelevant features, the risk of overfitting can be mitigated, which can enhance the generalizability of the model despite minor performance losses in specific metrics.

5. Discussion

Regarding individual results obtained by the methods and the data sets in Table 5, it can be seen that there is no method-data set pair that consistently produces the best results. The results are varied across different methods and data sets, indicating that there is no one-size-fits-all solution. Based on the results presented in Table 5, the highest scores had been obtained by ExTrees. Although there is no observed difference among the results of the ExTrees, outcomes of the data set reduced a third by BOA can be assumed as promising to predict mutagenicity via in-silico methods. To summarize the comparison, the average results of the tree-based classification methods had been calculated. In the light of these averages of the metrics, one can say that when the proposed approach given in Algorithm 1 had been used with BOA, considerably better results had been obtained with reduced data sets both for test and external data sets.

The results achieved by the reduced data set were compared with the results published in 2012 by Xu et al. (Xu et al. 2012) with the almost similar external data set for the sake of completeness of the study. While the external data set used in this study contained 599 mutagens and 117 non-mutagens chemicals, the original data set used in the study of Xu et al. had 614 mutagens

and 117 non-mutagens. The accuracies they obtained by these data sets with different fingerprints (descriptors) were from 90.4% to 98%. On the other hand, the range of the accuracy achieved for the external set used in this study by the proposed approach with BOA is from about 91.9% to 97.91%. It is essential to mention that both studies used the same chemicals with mutagenicity information, whereas their number of features and way of calculation is different. Seal et al. (Seal et al. 2012) have generated prediction models using RF classifier for predicting mutagenicity with the data set named Set3, similar to Baseline data set used in this study. The data set used in their study consists of the Bursi and Benchmark data sets explained in the Data Preparation section. According to the outcomes that they published in 2012, they have found the success rate of predicting mutagenicity as 85.15% and precision as 85.2% with 154 descriptors (Seal et al. 2012). Our study shows that results of the data set were reduced into 51 descriptors with BOA and mutagenicity prediction conducted by RF given in Table 4 yielded better results with 86.47% and 88.62% accuracy and precision, respectively.

Each optimization method uses different strategies and metrics. Therefore, selected features can be vary based on the search strategy of the algorithm. Through the hybridization of several optimization techniques with classification algorithms, the suggested method may be able to overcome this variability, eliminate the characteristics of the dataset, and lower the risk of overfitting. To evaluate the effectiveness of the proposed method in terms of overfitting, 5-fold cross-validation was applied alongside the dataset with an external validation test set. The test and validation sets results given in Table 4 and Table 5, respectively, are also a sign that there is no possible overfitting. Moreover, the proposed approach may effectively explore the complex and high-dimensional feature space of the drug toxicity datasets due to the stochastic, adaptive, and global search characteristics of the optimization algorithms.

It is worth mentioning that this approach can be used not only for mutagenicity prediction but also for different problems requiring feature selection and prediction. Since the presented approach can be conducted with any number and kind of metaheuristic optimization algorithms and classification methods, it could be considered a general and flexible framework and a wide range of application fields. It should be noted that, flexibility is the strongest property of proposed hybrid approach. However, the computational complexity is the weakness of the hybrid approach, and it is planned in future studies to overcome this with the new approaches even in high dimensional datasets.

6. Conclusions

Recently, because of the laborious and expensive nature of the drug discovery process, in-silico approaches have played crucial and indispensable roles in the drug approval process. Predicting mutagenicity, which can be defined as the most critical endpoint of toxicity at the early stages of the drug discovery process, is one of the essential steps. This study recommended a flexible approach as an in-silico method both for the early prediction of chemical mutagenicity and reducing the search space into the most effective descriptors. The proposed framework was designed as two sequential phases: feature selection phase through nature inspired optimization algorithms and prediction phase by several statistical and machine learning classification methods. Incorporating metaheuristic algorithms into in-silico studies is not commonly seen in the literature. One of the primary purposes of this study is to conclude whether using the metaheuristic algorithms can be suitable to search the descriptor space in the field of mutagenicity prediction. In order to reach this aim, the butterfly optimization algorithm (BOA) was hybridized with several statistical machine learning algorithms to select the most critical descriptors that are effective in predicting mutagenicity. As mentioned earlier, to the best of our knowledge, no studies are searching for an effective and representative subset of descriptors for mutagenicity estimation through metaheuristic optimization algorithms. To fair comparison, besides BOA, inspired by social butterfly foraging strategy, particle swarm optimization algorithm (PSO), inspired by not a single animal but swarm which is coordinated, were used (Arora and Singh 2019).

Two data sets were used to present the proposed approach: one for selecting the most informative descriptors and modelling; the other for validation. All descriptors were calculated by freely available PaDEL Descriptors software by using SMILES format of the molecules. The original data set contains 162 descriptors. The proposed approach with BOA reduced the number of descriptors to 51, whereas 87 were obtained with PSO. The experimental results present that the outcomes obtained by the BOA have yielded better results, especially with a smaller number of the descriptors sets. In the test data obtained from PaDEL, the highest ProbEx was obtained with the features selected with BOA. While 69.84% ProbEx was obtained with a baseline containing 162 variables in total, due to the model established using 51 variables, approximately 71% ProbEx value was reached with an increase of 1.15% with PaDEL. External Data also achieved the highest ProbEx with Baseline, but the 51 variables selected with BOA performed higher than the 87 variables selected with PSO. As a result, BOA and PSO methods were used for variable selection in the study, and the selected variables were classified using tree-based methods such as DT, ExTrees, RF, LightGBM, and XGBoost. Since no parameter optimization is performed specifically for the data sets, methods that perform highly in PaDEL data may have lower performance in the External data set. Another reason is that while the class distribution is balanced in PaDEL data, the proportion of classes that are non-mutagen in

the External data set is lower. Parameter optimization can be performed to increase model performance in future studies.

To ensure completeness of the study, the results were also compared with the results achieved by the studies Seal et al. (2012), Xu et al. (2012), which used similar chemicals. It could be concluded that our approach conducted with nature inspired BOA performed well in terms of accuracy and precisions.

As stated earlier, in-silico studies, i.e., the computational approach to toxicity, has started to gain more attention since predicting mutagenicity at the beginning of the drug design process has been inevitable and is a crucial step to shorten the process and thereby reduce the cost. This study was conducted to present highlight the importance of this approach. The findings in this study suggest that in-silico approaches have a significant role in the drug discovery process by predicting mutagenicity, reducing the search space, and ultimately saving time and resources. The use of metaheuristic optimization algorithms in this context represents a flexible approach that can potentially effective feature selection and prediction in various fields. Further research, including parameter optimization and multi-objective algorithms, can continue to refine and expand upon this methodology. As the part of the future works, a wider range of metaheuristic algorithms and machine learning algorithms can be evaluated to identify the best combination for different drug toxicity endpoints on a larger and more diverse dataset of compounds based on the experimental process, findings and also limitations of the study. It is worth pointing out that, although the computational complexity is a challenge, aiming to address this issue can be also one of the future studies.

In summary, this research demonstrates the potential of combining nature-inspired optimization algorithms with machine learning techniques for feature selection and mutagenicity prediction. The flexible framework presented here can be applied to a wide range of applications requiring feature selection and prediction.

References

Algamal, ZY, Qasim, MK, Lee, MH and Ali, HTM. 2020. High-dimensional QSAR/QSPR classification modelling based on improving pigeon optimization algorithm. Chemom. Intell. Lab. Syst, 206:104170, doi:10.1016/ j.chemolab.2020.104170.

Arora, S, Singh, S. 2019. Butterfly optimization algorithm: a novel approach for global optimization. Soft Comput. 23, 715–34 doi:10.1007/ s00500-018-3102-4.

Bakhtyari, N, Raitano, G, Benfenati, E, Martin, T and Young, D. 2013. Comparison of in silico models for prediction of mutagenicity. Carcinog. Ecotoxicol. Rev, 31(1):45–66, doi:10.1080/10590501.2013.763576

Breiman, L. 2001. Random forests. Mach. Learn, 45:5–32. doi:10.1023/ A:1010933404324.

Çakmak Pehlivanlı, A. and Çakmak, G. 2022. Genotoksik etkiyi belirlemeye yönelik in-silico yaklaşımlar. In Genetik Toksikoloji (Genetic Toxicology), ed. F. Ünal and D. Yüzbaşıoğlu, 475–92. Ankara: Nobel.

Cariello, NF, Wilson, JD, Britt, BH, Wedd, DJ, Burlinson, B and Gombar, V. 2002. Comparison of the computer programs DEREK and TOPKAT to predict bacterial mutagenicity. Mutagenesis 17(4):321-9, doi:10.1093/mutage/17.4.321.

Chen, T and Guestrin, C. 2016. XGBoost. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–94, doi:10.1145/2939672.2939785.

Chu, CSM, Simpson, JD, O'Neill, PM and Berry, NG. 2021. Machine learning predicting Ames mutagenicity of small molecules. Journal of Molecular Graphics and Modelling, 109. doi:10.1016/j.jmgm.2021.108011.

Cover, T. and Hart, P. 1967. Nearest neighbor pattern classification. IEEE Trans. Inf.Theory, 13(1):21–27. doi:10.1109/TIT.1967.1053964.

Fan, D., Yang, H., Li, F, Sun, L, Di, P, Li, W, Tang, Y and Liu, G. 2018. In silico prediction of chemical genotoxicity using machine learning methods and structural alerts. Toxicol, 7(2): 211–20. doi:10. 1039/c7tx00259a.

Geurts, P, Ernst, D and Wehenkel, L. 2006. Extremely randomized trees. Mach Learn, 63:3–42. doi:10.1007/s10994-006-6226-1.

Greene, N, Judson, P, Langowski, J and Marchant, C. 1999. Knowledge-based expert systems for toxicity and metabolism prediction: DEREK, StAR and METEOR. SAR QSAR Environ. Res. 10 2-3, 299-314. doi:10.1080/10629369908039182.

Guan, D, Fan, K, Spence, I and Matthews, S. 2018. QSAR ligand dataset for modelling mutagenicity, genotoxicity, and rodent carcinogenicity. Data Br, 17:876–84, doi: 10.1016/j.dib.2018.01.077.

Gupta, V and Rana, P. 2019. Toxicity prediction of small drug molecules of aryl hydrocarbon receptor using a proposed ensemble model. Turkish J. Electr. Eng. Comput. Sci, 27(4): 2833–49. doi:10.3906/elk-1809-9.

Hansch, C. 1980. Use of quantitative structure-activity relationships (QSAR) in drug design (review). Pharmaceutical Chemistry Journal, 14. doi: 10.1007/BF00765654.

Hansen, K, Mika, S, Schroeter, T, Sutter, A, Laak, AT, Steger-Hartmann, T, Heinrich, N and Müller, KR. 2009. Benchmark data set for in silico prediction of Ames mutagenicity. J. Chem. Inf. Model. 49, 9, 2077–81. doi:10.1021/ci900161g.

Haykin, S. 2011. Neural Networks and Learning Machines. Pearson Education, 3rd ed.

Ho, T. 1998. The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell, 20(8):832–44. doi: 10.1109/34.709601.

Honma M, Kitazawa, A, Cayley, A, Williams, RV, Barber, C, Hanser, T, Saiakhov, R, Chakravarti, S, Myatt, GJ, Cross, KP et. al.2019. Improvement of quantitative structure-activity relationship (QSAR) tools for predicting ames mutagenicity: Outcomes of the Ames/QSAR international challenge project. Mutagenesis, 34:41–48. doi:10. 1093/mutage/gey031.

Houssein, E, Honey, M, Oliva, D, Mohamed, W and Hassaballah, M. 2020. A novel hybrid harris hawks optimization and support vector machines for drug design and discovery. Comput. Chem. Eng, 133:106656. doi:10. 1016/j.compchemeng.2019.106656.

Ji, X, Tong, W, Liu, Z and Shi, T. 2019. Five-feature model for developing the classifier for synergistic vs. antagonistic drug combinations built by XGBoost. Front. Genet, 10(JUL):1–13. doi:10.3389/fgene.2019.00600.

John, L, Mahanta, HJ, Soujanya, Y, Narahari Sastry, G. 2023. Assessing machine learning approaches for predicting failures of investigational drug candidates during clinical trials. Computers in Biology and Medicine, Vol.153, 106494. doi: 10.1016/j.compbiomed.2022.106494.

Kazius, J, McGuire, R and Bursi, R. 2005. Derivation and validation of toxicophores for mutagenicity prediction. J. Med. Chem, 48. doi:10.1021/ jm040835a.

Ke, G, Meng, Q, Finley, T, Wang, T, Chen, W, Ma, W, Ye, Q and Liu, T. 2017. LightGBM: A highly efficient gradient boosting decision tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Vol. 30, 3149–57. Curran Associates Inc.

Kennedy, J and Eberhart, R. 1995. Particle swarm optimization. pages 1942–48. doi:10.1109/ICNN.1995.488968.

Liao, Q, Yao, J and Yuan, S. 2007. Prediction of mutagenic toxicity by combination of recursive partitioning and support vector machines. Mol. Divers 11, 59–72. doi:10.1007/s11030-007-9057-5.

Mazzatorta, P, Tran, L, Schilter, B and Grigorov, M. 2007. Integration of structure activity relationship and artificial intelligence systems to improve in silico prediction of ames test mutagenicity. J. Chem. Inf. Model. 47, 1, 34–38. doi: 10.1021/ci600411v.

Mirjalili, S and Lewis, A. 2013. S-shaped versus V-shaped transfer functions for binary particle swarm optimization. Swarm Evol. Comput, 9:1–14. 10.1016/j.swevo.2012.09.002.

Mitchell, T. 1997. Machine Learning. McGraw-Hill. New York

Moorthy, N, Kumar, S and Poongavanam, V. 2017. Classification of carcinogenic and mutagenic properties using machine learning method. Comput. Toxicol, 3:33–43. doi: 10.1016/j.comtox.2017.07.002.

Raghavan, N, Amaratunga, D, Nie, AY and McMillian, M. 2005. Class prediction in toxicogenomics, Journal of Biopharmaceutical Statistics, 15:2, 327-41, doi: 10.1081/BIP-200048836

Rifaioglu, AS, Atas, H, Martin, MJ, Cetin-Atalay, R, Atalay,V and Doğan, T. 2019. Recent applications of deep learning and machine intelligence on in silico drug discovery: Methods, tools and databases. Brief. Bioinform, 20(5):1878–1912. doi: 10.1093/bib/bby061.

Seal, A, Passi, A, Jaleel, U, Wild, D and Consortium, O. 2012. In-silico predictive mutagenicity model generation using supervised learning approaches. J. Cheminform. 4(1):10. doi:10.1186/1758-2946-4-10.

Sharma, A, Kumar, R, Varadwaj, P, Ahmad, A and Ashraf, G. 2011. A comparative study of support vector machine, artificial neural network and bayesian classifier for mutagenicity prediction. Interdiscip. Sci. Comput. Life Sci, 3(3):232–239. doi:10.1007/s12539-011-0102-9.

Stevens, S.S. 1986. Psychophysics: Introduction to Its Perceptual, Neural and Social Prospects. 1st ed. Routledge. doi.org/10.4324/9781315127675

Subaş, N and Çakmak Pehlivanlı, A. 2020. İkili parçacık sürü optimizasyonu ve destek vektör makinelerinin hibrit kullanımı ile ilaç keşfi için özellik seçimi. Gümüşhane Üniv. Fen Bilim. Enst. Derg., 11:169–78. doi:10. 17714/gumusfenbil.776329.

Tran, T T V, Surya Wibowo, A, Tayara, H and Chong, KT. 2023. Artificial intelligence in drug toxicity prediction: Recent advances, challenges, and future perspectives. Journal of Chemical Information and Modeling, 63(9):2628–43. doi: 10.1021/acs.jcim.3c00200.

Toropov, AA, Toropova, AP, Raska, I, Leszczynska, D, Leszczynski, J. 2014. Comprehension of drug toxicity: Software and databases. Computers in Biology and Medicine, 45: 20-25. doi: 10.1016/j.compbiomed.2013.11.013.

Tubishat, M, Alswaitti, M, Mirjalili, S, Al-Garage, M, Alrashdan, M and Rana, T. 2020. Dynamic butterfly optimization algorithm for feature selection. IEEE Access, 8:194303–14. doi:10.1109/access.2020.3033757.

Vapnik, V. 1995. The Nature of Statistical Learning Theory. Springer-Verlag, Berlin, Heidelberg.

Webb, SJ, Hanser, T, Howlin, B, Krause, P and Vessey, J. 2014a. Feature combination networks for the interpretation of statistical machine learning models: Application to Ames mutagenicity. J. Cheminform, 6(1):8. doi: 10.1186/1758-2946-6-8.

Webb, SJ, Hanser, T, Howlin, B, Krause, P and Vessey, J. 2014b. Interpretable Ames mutagenicity predictions using statistical learning techniques. In Handbook of abstracts, 6th Joint Sheffield Conference on Chemoinformatics. Qsar2012, 3–4.

White, A, Mueller, R, Gallavan, R, Aaron, A and Wilson, A. 2003. A multiple in silico program approach for the prediction of mutagenicity from chemical structure. Mutat. Res. - Genet. Toxicol. Environ. Mutagen, 539:77–89. doi:10.1016/S1383-5718(03)00135-9.

Wichard, J.D. 2017. In silico prediction of genotoxicity. Food and Chemical Toxicology, 106(Pt B):595-599. doi: 10.1016/j.fct.2016.12.013.

Xu, C, Cheng, F, Chen, L, Du, Z, Li, W, Liu, G, Lee, PW and Tang,Y. 2012. In silico prediction of chemical names mutagenicity. Journal of Chemical Information and Modeling, 52(11):2840–47. doi:10.1021/ci300400a.

Yang, Z, Thomson, R, Mcneil, P and Esnouf, R. 2005. Ronn: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinforma. Orig. Pap, 21(16):3369–3376. doi:10.1093/bioinformatics/bti534.

Yap, C. 2010. Padel-descriptor: an open-source software to calculate molecular descriptors and fingerprints. J. Comput. Chem, 32(4):1466–74. doi:10.1002/jcc.21707.

Zhang, H, Yu, P, Zhang, TG, Kang, YL, Zhao, X, Li, YY, He, JH and Zhang, J. 2015. In silico prediction of drug-induced myelotoxicity by using naïve Bayes method. Mol. Divers, 19(4): 945-53. doi: 10.1007/s11030-015-9613-3.

Zhang, H, Yu, P, Xiang, ML, Li, XB, Kong, WB, Ma, JY, Wang, JL, Zhang, JP and Zhang, J. 2016. Prediction of drug-induced eosinophilia adverse effect by using SVM and naïve Bayesian approaches. Med. Biol. Eng. Comput, 54(2–3):361–9. doi: 10.1007/s11517-015-1321-8.

Zhang, H, Kang, YL, Zhu, YY, Zhao, KX, Liang, JY, Ding, L, Zhang,TG and Zhang, J. 2017. Novel naïve Bayes classification models for predicting the chemical Ames mutagenicity. Toxicol. Vitr, 41:56–63. doi: 10.1016/j.tiv.2017.02.016.

Zhang, J, Mucs, D, Norinder, U and Svensson, F. 2019. LightGBM: An effective and scalable algorithm for prediction of chemical toxicity-application to the Tox21 and mutagenicity data sets. J. Chem. Inf. Model., 59(10):4150–58. doi: 10.1021/acs.jcim.9b00633.

Zheng, M, Liu, Z, Xue, C, Zhu, W, Chen, K, Luo, X and Jiang, H. 2006. Mutagenic probability estimation of chemical compounds by a novel molecular electrophilicity vector and support vector machine. Bioinformatics. 22(17):2099-106. doi: 10.1093/bioinformatics/btl352.