Journal of the American Medical Informatics Association, 27(11), 2020, 1667–1674
doi: 10.1093/jamia/ocaa127
Advance Access Publication Date: 4 September 2020
Research and Applications

Research and Applications

Synthetic minority oversampling of vital statistics data with generative adversarial networks

Aki Koivu,[1] **Mikko Sairanen**,[2] **Antti Airola**,[1] **and Tapio Pahikkala**[1]

[1]Department of Future Technologies, University of Turku, Turku, Finland, and [2]PerkinElmer, Turku, Finland

Corresponding Author: Aki Koivu, MSc, Department of Future Technologies, University of Turku, Turun Yliopisto, 20500 Turku, Finland; aki.i.koivu@utu.fi

Received 23 February 2020; Revised 26 May 2020; Editorial Decision 27 May 2020; Accepted 3 June 2020

ABSTRACT

Objective: Minority oversampling is a standard approach used for adjusting the ratio between the classes on imbalanced data. However, established methods often provide modest improvements in classification performance when applied to data with extremely imbalanced class distribution and to mixed-type data. This is usual for vital statistics data, in which the outcome incidence dictates the amount of positive observations. In this article, we developed a novel neural network-based oversampling method called actGAN (activation-specific generative adversarial network) that can derive useful synthetic observations in terms of increasing prediction performance in this context.
Materials and Methods: From vital statistics data, the outcome of early stillbirth was chosen to be predicted based on demographics, pregnancy history, and infections. The data contained 363 560 live births and 139 early stillbirths, resulting in class imbalance of 99.96% and 0.04%. The hyperparameters of actGAN and a baseline method SMOTE-NC (Synthetic Minority Over-sampling Technique-Nominal Continuous) were tuned with Bayesian optimization, and both were compared against a cost-sensitive learning-only approach.
Results: While SMOTE-NC provided mixed results, actGAN was able to improve true positive rate at a clinically significant false positive rate and area under the curve from the receiver-operating characteristic curve consistently.
Discussion: Including an activation-specific output layer to a generator network of actGAN enables the addition of information about the underlying data structure, which overperforms the nominal mechanism of SMOTE-NC.
Conclusions: actGAN provides an improvement to the prediction performance for our learning task. Our developed method could be applied to other mixed-type data prediction tasks that are known to be afflicted by class imbalance and limited data availability.

Key words: artificial intelligence, machine learning, deep learning, vital statistics, stillbirth

INTRODUCTION

Background and significance

Real-life vital statistics data commonly suffer from class imbalance, in which 1 or more of the predicted classes are highly underrepresented. This occurs naturally, as the amount of positive observations is controlled by the prevalence of a disease. This has been shown to have a negative effect on the learning process of probabilistic models.[1] Enough data to model from is a recurrent problem, and several methods have been proposed over the years to address it, ranging from method-specific[2] to more universal.[3] Simple methods such as minority oversampling and majority undersampling are still commonplace, as they provide straightforward and universal solution. Another common practice is to calculate class weights from the learning data and apply them during model training. This is called cost-sensitive learning,[4] in which the model is made aware of the class unbalance in the training data by first assigning higher weight value for the minority and lower for the majority, then incorporating those weights to the training process. For neural networks, one could use for example a weighted loss function such as the weighted cross-entropy with backpropagation.[5]

Cost-sensitive learning and sampling methods both try to solve the class imbalance problem, but it is not yet clear which is better in general. Weiss et al[6] compared cost-sensitive learning with sampling methods and found out that there was no clear winner when the testing was done using multiple different datasets with differing percentage of minority observations. Oversampling the minority class with duplicate observations was deemed useful with some datasets and useless with others.[6]

Synthetic data generation refers to a case of oversampling where instead of duplicating existing observations, completely new ones are created.[7] This is achieved by successfully modelling the distributions of each variable of the training data and then sampling from the created joint multivariate probability distribution. In the clinical domain, established methods for data generation, such as Synthetic Minority Over-sampling Technique (SMOTE),[8] have been successfully implemented.[9] Modern methods such as the generative adversarial networks (GANs)[10] have also made significant improvements to medical image generation tasks.[11] GAN methods for tabular data have also been proposed,[12] in which a long short-term memory recurrent neural network was used as the generator.

The success of a generator method is usually tied to the amount of feasible data available.[10] Proper domain expertise can also be utilized in a limited data setting. Real-life data can have constrains that are obvious to domain experts, but they can be missed by the modeling method if the used data does not represent this sufficiently. The problem is amplified with data that has a substantial class imbalance. To be able to generate valuable synthetic mixed-type observations in terms of improving prediction performance from highly imbalanced vital statistics data is therefore not an elementary task. Discovering a robust and generalizable method for this problem would be significant because deriving more value from this type of data could have a positive impact on public health in general.

Stillbirth is a serious pregnancy-related condition defined as a deadborn outcome of a delivery after 20 weeks of gestation.[13] Stillbirth can be further categorized as early and late stillbirth according to the gestational age thresholds. Maternal characteristics contributing to elevated risk for stillbirth have been identified as being high body mass index, advanced maternal age, maternal smoking, parity, ethnicity, education, and various preexisting conditions or comorbidities.[14–17] Currently, several screening and monitoring mechanisms have been proposed[18]; however a common golden standard has not been agreed upon, so most nations report, rather than screen. One of these mechanisms is predicting risk from maternal characteristics.

Recently, we proposed a novel model that assesses the risk of early stillbirth from maternal characteristics.[19] Also, models predicting risk of stillbirth in different gestational age intervals have been developed.[20–23] Performance of the models utilizing maternal characteristics has been modest, achieving area under the curve (AUC) from a receiver-operating characteristic (ROC) between 0.6 and 0.7, while models that had biophysical measurements as added features such as fetal presentation achieved an improved AUC of 0.8.[20–23]

OBJECTIVE

The aim of this research is to develop a novel GAN-based data generation method suitable for generating synthetic cases from vital statistics data. Tabular mixed-type data with a significant class imbalance problem pose challenges for modelling and limit the amount of applicable methods. Our goal is to introduce a new method capable of generating more value in this context. Prediction performance is the key optimizable parameter we want to focus on. The research question was the following: can we derive more predictive power from the existing data with generative methods while not changing the model that is responsible for the outcome prediction? Our new method could be applied to similar prediction tasks that involve highly imbalanced data. For demonstrating performance over established methods such as SMOTE, the prediction of early stillbirth pregnancies was chosen for our use case, as it provided the desired restrictions data-wise.

MATERIALS AND METHODS

Study data

The dataset was provided by the New York City Department of Health and Mental Hygiene and contained reported pregnancies in the New York City area from 2014 to 2016. The source of the data was collected birth and death certificates, and the data were deidentified of any variables that could be linked to a specific person. Use of the dataset for this research was granted an institutional review board approval by the ethics committee of the Hospital District of Southwest Finland.

The selected maternal characteristics feature variables were based on literature[14–17] and are listed in Table 1.

The dataset contains 364 124 pregnancies in total. In order to predict early stillbirth from a representative population, sample selection was conducted using the following exclusion criteria:

- Age of the mother was 18 years of age or older.
- Cases of maternal morbidity were excluded.
- All the feature variables and predicted class were present, and no value imputation was required.
- Pregnancies that concluded in fetal death because of external causes were excluded. These were identified with the International Classification of Diseases–Tenth Revision code values containing U, V, W, X, or Y characters.[13]
- Reported gestational age in pregnancies that concluded in live birth was 21 weeks or older. Because the earliest known preterm baby is 21 weeks and 6 days,[24] it is probable that these pregnancies were erroneously inputted during data collection.
- Postnatal death cases were excluded.
- Multiple birth pregnancies were excluded.

This decreased the amount of pregnancies to 363 560 live births and 139 early stillbirths. The prevalence for early stillbirth in this dataset is therefore 0.04%, or 1 in 2500. Because of this, the data are highly imbalanced. Data were randomly split into 2 equal-sized sets, 1 for hyperparameter optimization and 1 for model evaluation. This procedure was class-stratified so that the class imbalance would be present in both datasets.

Feature variables were preprocessed based on their type. Nominal features were one-hot encoded in order to accommodate modelling with neural networks.[25] Continuous variables were standardized by zero-mean normalization and unit-variance normalization, and the parameters were calculated using the training dataset only and then applied to both datasets.

Generator methods

SMOTE-Nominal Continuous

SMOTE is a well-established oversampling method based on k-nearest neighbors.[8] Given minority class data, the method utilizes Eu-

Table 1. Feature variables

Demographics	
Age	discrete
Race (white, black, American Indian, Alaskan native, Asian, or Pacific Islander)	nominal
Marital status	binary
Education (8th grade or less to doctorate)	nominal
Number of previous terminations	nominal
Special supplemental nutrition program	binary
Smoking before pregnancy	nominal
BMI	continuous
Height	continuous
Parity	nominal
Pregnancy history	
Prepregnancy diabetes	binary
Gestational diabetes	binary
Prepregnancy hypertension	binary
Gestational hypertension	binary
Hypertension eclampsia	binary
Previous preterm births	binary
Infertility treatment	binary
Infertility drugs	binary
Assisted reproductive technology	binary
Previous cesarean sections	binary
Infections	
Gonorrhea	binary
Syphilis	binary
Chlamydia	binary
Hepatitis B	binary
Hepatitis C	binary

BMI: body mass index.

clidean distances between data points to generate new ones. It features 2 hyperparameters: the number of closest neighbors to be included in calculation is controlled with k and N, which determines the number of generated observations. The impact of SMOTE and its variants on prediction performance has been studied, Blagus et al[26] demonstrated that SMOTE had no significant effect on performance when used with microarray data, and Van Hulse et al[27] showed that random undersampling of the majority class overperformed SMOTE. However, a variant of SMOTE called SMOTE-Nominal Continuous (SMOTE-NC) addresses an issue that is not addressed by other conventional oversampling methods, generating mixed-type data.[8]

Our study dataset contains nominal feature variables mixed with continuous features. SMOTE-NC takes this into consideration by having a separate logic for nominal features. First, the median of standard deviations of all continuous features of the minority class are calculated. When new data points are about to be created, if the nominal features differ when compared with the nearest neighbors, this median is added to the Euclidean distances. The mechanism penalizes differences in nominal features more effectively, which should in theory result in more accurate synthetic nominal features.

Activation-specific GAN

Generative adversarial networks use a competitive learning setting for 2 networks: a generator G and a discriminator D.[10] G is trained to map random noise derived from a distribution to data points in the training dataset, so its objective is to create synthetic observations that fool D. D is trained to classify the input as fake or real. This means either from the target distribution or not. The 2-objective training loss can be formalized as

$$\min_G \max_D L(D,G) = \mathbb{E}_{x\sim\mathbb{P}_r}[log(D(x))] + \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}[log(1-D(\tilde{x}))]$$

where P_r is the training data distribution and P_g is the model distribution, defined by $\tilde{x} = G(z), z \sim p(z)$, when z is the random noise input sampled from a distribution p.[10] Training this type of network has been shown to be fragile and unstable.[28] Nonconvergence of the 2 models, mode collapse, and diminishing gradient of the generator plagued the original GAN. In 2017, Wasserstein GAN (WGAN) was proposed as a more stable method that also provided more meaningful learning curves in terms of model fitting performance.[29] In WGAN, the discriminator network is replaced by a critic network that scores observations as being real or fake by learning a K-Lipschitz function to compute Wasserstein distance. In training, decreasing this distance is the objective loss function, so as it decreases, the resemblance of the generator output and the training data increases. The loss function used with the critic is now defined as

$$W(P_r,P_g) = \frac{1}{K} \sup_{\|f\|_L \le K} \mathbb{E}_{x\sim\mathbb{P}_r}[f(x)] - \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}[f(\tilde{x})]$$

where *sup* is supremum and K is the Lipschitz constant for function f, and is made to satisfy $\|f\|_L \le \mathrm{K}$, or K-Lipschitz continuity. The method was a clear improvement over GAN, but in order to preserve the K-Lipschitz continuity, the iterated weights of the model were limited to a small value range so that the K-Lipschitz function's lower and upper bounds could be obtained in a feasible manner. This was called weight clipping, which the model is highly sensitive to.

Later in 2017, Gulrajani et al[30] improved on this method and proposed WGAN with gradient penalty (WGAN-GP). In WGAN-GP, gradient penalty is used instead of weight clipping to impose the K-Lipschitz continuity. A differentiable function f is 1-lipschitz if and only if it has gradients with a norm of at most 1 everywhere.[30] The loss function was reworked to penalize if the gradient norm moved away from 1, so it was defined as

$$W(P_r,P_g) = \frac{1}{K} \sup_{\|f\|_L \le K} \mathbb{E}_{x\sim\mathbb{P}_r}[f(x)] - \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}[f(\tilde{x})] + \lambda\mathbb{E}_{\widehat{x}\sim\mathbb{P}_{\mathrm{x}}}\left[(\|\Delta_{\widehat{x}}D(\widehat{x})\|_2 - 1)^2\right]$$

where $\widehat{\mathrm{x}}$ is sampled from $\tilde{\mathrm{x}}$ and x with t uniformly sampled between 0 and 1 so that

$$\widehat{x} = t\tilde{x} + (1-t)x \text{ when } 0 \le t \le 1,$$

and λ is the penalty coefficient. This change in WGAN-GP stabilized the model training even further because there was no more need to parameterize weight clipping[30]; however, it introduced a new level of complexity to the model. Another popular variation of GAN called deep convolutional GAN[31] has also been proposed. The benefit of using convolution is data aggregation to a smaller space, which is something we do not want to do with mixed-type data, so WGAN-GP was chosen to be the starting point of our research.

In order to create synthetic positives that follow the variable-specific constrains of tabular mixed-type data, WGAN-GP needed to be altered to accommodate this. On the one hand, image data, the most common application of GANs, contains pixels that are represented as continuous values with no pixel-specific constraints.[11] On

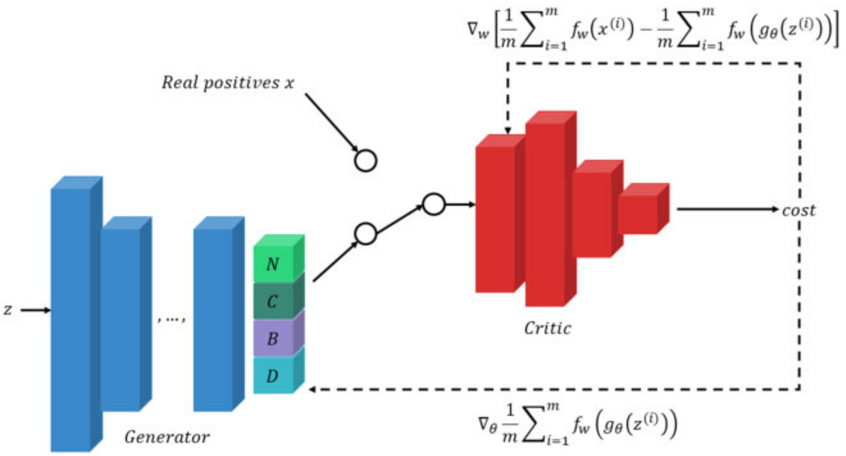
$\nabla_w\left[\frac{1}{m}\sum_{i=1}^{m} f_w(x^{(i)}) - \frac{1}{m}\sum_{i=1}^{m} f_w\left(g_\theta(z^{(i)})\right)\right]$
Real positives x
z
, ...,
N
C
B
D
Generator
Critic
cost
$\nabla_\theta \frac{1}{m}\sum_{i=1}^{m} f_w\left(g_\theta(z^{(i)})\right)$

Figure 1. Schematic view of the training cycle of the actGAN (activation-specific generative adversarial network) generator and critic models. Random noise z is given to the generator model as input, and the model produces vector representation of an observation based on the architecture defined in the output layer. The output layer of the generator is comprised of nominal (N), continuous (C), binary (B), and discrete (D) nodes. The critic is given real positive observations and fake generated observations as input, and it outputs a score of realness by approximating the Wasserstein distance. These 2 models are trained in adversarial manner. After training, the critic is discarded, and the generator is used to create fake positive observations.

the other hand, mixed-type data can contain nominal and ordinal variables that are affected by rules, such as non-negative integer values only. One-hot encoding also produces multidimensional representations of these integer values that have conditional properties like a stochastic vector; the representation should add up to 1. All these rules are learned by the generator model of a GAN when given enough data, but in a minority oversampling situation in which training data can be limited, this can become unfeasible. To solve this problem, we propose an output layer activation-specific GAN (actGAN).

Based on WGAN-GP architecture with Wasserstein loss and gradient penalty, actGAN was developed to specifically handle mixed-type data in a minority learning setting. Output layer neurons' activation functions of the generator model were selected based on the variable type they were generating. Continuous and discrete variables were generated using ReLU function[32]

$$ReLU(x) = \begin{cases} x \ \mathit{if} \ x > 0, \\ 0 \ \mathit{otherwise}, \end{cases}$$

which demands that only non-negative values are created for them. Binary variables were created with a logistic function

$$f(x) = \frac{1}{1 + e^{-x}}$$

which is suitable for creating a binary result. One-hot encoding representations of nominal features were generated with softmax[5]

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

for $i = 1, \ldots, K$ and $z = (z_1, \ldots, z_K) \in \mathbb{R}^K$. The internal normalization procedure of softmax ensures that the sum of the elements of the output vector $\sigma(z)$ is 1. This customized output layer enabled us to provide prior knowledge of the generated variables, which enabled actGAN to learn more sufficiently from a smaller number of observations. The hidden layer activations were chosen to be scaled exponential linear unit (SELU) functions accompanied with LeCun normal weight initialization.[33] The design of the generator is depicted in Figure 1.

The critic network architecture was kept simple on purpose, because discrimination as a task is simpler compared with generation. Two hidden layers of width 128 and 64 were used with Leaky ReLU activations[34] that were parameterized with 0.2 slope coefficient, and He normal initialization of weights.[35] The output would be linear activation of 1 node that was required by WGAN functionality. After training, the critic network would be discarded, and the finalized generator network would be used with random noise input to generate synthetic observations. The designs of the generator and critic models are depicted in Figure 1.

Classifiers

In order to compare data generation methods in terms of added prediction performance, 2 classifier algorithms were chosen that would utilize the generated data for prediction. They were the frequently used logistic regression (LR) and the neural network with a SELU network,[33] which represents state of the art of feed-forward networks. Hyperparameters for them were also tuned during hyperparameter optimization, with the aim of maximizing prediction performance.

For fitting the LR, limited-memory version of Broyden-Fletcher-Goldfarb-Shanno algorithm[36] was used with the L2 norm penalty. For SELU network, mini-batch adaptive moment estimation (Adam) gradient descent[37] with 0.001 learning rate, 0.9 β_1 decay rate, and 0.999 β_2 decay rate were used for updating weights with backpropagation.[38] Weight initialization was anchored with a random seed, and was used in all the experiments. The SELU network requires the use of LeCun normal weight initialization and alpha node dropout[33]; the amount of dropout after every hidden layer was set to 15%, which was deemed appropriate for preventing overfitting.

Cost-sensitive learning with class weights was deemed necessary for any feasible classifier model fit because of the magnitude of class imbalance in the data. Class weights w were calculated from the training dataset using the following equation:

$$w = s/(c * f(y)),$$

where s is the number of observations, c is the number of different classes, and $f(y)$ is the frequency of classes in data labels y.

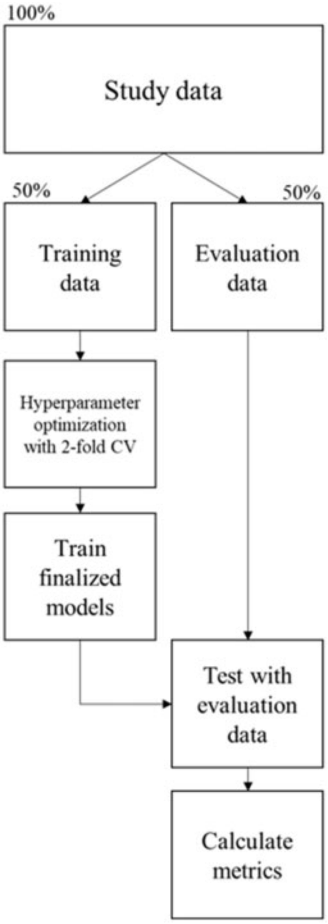

Figure 2. Data flow diagram of the experiments. Study data would be randomly divided in half, in a class-stratified way. Training data would be used to tune hyperparameters and train the final models, while evaluation data would be used to test their performance. CV: cross-validation.

Experimental overview

The study experiments were divided into 2 phases: hyperparameter optimization and model evaluation. Each of them had their own subset dataset. The model's hyperparameters were optimized in a 2-fold cross-validation procedure with the training dataset, and then finalized models were trained with the same data. After this, they were evaluated with the separate evaluation dataset. This guaranteed that there was no chance to overfit the model to the evaluation data, as both model fitting and hyperparameter selection were done on the separate training dataset. The data flow of our experiments is depicted in Figure 2. The performance of a generator or a classifier model is highly affected by used hyperparameters, so the hyperparameter spaces of LR, SELU network, SMOTE-NC, and actGAN were investigated iteratively. Hyperparameters for classifiers were first investigated independently, and the best found sets were used while optimizing hyperparameters for generator models. Bayesian optimization[39] was used to calculate optimal hyperparameters, and the process can be represented as a formula:

$$y^* = \arg \max_{x \in \mathbb{X}} f(x),$$

where $f(x)$ is the AUC score from an ROC curve to be maximized, and y^* are the optimal hyperparameters calculated from training dataset. Twofold class-stratified cross-validation was used for the optimization process, and the reported AUC was averaged over the folds. For measuring the effect of hyperparameters in terms of AUC, Pearson's correlation coefficient was used with the effect interpretation table[40] to determine linear correlation and its significance. The number of optimization iterations after 10 initial random iterations was chosen to be 500 because it was estimated that it would be enough for Bayesian optimization.

Tunable hyperparameters of LR were regularization parameter C and the number of iterations. For the SELU network, tunable hyperparameters were the number of hidden layers, nodes in hidden layers, number of epochs, and batch size. Number of synthetic observations and k were iterated for SMOTE-NC. For actGAN, the length of latent dimension, number of epochs, batch size, number of synthetic observations, hidden layers, and nodes in the generator network were iterated over. actGAN' s penalty coefficient λ was set to 10 and Adam optimizer parameters of learning rate and decay rates β_1 and β_2 for both networks were set to 0.0001, 0, and 0.9, respectively, according to the original publication of WGAN-GP.[30] Our method is WGAN-based, which is stated to be not sensitive to generator model architecture or chosen hyperparameters.[29] We wanted to verify this by including the network architecture parameters in our optimization experiment. Hyperparameters and their selected ranges for the optimization process are listed in Supplementary Table 1.

For designing an experiment that would test the data generator model's generalizability and prediction performance, the final models were selected based on 2-fold cross-validation on the training dataset. These models were then evaluated on the independent model evaluation dataset. The used metric was the AUC from an ROC curve, and true positive rate (TPR) at a clinically significant false positive rate (FPR), which would be TPR at 1%, 3%, and 5% FPR for early stillbirth, based on the real-world incidence. After hyperparameter optimization, LR and SELU network classifiers would serve as benchmarks, and all the possible combinations of classifier and generator models were experimented with in model evaluation. The full list of used software libraries and hardware are described in the Supplementary Appendix.

RESULTS

Model evaluation

After Bayesian hyperparameter optimization, the best sets of hyperparameters for the model and training datasets were used to train the final models. The optimization results of the classifiers and generators are listed in the Supplementary Appendix. The material also contains the learning curves of both final actGAN models, the first optimized with LR and the second with the SELU network. Six different prediction configurations were used to predict the evaluation dataset, and the results are presented in Table 2.

When comparing benchmark models, LR was able to achieve better AUC and improve TPR at 1% FPR by 2% over the SELU network. The usage of SMOTE-NC was minimized during SMOTE-NC and LR optimization, so its performance is identical to LR. SMOTE-NC and SELU was able to improve TPR at 3% and 5% FPR over the SELU network; however, AUC and TPR at 1% FPR decreased. actGAN and LR retained TPR at 1% and 3% FPR while improving TPR at 5% FPR but producing the worst AUC. This was caused by a drop in performance in higher FPRs, shown in Figure 3. actGAN and SELU improved every metric over the SELU network and achieved the best performance of the experiment. Variable

Table 2. Model evaluation results

Name	TPR at 1% FPR	TPR at 3% FPR	TPR at 5% FPR	AUC (95% CI)
LR	9%	16%	20%	0.688 (0.620-0.756)
SELU network	7%	16%	20%	0.659 (0.590-0.728)
SMOTE-NC and LR	9%	16%	20%	0.688 (0.620-0.756)
SMOTE-NC and SELU	6%	17%	23%	0.663 (0.594-0.733)
actGAN and LR	9%	16%	24%	0.637 (0.562-0.712)
actGAN and SELU	9%	**23%**	**27%**	**0.704 (0.635-0.772)**

actGAN: activation-specific generative adversarial network; AUC: area under the curve; CI: confidence interval; FPR: false positive rate; LR: logistic regression; SELU: scaled exponential linear unit; SMOTE-NC: Synthetic Minority Over-sampling Technique-Nominal Continuous; TPR: true positive rate.

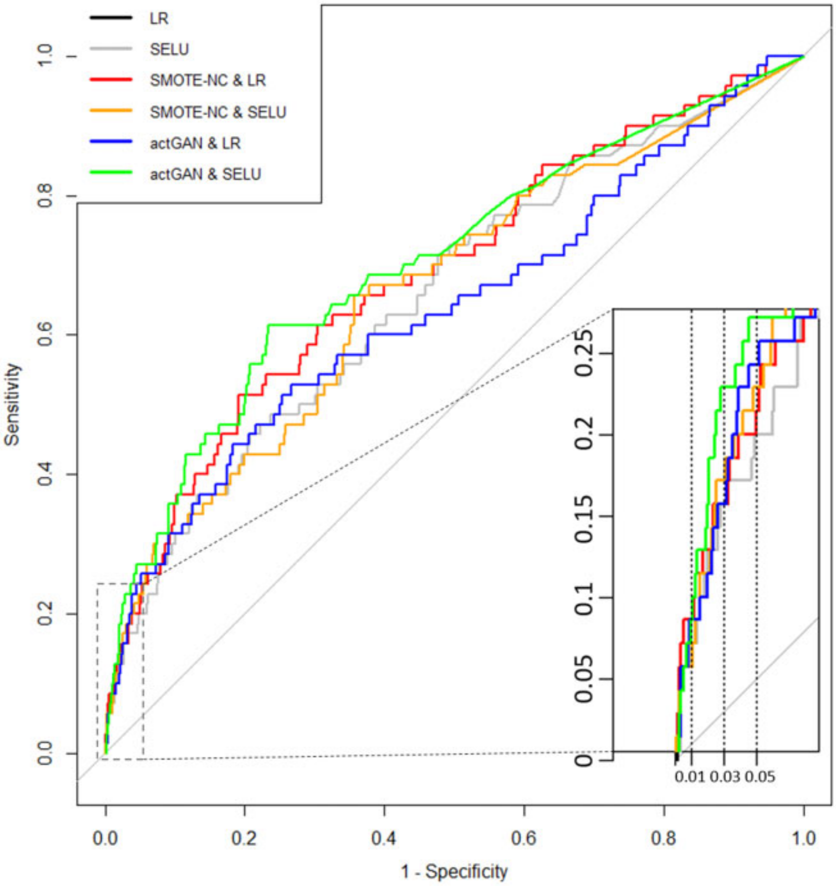

Figure 3. Receiver-operating characteristic curves of predicting the evaluation dataset with all the experimented configurations of models. The clinically significant false positive range of 1%, 3%, and 5% is presented in the magnified section, which illustrates the noteworthy performance of activation-specific generative adversarial network (actGAN) and scaled exponential linear units (SELU) in 3% and 5% false positive rate. The best area under the curve (AUC) of 0.704 was achieved by actGAN and SELU network. The usage of Synthetic Minority Over-sampling Technique-Nominal Continuous (SMOTE-NC) was minimized in SMOTE-NC and logistic regression (LR), resulting in overlapping receiver-operating characteristic curves with LR and same AUC of 0.688. The third-best AUC of 0.663 was achieved by SMOTE-NC and the SELU network. The worst AUCs of 0.659 and 0.637 were obtained by the SELU network and actGAN, and LR.

actGAN: activation-specific generative adversarial network; AUC:; CI: confidence interval; FPR: false positive rate; LR: logistic regression; SELU: scaled exponential linear unit; SMOTE-NC:; TPR: true positive rate.

importance results of the tested classifiers are listed in Supplementary Table 2.

DISCUSSION

Applying minority oversampling to vital statistics data can be challenging due to the limited amount of available positive case data, and mixed-type predictor variables. It limits the amount of feasible oversampling methods, and even when applied, the increase in predictive performance might not be substantial enough to justify the usage.

The best AUC of the 6 tested methods was achieved by actGAN and SELU, but confidence intervals in Table 2 reveal a notable overlap between all methods, so the statistical significance of the AUC cannot be stated. However, while AUC measures performance across the whole curve, the clinically significant performance in screening for rare disorders is achieved by either increasing TPR within the feasible FPR range of 1% to 5% or decreasing FPR while maintaining the same TPR.

SMOTE-NC provided mixed results in model evaluation, implying that the method could improve performance in specific FPRs while decreasing it at others. However, actGAN and SELU was able to derive observations from this data that provided consistent improvement to TPR at FPRs of 3% and 5%. The method was able to produce similar TPR with 3% FPR when compared with competing methods that used 5% FPR. This improvement of 2% to the FPR has a substantial clinical impact in a screening environment. For example, in our study data of 363 560 live births, the improvement from 5% FPR to 3% FPR would mean 7271 less false positive cases over 3 years of screening.

The total amount of hidden nodes in actGAN was optimized to 700 with LR and 944 with the SELU network, which would indicate that a more complex classifier paired with a more complex generator can create additional value from the same data in terms of prediction performance. An activation-specific output layer of the generator network enabled us to provide additional information about the underlying data structure of the task, which overperformed the nominal mechanism of SMOTE-NC.

While actGAN performed better when compared with base classifiers and SMOTE-NC, the generalizability of our proposed method should be investigated further. The method is not restricted to stillbirth prediction or even the clinical domain, and any prediction task that utilizes mixed-type data could in theory benefit from actGAN. External validation of stillbirth prediction and experimentation of several datasets from various domains would be the topic for future work.

Predicting an outcome of low incidence is bound to produce high FPRs.[41] In clinical risk modeling of a particular outcome, instead of maximizing AUC, the TPR at a clinically significant FPR indicates the performance a model would have in routine use. Developing actGAN further to take this performance metric into consideration during training would be the secondary topic for future work.

This work provides evidence that proposed synthetic data generation tools can significantly improve maternal characteristics–based risk prediction in rare conditions. Baseline performance received with the multivariate LR model in our experiment is in line with current published tools,[23] in which predicting stillbirth from maternal characteristics resulted in an AUC of 0.658 and 21.1% TPR at 5% FPR. However, synthetic data generation improved on this result in our experiments.

We believe that synthetic data generation of rare conditions combined with screening variables that are targeted toward detecting those conditions would be essential for improving clinical risk prediction in the future. Our study data did not contain first-trimester biophysical and biochemical measurements for stillbirth screening, prediction variables that have been suggested to improve detection.[42] Properly generating synthetic data of these continuous variables could be the key of improving stillbirth screening in general. At the population level, the improvement in prediction could change management of pregnancy in a number of women, resulting in fewer stillbirth outcomes.

CONCLUSION

Our demonstrated actGAN improved the clinically significant performance of the early stillbirth risk modelling. Activation-specific architecture could be designed to fit other clinical risk modelling tasks predicting outcomes from mixed-type data. This would improve development and validation of such clinical prediction tools. The benefits would be evident in models that rely on class-imbalanced population-level data, in which synthetic data generation of rare conditions would be valuable.

AUTHOR CONTRIBUTIONS

AK and MS conceived of the presented idea. AK developed the theory and performed the computations. AK and MS discussed the results and contributed to the first version of the manuscript. For the first revision of the manuscript, AA and TP reviewed the technical aspects and contributed to the grammar corrections, while MS contributed to the clinical aspects of the revision. AK reviewed the overall structure, wrote the necessary changes, and facilitated the revision process. For the second revision of the manuscript, all authors discussed the corrections and AK implemented the necessary changes.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Japkowicz N, Stephen S. The class imbalance problem: a systematic study. *Intell Data Anal* 2002; 6 (5): 429–49.
2. Zhou Z-H, Liu X-Y. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Trans Knowl Data Eng* 2006; 18: 63–77.
3. Ling CX, Li C. Data mining for direct marketing: problems and solutions. In: proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining; 1998: 73–9.
4. Ling C, Sheng V. Cost-sensitive learning and the class imbalance problem. In: Sammut C, ed. *Encyclopedia of Machine Learning*; New York, NY: Springer; 2010: 231–5.
5. Bishop CM. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Germany: Springer-Verlag; 2006.
6. Weiss G. McCarthy K. Zabar B. Cost-sensitive learning vs. sampling: which is best for handling unbalanced classes with unequal error costs? In: proceedings of the 2007 International Conference on Data Mining; 2007: 35–41.
7. Hoag JE. *Synthetic Data Generation: Theory, Techniques and Applications* [PhD thesis]. Fayetteville, AL, University of Arkansas; 2008.
8. Chawla N, Bowyer K, Hall L, *et al.* SMOTE: synthetic minority oversampling technique. *J Artif Intell Res* 2002; 16: 321–57.
9. Poolsawad N, Kambhampati C, Cleland JGF. Balancing class for performance of classification with a clinical dataset. In: proceedings of the World Congress on Engineering; 2014.
10. Goodfellow I, Pouget-Abadie J, Mirza M, *et al.* Generative adversarial nets. In: proceedings of the 27th International Conference on Machine Learning Processing Systems; 2014: 2672–80.
11. Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: a review. *Med Image Anal* 2019; 58: 101552.
12. Xu L, Veeramachaneni K. Synthesizing tabular data using generative adversarial networks. *arXiv*: 1811.11264; 2018.
13. World Health Organization. International Classification of Diseases 10th Revision (ICD-10). https://icd.who.int/browse10/2016/en Accessed December 12, 2019.

14. Flenady V, Koopmans L, Middleton P, *et al*. Major risk factors for stillbirth in high-income countries: a systematic review and meta-analysis. *Lancet* 2011; 377 (9774): 1331–40.
15. Gardosi J, Madurasinghe V, Williams M, *et al*. Maternal and fetal risk factors for stillbirth: population based study. *BMJ* 2013; 346: f108.
16. Little RE, Weinberg CA. Risk factors for antepartum and intrapartum stillbirth. *Am J Epidemiol* 1993; 137 (11): 1177–89.
17. McClure EM, Saleem S, Pasha O, *et al*. Stillbirth in developing countries: a review of causes, risk factors and prevention strategies. *J Matern Fetal Neonatal Med* 2009; 22 (3): 183–90.
18. Haws RA, Yakoob MY, Soomro T, *et al*. Reducing stillbirths: screening and monitoring during pregnancy and labour. *BMC Pregnancy Childbirth* 2009; 9 (Suppl 1): S5.
19. Koivu A, Sairanen M. Predicting risk of stillbirth and preterm pregnancies with machine learning. *Health Inf Sci Syst* 2020; 8 (1): 14.
20. Yerlikaya G, Akolekar R, McPherson K, *et al*. Prediction of stillbirth from maternal demographic and pregnancy characteristics. *Ultrasound Obstet Gynecol* 2016; 48 (5): 607–12.
21. Kayode GA, Grobbee DE, Amoakoh-Coleman M, *et al*. Predicting stillbirth in a low resource setting. *BMC Pregnancy Childbirth* 2016; 16 (1): 274.
22. Trudell AS, Tuuli MG, Colditz GA, *et al*. A stillbirth calculator: development and internal validation of a clinical prediction model to quantify stillbirth risk. *PLoS One* 2017; 12 (3): e0173461.
23. Akolekar R, Bower S, Flack N, *et al*. Prediction of miscarriage and stillbirth at 11-13 weeks and the contribution of chorionic villus sampling. *Prenat Diagn* 2011; 31 (1): 38–45.
24. Most-premature baby allowed home. BBC News. https://news.bbc.co.uk./2/hi/americas/6384621.stm Accessed December 12, 2019.
25. Harris DM, Harris SL. *Digital Design and Computer Architecture*. 2nd ed. Amsterdam, the Netherlands: Elsevier; 2013.
26. Blagus R, Lusa L. Evaluation of SMOTE for high-dimensional class-imbalanced microarray data. In: 2012 11th International Conference on Machine Learning and Applications; 2012: 89–94.
27. Van Hulse J, Khoshgoftaar T, Napolitano A. Experimental perspectives on learning from imbalanced data. In: proceedings of the 24th International Conference on Machine Learning; 2007: 937–42.
28. Arjovsky M, Bottou L. Towards principled methods for training generative adversarial networks. *arXiv*: 1701.04862; 2017.
29. Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. *arXiv*: 1701.07875; 2017.
30. Gulrajani I, Ahmed F, Arjovsky M, *et al*. Improved training of Wasserstein GANs. *arXiv*: 1704.00028; 2017.
31. Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv*: 1511.06434; 2015.
32. Hahnloser RHR, Sarpeshkar R, Mahowald MA, *et al*. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature* 2000; 405 (6789): 947–51.
33. Klambauer G, Unterthiner T, Mayr A, *et al*. Self-normalizing neural networks. In: proceedings of the 31st International Conference on Machine Learning Processing Systems; 2017: 972–81.
34. Maas AL. Rectifier nonlinearities improve neural network acoustic models. In: proceedings of the 30th International Conference on Machine Learning; 2013.
35. He K, Zhang X, Ren S, *et al*. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: proceedings of the 2015 IEEE International Conference on Computer Vision; 2015: 1026–34.
36. Fletcher R. *Practical Methods of Optimization*. Hoboken, NJ: Wiley-Interscience; 1987.
37. Kingma DP, Adam JB. A method for stochastic optimization. *arXiv*: 1412.6980; 2014.
38. Linnainmaa S. *The Representation of the Cumulative Rounding Error of an Algorithm as a Taylor Expansion of the Local Rounding Errors* [master's thesis]. Helsinki, Finland, Department of Computer Science, University of Helsinki.
39. Snoek J Larochelle HAdams RP. Practical Bayesian optimization of machine learning algorithms. In: proceedings of the 25th International Conference on Machine Learning Processing Systems; 2012: 2951–9.
40. Mukaka MM. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Med J* 2012; 24: 69–71.
41. Baldessarini RJ, Finklestein S, Arana GW. The predictive power of diagnostic tests and the effect of prevalence of illness. *Arch Gen Psychiatry* 1983; 40 (5): 569–73.
42. Mastrodima S, Akolekar R, Yerlikaya G, *et al*. Prediction of stillbirth from biochemical and biophysical markers at 11–13 weeks. *Ultrasound Obstetr Gynecol* 2016; 48: 613–7.