Incremental SMOTE with Control Coefficient for Classifiers in Data Starved Medical Applications

Wan D. Bae[1], Shayma Alkobaisi[2*], Siddheshwari Bankar[1], Sartaj Bhuvaji[1], Jay Singhvi[1], and Madhuroopa Irukulla[1] William McDonnell[1]

[1] Computer Science, Seattle University, Seattle, WA, USA
{baew,sbankar,sbhuvaji,jsinghvi,mirukulla,mcdonn11}@seattleu.edu
[2] College of Information Technology, United Arab Emirates University, Al Ain, UAE
{shayma.alkobaisi}@uaeu.ac.ae

Abstract. Prediction models for data starved medical applications lag behind general machine learning based solutions, despite the fact that the development of machine learning techniques has shown to improve early interventions. This is due to the fact that optimization approaches used to train models implicitly assume a balanced distribution of events in training data, yet medical data frequently has an imbalanced distribution of data samples within classes. The curse of dimensionality is further exacerbated by the fact that small samples and a high number of features make up the data for individual-based risk prediction models. In this paper, we proposed a data augmentation method to gradually create synthetic minority samples with control coefficient, which improves the quality of the generated data over time and consequently boosts prediction model performance. This is achieved by incrementally adjusting to the data distribution and thus avoiding overfitting the training data. We further employ four cutting-edge oversampling techniques to evaluate the proposed technique on real asthma patient data. Results of our study show that this method enhanced classifiers' overall performance across all four techniques. Specifically, by using the incremental data augmentation method on three oversampling methods, the transfer learning based classifiers led to an increase in sensitivity of 4.01% - 7.79%.

Keywords: class imbalance problem, synthetic minority oversampling technique, rare event prediction, data starved contexts, control coefficient

1 Introduction

Recent advances in machine learning play a pivotal role in informing clinical decision-making, enabling early disease detection, and enhancing patient care across a spectrum of medical conditions. For example, avoidable asthma exacerbation account for 63% of the total annual asthma costs; thus, preventive approaches to evaluate corresponding likelihood of symptom change and predict

* Corresponding author

the risk level of a patient's exacerbation can improve health care quality and achieve significant economical savings [14].

The construction of robust predictive models in the healthcare domain encounters a common challenge: the presence of class imbalance within medical datasets. Class imbalance arises when certain health conditions, demographics, or medical outcomes are rare or underrepresented, hence the class of interest (minority class) has significantly fewer samples than the other classes (majority classes). This leads to biases and the compromising of the performance of machine learning algorithms since accurately detecting minority class samples has a higher impact on the delivery of effective healthcare services. In response to this issue, our work aims to address the pervasive problem of class imbalance in healthcare datasets, with a focus on enhancing the prediction ability of machine learning models.

Data augmentation techniques are under active research to improve prediction models for various applications. While several sampling-based approaches have been proposed including various synthetic minority oversampling techniques (SMOTE) [1, 7, 8, 10, 17], little exists in the literature about their success in data starved contexts, which include many machine learning (ML) tasks in the medical domain where recording only several observations per day for patients is common and hence the performance of ML models is degraded.

In this paper, we propose an incremental synthetic data generation system to improve the quality of synthetic data generated by SMOTE variants and hence mitigate the class imbalance problem in individual-level health risk applications where the size of datasets is very small. We also propose to utilize control coefficient in the data generation process for data diversification.

Our research unfolds in several stages. First, we conduct data-level analysis to assess the quality of synthetic data generated by SMOTE variants within the incremental data generation system. This analysis leverages several metrics for data quality assessment. Next, we train baseline classifiers on the augmented training data and evaluate the performance of classifiers on the testing data. We assess the effectiveness of the proposed system for SMOTE variants to enhance classifiers for rare event prediction. Additionally, we explore the application of transfer learning (TL) for their potential benefits in improving the performance of classifiers, particularly their prediction ability for the minority class. TL is a deep learning technique focusing on retraining a large data model with a small amount of specialized training data and is therefore well-suited to data starved contexts.

2 Related Work

Synthetic Minority Oversampling Technique (SMOTE) is a common data augmentation method to solve the class imbalanced problem. The main idea of SMOTE is to generate synthetic samples similar to the minority class data to achieve a more balanced class distribution of samples. In [1], the authors introduced the first SMOTE method that generates synthetic samples between

two neighboring minority ckass samples by linear interpolation. The basic steps of the method are: (1) select a sample x from the minority class and find its k nearest neighbor samples within the minority class, (2) randomly select one neighboring sample y from the k nearest neighbors found in step 1, and (3) generate a new sample x_{new} by linear interpolation between x and y, $x_{new} = x + random(0, 1) * (y - x)$.

With SMOTE [1], it is likely that a lot of data will lie on the same line and therefore, researchers have explored different data distributions to reduce noise created by the random function in SMOTE. The authors in [10] proposed Gaussian SMOTE (G-SMOTE) that also generates synthetic samples between minority samples using k-nearest neighbors and linear interpolation but utilizes the Gaussian (normal) distribution to generate new samples that deviate from the line, but not so far that it degrades performance. The authors demonstrated that, while SMOTE generated a significant amount of duplicated data, G-SMOTE generated more widespread data. On the other hand, Gamma Distribution SMOTE (Gamma-SMOTE) [8] utilizes the gamma distribution to create new minority class points and produces data in a non-linear fashion, thus giving rich geometric structure. Since the Gamma distribution is asymmetric, new minority points are generated close to the existing minority data sample. Similar to SMOTE, Sample Density Distribution SMOTE (SDD-SMOTE) [17] generates synthetic samples similar to SMOTE [1] but considers the total dataset distribution and local sample density to reduce fuzzy classification boundaries and control the randomness of the SMOTE algorithm. It works by calculating the density of minority class samples, generating synthetic data points in regions of high density, and ensuring a balanced dataset. Specifically, it identifies the k-nearest neighbours of minority samples, measures their density, and generates synthetic samples with controlled coefficients to balance class distribution. SDD-SMOTE aims to improve the training of machine learning models by addressing the challenges posed by imbalanced data.

SMOTE and its variants proposed a wide range of solutions to the class imbalance problem. However, the SMOTE methods are not proven as a solution to data-starved contexts where datasets being used are very small and thurs a lot more synthetic data samples need to be generated to rebalance the classes; this is the case for our individual patient datasets for health risk prediction modeling where the average size of a patient's dataset is 170 and the imbalance ratio is 4.0 (the raio of majority class size and minority class size). It necessities an algorithmic level approach to enhance the quality of generated synthetic datasets in addition to data-level approaches that can be achieved by the SMOTE methods.

With the rise in popularity of deep learning, the methods for synthetic data generation utilizing deep learning have also expanded, which include Generative Adversarial Networks (GAN) and autoencoders. GAN and autoencoders are predominantly used to generate synthetic images [5, 4, 9], however, there are several recent work for tabular data using these methods, including conditional tabular GAN [19], SMOTified GAN technique [16] and variational autoencoders [3, 18], which demonstrated the potential of GAN and auto encoders to stretch limited

tabular data and improve the performance of prediction models. Yet, they have not been successful on small training data. Thus, in our experiments, we presented the performance of classifiers trained on the agumented data using an autoencoder [11] for addressing this limitation.

Boosting techniques for SMOTE, such as integrating AdaBoost in the oversampling process is a natural way to enhance the SMOTE methods. SMOTEBoost [2] is a combination of SMOTE and Boosting algorithm to improve the SMOTE algorithm [1]. The boosting algorithm combines the weak learners' predicted outcomes to convert them into strong learners by assigning weights to the dataset instances and stressing misclassified cases. In a given scenario, a weak learner, i.e., AdaBoost, decision tree is first trained on the augmented data created by the SMOTE algorithm, and then the weights of the instances are adjusted based on the misclassification rate until the predefined boosting iterations are finished.

Similar techniques of combining SMOTE and a boosting algorithm have been proposed to enhance SMOTE. SMOTEBoost was evaluated for the regression-based task in addition to its use for classification [12] using 30 different datasets with evaluation of its four variants of SMOTEBoost. The main difference between this approach from SMOTE with AdaBoost is that it introduces the preprocessing steps and some additional approaches before the weak learner. The ranking evaluation method was adopted, which showed that the proposed approach is a better rank than AdaBoost and other variants. In [13], authors implemented a SMOTEBoost method for binary classification on imbalanced microarray based on two datasets, i.e., colon cancer and myeloma, by applying SMOTEBoost and the support vector machine algorithm. The study results showed that SMOTEBoost with SVM outperforms SMOTE-SVM and AdaBoost-SVM in terms of geometric mean on both datasets. A similar work utilizing SMOTE and AdaBoost was proposed in [15]. The proposed technique used the weight adjustment on synthetic data to overcome the noise created by SMOTE. The study results demonstrated that SMOTE with a boosting method reduced the noisy area between the two classes.

While the existing SMOTEBoost borrowed general ML boosting techniques used for model performance improvement, our work focuses on a new algorithm-level approach to enhance the quality of synthetic data generated by the SMOTE methods, hence a general system that can be extended to use other synthetic data generation methods.

3 Methods

In this paper, we investigate the application of incremental synthetic data generation that can lead to improved results, given the performance of the SMOTE methods in literature. We first evaluate four existing SMOTE methods and then introduce an incremental data generation system with control coefficient. "Incremental boosting" typically refers to a technique where boosting algorithms are applied sequentially in multiple stages, each building on the results of the

previous stage. Each stage involves training a weak learner (e.g., decision tree or other simple model) on the dataset, with a focus on the misclassified or difficult-to-classify instances from the previous stage. The goal is to iteratively correct the mistakes made by previous learners and improve the overall accuracy of the ensemble model. In our context, incremental boosting is applied in multiple iterations to enhance the performance of the models. After each iteration, new synthetic samples are generated, and the boosting process is reapplied to further refine the model's ability to classify minority class instances. We implement the system with a simple boosting technique using baseline classifiers, focusing on the model improvement through incremental data generation. Figure 1 illustrates the overview of our proposed incremental synthetic data generation system.

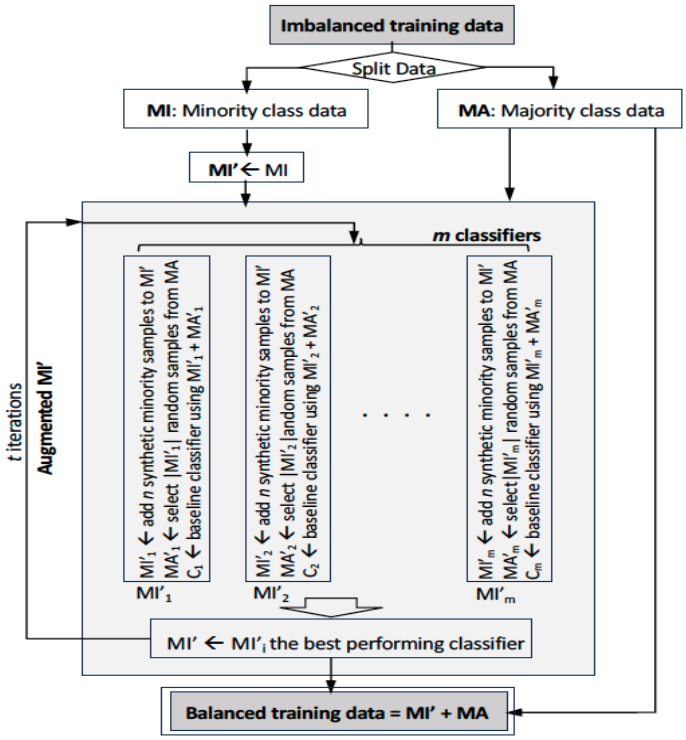


Fig. 1. Overview of an incremental synthetic data generation system

3.1 An Incremental Synthetic Data Generation Method

The incremental synthetic data generation method breaks the data generation process of SMOTE into several iterations. In each iteration, m number of subsets of minority class samples are generated and they are added to the current training data. Then the method trains and validates m number of classifiers, select a

subset of the synthetic data producing the best performing classifier and adds it into the current minority class dataset. In the next iteration, the method generates minority class samples by using the updated minority class dataset. The proposed method incrementally builds the augmented training data by adding n synthetic minority samples to the training data at each iteration, thus allowing more diverse data to be generated in the next iteration. Details of the proposed method are described in Algorithms 1 and 2.

The advantage of the incremental synthetic data generation method is that while SMOTE generates data based on the original dataset minority class, the proposed synthetic data generation method adds more samples of the minority class to the original dataset at each iteration. Adding and using similar data from previous iterations results in generating more diverse data with a moderate increase standard deviation and in turn, a better to handle the overfitting problem. The criteria to evaluate the quality of generated synthetic data are defined in Section 4.2.

Algorithm 1 $SMOTE_{incrCC}$ (D_{train}, A, k)

Input: D_{train} is a set of number of class-labeled training data points; A is a SMOTE method; k: k-nearest neighbor used in A
Output: $D_{balanced}$ is a balanced training dataset augmented by synthetic data
Method:
Split D_{train} into majority class dataset (MA) and minority class dataset (MI) and record their number of instances: $N_{MA} = |MA|$ and $N_{MI} = |MI|$
$N_{syn} \leftarrow N_{MA} - N_{MI}$
$t \leftarrow$ the number of iterations in the cremental data generation based on N_{syn}
$G \leftarrow$ Split N_{syn} into G. G is an array of size k, that will store the numbers of synthetic data that is generated for each iteration, roughly $n = \lceil \frac{N_{syn}}{t} \rceil$.
Initailization: $i \leftarrow 1$, $MI' \leftarrow MI$; MI' is the augmented minority class data
repeat
    $n \leftarrow G[i]$, the size of synthetic data for iteration i
    $N \leftarrow MI' + n$, N is the number of samples for both classes
    $MA' \leftarrow$ Randomly select N number of majority samples from MA
    $MI' \leftarrow generateData(MI', MA', m, A, k)$
    $i \leftarrow i + 1$
until i reaches t; all synthetic data specified in G are generated
$D_{balanced} \leftarrow MI' + MA$
return $D_{balanced}$

3.2 Incremental SMOTE with Control Coefficient

To further improve the SMOTE variants, we adopt the Control Coefficient (CC) from the SDD-SMOTE algorithm [17] where CC was used to solve the limitation of the uniform random function for synthesizing new samples. We use CC in the proposed incremental synthetic data generation method with consideration of

Algorithm 2 *generateData* (MI', MA', m, A, k)

Input: MI' and MA' are the current minority and majority class datasets ($|MI'| \approx |MA`| - n$); m is the number of classifiers; A is a SMOTE method, k is k nearest neighbor used in A
Output: A minority dataset MI' augmented by n synthetic data
Method:
$S \leftarrow S$ is an array of size m storing a synthetic dataset, initially assign {} for all element in S
$C \leftarrow C$ is an array of size m storing performance evaluation metrics, initially set to 0.
$i \leftarrow 1$
repeat
    $S[i] \leftarrow$ Generate n synthetic data using a SMOTE A with k and the precalculated control coefficient
    $D'_{train} \leftarrow MI' + S[i] + MA'$
    $C[i] \leftarrow$ Train a classifier and evaluate the model using D'_{train}
until i reachers m
$S_{best} \leftarrow$ Find a synthetic dataset in $S[i]$ that results in the performing classifier $C[i]$.
$MI' \leftarrow MI' + S_{best}$, augment MI' with the best synthetic dataset
return MI'

the distribution of synthetic data points related to the dense area of minority samples. The CC value is calculated in data preprocessing and applied when the SMOTE$_{incrCC}$ algorithm generates synthetic data points. The steps of the CC calculation are below:

1. Calculate the average Euclidean distance between all minority samples as D_{pos}.
2. Calculate the average Euclidean distance between all minority samples and majority samples as D_{neg}.
3. During the synthesis process of a new sample, calculate the average Euclidean distance D_1 between the selected minority and its K minority class neighbors.
4. Repeat Step (3) to calculate the average Euclidean distance D_2 for k majority class neighbors.
5. Calculate the relative distance ratio μ according to D_{pos}, D_{neg}, D_1, and D_2: $\mu = \frac{D_1 * D_{neg}}{D_2 * D_{pos}}$
6. According to μ, calculate the value range of CC. Below is an example of calculation used for SDD and SDD-SMOTE methods. Gausian and Gamma probability distribution functions are used in calculating CC values for G-SMOTE and Gamma-SMOTE.

$$CC = \begin{cases} random(0,1) & \text{if } \mu < 1 \\ 0.5 + 0.5 * random(0,1) & \text{if } 1 \leq \mu \leq 2 \\ 0.8 + 0.2 * random(0,1) & \text{if } \mu > 2 \end{cases}$$

Our proposed system incorporates with four state-of-the-art SMOTE methods, such as SMOTE [1], Gaussian SMOTE (G-SMOTE) [10], Gamma Distribu-

tion SMOTE (Gamma-SMOTE) [8], and Sample Density Distribution SMOTE (SDD-SMOTE) [17]. The proposed method can work with various SMOTE methods and enhance their ability to generate high quality synthetic data. The performance of the four SMOTE methods with the proposed incremental system is compared to the performance of the original SMOTE method using real asthma patients' datasets in Section 4.

4 Experiments

4.1 Datasets and Experiments Setup

Our datasets include 20 nonsmoking asthma patients' data consisting of 27 min-max normalized variables along with a binary label as class [0, 1] representing a risk zone or not. Patients' peak expiratory flow rate (PEFR), environmental exposure data (indoor and outdoor air quality), and behavioral data (home location, cooking habit and income level) were collected through a case study [double-blind]. Patients' exposures to environmental variables were estimated using 24-hour time window at each PEFR measurement. The high risk zone is defined as a PEFR less than a patient's critical cutoff, PEFR_C, which is suggested by medical practitioners. In our study, we set PEFR_C to the 20% quantile PEFR value of a patient's dataset and the samples below PEFR_C are the minority samples. The total number of data in each patient's dataset is small with varying sizes between 88 and 210 samples as shown in Table 1. Our datasets exhibit an imbalanced class distribution ranging from 2.32 to 5.52 (avg. 3.98). This imbalance ratio is much higher than that in the datasets used in SDD-SMOTE [17] where the imbalance ratio ranges from 1.21 to 2.60 (avg. 1.84).

Table 1. A summary of class imbalance in 20 asthma patients' datasets

# samples	# minority class (MI)	# majority class (MA)	Imbalance ratio
88 - 210 (avg. 168)	16 - 38 (avg. 35)	72 - 172 (avg. 132)	2.32 - 5.52 (avg. 3.98)

Our data analysis and experiments for classification models were developed in Python 3.8 and Keras framework. Model hyperparameters were selected through extended training and validation processes using k-fold cross validation (CV) to avoid overfitting while increasing the performance of the models. The models were tested using testing data that were not included in the training/validation phases.

4.2 Statistical Analysis on Synthetic Data

Data evaluation metrics Synthetic data generation for rebalancing classes focuses on three factors (1) maintaining similar probability density functions

of variables within the augmented training dataset, (2) keeping similar class boundaries, and (3) increasing data diversity to reduce overfitting. Data diversity refers to a robust synthetic data that is sufficiently representative of the data to prevent biasing, hence it works well with various classification algorithms. In our experiments, we used several statistical metrics for analyzing the data generated by the proposed method.

Factor (1) is roughly measured by the difference between the means of the original and generated datasets, with a smaller percent difference indicating better maintenance of the data distribution. Factor (2) is also measured by the difference in standard deviations with a very large increase (>15%) indicating possibly excessive boundary distortions. Factor (3) can roughly be measured by the difference between the standard deviations of the original and generated datasets, with a moderate increase/decrease indicating a healthy increase in diversity.

One method to assess the overall quality of synthetically generated data is Gretel [6], which is an open platform measuring a metric called "Gretel Score" that estimates how well the generated synthetic data maintains the same statistical properties as the original dataset. A synthetic data quality score is computed by comparing the distributional distance between the principal components in the original data and those in the synthetic data. The closer the principal components are, the higher the quality score will be. Finally, since the goal of minority class oversampling is to improve training by diversifying the minority class, a small fraction of the generated data duplicating original data is desirable.

Diversity of the generated data can be also measured by Kullback-Leibler (KL) divergence and Kernel density estimation (KDE). The KL divergence is a measure that quantifies the difference between two probability distributions, P and Q: $D_{KL}(P||Q) = \sum_{x \in X} P(x) log(\frac{P(x)}{Q(x)})$, where $P(x)$ is the probability of observing data point x in the original data distribution and $Q(x)$ is the probability of observing data point x in the synthetic data distribution. The summation is taken over all possible data points x in the dataset. Specifically, it calculates the information lost when using distribution Q to approximate distribution P. The KL divergence assesses the similarity between the distribution of real data and synthetically generated data. If the two distributions are identical, the KL divergence will be zero, indicating a perfect match between the real and synthetic data. Conversely, a higher KL divergence value suggests a greater dissimilarity between the distributions. The KDE measure is the process of estimating an unknown probability density function using the sum of a kernel function on every data point. By visualizing KDE distribution of the original data and the generated synthetic data, we can compare not only the sum of diverging area of the two distributions but also the shapes of data distributions.

Statistical Analysis Table 2 presents a summary of statistics on the generated data by the proposed incremental data generation method comparing to those generated by the existing SMOTE methods and an autoencoder, TVAE synthesizer [11]. Overall, the percentages of difference in mean values of the original

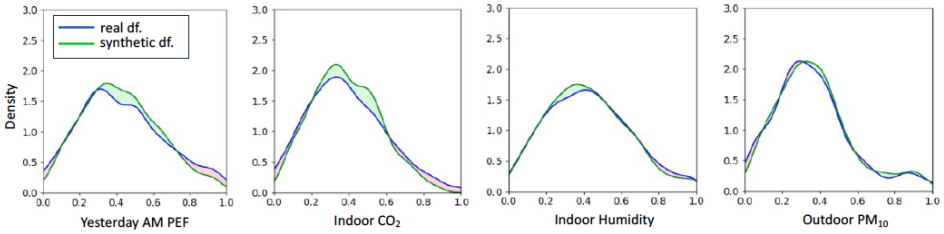
Table 2. Statistical summary and duplicate ratio on the generated synthetic data

Method	Mean diff.	STD diff.	Gretel score	KL	KDE area
`SMOTE`	0.27%	13.08%	92.63	0.008074	1.02
`SMOTE`$_{incrCC}$	0.33%	12.31%	92.95	0.006867	0.85
`G-SMOTE`	1.33%	14.17%	91.88	0.012217	1.32
`G-SMOTE`$_{incrCC}$	0.96%	10.32%	91.58	0.015719	1.44
`Gamma-SMOTE`	0.69%	14.73%	90.25	0.011598	1.17
`Gamma-SMOTE`$_{incrCC}$	0.56%	12.03%	92.84	0.004292	0.69
`SDD-SMOTE`	2.24%	12.95%	93.74	0.005129	0.82
`SDD-SMOTE`$_{incrCC}$	1.84%	11.03%	91.52	0.006220	0.85
TVAE synthesizer [11]	3.84%	33.21%	77.34	0.070124	2.84

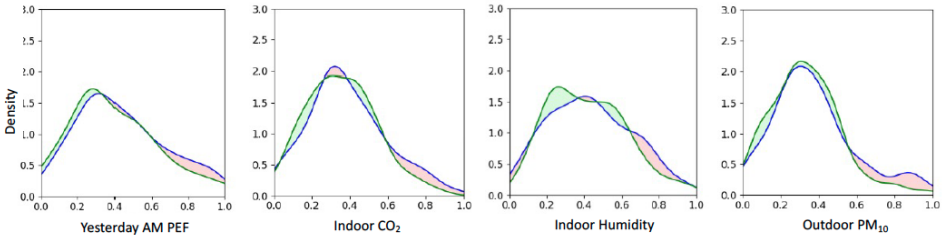
data and all generated data by the methods are relatively small, between 0.33% and 2.24%. On the other hand, standard deviation value differences between the original data and the generated data by the SMOTE methods and those with the incremental data generation range between 10.32% and 15.73%. Standard deviations in the data generated by the incremental data generation method were reduced comparing to the existing SMOTE methods.

This analysis shows the effectiveness of the proposed method at the data-level, increasing the diversification task while remaining competitive in terms of maintaining data distribution. For example, with 92.95 Gretel score and 0.0068 KL value, SMOTE$_{incrCC}$ provides a balance between preserving the statistical properties of the original data and providing reasonable diversity. Similar characteristics can be observed with the other incremental methods, which demonstrates the applicability of the proposed method. On the other hand, the data generated by TVAE synthesizer shows the highest mean and standard deviation difference from those in the original data. This resulted in the lowest Gretel score and the highest values in KL and KDE area.

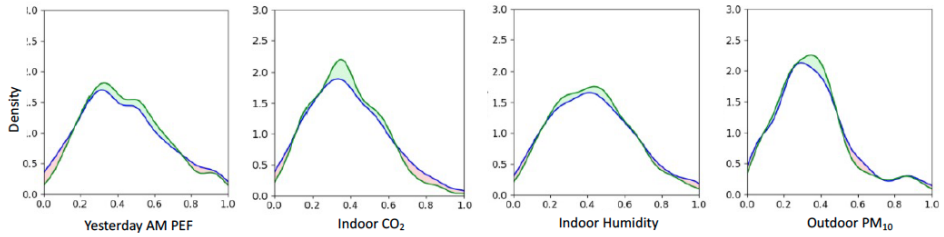
Data quality evaluation on the generated data can be also performed by visualizing KDE distributions. Figure 2 illustrates the KDE divergence of the real minority class samples and the synthetic minority samples generated by the four SMOTE variants for four selected variables; patients' yesterday morning PEFR, indoor CO_2, indoor humidity, and outdoor PM_{10}. Visually, we see that the data distributions of synthetic data by all SMOTE variants follow similar patterns to the real minority data distributions, while G-SMOTE$_{incrCC}$ and SDD-SMOTE$_{incrCC}$ presented moderate-level diverging distributions which can be seen as data diversity in the synthetic data, and thus reduced overfitting in classification. The two incremental SMOTE methods achieves data diversification in different ways, increasing or decreasing data density differently with respect to neighboring densities. The differences seem large enough that each algorithm might be expected to excel for particular data distributions or with particular classification algorithms, but further study is needed to provide guidelines and best practices applicable to specific cases.



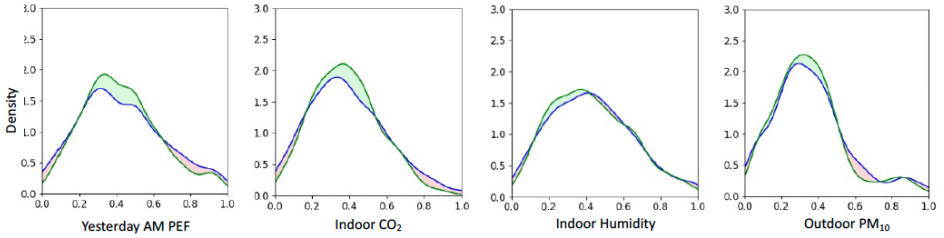
(a) SMOTE$_{incrCC}$



(b) G-SMOTE$_{incrCC}$



(c) Gamma-SMOTE$_{incrCC}$



(d) SDD-SMOTE$_{incrCC}$

Fig. 2. Comparisions of KDE diverging area
(selected variables: yesterday AM PEF, indoor CO_2 & humidity, and outdoor PM_{10})

Table 3. Improvement by SMOTE variants in classification performance

Classifier	Method	Accuracy	Sensitivity	Specificity	Precision	F_1 score
DT	No oversampling	0.5801	0.2780	0.8822	0.5821	0.5663
	SMOTE	0.5813	0.3663	0.7963	0.5726	0.5692
	G-SMOTE	0.5731	0.3781	0.7680	0.5668	0.5577
	Gamma-SMOTE	0.5849	0.3901	0.7797	0.5798	0.5676
	SDD-SMOTE	**0.5955**	**0.3979**	**0.7931**	**0.5849**	**0.5772**
KNN	No oversampling	0.5443	0.1279	0.9607	0.5575	0.5195
	SMOTE	0.5988	**0.5297**	0.6679	**0.5828**	**0.5673**
	G-SMOTE	**0.6016**	0.5241	**0.6791**	0.5798	0.5635
	Gamma-SMOTE	0.5901	0.5261	0.6541	0.5665	0.5452
	SDD-SMOTE	0.5974	0.5225	0.6722	0.5783	0.5615
LR	No oversampling	0.5303	0.0762	0.9844	0.4656	0.4864
	SMOTE	0.6168	0.5028	0.7308	0.5933	0.5852
	G-SMOTE	0.6199	0.5260	**0.7338**	**0.5992**	**0.5884**
	Gamma-SMOTE	**0.6215**	0.5195	0.7134	0.5918	0.5812
	SDD-SMOTE	0.6180	**0.5290**	0.7089	0.5894	0.5805
NB	No oversampling	0.5383	0.0987	0.9778	0.5024	0.5019
	SMOTE	0.5992	0.3783	**0.8201**	0.5995	**0.5910**
	G-SMOTE	0.5935	**0.3921**	0.7948	0.5880	0.5823
	Gamma-SMOTE	0.5892	0.3761	0.8023	0.5838	0.5753
	SDD-SMOTE	**0.6011**	0.3880	0.8141	**0.6016**	0.5876

4.3 Performance Evaluation on Classifiers

Classifier evaluation metrics In our experiments of health risk prediction on asthma patients' datasets, "positive" samples are the data tuples in which a patient's PEFR value is within the patient's high risk zone. We present the result using the most commonly used metrics for binary classification: (1) weighted accuracy, (2) sensitivity, (3) specificity, (4) precision average, and (5) F_1-score average, and (6) Receiver Operating Characteristic Area Under the Curve (ROCAUC). While all these metrics are equally important for evaluating classifiers, we focus on improving sensitivity scores that represent a model's ability to predict health risk correctly.

Performance Evaluation The effectiveness of the proposed method was tested through baseline classifiers where we implemented four conventional classification algorithms: (1) Decision Tree (DT), (2) K-Nearest Neighbors (KNN), (3) Logistic Regression (LR), and (4) Naive Bayes (NB). For each classifier, we generate and augment synthetic minority data samples to the original dataset.

First, we present the average performance analysis of the original SMOTE methods comparing to no oversampling with the baseline classifiers. As shown in Table 3, all four SMOTE methods significantly outperformed the no oversampling in weighted accuracy, sensitivity, precision and F_1 score for all base

Table 4. SDD-SMOTE vs. SDD-SMOTE$_{incrCC}$ in baseline classifier performance

Classifier	Method	Accuracy	Sensitivity	Specificity	Precision	F_1 score
DT	SDD-SMOTE	0.5955	0.3979	**0.7931**	0.5849	0.5772
	SDD-SMOTE$_{incrCC}$	**0.5987**	**0.4194**	0.7780	**0.5884**	**0.5783**
		(+0.54%)	(+5.40%)	(-1.90%)	(+0.60%)	(+0.19%)
KNN	SDD-SMOTE	0.5974	0.5255	0.6722	0.5783	0.5615
	SDD-SMOTE$_{incrCC}$	**0.6177**	**0.5287**	**0.7127**	**0.5904**	**0.5710**
		(+3.40%)	(+0.61%)	(+6.02%)	(+2.09%)	(+1.69%)
LR	SDD-SMOTE	0.6180	0.5290	0.7089	0.5894	0.5805
	SDD-SMOTE$_{incrCC}$	**0.6453**	**0.5447**	**0.7459**	**0.6097**	**0.6043**
		(+4.42%)	(+2.97%)	(+5.22%)	(+3.44%)	(+4.10%)
NB	SDD-SMOTE	0.6011	0.3880	**0.8141**	0.5881	0.5876
	SDD-SMOTE$_{incrCC}$	**0.6080**	**0.4095**	0.8065	**0.5951**	**0.5890**
		(+1.15%)	(+5.54%)	(-0.93%)	(+1.19%)	(+0.24%)

classifiers. When no oversampling was applied, the sensitivity scores of classifiers were extremely low, ranging from 0.0987 to 0.2790, while the specificity scores were high, ranging from 0.8822 to 0.9844, as expected in imbalanced datasets.

Second, we present the performance of the incremental data generation method with the SDD-SMOTE method. Overall, the proposed method improved classifiers in all evaluation metrics other than specificity when DT and NB were used as the classification algorithm. While the improvement by the proposed methods was not significant, such as 0.61% - 5.54% in sensitivity, its performance with all classifiers was robust. Table 4 shows the improvement in each performance metric.

Next, we present the performance improvement in TL models trained using the augmented datasets generated by four SMOTE variants within the proposed synthetic data generation method. TL models were trained with the following model hyper-parameters: Adam optimizer (learning rate = 0.001), the number of epochs for both source and target models were 100 - 1,000, and 5 CV for source model and 3 CV for target mode. The training-testing data ratio was 0.2.

The improvement of TL models by the incremental data generation method on the SMOTE methods over the original SMOTE methods ranged from +4.01% to +7.79% in sensitivity. G-SMOTE$_{incrCC}$ maintained the highest improvement in all evaluation metrics resulting in 5.79% in weighted accuracy, 7.79% in sensitivity, 4.37% in specificity, 3.81% in precision, 5.01% in F_1 score, and 5.79% in AUCROC. On the other hand, the sensitivity went down by -3.21% with the SMOTE$_{incrCC}$. Classifiers trained on the data by SMOTE variants and incremental SMOTE methods outperformed than those on the data by TVAE synthesizer [11], showing sensitivity values ranging from 6.08% to 17.5% higher. The detail of the performance analysis is shown in Table 5.

Table 5. TVAE vs. SMOTE vs. SMOTE$_{incrCC}$ in TL-based classifier performance

Method	Accuracy	Sensitivity	Specificity	Precision	F_1 score	AUC ROC
SMOTE	0.6697	**0.5449**	0.7846	0.6503	0.6489	0.6647
SMOTE$_{incrCC}$	**0.6762**	0.5371	**0.8153**	**0.6599**	**0.6572**	**0.6762**
	(+0.97%)	(-3.21%)	(+3.91%)	(+1.48%)	(+1.26%)	(+1.72%)
G-SMOTE	0.6592	0.5461	0.7723	0.6308	0.6282	0.6592
G-SMOTE$_{incrCC}$	**0.6973**	**0.5886**	**0.8060**	**0.6549**	**0.6597**	**0.6973**
	(+5.79%)	(+7.79%)	(+4.37%)	(+3.81%)	(+5.01%)	(+5.79%)
Gamma-SMOTE	0.6737	0.5505	0.7969	0.6508	0.6501	0.6770
Gamma-SMOTE$_{incrCC}$	**0.6975**	**0.5726**	**0.8225**	**0.6708**	**0.6704**	**0.6975**
	(+3.54%)	(+4.01%)	(+3.21%)	(+3.08%)	(+3.12%)	(+3.04%)
SDD-SMOTE	0.6853	0.5720	0.7986	0.6556	0.6543	0.6853
SDD-SMOTE$_{incrCC}$	**0.6982**	**0.5953**	**0.7998**	**0.6591**	**0.6598**	**0.6975**
	(+1.88%)	(+4.07%)	(+0.15%)	(+0.54%)	(+0.84%)	(+1.78%)
TVAE synthesizer [11]	0.6654	0.5063	0.8233	0.6474	0.6461	0.6653

5 Conclusions

In the medical environment, when rare occurrences of diseases or exacerbations occur, class imbalance is a prevalent concern that might impact the accuracy of prediction models. This study presented a systematic evaluation of the various SMOTE variants and proposed an incremental SMOTE based method that augments the training dataset with synthetically generated minority class samples, hence improving the minority dataset gradually. By employing control coefficient, the limitation of the uniform random function for creating new samples is resolved, greatly improving the process. Our proposed method was experimentally compared to the various SMOTE variants using data from actual asthma patients using the four classification algorithms: Decision Tree, K-Nearest Neighbors, Logistic Regression and Naive Bayes. The findings demonstrated the applicability of the incremental SMOTE based approach in data-starved medical applications, as the accuracy of asthma health risk prediction increased by 4.01% - 7.79% in sensitivity of TL models utilizing three SMOTE variants. One difficulty we foresee in applying this approach in practice is that, with extremely small number of minority samples, it may not generate enough non-duplicate data points to balance the dataset. Open challenges include the development of flexible and scalable SMOTE variants that are robust to different imbalanced ratios and data sizes and various architectures of TL with other classifiers.

References

1. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)

2. Chawla, N.V., Lazarevic, A., Hall, L.O., Bowyer, K.W.: Smoteboost: Improving prediction of the minority class in boosting. In: Knowledge Discovery in Databases: PKDD 2003: 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, Cavtat-Dubrovnik, Croatia, September 22-26, 2003. Proceedings 7. pp. 107–119. Springer (2003)
3. Fang, J., Tang, C., Cui, Q., Zhu, F., Li, L., Zhou, J., Zhu, W.: Semi-supervised learning with data augmentation for tabular data. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management. pp. 3928–3932 (2022)
4. Gong, X., Tang, B., Zhu, R., Liao, W., Song, L.: Data augmentation for electricity theft detection using conditional variational auto-encoder. Energies **13**(17), 4291 (2020)
5. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
6. Gretel: Gretel (accessed on October 4, 2023), https://gretel.ai/.
7. Hoens, T.R., Chawla, N.V.: Imbalanced datasets: from sampling to classifiers. Imbalanced Learning: Foundations, Algorithms, and Applications. Wiley (2013)
8. Kamalov, F., Denisov, D.: Gamma distribution-based sampling for imbalanced data. Knowledge-Based Systems **207**, 106368 (2020)
9. Khozeimeh, F., Sharifrazi, D., Izadi, N.H., Joloudari, J.H., Shoeibi, A., Alizadehsani, R., Gorriz, J.M., Hussain, S., Sani, Z.A., Moosaei, H., et al.: Combining a convolutional neural network with autoencoders to predict the survival chance of covid-19 patients. Scientific Reports **11**(1), 15343 (2021)
10. Lee, H., Kim, J., Kim, S.: Gaussian-based smote algorithm for solving skewed class distributions. International Journal of Fuzzy Logic and Intelligent Systems **17**(4), 229–234 (2017)
11. MIT: The synthetic data vault (accessed on October 4, 2023), https://sdv.dev.
12. Moniz, N., Ribeiro, R., Cerqueira, V., Chawla, N.: Smoteboost for regression: Improving the prediction of extreme values. In: 2018 IEEE 5th international conference on data science and advanced analytics (DSAA). pp. 150–159. IEEE (2018)
13. Pratama, R.F.W., Purnami, S.W., Rahayu, S.P.: Boosting support vector machines for imbalanced microarray data. Procedia computer science **144**, 174–183 (2018)
14. Raghupathi, W., Raghupathi, V.: Big data analytics in healthcare: promise and potential. Health information science and systems **2**(1), 1–10 (2014)
15. Sağlam, F., Cengiz, M.A.: A novel smote-based resampling technique trough noise detection and the boosting procedure. Expert Systems with Applications **200**, 117023 (2022)
16. Sharma, A., Singh, P.K., Chandra, R.: Smotified-gan for class imbalanced pattern classification problems. Ieee Access **10**, 30655–30665 (2022)
17. Wan, Q., Deng, X., Li, M., Yang, H.: Sddsmote: Synthetic minority oversampling technique based on sample density distribution for enhanced classification on imbalanced microarray data. In: The 6th International Conf. on Compute and Data Analysis. pp. 35–42 (2022)
18. Wewer, C.R., Iosifidis, A.: Improving online non-destructive moisture content estimation using data augmentation by feature space interpolation with variational autoencoders. In: 2023 IEEE 21st International Conference on Industrial Informatics (INDIN). pp. 1–7. IEEE (2023)
19. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional gan. Advances in neural information processing systems **32** (2019)