IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 13, No. 4, December 2024, pp. 4380∼4387
ISSN: 2252-8938, DOI: 10.11591/ijai.v13.i4.pp4380-4387

Character N-gram model for toxicity prediction

Eman Shehab[1], Hamada Nayel[2], Mohamed Taha[2]
[1]Department of Computer Science, Faculty of Computers and Artificial Intelligence, University of Sadat City, Sadat, Egypt
[2]Department of Computer Science, Faculty of Computers and Artificial Intelligence, Benha University, Benha, Egypt

Article Info

Article history:

Received Jan 24, 2024
Revised Mar 24, 2024
Accepted Apr 17, 2024

Keywords:

Feature extraction
Machine learning
Molecular toxicity prediction
N-gram
Simplified molecular-input line-entry system

ABSTRACT

Molecular toxicity prediction is a crucial step in the drug discovery process. It has a direct relationship with human health and medical destiny. Accurately assessing a molecule's toxicity can aid in the weeding out of low-quality compounds early in the drug discovery phase, avoiding depletion later in the drug development process. Computational models have been used automatically for molecular toxicity prediction. In this paper, a machine learning-based model has been proposed. TF/IDF representation scheme has been used for N-gram and integrated with simplified molecular-input line-entry system (SMILES). Multiple machine learning classifiers such as logistic regression (LR), support vector machine (SVM), random forest (RF), decision tree (DT), k-nearest neighbors (KNN), AdaBoost, multi-layer perceptron (MLP), and stochastic gradient descent (SGD) classifiers have been implemented. A wide range of N-gram models have been implemented and trigram reported the best results. RF and SVM achieved 85% and 84% accuracy respectively. Comparable to state-of-the-art models, our results are acceptable as we used minimum available resources.

Corresponding Author:

Eman Shehab
Department of Computer Science, Faculty of Computers and Artificial Intelligence, University of Sadat City
Sadat, Egypt
Email:eman.shehab@fcai.usc.edu.eg

1. INTRODUCTION

Numerous therapeutic candidates have recently, failed in late-stage clinical trials [1]. Chemical respiratory toxicity is a primary cause of clinical trial failures, and it has also resulted in the withdrawal of numerous medications from the market [2], [3]. Since there are no known adverse effects of medicinal products in human respiratory systems [4], [5], it is important to analyze potential for respiratory toxicity from compounds as soon as possible during drug discovery. Rapid screening of drug candidates is made possible by accurately anticipating the properties of drug molecules, which helps to save both time and money. Pharmacokinetic characteristics absorption, distribution, metabolism, excretion, and toxicity (ADMET) are major concerns during the screening stage of drug candidate [6], [7].

At an early stage of the drug development process, the ADMET property assessment approach can effectively address the issue of species differences, significantly increase the success rate and reduce the cost of drug discovery. The process of bringing Food and Drug Administration (FDA) medicine to market requires more than ten years and $200 million [8], [9]. The main cause of such high costs is medication safety, which accounts for 96% of therapeutic failures [10]. In the final stages of drug research, drug toxicity, and side effects are a crucial practiced challenge [11]–[14]. Therefore, to avoid high-cost consumption, predicting molecular toxicity should be performed as soon as possible during the development stage of a drug.

Machine learning techniques are becoming very popular in the pharmaceutical sector, which makes it possible to analyze a great deal of data available faster and easier [15]. Classification, regression, clustering, and pattern recognition are some of the main tasks carried out by artificial intelligence algorithms in a wide data set. New molecular characteristics, interactions, biological activities, and side effects of medicines are predicted by a wide variety of machine learning methods in the pharmaceutical industry [16]–[19].

According to Wang *et al.* [20], for predicting the chemical respiratory toxicity, six machine learning methods with nine types of molecular fingerprints have been used. According to Peng *et al.* [21], a novel method of molecular representation and developed the corresponding deep neural network framework, integrates designed data preprocessing techniques, a recurrent neural network (RNN) based on the bidirectional gated recurrent unit and fully connected neural networks for end-to-end molecular representation learning and chemical toxicity prediction. For predicting chemical toxicity, a graph convolution neural network have been developed and trains by mean teacher algorithm, based on the success of semi-supervised learning (SSL) algorithm [22]. According to Jaganathan *et al.* [23], using machine learning algorithms, and systematic tool selection methods to select features of the molecular describer sets, authors set up quantitative structure-activity relationship models. According to Huang [24], different conventional machine learning and deep learning algorithms have been applied to an online chemical database and model environment, creating a series of computational models. According to Feng *et al.* [25], to predict the compound's reproductive toxicity, ensembles of learning models have been developed using 9 molecular fingerprints, random forest (RF), extreme gradient boosting methods and support vector machine (SVM). Zhang *et al.* [26] studied a combination of deep neural networks with predictors based on the coherence forecasting framework in order to generate high probability models with clear uncertainties. Hua *et al.* [27] focusing on using machine learning and deep learning techniques to predict chemically induced hematotoxicity in silico, using QNPR descriptors and the random forest regression (RFR) and classification method.

The rest of this paper is shown as follows. Section 2 presented materials and data preparation, section 3 introduced method, section 4 presented results and discussion. Finally, the conclusion is presented in section 5.

2. MATERIALS AND DATA PREPARATION

2.1. Dataset

The dataset used in this study was collected by gathering positive data from the three datasets [20]. The first database is the Pneumotox database, which includes medicines that cause respiratory disorders [28]. Adverse drug reaction classification system (ADReCS) database that contains a wide range of undesirable reactions [29]. We've been focusing on the negative effects of medicines on the respiratory system. From the hazardous chemical information system, we selected substances that have a detrimental effect on respiratory systems (published on May 9, 2018). In addition, from the relevant literature we have obtained positively and negatively charged chemicals [30]–[35]. In the ChemIDplus database, all chemicals were matched to simplified molecular-input line-entry system (SMILES) [36].

2.2. Molecular representation

In this section, we discussed the most common techniques that are used for molecule representation, string representation and molecular graphs. String representations, with a large selection of sequence modeling techniques, e.g. RNN, attention mechanisms, and dilated convolutions, have been quickly adopted for generative models that represent chemical structures as a string. SMILES is the most often used string encoding for generative machine learning models [37]. The SMILES technique preserves atom and edge tokens while traversing the spanning tree of a chemical graph in depth-first order. Specific tokens for branching and edges that are not part of a spanning tree shall also be used by SMILES. Because there are multiple spanning trees on a molecule, numerous SMILES strings could be representing the same molecule. Is it possible to create the SMILES string uniquely from a molecule, its ambiguity can also enrich and improve generative models [38].

Molecular graphs, graph representations have been used in chemoinformatics for a long time to store and analyze biological data. Each node on the molecular graph represents the atom, and all edges represent a connection. In such a graph, the hydrogens can be specified directly or implicitly. The number of hydrogens can then be calculated based on their atomic value in this case.

3. METHOD

Figure 1 represents the general structure of the proposed model. It comprises of six phases: dataset gathering, data pre-processing, feature extraction, data splitting, learning models and evaluation. Details of dataset gathering and pre-processing aforementioned. The feature extraction phase has been implemented using the ngram term frequency-inverse document frequency (TF-IDF) based model. The dataset must be randomly split into train and test sets with a ratio of 8:2 respectively. Logistic regression (LR), SVM, RF, decision tree (DT), k-nearest neighbors (KNN), AdaBoost, multi-layer perceptron (MLP), and stochastic gradient descent (SGD) classifiers have been used to train the proposed model. The last phase is the evaluation of the performance of the proposed model. These procedures were put into practice in scikit-learn 0.24.1 [39].

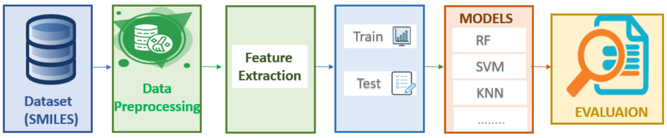


Figure 1. Shows the proposed model for predicting the chemical respiratory toxicity

3.1. Feature extraction

In machine learning algorithms, the TF-IDF with n-gram provides numerical weights to textual content for mining. It measures how important a term is within a compound relative to a collection of compounds. Choosing the right feature parameters can help to increase the classification accuracy of predecting toxicity [40].

3.2. Logistic regression

LR is a multivariate analytic discrete choice model. This approach is most commonly used for statistical analysis of biostatistics, sociology, quantitative psychology, clinical practice, econometrics, and marketing as well as to compare it with machine learning studies. It offers numerous benefits such as accuracy and strong power [41].

3.3. Support vector machine

SVM is a very commonly used algorithm of machine learning. It is capable of both linear and nonlinear classification [42]. We set the kernel parameter to Gaussian radial basis kernel function (RBF). By using a grid search, the parameters C and gamma have been determined.

3.4. Random forest

RF are an ensemble method of machine learning. RF generates a large number of DT randomly from the training set and can infer values for various DT to forecast their overall severity deficit [41]. The parameters used in the dataset are criterion (entropy), class weight (balanced), and n estimators (n=200).

3.5. Decision tree

This classifier has a tree structure, with internal nodes that represent features, branches that reflect decision mechanisms, and leaves representing the outcome. It is easy to understand since it mirrors the human thought process. This classifier's internal decision-making process is recognized for its tree structure. This can easily process multidimensional data and requires very little training time [43].

3.6. K-nearest neighbor

The KNN classification method is an easy and clean way of classification [44]. The parameters considered for KNN are n_neighbors(n=4) with a step size of 1. We set the parameter weights to distance.

3.7. AdaBoosting

AdaBoosting classifier is an ensemble strategy for machine learning that combines weak learner models and produces strong learners. When a machine learning system receives weights from training data, it acts as a base learner. We used sklearn.ensemble import AdaBoostClassifier to implement AdaBoost classifier. At first, a randomly training set will be used, and the model will be trained continually. In next iteration, misclassification observations are considered to have more weight and a greater likelihood. This technique will be continued as long as the data in the training database are not entirely aligned with the model [45].

3.8. Multi-layer perceptron

MLP is a powerful artificial neural network that mimics the operation of the human brain and map its collection of inputs and outputs accordingly. For the experiment, we import MLP classifier via sklearn.neural_network. It's parameters include activation function, distinct classes of deficit severity, input/output layers, learning rate, and iterations [45].

3.9. Stochastic gradient descent

SGD is a machine learning algorithm that performs the discriminative analysis of differential classification with loss functions. The benefit of adopting SGD is that it is simple to implement and increases efficiency. In order to assist in determining different loss functions and penalties, a SGD classifier is used for the implementation of SGD. The SGD classifier is created by importing SGD classifier from sklearn.linear_model [45].

3.10. Performance evaluation

This section explores a set of evaluation metrics for assessing the quality of generative models. Various metrics have been calculated such as false positives (FP), true positives (TP), false negatives (FN), true negatives (TN), recall, precision, F1-score, accuracy, and area under the curve (AUC) [46]. An additional instrument for assessing the effectiveness of machine learning classification the confusion matrix.

4. RESULTS AND DISCUSSION

This section display the results of the proposed models. Machine learning algorithms have been implemented using bigram and trigram TF-IDF based models. The RF and SVM achieved the highest results with trigram representation as shown in Table 1. All performance metrics have been calculated for each classifier. Precision values range from 0.73 to 0.84, F1-score values vary from 0.78 to 0.87, recall varies from 0.81 to 0.90 and the accuracy value is between 0.74 and 0.85. For bigram TF-IDF based models, RF, KNN, and SVM reported accuracy of 0.85, 0.81, and 0.80 respectively. As shown in Table 2, AUC scores for bigram models range from 0.72 to 0.84 for all algorithms. RF, KNN, and SVM outperformed all other models resulting AUC of 0.84, 0.80, and 0.79 respectively.

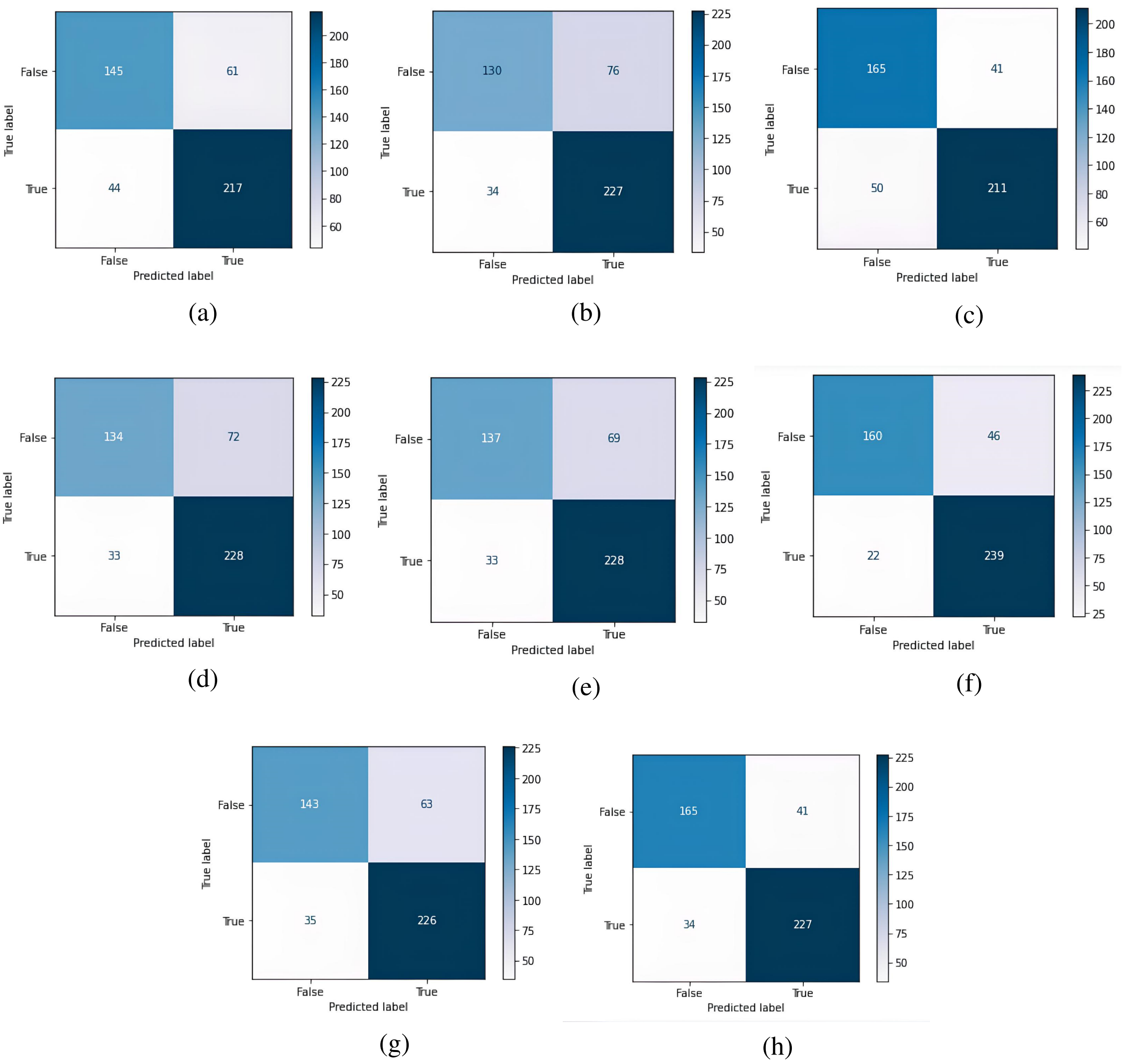
As shown in Table 1, different evaluation metrics have been calculated for all algorithms and trigram models. The precision score varies from 0.75 to 0.85, F1-score ranges from 0.80 to 0.87, recall varies from 0.81 to 0.92 and accuracy ranges from 0.76 to 0.85. RF and SVM with trigram model reported accuracy of 0.85 and 0.84 respectively. As shown in Table 2, AUC scores for trigram models range from 0.74 to 0.84 for all algorithms. RF and SVM outperformed all other models resulting in AUC of 0.84 and 0.83 respectively. Figures 2 and 3 show the confusion matrices and AUC for all algorithms with trigram model. Figures 2(a) to 2(h) represent the confusion matrix for Adaboost, DT, KNN, LR, MLP, RF, SGD, and SVM respectively. Figures 3(a) to 3(h) represent the AUC for Adaboost, DT, KNN, LR, MLP, RF, SGD, and SVM respectively.

Table 1. Results of the proposed system with eight classifiers in percentage with bigram and trigram

N-gram	Algorithms	Accuracy	Precision	recall	F1-Score
Trigram	Random forest	0.85	0.83	0.92	0.87
Bigram		0.85	0.84	0.90	0.87
Trigram	SVM	0.84	0.85	0.87	0.86
Bigram		0.80	0.81	0.85	0.83
Trigram	K-neighbors	0.81	0.84	0.81	0.82
Bigram		0.81	0.83	0.82	0.83
Trigram	AdaBoost	0.78	0.79	0.84	0.81
Bigram		0.79	0.79	0.84	0.81
Trigram	SGD	0.81	0.81	0.85	0.83
Bigram		0.77	0.76	0.84	0.80
Trigram	Decision tree	0.76	0.75	0.86	0.80
Bigram		0.75	0.75	0.81	0.78
Trigram	Logistic regression	0.78	0.76	0.87	0.81
Bigram		0.74	0.73	0.84	0.78
Trigram	MLP	0.77	0.75	0.87	0.81
Bigram		0.74	0.74	0.84	0.79

Table 2. Results of AUC of the proposed system with eight classifiers with bigram and trigram

N-gram	Algorithms	AUC
Trigram	Random forest	0.8408
Bigram		0.84428
Trigram	SVM	0.8353
Bigram		0.7947
Trigram	K-neighbors	0.8046
Bigram		0.80265
Trigram	AdaBoost	0.7763
Bigram		0.77876
Trigram	SGD	0.7985
Bigram		0.756
Trigram	Decision tree	0.7489
Bigram		0.728
Trigram	Logistic regression	0.76202
Bigram		0.722
Trigram	MLP	0.7523
Bigram		0.73073

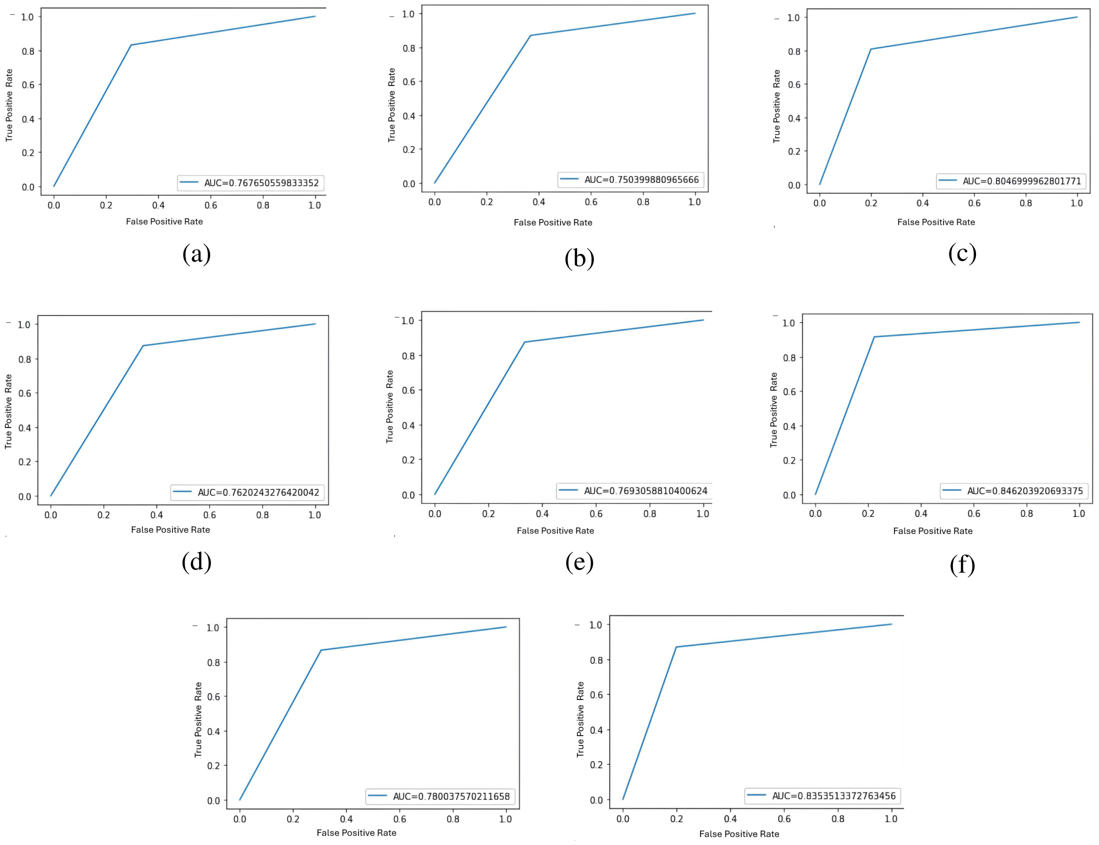


(a) (b) (c)

(d) (e) (f)

(g) (h)

Figure 2. Confusion matrix of (a) Adaboost, (b) DT, (c) KNN, (d) LR, (e) MLP, (f) RF, (g) SGD, and (h) SVM



(a) (b) (c)

(d) (e) (f)

(g) (h)

Figure 3. AUC of (a) Adaboost, (b) DT, (c) KNN, (d) LR, (e) MLP, (f) RF, (g) SGD, and (h) SVM

The results revealed that trigram produced higher performance than bigram. RF, SVM, and KNN classifiers outperformed all other classifiers. To our knowledge, the proposed model resulted in a considerable performance according to the resources that used to train the models. The reported results of the proposed models are close to the reported results in [20], [47].

5. CONCLUSION

In this research, machine learning techniques and N-gram model were integrated and used to classify chemical respiratory toxicity. For this purpose, different standard models were used for detailed analysis. We gathered chemicals associated with respiratory toxicity from various databases and the literature. This study used SVM, RF, LR, AdaBoost, DT, MLP, SGD, and KNN as classification algorithms. The proposed model have been evaluated using accuracy, confusion matrix, F1-score precision, and sensitivity. Results showed that the trigram model outperforms the bigram model. On the other hand, RF and SVM achieved the highest results in terms of accuracy with trigram representation. For future work, we will study to improve the model interpretability and prediction performance using different AI techniques and molecular properties. In addition, different representation models can be experimented with other learning models.

REFERENCES

[1] D. D. Martini, "Empowering phase II clinical trials to reduce phase III failures," *Pharmaceutical Statistics*, vol. 19, no. 3, pp. 178–186, May 2020, doi: 10.1002/pst.1980.

[2] R. K. Harrison, "Phase II and phase III failures: 2013–2015," *Nature Reviews Drug Discovery*, vol. 15, no. 12, pp. 817–818, Dec. 2016, doi: 10.1038/nrd.2016.184.

[3] G. Biala *et al.*, "Research in the field of drug design and development," *Pharmaceuticals*, vol. 16, no. 9, 2023, doi:10.3390/ph16091283.

[4] F. M. I. Hunter, A. P. Bento, N. Bosc, A. Gaulton, A. Hersey, and A. R. Leach, "Drug safety data curation and modeling in ChEMBL: boxed warnings and withdrawn drugs," *Chemical Research in Toxicology*, vol. 34, no. 2, pp. 385–395, Feb. 2021, doi: 10.1021/acs.chemrestox.0c00296.

[5] H. Zhang, J.-X. Ma, C.-T. Liu, J.-X. Ren, and L. Ding, "Development and evaluation of in silico prediction model for drug-induced respiratory toxicity by using naïve Bayes classifier method," *Food and Chemical Toxicology*, vol. 121, pp. 593–603, Nov. 2018, doi: 10.1016/j.fct.2018.09.051.

[6] V. Venkatraman, "FP-ADMET: a compendium of fingerprint-based ADMET prediction models," *Journal of Cheminformatics*, vol. 13, no. 1, Dec. 2021, doi: 10.1186/s13321-021-00557-5.

[7] H. Zhang, L. Zhang, C. Gao, R. Yu, and C. Kang, "Pharmacophore screening, molecular docking, ADMET prediction and MD simulations for identification of ALK and MEK potential dual inhibitors," *Journal of Molecular Structure*, vol. 1245, Dec. 2021, doi: 10.1016/j.molstruc.2021.131066.

[8] Y. Hua *et al*., "Drug repositioning: Progress and challenges in drug discovery for various diseases," *European Journal of Medicinal Chemistry*, vol. 234, Apr. 2022, doi: 10.1016/j.ejmech.2022.114239.

[9] A. B. Deore, J. R. Dhumane, R. Wagh, and R. Sonawane, "The stages of drug discovery and development process," *Asian Journal of Pharmaceutical Research and Development*, vol. 7, no. 6, pp. 62–67, Dec. 2019, doi: 10.22270/ajprd.v7i6.616.

[10] B. Shaker, S. Ahmad, J. Lee, C. Jung, and D. Na, "In silico methods and tools for drug discovery," *Computers in Biology and Medicine*, vol. 137, Oct. 2021, doi: 10.1016/j.compbiomed.2021.104851.

[11] O. Silakari and P. K. Singh, "ADMET tools: Prediction and assessment of chemical ADMET properties of NCEs," *in Concepts and Experimental Protocols of Modelling and Informatics in Drug Design*, Elsevier, 2021, pp. 299–320, doi: 10.1016/B978-0-12-820546-4.00014-3.

[12] W. Zhang, H. Zou, L. Luo, Q. Liu, W. Wu, and W. Xiao, "Predicting potential side effects of drugs by recommender methods and ensemble learning," *Neurocomputing*, vol. 173, pp. 979–987, Jan. 2016, doi: 10.1016/j.neucom.2015.08.054.

[13] W. Zhang, X. Yue, F. Liu, Y. Chen, S. Tu, and X. Zhang, "A unified frame of predicting side effects of drugs by using linear neighborhood similarity," *BMC Systems Biology*, vol. 11, no. S6, Dec. 2017, doi: 10.1186/s12918-017-0477-2.

[14] W. Zhang, X. Liu, Y. Chen, W. Wu, W. Wang, and X. Li, "Feature-derived graph regularized matrix factorization for predicting drug side effects," *Neurocomputing*, vol. 287, pp. 154–162, Apr. 2018, doi: 10.1016/j.neucom.2018.01.085.

[15] P. C.-Reboredo *et al*., "A review on machine learning approaches and trends in drug discovery," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 4538–4558, 2021, doi: 10.1016/j.csbj.2021.08.011.

[16] Y. Liu *et al*., "Experimental study and random forest prediction model of microbiome cell surface hydrophobicity," *Expert Systems with Applications*, vol. 72, pp. 306–316, Apr. 2017, doi: 10.1016/j.eswa.2016.10.058.

[17] P. R. -Fernandez, C. R. Munteanu, J. Dorado, R. M.-Romalde, A. D.-Sanchez, and H. G.-Diaz, "From chemical graphs in computer-aided drug design to general markov-galvez indices of drug-target, proteome, drug-parasitic disease, technological, and social-legal networks," *Current Computer Aided-Drug Design*, vol. 7, no. 4, pp. 315–337, Dec. 2011, doi: 10.2174/157340911798260340.

[18] P. Shirvani and A. Fassihi, "Molecular modelling study on pyrrolo[2,3- b]pyridine derivatives as c-Met kinase inhibitors: a combined approach using molecular docking, 3D-QSAR modelling and molecular dynamics simulation," *Molecular Simulation*, vol. 46, no. 16, pp. 1265–1280, Nov. 2020, doi: 10.1080/08927022.2020.1810853.

[19] B. S. -Garcia, J. I. B. -Bordils, A. Falcó, M. T. P. -Gracia, G. A. -Fos, and P. A.-López, "Quantitative structure–activity relationship methods in the discovery and development of antibacterials," *WIREs Computational Molecular Science*, vol. 10, no. 6, Nov. 2020, doi: 10.1002/wcms.1472.

[20] Z. Wang *et al*., "In silico prediction of chemical respiratory toxicity via machine learning," *Computational Toxicology*, vol. 18, May 2021, doi: 10.1016/j.comtox.2021.100155.

[21] Y. Peng, Z. Zhang, Q. Jiang, J. Guan, and S. Zhou, "TOP: A deep mixture representation learning method for boosting molecular toxicity prediction," *Methods*, vol. 179, pp. 55–64, Jul. 2020, doi: 10.1016/j.ymeth.2020.05.013.

[22] J. Chen, Y.-W. Si, C.-W. Un, and S. W. I. Siu, "Chemical toxicity prediction based on semi-supervised learning and graph convolutional neural network," *Journal of Cheminformatics*, vol. 13, no. 1, Dec. 2021, doi: 10.1186/s13321-021-00570-8.

[23] K. Jaganathan, H. Tayara, and K. T. Chong, "Prediction of drug-induced liver toxicity using SVM and optimal descriptor sets," *International Journal of Molecular Sciences*, vol. 22, no. 15, Jul. 2021, doi: 10.3390/ijms22158073.

[24] X. Huang, F. Tang, Y. Hua, and X. Li, "In silico prediction of drug-induced ototoxicity using machine learning and deep learning methods," *Chemical Biology & Drug Design*, vol. 98, no. 2, pp. 248–257, Aug. 2021, doi: 10.1111/cbdd.13894.

[25] H. Feng *et al*., "Predicting the reproductive toxicity of chemicals using ensemble learning methods and molecular fingerprints," *Toxicology Letters*, vol. 340, pp. 4–14, Apr. 2021, doi: 10.1016/j.toxlet.2021.01.002.

[26] J. Zhang, U. Norinder, and F. Svensson, "Deep learning-based conformal prediction of Toxicity," *Journal of Chemical Information and Modeling*, vol. 61, no. 6, pp. 2648–2657, Jun. 2021, doi: 10.1021/acs.jcim.1c00208.

[27] Y. Hua, Y. Shi, X. Cui, and X. Li, "In silico prediction of chemical-induced hematotoxicity with machine learning and deep learning methods," *Molecular Diversity*, vol. 25, no. 3, pp. 1585–1596, Aug. 2021, doi: 10.1007/s11030-021-10255-x.

[28] P. Camus, "The drug-induced respiratory disease website," *Pneumotox*, 2024. [Online]. Available: https://www.pneumotox.com/drug/index/

[29] M.-C. Cai *et al*., "ADReCS: an ontology database for aiding standardization and hierarchical classification of adverse drug reaction terms," *Nucleic Acids Research*, vol. 43, no. D1, pp. D907–D913, Jan. 2015, doi: 10.1093/nar/gku1066.

[30] S. Dik, J. Ezendam, A. R. Cunningham, C. A. Carrasquer, H. V. Loveren, and E. Rorije, "Evaluation of in silico models for the identification of respiratory sensitizers," *Toxicological Sciences*, vol. 142, no. 2, pp. 385–394, Dec. 2014, doi: 10.1093/toxsci/kfu188.

[31] J. Jarvis, M. J. Seed, S. J. Stocks, and R. M. Agius, "A refined QSAR model for prediction of chemical asthma hazard," *Occupational Medicine*, vol. 65, no. 8, pp. 659–666, Nov. 2015, doi: 10.1093/occmed/kqv105.

[32] G. R. Verheyen, E. Braeken, K. V. Deun, and S. V. Miert, "Evaluation of in silico tools to predict the skin sensitization potential of chemicals," *SAR and QSAR in Environmental Research*, vol. 28, no. 1, pp. 59–73, Jan. 2017, doi: 10.1080/1062936X.2017.1278617.

[33] S. J. Enoch, D. W. Roberts, and M. T. D. Cronin, "Mechanistic category formation for the prediction of respiratory sensitization," *Chemical Research in Toxicology*, vol. 23, no. 10, pp. 1547–1555, Oct. 2010, doi: 10.1021/tx100218h.

[34] M. A. Warne, J. K. Nicholson, J. C. Lindon, P. D. Guiney, and K. P. R. Gartland, "A QSAR investigation of dermal and respiratory

chemical sensitizers based on computational chemistry properties," *SAR and QSAR in Environmental Research*, vol. 20, no. 5–6, pp. 429–451, Jul. 2009, doi: 10.1080/10629360903278768.

[35] R. C. Braga *et al.*, "Pred-Skin: A fast and reliable web application to assess skin sensitization effect of chemicals," *Journal of Chemical Information and Modeling*, vol. 57, no. 5, pp. 1013–1017, May 2017, doi: 10.1021/acs.jcim.7b00194.

[36] P. Tomasulo, "ChemIDplus-super source for chemical and drug information," *Medical Reference Services Quarterly*, vol. 21, no. 1, pp. 53–59, Jan. 2002, doi: 10.1300/J115v21n01_04.

[37] F. Grisoni, "Chemical language models for de novo drug design: Challenges and opportunities," *Current Opinion in Structural Biology*, vol. 79, Apr. 2023, doi: 10.1016/j.sbi.2023.102527.

[38] J. A. -Pous *et al.*, "Randomized SMILES strings improve the quality of molecular generative models," *Journal of Cheminformatics*, vol. 11, no. 1, Dec. 2019, doi: 10.1186/s13321-019-0393-0.

[39] "Supervised learning," *Scikit Learn*, 2023. [Online]. Available: https://scikit-learn.org/stable/supervised_learning.html

[40] M. Taha, H. H. Zayed, M. Azer, and M. Gadallah, "Automated COVID-19 misinformation checking system using encoder representation with deep learning models," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 12, no. 1, pp. 488-495, Mar. 2023, doi: 10.11591/ijai.v12.i1.pp488-495.

[41] C.-C. Wu *et al.*, "Prediction of fatty liver disease using machine learning algorithms," *Computer Methods and Programs in Biomedicine*, vol. 170, pp. 23–29, Mar. 2019, doi: 10.1016/j.cmpb.2018.12.032.

[42] J. Cervantes, F. G.-Lamont, L. R.-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189–215, Sep. 2020, doi: 10.1016/j.neucom.2019.10.118.

[43] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 1, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.

[44] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of k-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning," *Decision Analytics Journal*, vol. 3, Jun. 2022, doi: 10.1016/j.dajour.2022.100071.

[45] G. Sambasivam, J. Amudhavel, and G. Sathya, "A predictive performance analysis of vitamin D deficiency severity using machine learning methods," *IEEE Access*, vol. 8, pp. 109492–109507, 2020, doi: 10.1109/ACCESS.2020.3002191.

[46] M. Morgan, C. Blank, and R. Seetan, "Plant disease prediction using classification algorithms," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 10, no. 1, pp. 257-264, Mar. 2021, doi: 10.11591/ijai.v10.i1.pp257-264.

[47] K. Jaganathan, H. Tayara, and K. T. Chong, "An explainable supervised machine learning model for predicting respiratory toxicity of chemicals using optimal molecular descriptors," *Pharmaceutics*, vol. 14, no. 4, Apr. 2022, doi: 10.3390/pharmaceutics14040832.

BIOGRAPHIES OF AUTHORS

Eman Shehab is currently a Lecturer Assistant in the Department of Computer Science, Faculty of Computers and Artificial Intelligence, University of Sadat City, Egypt, since 2020. She hols a bachelor of computer science. She achieved her master's degree from Menoufia University. She has worked on several research topics, her research interests are neural network and machine learning. She can be contacted at email: eman.shehab@fcai.usc.edu.eg.

Hamada Nayel is an Assistant Professor at the Department of Computer Science, Faculty of Computers and Artificial Intelligence, Benha University, Benha, Egypt. In 2019, he received his Ph.D. from Mangalore University, India. His research interests include Arabic NLP, biomedical NLP, and social media analysis. He can be contacted at email: hamada.ali@fci.bu.edu.eg.

Mohamed Taha is an Associate Professor at Department of Computer Science, Faculty of Computers and Artificial intelligence, Benha University, Egypt. He received his M.Sc. degree and his Ph.D. degree in computer science at Ain Shams University, Egypt, in February 2009 and July 2015. He is the founder and coordinator of "Networking and Mobile Technologies" program, Faculty of Computers and Artificial Intelligence, Benha University. His research interest's concern: computer vision (object tracking and video surveillance systems), digital forensics (image forgery detection, document forgery detection, and fake currency detection), image processing (OCR), computer network (routing protocols and security), augmented reality, cloud computing, and data mining (association rules mining and knowledge discovery). He has contributed more than 30+ technical papers in international journals and conferences. He can be contacted at email: mohamed.taha@fci.bu.edu.eg.