ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

Identifying High-Quality Leads among Screened Anticancerous Compounds Using SMILES Representations

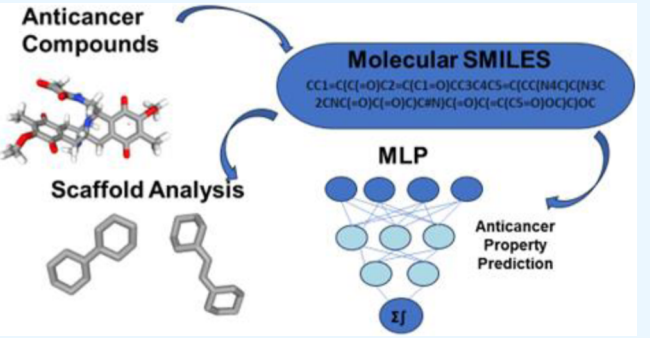
Swathik Clarancia Peter, Yogesh Kalakoti, and Durai Sundar*

Cite This: *ACS Omega* 2024, 9, 30645−30653

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Cancer is a lethal disease that affects numerous people worldwide. Chemotherapy stands as one of the most effective treatment regimens to combat cancer. Nevertheless, anticancer drugs face a high failure rate due to safety and efficacy issues. Drug failure could be subdued by instigating drug leads with reduced toxicity and enhanced efficacy. Computer-aided drug discovery endorses drug leads in manoeuvring protein and ligand structures or representations. Simplified molecular input line entry system (SMILES) is a linear notation representing the three-dimensional structure of a molecule using symbols and alphanumeric characters. SMILES representation hoards rings and scaffold structures in its depiction. Mining ring and scaffold patterns from molecular SMILES would assist in ascertaining biological properties based on molecular patterns. Moreover, the emergence of artificial intelligence (AI) technologies would accelerate identification of efficient anticancer drug leads. AI algorithms proclaimed for their pattern recognition ability could be employed for identifying molecular patterns from SMILES representation, thereby enabling property prediction. Consequently, we developed a multilayer perceptron (MLP) model for the prediction of anticancer activity using SMILES of NCI-60 cancer growth inhibition data. Furthermore, the top 8 frequent scaffolds were identified on preliminary analysis of cancer growth inhibition data and ChEMBL drugs. The developed MLP model classified anticancer and nonanticancer compounds with a classification accuracy of 0.92. Also, benchmarking of the developed model with machine learning algorithms exhibited better performance of the MLP model.



1. INTRODUCTION

Cancer is a genetic disease characterized by uncontrolled proliferation of cells due to aberrations in genes,[1,2] mRNAs, miRNAs,[3] proteins,[4] and metabolites.[5] The key processes that are involved in tumor pathogenesis are referred to as hallmarks of cancer that include cell death resistance, deregulation of cellular metabolism, sustenance of proliferative signaling, obstructing immune destruction, inducing vasculature, replicative immortality, evading growth suppressors, activation of invasion and metastasis, and so forth. Treatment strategies include chemotherapy, which is one of the most common therapies recommended to cancer patients by clinicians. Chemotherapy mainly involves targeting the hallmarks of cancer with chemical molecules to inhibit tumor pathogenesis. Drug molecules inhibiting these key processes include cyclin-dependent kinase inhibitors, epidermal growth factor receptor inhibitors, telomerase inhibitors, vascular endothelial growth factor signaling inhibitors, and so forth. Despite numerous types of inhibitors, chemotherapy faces a high drug failure rate due to low efficacy and safety issues. Nonspecific activity and side effects are some of the major reasons for anticancer drug attrition at different phases of clinical trials. Hence, drug leads with more specificity and less toxicity need to be explored.

Moreover, the conventional drug development process is time-consuming and laborious. To accelerate the identification of potential anticancer drug leads, robust drug discovery methods must be implemented. After the boom of artificial intelligence (AI) technologies, various aspects of the drug development process have been impacted by AI algorithms, viz., lead identification, target identification, survival prediction, and so forth. AI algorithms known for their robust pattern identification through abstract and hierarchical data representations have been deployed in drug discovery research.[6] Multilayer perceptron (MLP) is a form of deep neural network with one or more hidden layers. It is primarily composed of three layers, viz., an input layer, hidden layer(s), and an output layer. MLP is a feed-forward neural network that uses a backpropagation algorithm to minimize the error in prediction.

Received: March 22, 2024
Revised: June 10, 2024
Accepted: June 13, 2024
Published: June 28, 2024

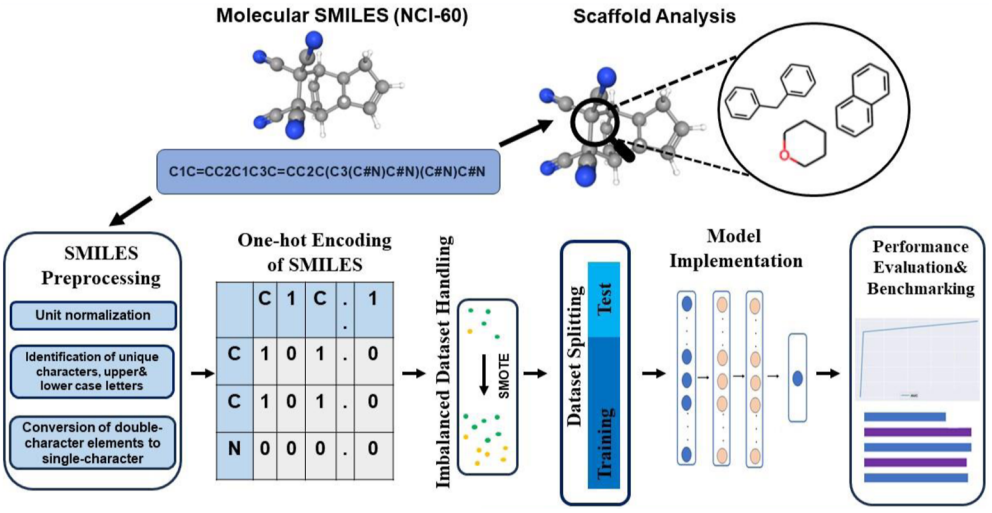


Figure 1. Schematic representation of molecular scaffold analysis and MLP model development.

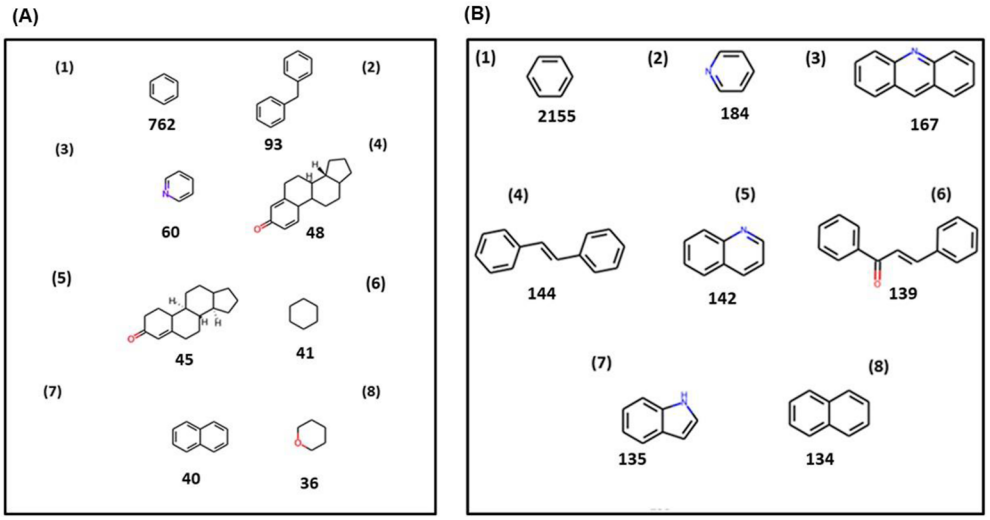


Figure 2. Top 8 frequent scaffolds occurring in (A) ChEMBL drug molecules and (B) NCI-60 screened compounds. Number in the bracket refers to as the order of frequency of occurrence. Number below each substructure refers to as the frequency of scaffold occurring in the data set.

MLP models could maneuver complex nonlinear problems and are also robust for large data sets.

In this study, we propose an MLP-based model to predict the anticancer properties of chemical compounds using simplified molecular input line entry system (SMILES). Chemical SMILES are linear notation of molecules and are represented by a combination of alphabets (symbols of atomic elements), numbers, and special characters. Atoms, bonds, branches, and rings and cyclic structures are depicted in SMILES. These linear notations are rendered by applying graph theory on molecules.[7,8] Rings, cyclic structures, and other molecular patterns are depicted in the SMILES representation.

Molecular SMILES have been employed in substructure searching and quantitative structure−activity relationship (QSAR) predictions[9−16] SMILES-based descriptors have also been used for property prediction. For instance, SMILES representation of quantum-chemical data set was used for prediction of nine molecular properties including heat capacity, HOMO energies, dipole moment, electronic polarizability, energy separation between HOMO and LUMO states and four other thermodynamic properties.[17] Similarly, the quantitative structure−activity (QSAR) model for toxicological carcinogenicity prediction of drugs using TD_{50} values has been developed using SMILES-based descriptors.[18] Furthermore, QSAR models for bioactivity prediction of aromatase inhibitors using SMILES have also been developed.[9] Employing SMILES alleviates the development of property prediction models devoid of using descriptor selection and molecular geometrical optimization methods. Developed MLP model would predict anticancer activity using SMILES information on chemical compounds. Also, benchmarking of the developed MLP model depicted better performance than compared machine learning algorithms. Schematic representation of scaffold analysis and model development has been summarized in Figure 1.

2. RESULTS

2.1. Frequent Scaffold Analysis. Observing the vast chemical space of drugs reveals that certain rings and cyclic structures are privileged and also that entry of new ring systems in drug space is less common.[19] Likewise, analysis of

https://doi.org/10.1021/acsomega.4c02801
ACS Omega 2024, 9, 30645−30653

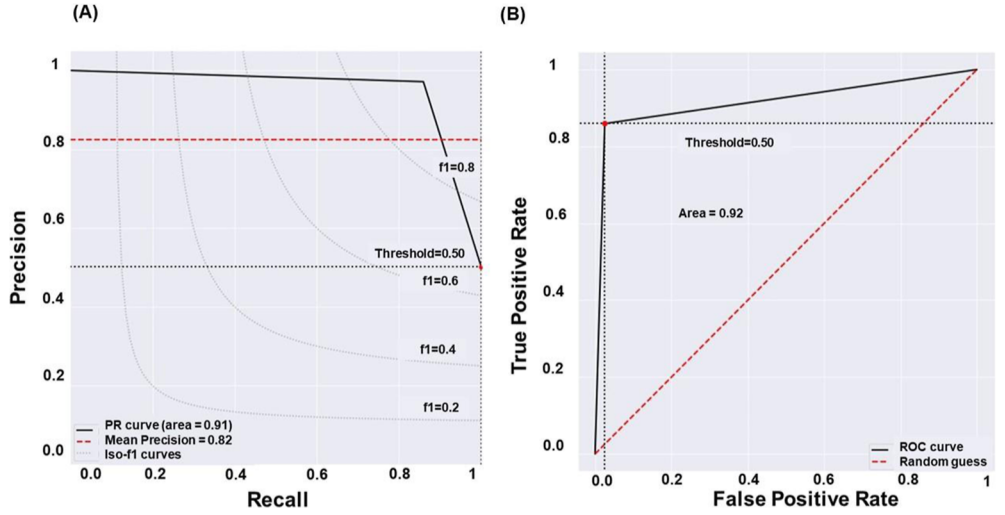


Figure 3. Precision−recall (PR) plot and ROC curve. (A) PR curve plotted with precision and recall values of the MLP model and (B) ROC curve showing AUC of 0.92.

ring replications in the drug space emphasizes novel configurations of existing drug molecules rather than a completely novel molecule.[20] Analysis of frequent scaffolds occurring in the ChEMBL drug molecules and NCI-60 cancer data set revealed top 8 frequent scaffolds occurring in both the data sets (Figure 2). The frequent occurrence of these substructures indicated a recurrence of molecular scaffolds in drug-like molecules. Frequency of the most common scaffold occurring in CHEMBL data set was 762, whereas it was 2155 in the NCI-60 data set. Similarly, frequency of the eighth most commonly occurring scaffold was 36 and 134 in CHEMBL and NCI-60 data sets, respectively. Replication of rings and scaffolds implied inclination of drug space to certain structures. In other words, there was preference in molecular patterns in drug space.

2.2. MLP Model for Anticancer Property Prediction. The proposed multi-layer perceptron model has one input layer, two hidden layers, and one output layer for training of molecular SMILES to identify patterns and predict anticancer property. SMILES notation represents molecules in a linear string of alphabets and symbols. One-hot encoded SMILES matrix of the NCI-60 cancer growth inhibition data set was padded with a maximum length of 421 characters (i.e., length of the longest SMILES in the NCI-60 data set). Prior to training, data set was processed to harmonize class imbalance. NCI-60 data set was imbalanced with a greater number of active compounds. Imbalance in the data set was handled using the Synthetic Minority Oversampling TEchnique (SMOTE) oversampling technique. SMOTE generates synthetic data points for minor class (i.e., nonanticancer), thereby reducing the class imbalance and improving the model performance. The MLP model was optimized by hyperparameter tuning. Extensive and grid search methods were used for the hyperparameter search. Kernel initialization assigns random weights to initialize the network. HeNormal was used for kernel initialization, which initializes weights using a truncated normal distribution. The learning rate scheduler tunes the learning rate during the training iterations. Decay learning rate set reduces the learning rate by a factor of 0.96. Early stopping regularizes the network based on the monitored metric, which avoids overfitting. During early stopping, validation loss was monitored across training iterations. ReLU and Sigmoid activation functions were used for hidden and output layers, respectively. Adam optimization algorithm was used to minimize the loss function. Cross-entropy function calculates loss between actual and predicted labels. Binary cross-entropy loss function was used in the developed MLP model. The proposed MLP model classifies the data into binary labels anticancer and nonanticancer.

2.3. MLP Model Performance. The MLP model trained with NCI-60 data set had an accuracy of 92% and Matthew's correlation coefficient (MCC) of 0.84. Precision−recall (PR) and receiver operating characteristic (ROC) curves were generated (Figure 3). The PR curve (Figure 3A) gives a trade-off between precision and recall at the threshold of 0.5. Area under the PR curve was 0.91 that showed higher precision with low false positive rate (FPR) and higher recall with a low false negative rate and had a mean precision of 0.82. The ROC curve (Figure 3B) shows an AUC of 0.92. The ROC curve was plotted with true positive rate, TPR (recall) and FPR. Area under the ROC curve (AUROC) gives an aggregate performance measure of the model. AUC ranges from 0 to 1, where 0 refers to as the incorrect classifier and 1 refers to as the perfect classifier. AUC of the MLP model was 0.92, confirming better performance in classifying anticancer and nonanticancer classes.

Performance metrics like precision, recall, and F1-score were calculated for developed MLP and compared to other machine learning algorithms (Table 1). The performance metrics for anticancer and nonanticancer class are shown in (Figure 4). The precision score is given as the ratio of true positives (TP)

Table 1. Overall Performance Metrics for Both Classes of MLP and Compared Algorithms

Algorithm	precision	recall	F1-score	MCC	accuracy
MLP	0.97	0.87	0.92	0.84	0.92
SVC	0	0	0	0	0.50
logistic regression	0	0	0	0	0.50
decision tree	1	0.07	0.13	0.18	0.54
AdaBoost	0.86	0.83	0.85	0.70	0.84
CatBoost	0.90	0.89	0.91	0.81	0.90
XGBoost	0	0	0	0	0.50

https://doi.org/10.1021/acsomega.4c02801
ACS Omega 2024, 9, 30645−30653

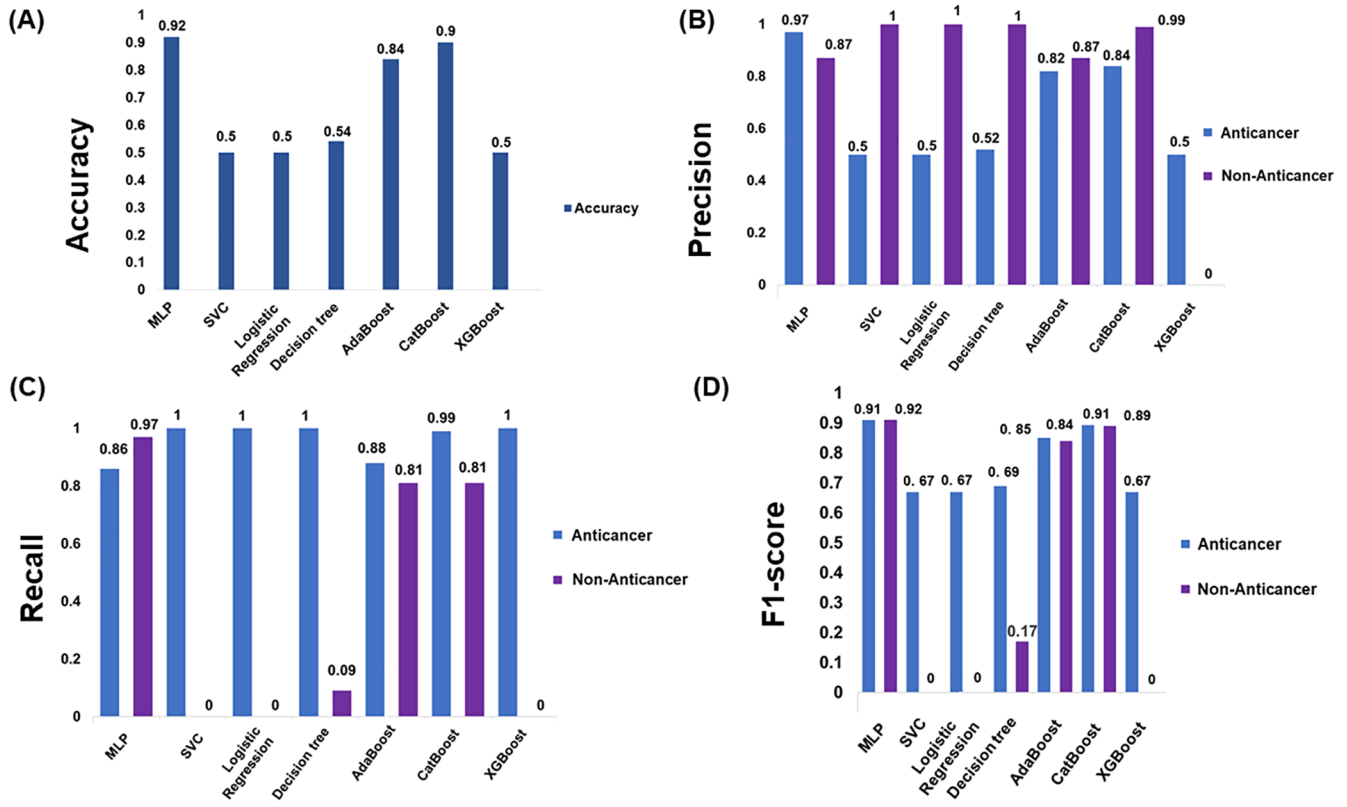


Figure 4. Performance comparison of the MLP model with other machine learning algorithms (A) accuracy, (B) precision, (C) recall, and (D) F1-score.

to the sum of TP and false positives (FP). Higher precision implies that the classifier makes accurate predictions. Precision scores of anticancer and nonanticancer class were 0.97 and 0.87, respectively. Recall score is given as the ratio of TP to the sum of TP and false negatives (FN). The MLP model had recall scores of 0.86 and 0.97 for anticancer and nonanticancer classes, respectively. F1-score is calculated as the harmonic mean of precision and recall scores. Unlike accuracy, F1-score measures the predictive ability of the model based on the class-wise performance. F1 score values range from 0 to 1, where 0 represents poor and 1 represents good classifier, respectively. The MLP model had F1-score of 0.91 and 0.92 for anticancer and nonanticancer classes, respectively.

Confusion matrix gives information about correct predictions and model errors. It is a 2 × 2 matrix calculated based on TPs, FPs, true negatives (TN), and FNs. The matrix represents actual and predicted values or labels. The MLP model classified 22,804 compounds (0.97) correctly as positive, which are truly anticancer (TPs). Similarly, 20,218 compounds (0.86) were correctly classified by the model as negative which are truly nonanticancer (TNs). FPs (type I error) and FNs (type II error) classified by the model were 705 (0.3) and 3291 (0.14), respectively (Figure 5).

2.4. Model Comparison with Machine Learning Algorithms and Cross-Validation. The MLP model was benchmarked against the support vector machine, logistic regression, decision tree, AdaBoost, CatBoost, and XGBoost algorithms. Performance comparison of the MLP model with other algorithms are given in Table 1. Comparison of the MLP model against boosting and other classification algorithms depicted better performance of the MLP model. Performance metrics viz., precision, recall, and F1-score were compared

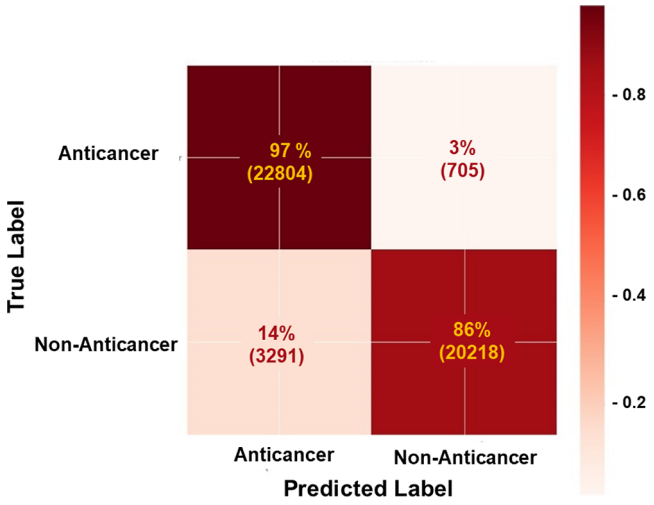


Figure 5. Confusion matrix for the MLP model showing true labels and predicted labels classified by the MLP model.

(Figure 4). Among the algorithms compared with MLP, CatBoost had good performance with a classification accuracy of 0.90 followed by AdaBoost with an accuracy of 0.84. However, MLP outperformed all other compared algorithms with a classification accuracy of 0.92. Precision scores of MLP for anticancer and nonanticancer class were 0.97 and 0.87 which were higher to all compared algorithms, except for AdaBoost and CatBoost algorithms. AdaBoost had similar precision score (0.87) for negative class, whereas CatBoost had a higher precision score of 0.99 for nonanticancer class that is higher than MLP. XGBoost had the lowest precision score of 0

https://doi.org/10.1021/acsomega.4c02801
ACS Omega 2024, 9, 30645−30653

for nonanticancer class. The MLP model had recall scores 0.86 and 0.97 for anticancer and nonanticancer classes. F1-score of MLP for both classes were higher (0.91 and 0.92). F1-scores of nonanticancer classes of SVC, logistic regression, and XGBoost were 0. Performance metrics were also calculated separately for anticancer and nonanticancer classes of MLP and compared algorithms (Table S1). Furthermore, MLP was cross-validated to avoid overfitting and to evaluate the generalization capacity of the model on unseen data. k-fold cross-validation of the MLP model was carried out in five validation folds ($k = 5$). Performance metrics of MLP K-fold cross-validation were computed. Five-fold cross-validation carried out resulted with mean training and validation accuracy of 0.89 (89.44) and 0.89 (89.38), respectively (Table S2). Also, the MLP-based model showed better performance compared to existing methods. CDRUG server was not accessible. Hence, AUROC reported earlier was taken for performance comparison.[21] MLP had an AUC of 0.92 higher than the reported CDRUG AUROC of 0.88. pdCSM when tested with compounds screened for breast cancer had an AUROC of 0.49. Further evaluation of the MLP model carried out with independent ChEMBL data set resulted in AUROC of 0.60. Despite lower performance on independent data set, AUROC of the MLP model is better than AUROC of pdCSM.

3. DISCUSSION AND CONCLUSIONS

Computational techniques for pattern identification, substructure searching, and molecular property prediction have been widely applied in computer-aided drug discovery and QSAR. Scaffold analysis identified frequently occurring scaffolds in the NCI-60 growth inhibition data and the ChEMBL32. Scaffolds are core structures and building blocks of chemical molecules. Identification of scaffolds assists in an intuitive understanding of shape features or molecular patterns. Scaffold analysis leverages frequent occurrence of certain molecular patterns in a particular pharmacological or bioactivity space. Inevitably this suggests the functional relevance of molecular patterns as well as the need to be identified and privileged for the clinical success of drugs. AI algorithms recognize and learn patterns from trained data and predict similar patterns in newly trained or untrained data. MLPs have been widely implicated to solve nonlinear problems and known for its application in pattern identification studies.[22] Multilayer artificial neural networks with back-propagation algorithm for error rate optimization were implemented for pattern recognition.[23] In MLPs, pattern recognition tasks are carried out in three layers, viz., input, hidden, and output layers. Hidden layers of MLP allow deciphering complicated patterns via nonlinear transformation of the data set. Besides hierarchical representations of data learned in hidden layers assist in capturing interpretable intermediate patterns in the data set.[24] The foremost hidden layer receives inputs (I_n) from the input layer with associated weights (w_n) and is followed by inclusion of bias (b) to generate output signals to be forwarded in the network. Output of the hidden layer (H_1) can be given as

$$H_1 = I1 \times w1 + I2 \times w2 + I3 \times w3 + \cdots + In \times wn + b \quad (1)$$

Propagation of output signals from hidden layers and output prediction based on the recognized pattern in the output layer are determined by activation functions used in these layers.[25] In QSAR, MLPs were used in the prediction of toxicity,[26] bioactivity for antibreast cancer drug development,[27] blood–brain barrier permeability for discovery of CNS (central nervous system) therapeutic drugs,[28] and so forth. Rings, cyclic structures, and other molecular patterns are encompassed in the SMILES representation. Like the word syntax in natural languages, SMILES harbors patterns or scaffolds syntactically in terms of molecular organization, symmetry, and topology.[7] These patterns might be extrapolated to the pharmacological or bioactivity space. Learned patterns via molecular SMILES in relevant anticancer drug space could assist in identifying inherently favored anticancer leads. Consequently, identified high-quality anticancer leads would assist in the development of efficient drugs with potent anticancer activity.

Pattern recognition can be performed in different types of data representations, viz., text, image, video, and so on. SMILES are textual notation of molecules obtained based on molecular graph theory and harbor ring and cyclic structures in its depiction. Exploiting different learning algorithms on molecular SMILES enables the identification of underlying patterns in the molecules. Implying the similarity structure principle, the properties of unknown molecules could be unveiled using machine learning algorithms. SMILES of chemical compounds were one-hot encoded to train the MLP model for prediction of anticancer activity. One-hot encoding transforms string data into numerical values, easing application of learning algorithms for training. Encoding data using multidimensional binary vectors becomes advantageous when the relationship between data points is not ordinal. One-hot encoding is a vital text preprocessing step in many NLP-related tasks. Simple one-hot vector representations enable straightforward implementation providing nuance predictions by including all unique categories or texts or strings. Hence, distinctive patterns could be ideally identified, allowing precise predictions. Moreover, training using one-hot vector representations is computationally less expensive compared with other embedding representations. Most commonly, biological data sets are imbalanced. Training of machine and deep learning models using imbalanced data might lead to classification bias.[29,30] Hence, to subdue class imbalance different sampling techniques were being implemented viz., oversampling,[31] undersampling,[32] hybrid sampling,[33] and so forth. SMOTE is a robust oversampling technique used to overcome class imbalance and depends on the pattern augmentation.[34] Synthetic data instances generated by SMOTE in minority class intensifies the class boundaries by boosting minority patterns, thereby increasing classification accuracy.[35] SMOTE was used in combination with different algorithms to handle imbalance data sets so as to improve the classification accuracy.[36,37] Employing SMOTE oversampling enriches the data set to be variegated, thereby reducing overfitting and improving the generalization capacity of the model with better performance metrics.[38] Moreover, oversampling the data sets using SMOTE reduces classification bias toward the highly populated class in the real data. However, noise introduced due to synthetic data points generated during SMOTE oversampling might also lead to overfitting, furthermore affecting the generalization ability of the model. Accordingly, hyperparameter optimization and k-fold cross-validation were implemented to curb overfitting and improve the generalization ability. Consequently, the MLP model developed in this study for anticancer property prediction had good performance with a classification accuracy of 92%. According to the similar property principle, molecules which

https://doi.org/10.1021/acsomega.4c02801
ACS Omega 2024, 9, 30645−30653

are structurally similar would possess similar biological property.[39] SSP has been implied in various property prediction studies using fingerprints, substructures, and other parameters. We suggest that the proposed MLP model would predict the anticancer activity of any other molecules by identifying the presence of patterns in SMILES pertinent to anticancer activity. As the model is built using SMILES representation, it could handle diverse molecular structures represented in the textual form. Generalization ability evaluated using k-fold cross-validation highlights the propitious performance of the MLP model on unseen data. Also, comparative assessment indicated better performance of the developed MLP model over other similar algorithms and existing methods. However, evaluation of the MLP model on independent data set showed that performance of MLP to be not exorbitant suggesting for improvement to achieve better model performance. We recommend exploring more robust algorithms for data set sampling and model training for enhancement of the model performance. Additionally, different descriptors could be utilized for anticancer property prediction, exploiting varying types of molecular representations using machine and deep learning algorithms. This would accelerate anticancer drug discovery research by assisting in the identification of prominent anticancer drug leads. Hence, we also suggest that anticancer property prediction can become more effective by deploying AI learning algorithms on various types of molecular data representations or descriptor information.

4. METHODS

4.1. Data Set Retrieval and Preprocessing. Molecular SMILES and bioactivity values (IC_{50}) of chemical compounds tested against different cancer cell lines were retrieved from NCI-60 growth inhibition data [NCI-60 Growth Inhibition Data, NCI Developmental Therapeutics Program (DTP) Data, NCI Wiki (nih.gov)] (accessed on 24/06/2023). NCI-60 data embody compounds screened against various cancer cell lines to identify molecules with potential anticancer activity either by inhibiting or killing cancer cells. Human cancer cell lines screened are representatives of different types of cancers, which include blood cell (leukemia), skin (melanoma), breast, lung, brain, kidney, ovary, prostate, and colon cancers. A list of NCI-60 human cancer cell lines used for in vitro screening of compounds can be found here [Cell Lines in the In Vitro Screen | NCI-60 Human Tumor Cell Lines Screen | Discovery & Development Services | DTP (cancer.gov)]. End points of anticancer activity screens of 47,196 compounds measured in terms of IC_{50} and their SMILES were used for training the MLP model. IC_{50} is a measure of molecular activity that is defined as the half-maximal inhibitory concentration and is used to measure drug efficacy. The retrieved IC_{50} units were normalized to μg/mL. Compounds with bioactivity value of $\leq$500 μg/mL were labeled as active, implying to exhibit cancerous activity. Similarly, compounds with activity values $\geq$500 μg/mL were labeled as inactive, implying non-anticancerous nature. Subsequently, drug molecules from ChEMBL database (https://www.ebi.ac.uk/chembl/) (accessed on 04/03/2023) were downloaded. A total of 14,293 drug molecules screened against different ailments along with SMILES information were obtained. This data set was preprocessed by removing null values prior to the model training and scaffold analysis. Furthermore, SMILES of 3144 small molecule drugs with anticancer activity were curated to perform evaluation of the MLP model on independent data set.

4.2. Scaffold Analysis of NCI-60 and ChEMBL Molecules. Scaffolds are core structures of chemical molecules obtained after removing R-group substituents with only aliphatic linkers between ring systems preserved.[40] Molecular scaffolds are privileged in drug space, and entry of new ring systems into chemical space of drugs is low.[19] Scaffold analysis of NCI-60 and ChEMBL molecules was carried out. Most frequently occurring Murcko scaffolds were identified using molecular SMILES data. Scaffold frequency is given as the number of unique scaffolds present in the data set along with number of molecules in which it is present.

4.3. One-Hot Encoding of Molecular SMILES. Encoding the SMILES representation of molecules in the form of ones and zeros is termed as a one-hot encoding of SMILES (Figure 6). To enable one-hot encoding, NCI-60 molecular SMILES

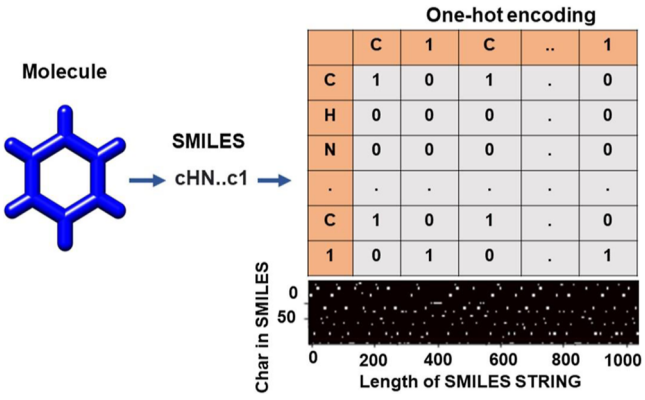


Figure 6. One-hot encoding representation of molecular SMILES.

were identified for unique characters. Upper-case and lower-case letters were also identified, and linear combination was generated to identify the double character elements. Double character elements in the data set were converted to single character. All double characters in the data set were replaced to single character except for "Cn" and "Sc". The aromatic carbon and nitrogen could be depicted as Cn, so it is not replaced. "Sc" represents Scandium which is rarely present in the drug molecules. Besides "Sc" could represent sulfur and carbon. Hence, "Cn" and "Sc" were not replaced. The list of replaced characters in NCI growth inhibition data is given in Table 2. Length of longest SMILES in NCI-60 compounds was identified for padding which is 421 characters. One-hot matrix was constructed with unique characters in the NCI-60 SMILES data set.

4.4. Imbalanced Data Set Handling and Data Set Splitting. The NCI-60 data set has 47,017 compounds with anticancer activity (anticancer) and only 179 molecules with no anticancer activity (nonanticancer) (Table S3). Number of active compounds is higher compared to that of inactive compounds. Hence, there is a class imbalance between active and inactive classes. Training the model with an imbalanced data set would result in a biased model by favoring active class and overlooking inactive class. To overcome imbalance in the data set, oversampling technique SMOTE was employed. k_neighbors parameter was set to 5. After over sampling, the NCI-60 data set was trained and evaluated with a 25% independent test split.

https://doi.org/10.1021/acsomega.4c02801
ACS Omega 2024, 9, 30645−30653

Table 2. List of Double Character Elements Replaced by Single Character in NCI Growth Inhibition Data

double character	replaced single character	double character	replaced single character
Ac	J	Mg	ρ
Ag	Q	Mn	σ
As	X	Mo	τ
Au	J	Na	υ
Ba	K	Nb	φ
Bi	P	Nd	χ
Br	Q	Ni	ψ
Ca	V	Os	ω
Cd	W	Pb	ς
Ce	X	Pd	Γ
Cl	Z	Pt	Θ
Co	A	Re	Ξ
Cr	B	Rh	Π
Cu	Γ	Ru	Σ
Dy	Δ	Sb	Φ
Er	E	Se	Ψ
Eu	Z	Si	Ω
Fe	H	Sm	Á
Ga	Θ	Sn	É
Gd	I	Ta	Í
Ge	K	Te	Ó
Hf	Λ	Th	Ö
Hg	M	Ti	Ő
In	N	Tl	Ú
Ir	Ξ	Zn	Ü
La	É	Zr	Ű
Li	Π		

4.5. MLP Model Implementation and Evaluation. The MLP model was constructed with one input layer, two hidden layers, and output layer. Flowgraph of the MLP model is given in Figure 3. The input layer comprises 421 neurons, whereas two hidden layers comprise 421 and 25 neurons by first and second hidden layers, respectively. The output layer has one neuron (Figure 7). The activation function used in hidden layers was ReLU, and in the output layer, the Sigmoid function was used for activation. Kernel initialization was performed using the He normal initializer. The learning rate decay schedule was used for tuning the learning rate during training of the network, and the decay rate was set to 0.96. Early stopping was used to avoid overfitting. Adam optimization function was used, and the epsilon was set to 1. Hyperparameters were tuned by using manual and grid search methods. Hyperparameter sets for the MLP network are given in Table 3.

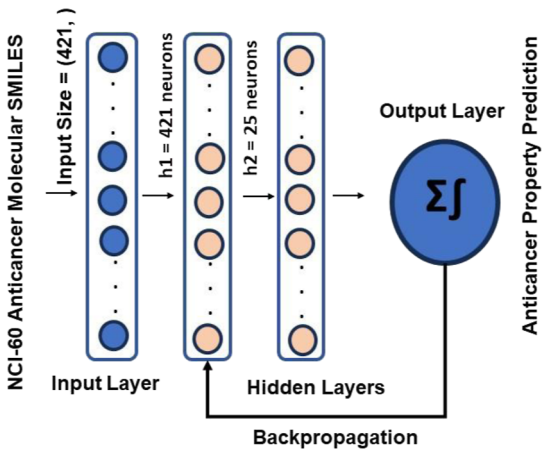


Figure 7. MLP model architecture for anticancer property prediction comprising an input layer, two hidden layers, and an output layer. Error minimization is performed using backpropagation algorithm.

Table 3. Hyperparameters of the MLP Model

hyperparameter	setting
batch size	auto
epochs	200
initial learning rate	0.1
decay rate	0.96
activation function	ReLU, Sigmoid
optimization function	Adam
epsilon	1
decay steps	100,000

Performance metrics like accuracy, precision, recall, f1, and MCC scores were calculated to assess the performance of the model. Confusion matrix was computed. PR and ROC curves were generated. To avoid overfitting and to evaluate the generalization capacity of the model, k-fold cross-validation was carried out with a k value of 5. Furthermore, the MLP model was benchmarked by comparing with different algorithms, viz., support vector machine, logistic regression, decision tree, AdaBoost, CatBoost, and XGBoost algorithms. Hyperparameter sets for the compared machine learning algorithms are given in Table S4. All the codes implemented for the model training and evaluation are available on the GitHub repository (https://github.com/TeamSundar/Anticancer-Activity-Prediction). Performance comparison of the MLP model with existing methods CDRUG[21] and pdCSM-cancer[41] was also carried out.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.4c02801.

Performance metrics calculated for anticancer and nonanticancer classes; performance metrics of MLP k-fold cross-validation; NCI-60 data statistics; and hyperparameters of compared algorithms (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Durai Sundar − *Department of Biochemical Engineering & Biotechnology, Indian Institute of Technology (IIT) Delhi, New Delhi 110016, India; Yardi School of Artificial Intelligence, Indian Institute of Technology (IIT) Delhi, New Delhi 110016, India; Institute of Bioinformatics and Applied Biotechnology (IBAB), Bengaluru 560100, India;* orcid.org/0000-0002-6549-6663; Email: sundar@dbeb.iitd.ac.in

Authors

Swathik Clarancia Peter − *Regional Centre for Biotechnology (RCB), Faridabad, Haryana 121001, India; Department of Biochemical Engineering & Biotechnology, Indian Institute of Technology (IIT) Delhi, New Delhi 110016, India*

https://doi.org/10.1021/acsomega.4c02801
ACS Omega 2024, 9, 30645−30653

Yogesh Kalakoti − *Department of Biochemical Engineering & Biotechnology, Indian Institute of Technology (IIT) Delhi, New Delhi 110016, India*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c02801

Author Contributions

S.C.P.: conceptualization, methodology, software, formal analysis, and writing—original draft; Y.K.: conceptualization, methodology, and review; and D.S.: supervision, investigation, and writing—review and editing.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

S.C.P. is supported by GSK's "Trust in Science" initiative in partnership with RCB.

■ REFERENCES

(1) Ma, X. J.; Salunga, R.; Tuggle, J. T.; Gaudet, J.; Enright, E.; McQuary, P.; Payette, T.; Pistone, M.; Stecker, K.; Zhang, B. M.; Zhou, Y. X.; Varnholt, H.; Smith, B.; Gadd, M.; Chatfield, E.; Kessler, J.; Baer, T. M.; Erlander, M. G.; Sgroi, D. C. Gene Expression Profiles of Human Breast Cancer Progression. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (10), 5974−5979.

(2) Cho, J. H.; Kim, W. H. Altered Topographic Expression of P21(WAF1/CIP1/SDI1), Bcl2 and P53 during Gastric Carcinogenesis. *Pathol., Res. Pract.* **1998**, *194* (5), 309−317.

(3) Wang, J.; Gao, X.; Wang, M.; Zhang, J. Clinicopathological Significance and Biological Role of TCF21 MRNA in Breast Cancer. *Tumor Biol.* **2015**, *36* (11), 8679−8683.

(4) Rohde, M.; Daugaard, M.; Jensen, M. H.; Helin, K.; Nylandsted, J.; Jäättelä, M. Members of the Heat-Shock Protein 70 Family Promote Cancer Cell Growth by Distinct Mechanisms. *Genes Dev.* **2005**, *19* (5), 570−582.

(5) Oermann, E. K.; Wu, J.; Guan, K. L.; Xiong, Y. Alterations of Metabolic Genes and Metabolites in Cancer. *Semin. Cell Dev. Biol.* **2012**, *23* (4), 370−380.

(6) Inglese, P.; McKenzie, J. S.; Mroz, A.; Kinross, J.; Veselkov, K.; Holmes, E.; Takats, Z.; Nicholson, J. K.; Glen, R. C. Deep Learning and 3D-DESI Imaging Reveal the Hidden Metabolic Heterogeneity of Cancer. *Chem. Sci.* **2017**, *8* (5), 3500−3511.

(7) Weininger, D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31−36.

(8) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29* (2), 97−101.

(9) Worachartcheewan, A.; Mandi, P.; Prachayasittikul, V.; Toropova, A. P.; Toropov, A. A.; Nantasenamat, C. Large-Scale QSAR Study of Aromatase Inhibitors Using SMILES-Based Descriptors. *Chemom. Intell. Lab. Syst.* **2014**, *138*, 120−126.

(10) Ghaedi, A. Predicting the Cytotoxicity of Ionic Liquids Using QSAR Model Based on SMILES Optimal Descriptors. *J. Mol. Liq.* **2015**, *208*, 269−279.

(11) Werner, J. E.; Swift, J. A. Data Mining the Cambridge Structural Database for Hydrate-Anhydrate Pairs with SMILES Strings. *CrystEngComm* **2020**, *22* (43), 7290−7297.

(12) Kimber, T. B.; Gagnebin, M.; Volkamer, A. Maxsmi: Maximizing Molecular Property Prediction Performance with Confidence Estimation Using SMILES Augmentation and Deep Learning. *Artif. Intell. Life Sci.* **2021**, *1* (October), 100014.

(13) Costa, A. S.; Martins, J. P. A.; de Melo, E. B. SMILES-Based 2D-QSAR and Similarity Search for Identification of Potential New Scaffolds for Development of SARS-CoV-2 MPRO Inhibitors. *Struct. Chem.* **2022**, *33* (5), 1691−1706.

(14) Jiang, J.; Zhang, R.; Zhao, Z.; Ma, J.; Liu, Y.; Yuan, Y.; Niu, B. MultiGran-SMILES: Multi-Granularity SMILES Learning for Molecular Property Prediction. *Bioinformatics* **2022**, *38* (19), 4573−4580.

(15) Winter, B.; Winter, C.; Schilling, J.; Bardow, A. A Smile Is All You Need: Predicting Limiting Activity Coefficients from SMILES with Natural Language Processing. *Digital Discovery* **2022**, *1* (6), 859−869.

(16) Zhang, X.-C.; Wu, C.-K.; Yi, J.-C.; Zeng, X.-X.; Yang, C.-Q.; Lu, A.-P.; Hou, T.-J.; Cao, D.-S. Pushing the Boundaries of Molecular Property Prediction for Drug Discovery with Multitask Learning BERT Enhanced by SMILES Enumeration. *Research* **2022**, *2022* (Dl), 1−15.

(17) Pinheiro, G. A.; Mucelini, J.; Soares, M. D.; Prati, R. C.; Da Silva, J. L. F.; Quiles, M. G. Machine Learning Prediction of Nine Molecular Properties Based on the SMILES Representation of the QM9 Quantum-Chemistry Dataset. *J. Phys. Chem. A* **2020**, *124* (47), 9854−9866.

(18) Toropov, A. A.; Toropova, A. P.; Benfenati, E. Additive SMILES-Based Carcinogenicity Models: Probabilistic Principles in the Search for Robust Predictions. *Int. J. Mol. Sci.* **2009**, *10* (7), 3106−3127.

(19) Shearer, J.; Castro, J. L.; Lawson, A. D. G.; MacCoss, M.; Taylor, R. D. Rings in Clinical Trials and Drugs: Present and Future. *J. Med. Chem.* **2022**, *65* (13), 8699−8712.

(20) Taylor, R. D.; Maccoss, M.; Lawson, A. D. G. Rings in Drugs. *J. Med. Chem.* **2014**, *57* (14), 5845−5859.

(21) Li, G. H.; Huang, J. F. CDRUG: A Web Server for Predicting Anticancer Activity of Chemical Compounds. *Bioinformatics* **2012**, *28* (24), 3334−3335.

(22) Zhang, Y. X. Artificial Neural Networks Based on Principal Component Analysis Input Selection for Clinical Pattern Recognition Analysis. *Talanta* **2007**, *73* (1), 68−75.

(23) Kalyan, K.; Jakhia, B.; Lele, R. D.; Joshi, M.; Chowdhary, A. Artificial Neural Network Application in the Diagnosis of Disease Conditions with Liver Ultrasound Images. *Adv. Bioinf.* **2014**, *2014*, 1−14.

(24) Choi, E.; Bahadori, M. T.; Searles, E.; Coffey, C.; Thompson, M.; Bost, J.; Tejedor-Sojo, J.; Sun, J. Multi-Layer Representation Learning for Medical Concepts. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2016, 13−17-Aug, pp 1495−1504..

(25) Isa, I. S.; Saad, Z.; Omar, S.; Osman, M. K.; Ahmad, K. A.; Sakim, H. M. Suitable MLP Network Activation Functions for Breast Cancer and Thyroid Disease Detection. *2010 Second International Conference on Computational Intelligence, Modelling and Simulation* 2010, pp 39−44..

(26) Hamadache, M.; Amrane, A.; Hanini, S.; Benkortbi, O. Multilayer Perceptron Model for Predicting Acute Toxicity of Fungicides on Rats. *Int. J. Quant. Struct.-Prop. Relat.* **2018**, *3* (1), 100−118.

(27) Qin, Y.; Li, C.; Shi, X.; Wang, W. MLP-Based Regression Prediction Model For Compound Bioactivity. *Front. Bioeng. Biotechnol.* **2022**, *10* (July), 1−10.

(28) Sharma, A.; Selvam, S.; Balaji, P. D.; Madhavan, T. ANN Multi-Layer Perceptron for Prediction of Blood-Brain Barrier Permeable Compounds for Central Nervous System Therapeutics. *J. Biomol. Struct. Dyn.* **2024**, 1−6.

(29) Li, Z.; Kamnitsas, K.; Glocker, B. Analyzing Overfitting under Class Imbalance in Neural Networks for Image Segmentation. *IEEE Trans. Med. Imag.* **2021**, *40* (3), 1065−1077.

(30) Walsh, R.; Tardy, M. A Comparison of Techniques for Class Imbalance in Deep Learning Classification of Breast Cancer. *Diagnostics* **2022**, *13* (1), 67.

(31) Cao, H.; Li, X. L.; Woon, D. Y. K.; Ng, S. K. Integrated Oversampling for Imbalanced Time Series Classification. *IEEE Trans. Knowl. Data Eng.* **2013**, *25* (12), 2809−2822.

(32) Krawczyk, B.; Galar, M.; Jeleń, Ł.; Herrera, F. Evolutionary Undersampling Boosting for Imbalanced Classification of Breast Cancer Malignancy. *Appl. Soft Comput.* **2016**, *38*, 714−726.

(33) Cao, P.; Yang, J.; Li, W.; Zhao, D.; Zaiane, O. Ensemble-Based Hybrid Probabilistic Sampling for Imbalanced Data Learning in Lung Nodule CAD. *Comput. Med. Imaging Graph.* **2014**, *38* (3), 137−150.
(34) Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16* (1), 321−357.
(35) Elreedy, D.; Atiya, A. F. A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for Handling Class Imbalance. *Inf. Sci* **2019**, *505*, 32−64.
(36) Rupapara, V.; Rustam, F.; Shahzad, H. F.; Mehmood, A.; Ashraf, I.; Choi, G. S. Impact of SMOTE on Imbalanced Text Features for Toxic Comments Classification Using RVVC Model. *IEEE Access* **2021**, *9*, 78621−78634.
(37) Abdoh, S. F.; Abo Rizka, M.; Maghraby, F. A. Cervical Cancer Diagnosis Using Random Forest Classifier with SMOTE and Feature Reduction Techniques. *IEEE Access* **2018**, *6*, 59475−59485.
(38) Jeatrakul, P.; Wong, K. W.; Fung, C. C. Classification of Imbalanced Data by Combining the Complementary Neural Network and SMOTE Algorithm. *Neural Information Processing. Models and Applications*; *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) LNCS* 2010; Vol *6444* (PART 2), pp 152−159.
(39) Kubinyi, H. Similarity and Dissimilarity: A Medicinal Chemist's View. *Perspect. Drug Discovery Des.* **1998**, *9/11*, 225−252.
(40) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39* (15), 2887−2893.
(41) Al-Jarf, R.; de Sá, A. G. C.; Pires, D. E. V.; Ascher, D. B. PdCSM-Cancer: Using Graph-Based Signatures to Identify Small Molecules with Anticancer Properties. *J. Chem. Inf. Model.* **2021**, *61* (7), 3314−3322.

https://doi.org/10.1021/acsomega.4c02801
ACS Omega 2024, 9, 30645−30653