Hindawi
Complexity
Volume 2021, Article ID 6675279, 15 pages
https://doi.org/10.1155/2021/6675279

Research Article

Handling Imbalance Classification Virtual Screening Big Data Using Machine Learning Algorithms

Sahar K. Hussin,[1] **Salah M. Abdelmageid**,[2] **Adel Alkhalil**,[3] **Yasser M. Omar**,[4] **Mahmoud I. Marie**,[5] **and Rabie A. Ramadan**[3,6]

[1]*Communication and Computers Engineering Department Alshrouck Academy, Cairo, Egypt*
[2]*Computer Engineering Department, Collage of Comp. Science and Engineering, Taibah University, Medina, Saudi Arabia*
[3]*College of Computer Science and Engineering, University of Hai'l, Hai'l, Saudi Arabia*
[4]*Arab Academy for Science Technology and Maritime Transport, Cairo, Egypt*
[5]*Computer and System Engineering Department, Al-Azhar University, Cairo, Egypt*
[6]*Computer Engineering Department, Cairo Universality, Cairo, Egypt*

Correspondence should be addressed to Rabie A. Ramadan; rabie@rabieramadan.org

Received 28 November 2020; Revised 19 December 2020; Accepted 2 January 2021; Published 28 January 2021

Academic Editor: Abd E.I.-Baset Hassanien

Copyright © 2021 Sahar K. Hussin et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Virtual screening is the most critical process in drug discovery, and it relies on machine learning to facilitate the screening process. It enables the discovery of molecules that bind to a specific protein to form a drug. Despite its benefits, virtual screening generates enormous data and suffers from drawbacks such as high dimensions and imbalance. This paper tackles data imbalance and aims to improve virtual screening accuracy, especially for a minority dataset. For a dataset identified without considering the data's imbalanced nature, most classification methods tend to have high predictive accuracy for the majority category. However, the accuracy was significantly poor for the minority category. The paper proposes a *K*-mean algorithm coupled with Synthetic Minority Oversampling Technique (SMOTE) to overcome the problem of imbalanced datasets. The proposed algorithm is named as KSMOTE. Using KSMOTE, minority data can be identified at high accuracy and can be detected at high precision. A large set of experiments were implemented on Apache Spark using numeric PaDEL and fingerprint descriptors. The proposed solution was compared to both no-sampling method and SMOTE on the same datasets. Experimental results showed that the proposed solution outperformed other methods.

1. Introduction

The discovery of new medication to cure human illnesses is progressively hard, expensive, and tedious [1]. A wide variety of atoms and molecules must be chosen and prepared to generate a set of predetermined drugs. The drug discovery process can take between 12 and 15 years, with a possibility of failure, and expenses are worth more than one billion dollars [2]. Virtual screening is the most critical process in drug discovery. It is employed to search for small chemical compounds (molecules) in libraries to identify structures that have an affinity to bind to a drug target or protein receptor [3]. Up to 1010 libraries of virtual screening exist, and since this record keeps increasing, traditional classification methods have become insufficient to manage such large amounts of datasets [4]. One of the well-known repositories is PubChem [5] for small molecules and their biological properties. It offers several resources that are unfortunately constrained by the unbalanced nature of high-throughput screening (HTS) data. These data usually contain a few hundred active compounds, excluding many inactive compounds.

Dataset imbalances occur when one of the classes is described by a minimal number of samples, typically of major importance, compared to the other classes [6]. This problem may distort prediction accuracy in the used

classification models, which leads to poor classification performance. In HTS experiments, thousands of compounds are usually screened; nevertheless, a small fraction of the tested compounds are classified to be active while other classes are recognized inactive [7]. Such data imbalance affects the accuracy and precision of activity predictions in individual virtual screening datasets. Using the binary classification of an imbalanced dataset, an instance of one class became fewer compared to another class. The minority class is known as the class with fewer cases, and the other is called the majority class [8].

Current classification models such as k-nearest neighbor (KNN), random forest (RF), multilayer perceptron (MLP), support vector machine (SVM), decision tree (DT), logistic regression (LG), and gradient boosting (GBT) depend on a sufficient, representative, and reasonably balanced collection of training data to draw an approximate boundary for decision-making between different groups. These learning algorithms are utilized in a variety of fields, including financial forecasting and text classification [9]. Despite current advances in machine learning (ML), developing successful algorithms that learn from unbalanced datasets remains a daunting task. In ML, many approaches have been created to deal with imbalanced data. However, very few algorithms have been able to handle problems related to negatives and false positives. Positive or negative states usually dominate the unbalanced dataset. Therefore, specificity and recall (sensitivity) are very vital when processing an imbalanced dataset [10]. The increase in sensitivity increases the true-positive expectations of the model and reduces false negatives. Likewise, an upgrade in specificity increases true-negative expectations and thus reduces wrong responses. Therefore, it is critical that, for a good model, the gap between sensitivity and specificity metrics should be as small as possible [11].

Although some of the researchers highlighted the problem of negatives and false positives when using PubChem data, to the best of our knowledge, no technique has been reported to address the problem effectively. It was also reported that the imbalance problem obstructed the classification accuracy of bioactivity [12]. In another study [13], the authors compared the performance of seven different descriptive classifiers based on their ability to deal with unbalanced datasets. Another research stated that multilevel SVM-based algorithms outperformed certain algorithms such as (1) traditional SVM, (2) weighted SVM, (3) neural networks, (4) linear regression, (5) Naïve Bayes (NB), and (6) C4.5 tree with imbalanced groups, missing values, and real health-related data [14].

Besides, the main concept of resampling is to reduce variance between class samples by preprocessing the training data. In other words, the resampling approach is used in training samples in order to achieve the same number of samples for each class to adjust the previous majority and minority sample distributions. There are two basic methods in the traditional resampling methodology, namely, undersampling and oversampling [15]. Undersampling produces a smaller number of majority samples while all minority samples are retained. The predominant class samples will be eliminated randomly before a satisfactory ratio has been accomplished for all groups. Undersampling is ideal for applications where there is an enormous number of majority samples, and the reduction of training samples would minimize the training time for the model. However, a drawback with undersampling that it discards samples contributes to the loss of majority class information [16].

Another approach to address the imbalanced data is oversampling. By replicating samples, it raises the number of samples in the minority groups [17]. The benefit from oversampling is that because all samples are used, no knowledge is lost. Oversampling, however, has its own drawback. It contributes to higher processing costs by generating extra training samples. Therefore, to compensate for that limitation, more efficient algorithms are needed.

Although resampling methods are usually used to solve problems with imbalances in the class, there is little defined strategy to identify the acceptable class distribution for a particular dataset [18]. As a result, the optimal class distribution differs from one dataset to another. Recent variants of resampling approaches derived from oversampling and undersampling overcome some of the shortcomings of current technologies, including SMOTE (Synthetic Minority Oversampling Techniques). SMOTE is one of the most important oversampling approaches that generate interpolation instances, which is added to the training samples without duplicating the samples in the class of the minority. The SMOTE approach examines the KNN of the minority class test that will be utilized as a base for the new synthetic sample [19]. If created instances are smaller than the size of the initial dataset, the approach randomly selects the original instances utilized to create the artificial ones. If instances are larger than the size of the original dataset, the approach iterates over the dataset, creating an artificial instance per original instance until it reaches the previous scenario [20]. SMOTE is considered as an oversampling technique that produces synthetic minority class samples. This is theoretically performing better than simple oversampling, and it is commonly used. For example, SMOTE was utilized to detect network intrusions [21] or speech boundary sentence, to predict species distribution [22]. SMOTE will be utilized in this research.

Data mining techniques can help to reduce promising candidate chemicals for interaction with specific molecular targets before they are experimentally evaluated [23]. In theory, this can help to speed up the drug development process. However, the improvement of accurate prediction models for HTS is difficult. For datasets such as those taken from HTS experiments, the achievement of high predictability accuracy may be misleading since this may be accompanied by an unacceptable false-positive rate [24] as high accuracy does not always imply a small proportion of false predictions.

In the event of a large class imbalance, this paper attempts to address the most effective variant of data preprocessing to enhance data imbalance accuracy, which favors the collection of interactions that increase the overall accuracy of a learning model. We propose a SMOTE coupled with k-mean method to classify several imbalanced

PubChem datasets with the goal of (1) validating whether *k*-mean with SMOTE affects the output of established models and (2) exploring if KSMOTE is appropriate and useful in finding interesting samples from broad datasets. Our model is also applied to different ML algorithms (random forest, decision tree, multilayer perceptron, logistic regression, and gradient boosting) for comparison purposes with three different datasets. The paper also introduces a procedure for data sampling to increase the sensitivity and specificity of predicting several molecules' behavior. The proposed approach is implemented on standalone clusters for Apache Spark 2.4.3 in order to address the imbalance in a big dataset.

The remainder of this paper is structured as follows. Section 2 provides a general description of class imbalance learning concepts and reviews the related research conducted on the subject matter. Section 3 explains how the proposed approach for VS in drug discovery was developed. Section 4 presents performance evaluations and experimental results. Section 5 presents the discussion of our proposal. Finally, Section 6 highlights the conclusions and topics of study for future research.

2. Related Work

For the paper to be self-contained, this section reviews the most related work to the VS research, techniques, problems, and state-of-the-art solutions. It also examines some of the big data frameworks that could help solve the problem of imbalanced datasets.

Since we live in the technological era where older storage and processing technologies are not enough, computing technologies must be scaled to handle a massive amount of data generated by different devices. The biggest challenge in handling such volumes of data is the speed at which they will grow much faster than the computer resources. One of the research areas that generate huge data to be analyzed is searching and discovering medicines. The proposed methods aim to find a molecule capable of binding and activating or inhibiting a molecular target. The discovery of new drugs for human diseases is exceptionally complicated, expensive, and time-consuming. Drug discovery uses various methods [25] based on a statistical approach to scan for small molecule libraries and determines the structures most likely to bind to a compound. However, the drug target is a protein receptor that is involved in a metabolic cycle or signaling pathway by which a particular disease disorder is established or another anatomy.

There are two VS approaches, which are ligand-based VS (LBVS) and structure-based virtual screening (SBVS) [26]. LBVS depends on the existing information about the ligands. It utilizes the knowledge about a set of ligands known to be active for the given drug target. This type of VS uses the mining of big data analytics. Training binary classifiers by a small part of a ligand can be employed, and very large sets of ligands can be easily classified into two classes: active and nonactive ones. SBVS, on the other side, is utilized to dock experimentally. However, 3D protein structural information is required [27], as shown in Figure 1.

K-mean clustering is one of the simplest nonsupervised learning algorithms, which was first proposed by Macqueen in 1967. It has been applied by many researchers to solve some of the problems of known groups [28]. This technique classifies a particular dataset into a certain number of groups. The algorithm randomly initializes the center of the groups. Then, it calculates the distance between an object and the midpoint of each group. Next, each data instance is linked to the nearest center, and the cluster centers are recalculated. The distance between the center and each sample is calculated by the following equation:

$$\text{Euclidean distance} = \sum_{i=1}^{c} \sum_{j=1}^{ci} \|Xi - Yj\|, \tag{1}$$

where the Euclidean distance between the data point Xi and cluster center y is d, Ci is the total number of data points i in cluster, and c is the total number of cluster centers. All of the training samples are first grouped into K groups (the experiment with diverse K values runs to observe the result). Suitable training samples from the derived clusters are selected. The key idea is that there are different groups in a dataset, and each group appears to have distinct characteristics. When a cluster includes samples of large majority class and samples of low minority class, it functions as a majority class sample. If on the other side, a cluster has extra minority samples and fewer majority samples, it acts more like a minority class. Therefore, by considering the number of majority class samples to that of minority class samples in various clusters, this method oversamples the required number of minority class samples from each cluster.

Several approaches have been proposed in the literature to handle big data classification including classification algorithms, random forest, decision tree, multilayer perceptron, logistic regression, and gradient boosting.

Classification algorithms (CA) are mainly depending on machine learning (ML) algorithms, where they play a vital role in VS for drug discovery. It can be considered as an LBVS approach. Researchers widely used the ML approach to create a binary classification model that is a form of filter to classify ligands as active or inactive in relation to a particular protein target. These strategies need fewer computational resources, and because of their ability to generalize, they find more complex hits than other earlier approaches. Based on our experience, we believe that many classification algorithms can be utilized for dealing with unbalanced datasets in VS, such as SVM, RF, Naïve Bayes, MLP, LG, ANN, DT, and GBT. Five ML algorithms are applied in this paper RF, DT, MLP, LG, and GBT [29].

Random forest (RF) is an ensemble learning approach in which multiple decision trees are constructed based on training data and a majority voting mechanism. Like KNN, it is utilized to predict classification or regression for new inputs. The key advantage of RF is that it can be utilized for problems that need classification and regression. Besides, RF is the ability to manage many higher-dimensional datasets. It has a powerful strategy for determining the lack of information and preserving accuracy when much of the information is missing [29].

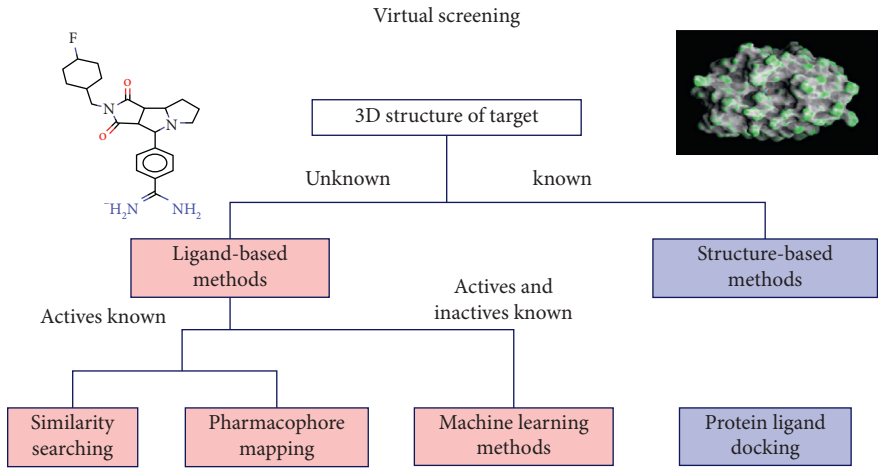

Figure 1: Taxonomy for the 3D structure of VS methods [25].

Decision tree (DT) is represented as an actual tree with its root at the top and the leaves at the bottom. The root of the tree is divided into two or more branches. The branch may be broken down into two branches or more. This process continues until the leaf is reached, meaning no more split remains.

Multilayer perceptron (MP) has two main types of artificial neural networks (ANNs), which are supervised and unsupervised networks. Every network consists of a series of linked neurons. A neuron takes multiple numerical inputs and outputs of values depending on the number of inputs weighted. Popular functions of transformation embody the functions of tanh and sigmoid. Neurons are formed into layers. ANN can contain several hidden layers, and the neurons will only be linked to those in the next layers, known as forwarding feed networks, multiplayer perceptions (MLPs), or functional radial base network (RBN) [29].

Logistic regression (LR) is one of the simplest and frequently utilized ML algorithms for two-class classification. It is simple to implement and can be utilized as the baseline for any binary classification problem. In deep learning, basic principles are also constructive. The relationship between a single dependent relative binary variable and independent variables is defined and estimated by logistic regression. It is also a mathematical method for predicting binary classes. The effect or target variable is dichotomous in nature. Dichotomous means that only two possible groups can be used for cancer detection issues, for instance [30].

Gradient boosting is a type of ML boosting. It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. The name gradient boosting arises because target outcomes for each case are set based on the gradient of the error with respect to the prediction. Each new model takes a step in the direction that minimizes prediction error in the space of possible predictions for each training case [30].

In [31], the authors compared four Weka classifiers, including SVM, J48 tree, NB, and RF. From a completed cost-sensitive survey, SVM and Tree C4.5 (J48) performed well, taking minority group sizes into account. It shows that a hybrid of majority class undersampling and SMOTE can improve overall classification performance in an imbalanced dataset. In addition, the authors in [32] have employed a repetitive SVM as a sample method that is used for SVM processing from bioassay information on luciferase inhibition that has a high active/inactive imbalance ratio (1/377). The models' highest performance was 86.60 and 88.89 percent for active compounds and inactive compounds, respectively, associated with a combined precision of 87.74 percent when using validation and blind test. These findings indicate the quality of the proposed approach for managing the intrinsic imbalance problem in HTS data used to cluster possible interference compounds to virtual screening utilizing luciferase-based HTS experiments.

Similarly, in [33], the authors analyzed several common strategies for modeling unbalanced data, including multiple undersampling, threshold, ratio 1 : 3 undersampling, one-sided undersampling, similarity undersampling, cluster undersampling, diversity undersampling, and only threshold selection. In total, seven methods were compared using HTS datasets extracted from PubChem. Their analysis led to the proposal of a new hybrid method, which includes both low-cost learning and less sampling approaches. The authors claim that the multisample and hybrid methods provide accurate predictions of results more than other methods. Besides, some other research studies used toxicity datasets in imbalance algorithms as in [34]. The model was based on a data ensemble, where each model sees an equal distribution of the two toxic and nontoxic classes. It considers various aspects of creating computational models to predict cell toxicity based on cell proliferation screening dataset. Such predictive models can be helpful in evaluating cell-based screening outcomes in general by bringing feature-based data into such datasets. It also could be utilized as a method to recognize and remove potentially undesirable compounds. The authors concluded that the issue of data imbalance hindered the accuracy of the critical activity classification. They also developed an artificial random forest group model that was designed to mitigate dataset misalignment in predicting cell toxicity.

The authors in [35] used a simple oversampling approach to build an SVM model classifying compounds based on the expected cytotoxic versus Jurkat cell line. Oversampling with the minority has been shown to contribute to better predictive SVM models in training groups and external test groups. Consequently, the authors in [36] analyzed and validated the importance of different sampling methods over the nonsampling method in order to achieve well-balanced sensitivity and specificity of the ML model that has been created for unbalanced chemical data. Additionally, the study conducted an accuracy of 93.00% under the curve (AUC) of 0.94, a sensitivity of 96.00%, and specificity of 91.00% using SMOTE sampling and random forest classification to predict drug-induced liver injury. Although it was presented in the literature that some of the proposed approaches have succeeded somehow in responding to the issues of unbalanced PubChem datasets, there is still a lack of time efficiency during calculations.

3. Proposed KSMOTE Framework

K-Mean Synthetic Minority Oversampling Technique (KSMOTE) is proposed in this paper as a solution for virtual screening to drug discovery problems. KSMOTE combines *K*-mean and SMOTE algorithms to avoid the imbalanced original datasets, ensuring that the number of minority samples is as close as possible to the majority of the population samples. As shown in Figure 2, the data were first separated into two sets—one set contains majority samples and the other set contains the entire minority sample. First, majority samples were clustered into *K* clusters and minority samples, where *K* is greater than one in both cases. The number of clusters for each class is chosen according to the elbow method. The Euclidean distance was employed to calculate the distance between the center of each majority cluster and the center of each minority cluster. Each majority cluster sample was combined with the minority cluster sample subset to make *K* separate training datasets. This combination was done based on the largest distance between each majority and minority cluster. SMOTE was then applied to each combination of the clusters. It generates an instance of synthetic minority, oversampling minority class. For any minority example, the *k* (5 in SMOTE) is the nearest. Neighbors of the same class are determined, and then some instances are randomly selected according to the oversampling factor. After that, new synthetic examples are generated along the line between the minority example and its nearest chosen example.

3.1. Environment Selection and the Dataset. Since we are dealing with big data, a big data framework must be chosen. One of the most powerful frameworks that have been used in many data analytics is Spark. Spark is a well-known cluster computing engine that is very reliable. It presents application programming interfaces in various programming languages such as Java, Python, Scala, and R. Spark supports in-memory computing, allowing handling records much faster than disk-based engines Hadoop. Spark engine is advanced for in-reminiscence processing as well as disk-based totally processing [37]. It has been installed on different operating systems such as Windows and Ubuntu.

This paper implements the proposed approach using PySpark version 2.4.3 [38] and Hadoop version 2.7, installed on Ubuntu 18.04, and Python is used as a programming language. A Jupyter notebook version 3.7 was used. The computer configuration for experiments was a local machine Intel Core i7 with 2.8 GHz speed and 8 GB of RAM. To illustrate the performance of the proposed framework, three datasets are chosen. They are carefully chosen where each of them differs in its imbalance ratio. They are also large enough to illustrate the big data analysis challenge. The three datasets are AID 440 [39], AID 624202 [40], and AID 651820 [41]. All of them are retrieved from the PubChem database Library [5]. The three datasets are summarized in Table 1 and briefly described in the following paragraphs. All of the data exist in an unstructured format as SDF files. Therefore, they require some preprocessing to be accepted as input to the proposed platform:

(1) AID 440 is a formylpeptide receptor (FPR). The G-protein, coupled with the formylpeptide receptor, was one of the originating chemo-attracting receptor members [39]. It consists of 185 active and 24,815 nonactive molecules.

(2) AID 624202 is a qHTS test to identify small molecular stimulants for BRCA1 expression. BRCA1 has been involved in a wide range of cellular activities, including repairing DNA damage, cell cycle checkpoint control, growth inhibition, programmed cell death, transcription regulation, chromatin recombination, protein presence, and autogenously stem cell regeneration and differentiation. The increase in BRCA1 expression would enable cellular differentiation and restore tumor inhibitor function, leading to delayed tumor growth and less aggressive and more treatable breast cancer. Promising stimulants for BRCA1 expression could be new preventive or curative factors against breast cancer [40].

(3) AID 651820 is a qHTS examination for hepatitis C virus (HCV) inhibitors [41]. About 200 million people globally are hepatitis C (HCV) contaminated. Many infected individuals progress to chronic liver disease, including cirrhosis, with the risk of developing hepatic cancer. There is no effective hepatitis C vaccine available to date. Current interferon-based therapy is effective only in about half of patients and is associated with significant adverse effects. It is estimated that the fraction of people with HCV who can complete treatment is no more than 10 percent. The recent development of direct-acting antivirals against HCV, such as protease and polymerase inhibitors, is promising. However, it still requires a combination of peginterferon and ribavirin for maximum efficacy. Moreover, these agents are associated with a high resistance rate, and many have significant side effects (Figure 3)

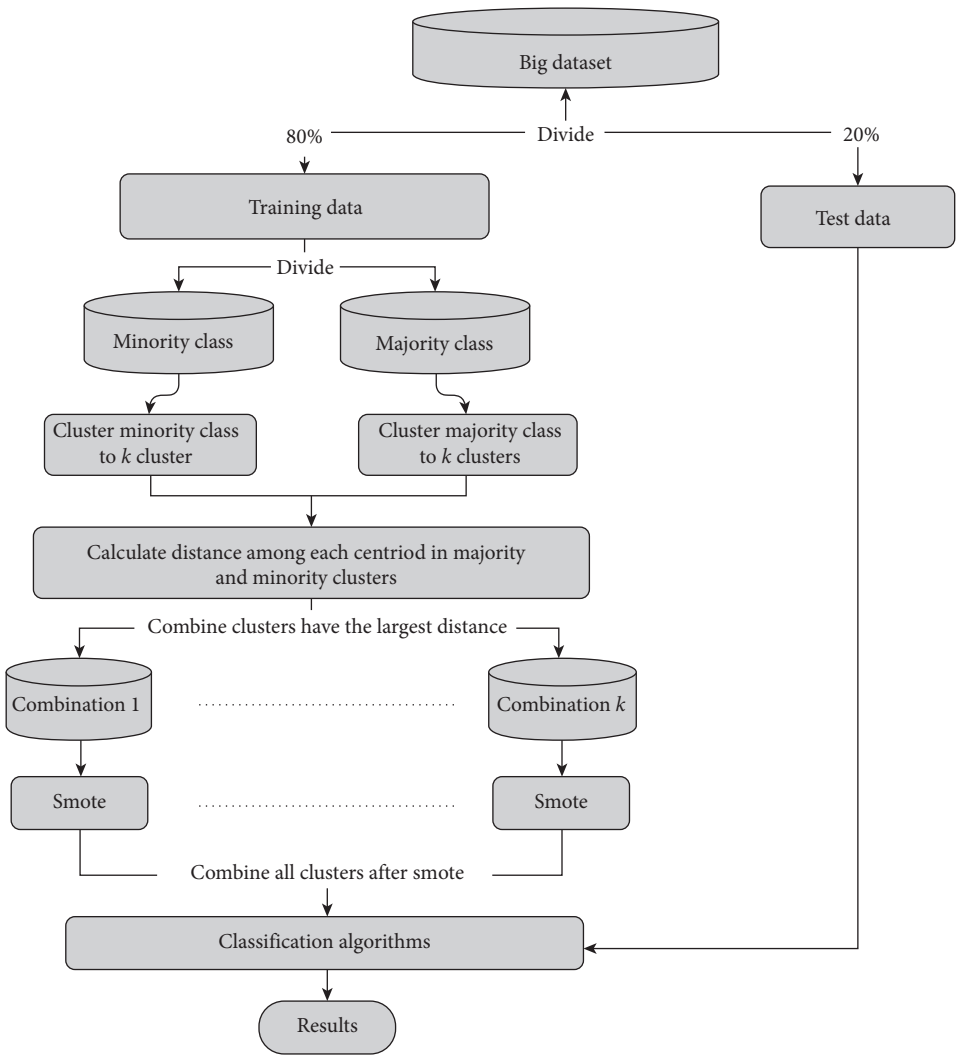

Figure 2: The proposed model.

Table 1: Benchmark imbalanced dataset.

Datasets	Total records	Inactive	Active	Balance ratio
AID 440	25,000	24,815	185	1 : 134
AID 624202	377,550	364,035	3980	1 : 91.5
AID 651820	283,005	271,341	11,664	1 : 23

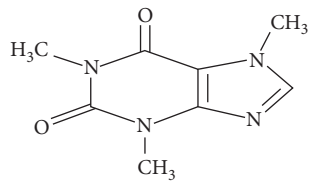

Figure 3: Molecules (chemical compound) [42].

3.2. Preprocessing. A chemical compound (molecules) is stored in SDF files, so those files have to be converted to feature vectors [$d1$, $d2$, and $d3$... $d1444$] with 1444 dimensions, and it was suitable to use PaDEL cheminformatics software [42] for this process. There are two types of PaDEL cheminformatics software, numeric and fingerprint descriptors. A PaDEL numeric descriptor gives information about the quantity of a feature in each compound. Molecules are represented based on constitutional, topological, and geometrical descriptors as well as other molecular properties. This includes aliphatic ring count, aromatic ring count, logP, donor count, polar surface area, and Balaban index. The Balaban index is a real number and can be either positive or negative. PaDEL is also used as a fingerprint descriptor, and it gives 881 attributes. The fingerprint was also calculated in order to compare the model performance derived from PaDEL descriptors. Molecules are referred to as instances and labeled as 1 or 0, where 1 means active and 0 means not active. The reader is directed to [43] for further information about the descriptors and fingerprints.

4. Evaluation Metrics

Instead of utilizing complicated metrics, four intuitive and functional metrics (specificity, sensitivity, G-mean, and accuracy) were introduced according to the following

reasons: First, the predictive power of the classification method for each sample, particularly the predictive power of the minority group (i.e., active power), is demonstrated by measuring performance for both sensitivity and specificity. Second, G-mean is a combination of sensitivity and specificity, indicating a compromise between the majority and minority output of the classification. Poor quality in predicting positive samples reduces the G-mean value, whereas negative samples are classified with accuracy with a high percentage. This is a typical state for imbalanced data collection. It is strongly recommended that external predictions be used to build a reliable model of prediction [41, 44]. The four statistical assessment methods are described as follows:

(1) Sensitivity: the proportion of positive samples appropriately classified and labeled, and it can be determined by the following equation:

$$\text{sensitivity} = \frac{\text{TP}}{(\text{TP} + \text{FN})}, \tag{2}$$

where true positive (TP) corresponds to the right classification of positive samples (e.g., in this work, active compounds); true negative (TN) corresponds to the correct classification (i.e., inactive compounds) of negative samples; false positive (FP) means that negative samples have been incorrectly identified in positive samples; and false negative (FN) is an indicator to incorrectly classified positive samples.

(2) Specificity: the proportion of negative samples that are correctly classified; its value indicates how many cases that are predicted to be negative and that are truly negative as stated in the following equation:

$$\text{specificity} = \frac{\text{TN}}{(\text{TN} + \text{FP})}, \tag{3}$$

(3) G-mean: it offers a simple way to measure the model's capability to correctly classify active and inactive compounds by combining sensitivity and specificity in a single metric. G-mean is a measure of balance accuracy [45] and is defined as follows:

$$\text{G-mean} = \sqrt{\text{specificity} \times \text{sensitivity}}. \tag{4}$$

G-mean is a hybrid of sensitivity and specificity, indicating a balance between majority and minority rating results. Low performance in forecasting positive samples also contributes to reducing G-mean value, even though negative samples are highly accurate. This is a typical unbalanced dataset condition [45].

(4) Accuracy: it shows the capability of a model to correctly predict the class labels as given in the following equation:

$$\text{accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}. \tag{5}$$

4.1. Experimental Results. This section presents the results based on extensive sets of experiments. The average of the experiments is concluded and presented with discussion. Tenfold cross-validation is used to evaluate the performance of the proposed model. Besides, the original sample retrieved from the datasets is randomly divided into ten equal-sized subsamples. Out of the ten subsamples, one subsample is maintained as validation data for model testing, while the remaining nine subsamples are used as training data. The validation process is then replicated ten times (folds), using each of the ten subsamples as validation data exactly once. It is then possible to compute the average of the ten outcomes from the folds. To validate the effectiveness of KSMOTE, the performance of the proposed model is compared with that of SMOTE only and no-sampling models. Furthermore, two types of descriptors, numeric PaDEL and fingerprint, are used to validate their impact on the model's performance for all compilations. The results presented in this work are based on original test groups only without oversampling. Two types of descriptors, PaDEL and fingerprint, are used to generate five algorithms (RF, DT, MLP, LG, and GBT) to validate their effect on model output. Therefore, the performance of applying these algorithms is examined.

4.2. G-Mean-Based Performance. In this section, G-mean results are presented for the three selected datasets. Figure 4 describes G-mean for the AID 440 dataset based on both PaDEL and fingerprint descriptors. This figure demonstrates the performance of the various PaDEL descriptor and fingerprint sets. It also shows a comparison between three different approaches, which are no-sample, SMOTE, and KSMOTE. Besides the performance of employing RF, DT, MLP, LG, and GBT classifiers with the three different approaches examined, 20% of the datasets are used for testing, as mentioned before.

As shown in Figure 4, the best G-mean gained in the case of PaDEL fingerprint descriptor was by the KSMOTE, where it reaches, on average, 0.963. However, utilizing KSMOTE with LG gives almost G-means of 0.97, which is the best value over other classifiers. Based on our experiments, SMOTE- and no-sample-based PaDEL fingerprint descriptors are not recommended for virtual screening and classification where their G-mean is too small.

With the same settings applied to the fingerprint descriptor, the PaDEL numeric descriptor performance is examined. Again, the results are almost similar, where the KSMOTE shows higher performance than SMOTE and no-sample with nearly 55%. However, KSMOTE with DT and GBT classifiers are not up to other classifiers' level in this case. Therefore, it is recommended to utilize RF, MLP, and LG when a numeric descriptor is used. On the contrary, although the performance of SMOTE and no-sample is not that good, they show enhancement over the PaDEL fingerprint with an average of 10%. Among the overall results, it has been noticed that the worst G-mean value is produced from applying RF classifier with no-sample approach.

The performance of KSMOTE produced from the AID 440 dataset is confirmed using the AID 624202 dataset. As shown in Figure 5, KSMOTE gives the best G-mean using all

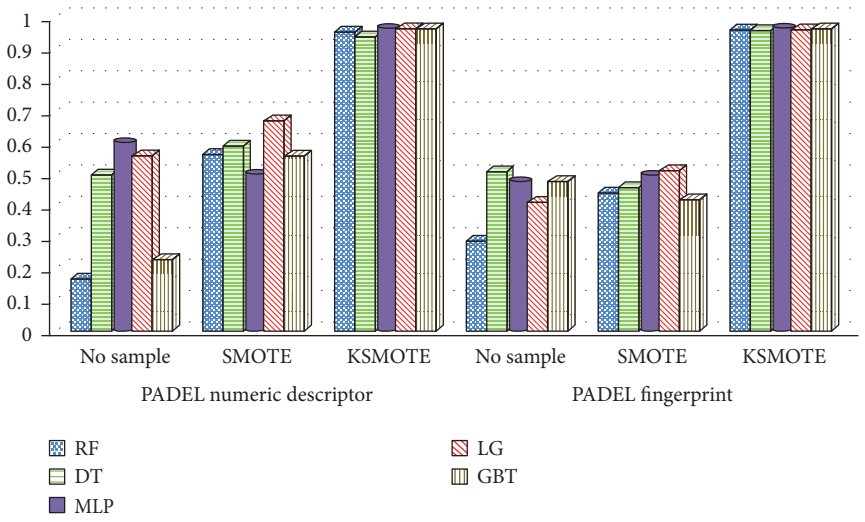

Figure 4: G-mean of PaDEL descriptor and fingerprint for AID 440.

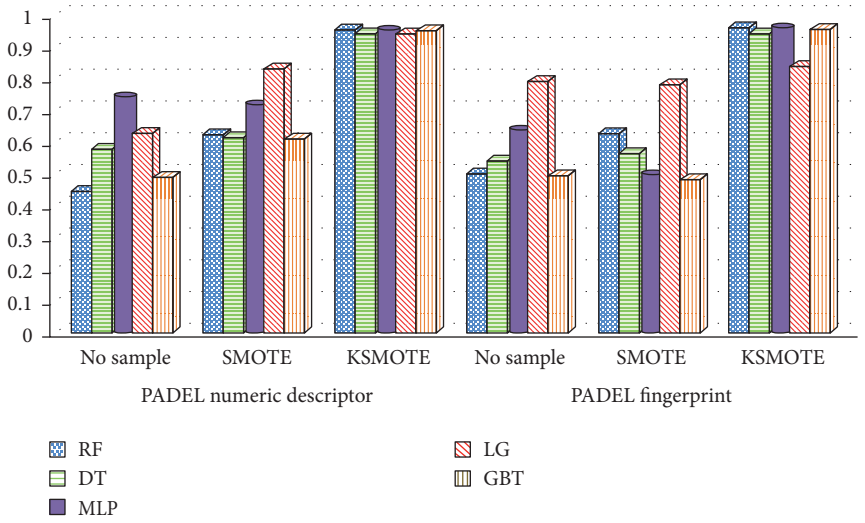

Figure 5: G-mean of PaDEL descriptor and fingerprint for AID624202.

of the classifiers. The only drawback shown is the G-mean of LG performance by almost 8% from other classifier results in case of PaDEL fingerprint classifier is used. On the contrary, LG classifier shows much more progress than before, where its average G-mean reached 0.8, which is a good achievement. For the rest of the classifiers in both descriptors, G-mean values are enhanced from the previous dataset. Still, among all the classifiers, G-mean value for the RF classifier with no-sample approach is worst. The overall conclusion is that KSMOTE is recommended to be used with both AID 440 and AID 624202 datasets.

Table 2 summarizes Figures 4–6 collecting all of the results in one place. It shows the complete set of experiments and average G-mean results in values to see the overall picture. Again, as can be exerted from the table, the proposed KSMOTE approach gives the best results of G-mean on the three datasets. It is believed that partitioning active and nonactive compounds to K clusters and then combining pairs that have large distances led to an accurate rate of oversampling instances in the SMOTE algorithm. This explains why the proposed model produces the best results.

4.3. Sensitivity-Based Performance. Sensitivity is another metric to measure the performance of the proposed approach compared to others. Here, the sensitivity performance presentation is a little bit different where the performance of the three approaches is displayed for the three datasets. Figure 7 shows the sensitivity of all datasets based on PaDEL numeric descriptor, while Figure 8 presents the sensitivity results for the fingerprint descriptor. It is obvious that the KSMOTE sensitivity values are superior to other approaches using both descriptors. In addition, SMOTE is overperforming the no-sample approach in almost all of the cases. For the AID 440 dataset, a low sensitivity value of 0.37 for the minority class is shown by the MLP model from the initial dataset (i.e., without SMOTE resample). The SMOTE algorithm was introduced to

Table 2: Complete set of experiments and average G-mean results.

Algorithm		PaDEL numeric descriptor				PaDEL fingerprint			
		No-sample	SMOTE	KSMOTE	Time	No-sample	SMOTE	KSMOTE	Time
AID 440	RF	0.167	0.565	0.954	23	0.29	0.442	0.96	12
	DT	0.5	0.59	0.937	**9.3**	**0.51**	0.459	0.958	4.9
	MLP	**0.6**	0.5	**0.963**	20	0.477	0.498	**0.964**	**9.6**
	LG	0.56	**0.67**	**0.963**	11	0.413	**0.512**	0.96	5.6
	GBT	0.23	0.56	**0.963**	33	0.477	0.421	0.963	17.1
AID624202	RF	0.445	0.625	**0.952**	29.7	0.5	0.628	**0.96**	**15.3**
	DT	0.576	0.614	0.94	10	0.54	0.564	0.94	5
	MLP	**0.74**	0.715	0.95	25.2	0.636	0.497	0.958	13.5
	LG	0.628	**0.83**	0.94	**26.8**	**0.791**	**0.78**	0.837	13.25
	GBT	0.489	0.61	0.95	45	0.495	0.482	0.954	22.36
AID 651820	RF	0.722	0.792	0.956	41	0.741	0.798	**0.92**	19.25
	DT	0.725	0.72	0.932	8.78	0.765	0.743	0.89	4.44
	MLP	**0.82**	0.817	0.915	**35**	**0.788**	**0.8**	0.91	17.3
	LG	0.779	**0.8357**	**0.962**	19	0.75	0.768	0.89	**9.36**
	GBT	0.714	0.742	0.9	60.5	0.762	0.766	0.905	29.9

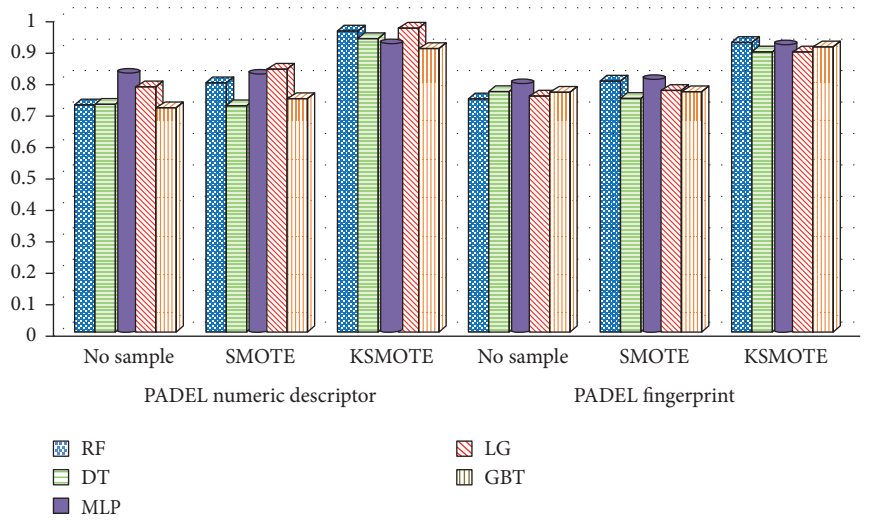

Figure 6: G-mean of numeric PaDEL descriptor and fingerprint for AID 65182.

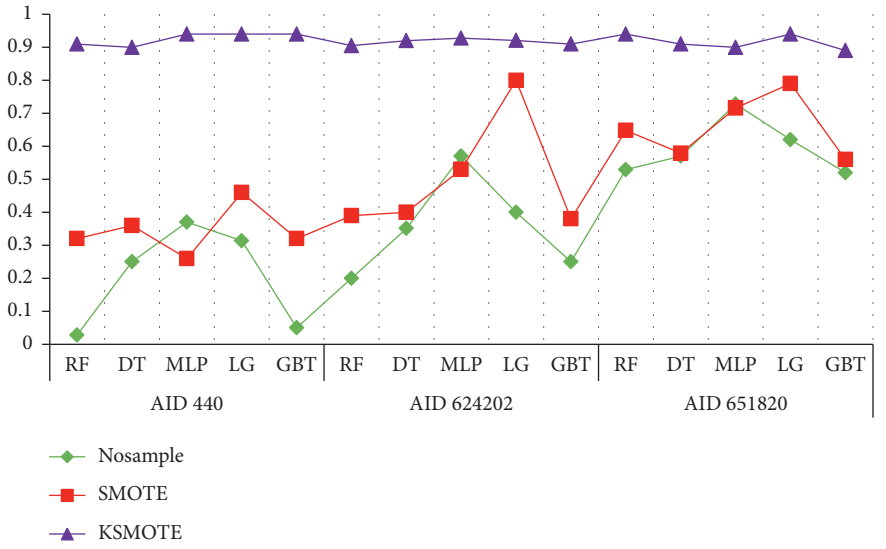

Figure 7: Sensitivity of all datasets for the PaDEL numeric descriptor.

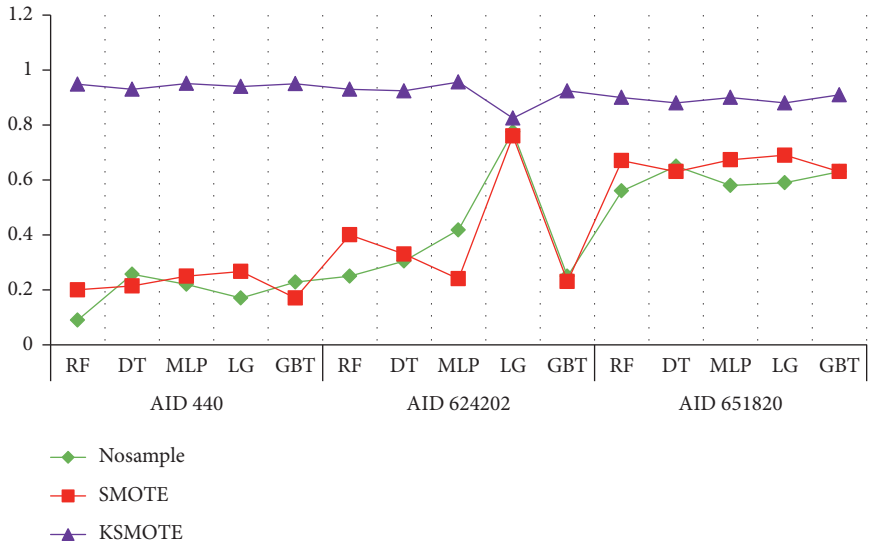

Figure 8: Sensitivity of all datasets for the fingerprint descriptor.

oversampling this minority class to significantly improve perceptibility, with LG jumping from 0.314 to 0.46%. The KSMOTE algorithm has been used to sample this minority class to boost the predictability of the interesting class, which represents active compounds. Besides, the sensitivity increases have been shown in the LG, with the KSMOTE sensitivity value jumped from 0.46 to 0.94 percent.

For the AID 624202 dataset, where the original model hardly recognizes the rare class (active compounds) with exceptionally weak sensitivity, a 0.2 RF is more significantly improved. However, the sensitivity value improves dramatically from 0.4 to 0.8 in LG with the incorporation of SMOTE. In KSMOTE, however, the sensitivity value in MLP rises considerably from 0.8 to 0.928. The model classification of AID 651820 is similarly enhanced, with sensitivity in the majority class in MLP 0.728 (inactive compounds).

Figure 8 presents the sensitivity of the three approaches using fingerprint descriptors. It is obvious that KSMOTE sensitivity values are superior to other approaches. As can be seen, the performance differs based on the type of the used dataset; on the contrary, KSMOTE has a stable performance using different classifiers. In other words, the difference in the KSMOTE is not that noticeable. However, it is clear from the figure that, on average, SMOTE and no-sample approaches have the same performance as well as behavior when applied to all datasets. Besides, the sensitivity results became much better when they are applied on the AID651820 dataset than when AID 440 and AID624202 were used. Again, the results go along with the previous measurement.

4.4. Specificity-Based Performance. Specificity is another important performance measure where it measures the percentage of negative classified classes that are correctly classified. Figures 9 and 10 show the specificity of all classifiers using PaDEL and fingerprint descriptors, respectively. Those figures illuminate two points as follows:

(A) All algorithms, on average, are correctly identifying the negative classes, except SMOTE, LG classifier in both AID 624202 and AID 651820 datasets

(B) Fingerprint descriptor results are more stable than PaDEL descriptor results

To summarize the sensitivity and specificity results, Table 3 shows the produced results using different classifiers. KSMOTE has a superior result in most of the experiments. Sensitivity and specificity results of the three datasets in numeric and fingerprint descriptors are shown, and the values marked in bold are the highest gained values among the results. Those values show the efficiency of the proposed method, KSMOTE.

4.5. Computational Time Comparison. One of the issues that the algorithms always face is the computational time, especially if those algorithms are designed to work on limited-resource devices. The models without SMOTE, for samples with minority classes, cannot achieve adequate performance. On the other hand, the five classifiers' computational time when KSMOTE is used has been proven to be accurate using sensitivity and G-mean values in almost all of the three PubChem datasets (see Figures 4–8). The computed computational time is reported in Figures 11 and 12. It is interesting to note that both PaDEL descriptors and fingerprint PaDEL descriptors produce similar computational efficiency among the five classifiers. From the results, DT and LG give the best computation time among all classifiers, followed by MLP. The computational time values of fingerprint PaDEL are much smaller than the numeric PaDEL descriptor's computational time in most cases. However, looking at the maximum computational time among the classifiers in both Figures 11 and 12, it turns out to be a GBT classifier with a value of 29 and 60 seconds in numeric and fingerprint descriptors.

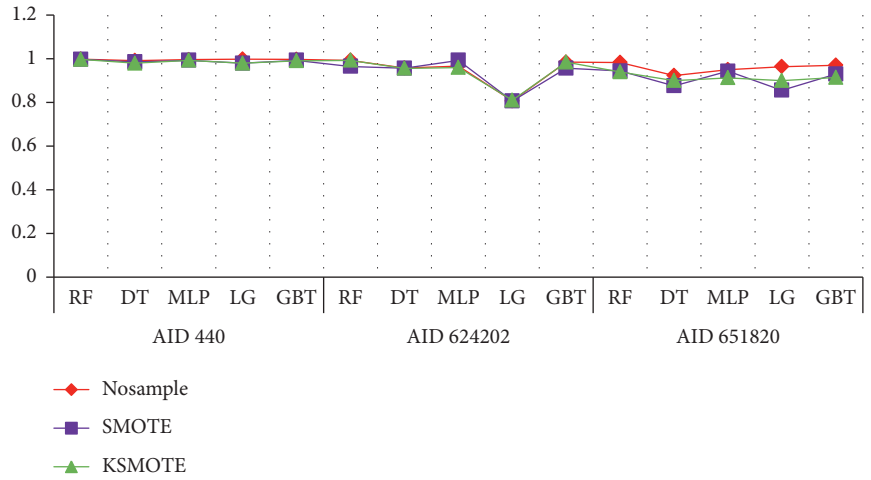

Figure 9: Specificity of all datasets for the PaDEL descriptor.

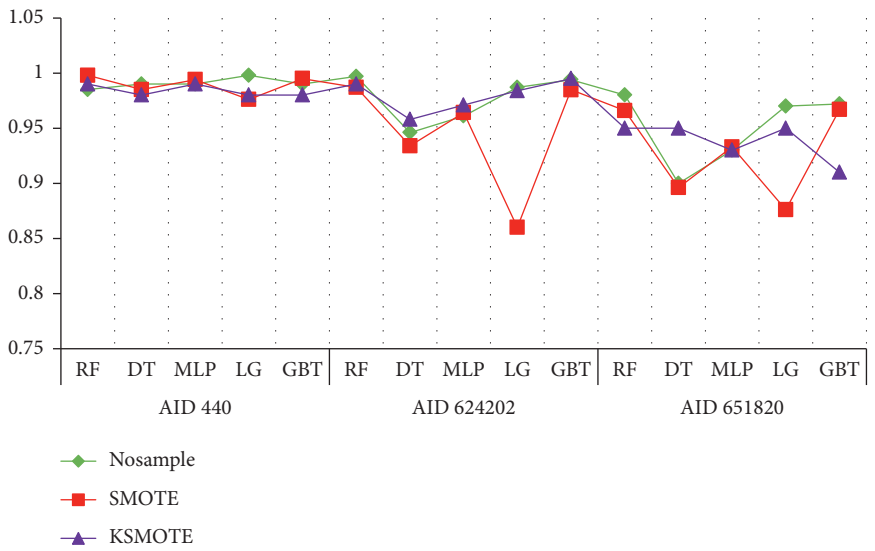

Figure 10: Specificity of all datasets for the fingerprint descriptor.

5. Discussion

It has been considered that the key problem of HTS data is its extreme imbalance, with only a few hits, often identified from a wide variety of compounds analyzed. The imbalance ratio distribution of AID 440 is 1/134, that of AID 624202 is 1/91.5, and for AID 651820, it is 1/23, as shown in Figure 1. Based on the conducted experiments, our proposed model successfully distinguished the active compounds with an average accuracy of 97% and the inactive compounds with an accuracy of 98%, with a G-mean of 97.5%.

Moreover, HTS data size, which typically comprises hundreds of thousands of compounds, poses another challenge. A statistical model may be trained and optimized on such a highly time-intensive dataset. Big data platforms, such as Spark in this study, were computationally effective and dramatically decreased computing costs in the optimized phase and substantially improved the KSMOTE model's performance.

Ideally, the KSMOTE model separates active (minority) dataset from inactive (majority) data with maximum distance. But the KSMOTE model, constructed from an imbalanced dataset, appears to move the hyperplane away from the optimal location to the minority side. Thus, most items are likely to be categorized into the majority class by both no-sample and SMOTE models, leading to a broad difference between specificity and sensitivity. Therefore, such a model's predictability can be significantly weak. We not only rely on cluster sampling to investigate the progress of the KSMOTE model but also built a SMOTE model for each sampling round.

We checked the KSMOTE model's performance with the blind dataset, which included 37 active compounds and 4963 inactive compounds for AID 440. KSMOTE in AID 440 was able to classify the inactive compounds very well with an overall accuracy of >98 percent, while it correctly classifies the active compounds at an accuracy of 95%. However, AID 624202 contains 796 active compounds and 72807 inactive

Table 3: Sensitivity and specificity results of the three datasets in numeric and fingerprint descriptors.

Algorithm		PaDEL numeric descriptor						PaDEL fingerprint					
		No-sample		SMOTE		KSMOTE		No-sample		SMOTE		KSMOTE	
		Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
AID 440	RF	0.028	**0.998**	0.32	**0.997**	0.91	**0.997**	0.09	0.985	0.2	0.998	0.948	**0.99**
	DT	0.25	0.992	0.36	0.986	0.9	0.98	**0.257**	0.99	0.214	0.985	0.93	0.98
	MLP	**0.37**	0.996	0.26	0.993	0.94	0.993	0.22	0.99	0.25	0.994	**0.951**	**0.99**
	LG	0.314	**0.998**	**0.46**	0.98	0.94	0.98	0.17	**0.998**	**0.267**	0.976	0.94	0.98
	GBT	0.05	0.997	0.32	0.993	0.94	0.991	0.228	0.99	0.17	**0.995**	0.95	0.98
AID 624202	RF	0.2	**0.993**	0.39	0.965	0.905	**0.993**	0.25	0.997	0.4	**0.987**	0.93	0.99
	DT	0.351	0.958	0.4	0.957	0.92	0.956	0.305	0.946	0.33	0.934	0.924	0.958
	MLP	**0.57**	0.967	0.53	**0.993**	**0.928**	0.96	0.418	0.961	0.24	0.964	**0.956**	0.971
	LG	0.4	0.807	**0.8**	0.806	0.921	0.81	**0.775**	0.987	**0.76**	0.86	0.825	0.984
	GBT	0.25	0.985	0.38	0.957	0.91	0.985	0.249	**0.994**	0.23	0.9848	0.924	**0.995**
AID 651820	RF	0.529	0.983	0.648	**0.944**	**0.94**	**0.94**	0.56	**0.98**	0.67	0.966	0.9	**0.95**
	DT	0.57	0.923	0.579	0.876	0.91	0.9	**0.65**	0.9	0.63	0.896	0.88	**0.95**
	MLP	**0.728**	0.95	0.716	0.943	0.9	0.913	0.58	0.93	0.673	0.933	0.9	0.93
	LG	0.62	0.964	**0.79**	0.856	**0.94**	0.9	0.59	0.97	**0.69**	0.876	0.88	**0.95**
	GBT	0.52	**0.97**	0.56	0.93	0.89	0.914	0.63	0.972	0.63	**0.967**	**0.91**	0.91

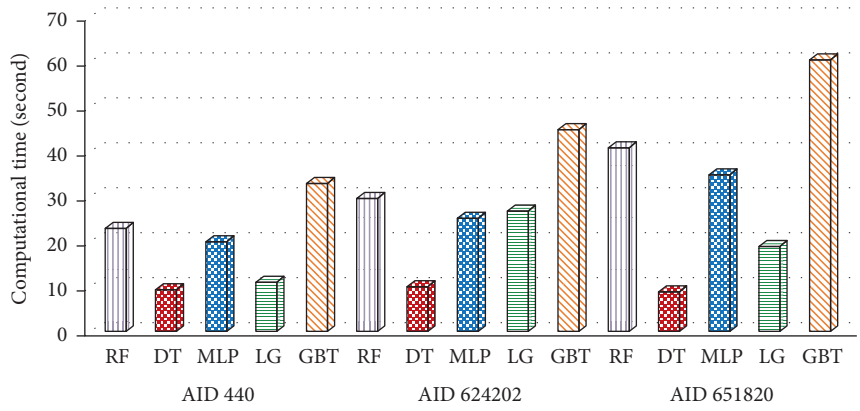

Figure 11: Comparison of computational time (seconds) in KSMOTE for the numeric PaDEL descriptor.

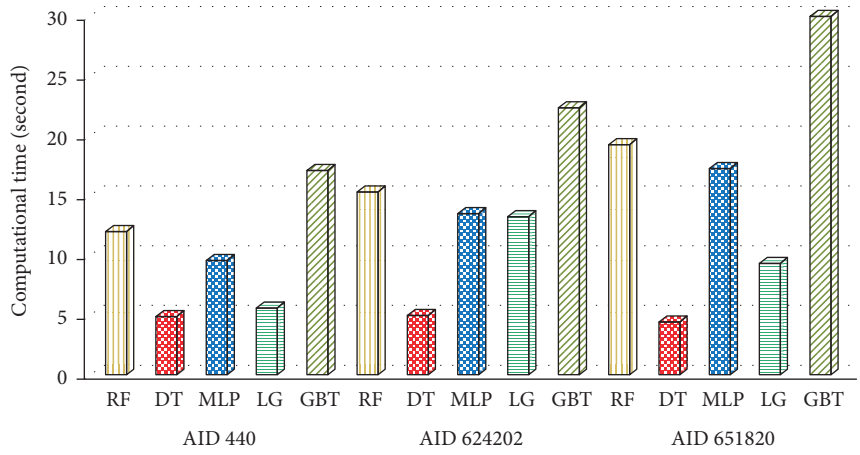

Figure 12: Comparison of computational time (seconds) in KSMOTE for the fingerprint PaDEL descriptor.

compounds, and AID 651820 contains 2332 active compounds and 54269 inactive compounds. Thus, AID 624202 is able to identify inactive compounds very well with an average accuracy of >96 percent, while it correctly classifies the active compounds at an accuracy of 94%.

Also, AID 651820 is able to identify inactive compounds quite well with an average accuracy of >94 percent, while it correctly classifies the active compounds at an accuracy of 92%. KSMOTE is considered better than the other systems (SMOTE only and no-sample) because it depends at the beginning on the lack of similarity in taking the samples of clusters. This dissimilarity between samples increases classifiers' accuracy, and besides that, it used the SMOTE to increase the number of minority samples (active compound) by generating a new sample for the minority.

The strength of KSMOTE lies in the fact that, in addition to the oversampling minority class accurately, CBOS produced new samples that do not affect majority class space in any way. We use the randomness in an effective way by restraining the maximum and minimum values of the newly generated samples.

Recent developments in technology allow for high-throughput scanning facilities, large-scale hubs, and individual laboratories that produce massive amounts of data at an unprecedented speed. The need for extensive information management and analysis attracts increasing attention from researchers and government funding agencies. Hence, computational approaches that aid in the efficient processing and extraction of large data are highly valuable and necessary. Comparison among the five classifiers (RF, DT, MLP, LG, and GBT) showed that DT and LG not only performed better but also had higher computational efficiency. Detecting active compounds by KSMOTE makes it a promising tool for data mining applications to investigate biological problems, mostly when a large volume of imbalanced datasets is generated. Apache Spark improved the proposed model and increased its efficiency. It also enabled the system to be more rapid in data processing compared to traditional models.

6. Conclusion

Building accurate classifiers from a large imbalanced dataset is a difficult task. Prior research in the literature focused on increasing overall prediction accuracy; however, this strategy leads to a bias towards the majority category. Given a certain prediction task for unbalanced data, one of the relevant questions to ask is what kind of sampling method should be used? Although various sampling methods are available to address the data imbalance problem, no single sampling

method works best for all problems. The choice of data sampling methods depends, to a large extent, on the nature of the dataset and the primary learning objective. The results indicate that, regardless of the datasets used, sampling approaches substantially affect the gap between the sensitivity and the specificity of the model trained in the nonsampling method. This study demonstrates the effectiveness of three different models for balanced binary chemical datasets. This work implements both *K*-mean and SMOTE on Apache Spark to classify unbalanced datasets from PubChem bioassay. To test the generalized application of KSMOTE, both PaDEL and fingerprint descriptors were used to construct classification models. An analysis of the results indicated that both sensitivity and G-mean showed a significant improvement after KSMOTE was employed. Minority group samples (active compounds) were successfully identified, and pathological prediction accuracy was achieved. In addition, models created with PaDEL descriptors showed better performance. The proposed model achieved high sensitivity and G-mean, up to 99% and 98.3%, respectively.

For future research, the following points are identified based on the work described in this paper:

(1) It is necessary to find solutions to other similar problems in chemical datasets, such as using semi-supervised methods to increase labeled chemical datasets. There is no doubt that the topic needs to be studied in depth because of its importance and its relationship with other areas of knowledge, such as biomedicine and big data.

(2) It is suggested to study deep learning algorithms for the treatment of class imbalance. Utilizing deep learning may increase the accuracy of the classification overcoming the deficiencies of existing methods.

(3) One more open area is the development of an online tool that can be used to try different methods and decide on the best results instead of working with only one method at a time.

Data Availability

The dataset used is publicly available at (1) AID 440 (https://pubchem.ncbi.nlm.nih.gov/bioassay/440), (2) AID 624202 (https://pubchem.ncbi.nlm.nih.gov/bioassay/624202), and (3) AID 651820 (https://pubchem.ncbi.nlm.nih.gov/bioassay/651820).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] J. Kim and M. Shin, "An integrative model of multi-organ drug-induced toxicity prediction using gene-expression data," *BMC Bioinformatics*, vol. 15, no. S16, p. S2, 2014.

[2] N. Nagamine, T. Shirakawa, Y. Minato et al., "Integrating statistical predictions and experimental verifications for enhancing protein-chemical interaction predictions in virtual screening," *PLoS Computational Biology*, vol. 5, no. 6, Article ID e1000397, 2009.

[3] P. R. Graves and T. Haystead, "Molecular biologist's guide to proteomics," *Microbiology and Molecular Biology Reviews*, vol. 66, no. 1, pp. 39–63, 2002.

[4] B. Chen, R. F. Harrison, G. Papadatos et al., "Evaluation of machine-learning methods for ligand-based virtual screening," *Journal of Computer-Aided Molecular Design*, vol. 21, no. 1-3, pp. 53–62, 2007.

[5] NIH, "PubChem," 2020, https://pubchem.ncbi.nlm.nih.gov/.

[6] R. Barandela, J. S. Sánchez, V. Garćia, and E. Rangel, "Strategies for learning in class imbalance problems," *Pattern Recognition*, vol. 36, no. 3, pp. 849–851, 2003.

[7] L. Han, Y. Wang, and S. H. Bryant, "Developing and validating predictive decision tree models from mining chemical structural fingerprints and high-throughput screening data in PubChem," *BMC Bioinformatics*, vol. 9, no. 1, p. 401, 2008.

[8] Q. Li, T. Cheng, Y. Wang, and S. H. Bryant, "PubChem as a public resource for drug discovery," *Drug Discovery Today*, vol. 15, no. 23-24, pp. 1052–1057, 2010.

[9] L. Cao and F. E. H. Tay, "Financial forecasting using support vector machines," *Neural Computing & Applications*, vol. 10, no. 2, pp. 184–192, 2001.

[10] A. Estabrooks, T. Jo, and N. Japkowicz, "A multiple resampling method for learning from imbalanced data sets," *Computational Intelligence*, vol. 20, no. 1, pp. 18–36, 2004.

[11] C.-Y. Chang, M.-T. Hsu, E. X. Esposito, and Y. J. Tseng, "Oversampling to overcome overfitting: exploring the relationship between data set composition, molecular descriptors, and predictive modeling methods," *Journal of Chemical Information and Modeling*, vol. 53, no. 4, pp. 958–971, 2013.

[12] N. Japkowicz and S. Stephen, "The class imbalance problem: a systematic study1," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[13] G. M. Weiss and F. Provost, "Learning when training data are costly: the effect of class distribution on tree induction," *Journal of Artificial Intelligence Research*, vol. 19, pp. 315–354, 2003.

[14] K. Verma and S. J. Rahman, "Determination of minimum lethal time of commonly used mosquito larvicides," *The Journal of Communicable Diseases*, vol. 16, no. 2, pp. 162–164, 1984.

[15] S. Q. Ye, *Big Data Analysis for Bioinformatics and Biomedical Discoveries*, Chapman and Hall/CRC, Boca Raton, FL, USA, 2015.

[16] S. del Río, V. López, J. M. Benítez, and F. Herrera, "On the use of mapreduce for imbalanced big data using random forest," *Information Sciences*, vol. 285, pp. 112–137, 2014.

[17] A. Fernández, M. J. del Jesus, and F. Herrera, "On the influence of an adaptive inference system in fuzzy rule based classification systems for imbalanced data-sets," *Expert Systems with Applications*, vol. 36, no. 6, pp. 9805–9812, 2009.

[18] K. Sid and M. Batouche, *Ensemble Learning for Large Scale Virtual Screening on Apache Spark*, Springer, Cham, Switzerland, 2018.

[19] V. López, A. Fernández, J. G. Moreno-Torres, and F. Herrera, "Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics," *Expert Systems with Applications*, vol. 39, no. 7, pp. 6585–6608, 2012.

[20] B. Hemmateenejad, K. Javadnia, and M. Elyasi, "Quantitative structure-retention relationship for the kovats retention indices of a large set of terpenes: a combined data splitting-feature

selection strategy," *Analytica Chimica Acta*, vol. 592, no. 1, pp. 72–81, 2007.

[21] S. Jain, E. Kotsampasakou, and G. F. Ecker, "Comparing the performance of meta-classifiers-a case study on selected imbalanced data sets relevant for prediction of liver toxicity," *Journal of Computer-Aided Molecular Design*, vol. 32, no. 5, pp. 583–590, 2018.

[22] A. V. Zakharov, M. L. Peach, M. Sitzmann, and M. C. Nicklaus, "QSAR modeling of imbalanced high-throughput screening data in PubChem," *Journal of Chemical Information and Modeling*, vol. 54, no. 3, pp. 705–712, 2014.

[23] B. X. Wang and N. Japkowicz, "Boosting support vector machines for imbalanced data sets," *Knowledge and Information Systems*, vol. 25, no. 1, pp. 1–20, 2010.

[24] Y. Xiong, Y. Qiao, D. Kihara, H.-Y. Zhang, X. Zhu, and D.-Q. Wei, "Survey of machine learning techniques for prediction of the isoform specificity of cytochrome P450 substrates," *Current Drug Metabolism*, vol. 20, no. 3, pp. 229–235, 2019.

[25] B. K. Shoichet, "Virtual screening of chemical libraries," *Nature*, vol. 432, no. 7019, pp. 862–865, 2004.

[26] L. T. Afolabi, F. Saeed, H. Hashim, and O. O. Petinrin, "Ensemble learning method for the prediction of new bio-active molecules," *PLoS One*, vol. 13, no. 1, Article ID e0189538, 2018.

[27] A. C. Schierz, "Virtual screening of bioassay data," *Journal of Cheminformatics*, vol. 1, no. 1, p. 21, 2009.

[28] S. K. Hussin, Y. M. Omar, S. M. Abdelmageid, and M. I. Marie, "Traditional machine learning and big data analytics in virtual screening: a comparative study," *International Journal of Advanced Research in Computer Science*, vol. 10, no. 47, pp. 72–88, 2020.

[29] l. E. Frank and J. H. Friedman, "A statistical view of some chemometrics regression tools," *Technometrics*, vol. 35, no. 2, pp. 109–135, 1993.

[30] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[31] Q. Li, Y. Wang, and S. H. Bryant, "A novel method for mining highly imbalanced high-throughput screening data in PubChem," *Bioinformatics*, vol. 25, no. 24, pp. 3310–3316, 2009.

[32] R. Guha and S. C. Schürer, "Utilizing high throughput screening data for predictive toxicology models: protocols and application to MLSCN assays," *Journal of Computer-Aided Molecular Design*, vol. 22, no. 6-7, pp. 367–384, 2008.

[33] P. Banerjee, F. O. Dehnbostel, and R. Preissner, "Prediction is a balancing act: importance of sampling methods to balance sensitivity and specificity of predictive models based on imbalanced chemical data sets," *Frontiers in Chemistry*, vol. 6, 2018.

[34] P. W. Novianti, V. L. Jong, K. C. B. Roes, and M. J. C. Eijkemans, "Factors affecting the accuracy of a class prediction model in gene expression data," *BMC Bioinformatics*, vol. 16, no. 1, p. 199, 2015.

[35] Spark, "Apache spark is a unified analytics engine for big data processing," 2020.

[36] P. AID440, "Primary HTS assay for formylpeptide receptor (FPR) ligands and primary HTS counter-screen assay for formylpeptide-like-1 (FPRL1) ligands," 2007, https://pubchem.ncbi.nlm.nih.gov/bioassay/440.

[37] P. 624202 AID, "qHTS assay to identify small molecule activators of BRCA1 expression," 2012, https://pubchem.ncbi.nlm.nih.gov/bioassay/624202.

[38] P. 651820 AID, "qHTS assay for inhibitors of hepatitis C virus (HCV)," 2012, https://pubchem.ncbi.nlm.nih.gov/bioassay/651820.

[39] Chem.libretexts.org, "Molecules and molecular compounds," 2020, https://chem.libretexts.org/Bookshelves/General_Chemistry/Map%3A_Chemistry_-_The_Central_Science_(Brown_et_al.)/02._Atoms_Molecules_and_Ions/2.6%3A_Molecules_and_Molecular_Compounds.

[40] X. Wang, X. Liu, S. Matwin, and N. Japkowicz, "Applying instance-weighted support vector machines to class imbalanced datasets," in *Proceedings of the 2014 IEEE International Conference on Big Data (Big Data)*, pp. 112–118, IEEE, Washington, DC, USA, December 2014.

[41] V. H. Masand, N. N. E. El-Sayed, M. U. Bambole, V. R. Patil, and S. D. Thakur, "Multiple quantitative structure-activity relationships (QSARs) analysis for orally active trypanocidal N-myristoyltransferase inhibitors," *Journal of Molecular Structure*, vol. 1175, pp. 481–487, 2019.

[42] S. W. Purnami and R. K. Trapsilasiwi, "SMOTE-least square support vector machine for classification of multiclass imbalanced data," in *Proceedings of the 9th International Conference on Machine Learning and Computing - ICMLC*, pp. 107–111, Singapore, Singapore, Febuary 2017.

[43] T. Cheng, Q. Li, Y. Wang, and S. H. Bryant, "Binary classification of aqueous solubility using support vector machines with reduction and recombination feature selection," *Journal of Chemical Information and Modeling*, vol. 51, no. 2, pp. 229–236, 2011.

[44] B. X. Wang and N. Japkowicz, "Boosting support vector machines for imbalanced datasets Knowledge and Information Systems," *Knowledge and Information Systems*, vol. 25, no. 1, p. 25, 2010.

[45] S. P. R. Trapsilasiwi, "SMOTE-least square support vector machine for classification of multiclass imbalanced data," in *Proceedings of the 9th International Conference on Machine Learning and Computing*, pp. 107–111, ACM, Singapore, February 2017.