Experimental Evaluation of the Effectiveness of ANN-based Numerical Data Augmentation Methods for Diagnostics Tasks

Ivan Izonin[a], Roman Tkachenko[a], Roman Pidkostelnyi[a], Olena Pavliuk[a], Viktor Khavalko[a] and Anatoliy Batyuk[a]

[a] *Lviv Polytechnic National University, S. Bandera str., 12, 79013, Lviv, Ukraine*

Abstract

Improving the accuracy of diagnostics tasks is essential in various medical fields. When there are small data for training, there are high risks of overfitting or underfitting the machine learning model. This makes it impossible to apply it in practice. To solve such a problem, we can use various data augmentation methods. This paper focuses on neural network methods of data augmentation. The authors have investigated a variational autoencoder and approach based on GAN to generate artificial numerical data and then use it by machine-learning-based classifiers. The authors examined the proposed method for diagnosing diabetes mellitus development task. Experiments confirmed that autoencoders generated a dataset similar to an initial one, with a similarity score being 0.93. The authors established a significant accuracy improvement of Random Forest, AdaBoost, and Logistic regression classifiers based on processing an extended dataset. The application of the new dataset obtained using GAN does not ensure satisfactory accuracy. Such an issue may be due to a lack of samples for the training of this neural networks class. Further research is likely to be carried out into ensembles based on a single machine learning method, which will process decorrelated samples acquired by methods investigated in this paper.

Keywords [1]

Tabular data, classification, overfitting risk, underfitting risk, data augmentation, ANN, GAN, autoencoder, small data approach

1. Introduction

The development of modern medicine has been marked by digitizing a wide variety of information and the automation of many processes [1]. This makes it possible to collect a large amount of data for analysis. It also opens up new opportunities for applying data mining techniques to intellectualizing specific diagnostics or treatment processes.

However, the scarce data may impede the implementation of machine learning. Alternatively, abnormal data may lead to increased accuracy, which is a critical point in this area.

One possible solution to this problem lies in adopting data augmentation methods. This approach can allow synthesizing of enough data to train the selected artificial intelligence tool.

Nowadays, there are quite a few simple methods for manipulating an available sample of data to increase its size. The data are enlarged both by rows and by columns [1]. However, these methods do not always introduce helpful information into the expanded dataset and, consequently, only increase the learning time of the selected model. The accuracy of the chosen classifier or regressors is not affected here.

Many neural network methods have been developed to increase a dataset today. A wide variety of artificial neural network topologies are employed here. The augmentation is performed using a variety of information - from time series to images. Generative adversarial network [2] is among the most used

IDDM-2021: 4th International Conference on Informatics & Data-Driven Medicine, November 19–21, 2021 Valencia, Spain

EMAIL: ivanizonin@gmail.com (A. 1); roman.tkachenko@gmail.com (A. 2); roman.pidkostelnii@gmail.com (A. 3) olena.m.pavliuk@lpnu.ua (A. 4); khavalkov@gmail.com (A. 5); abatyuk@gmail.com (A. 6)
ORCID: 0000-0002-9761-0096 (A. 1); 0000-0002-9802-6799 (A. 2); 0000-0003-4561-3874 (A. 4); 0000-0002-9585-3078 (A. 5)

methods for artificial augmentation of datasets, particularly in the field of image processing. This type of neural network is most commonly used to synthesize new images for further use by deep learning neural networks. Another type of neural network is autoencoders, which is often and successfully applied in time series analysis. However, developing and researching a methodology for effective artificial augmentation of numerical datasets remains to be solved. On the one hand, neural network methods are more sophisticated and should reveal patterns in the dataset that are difficult to detect with simple methods [3]. Such information can serve as a basis for the synthesis of new patterns in the dataset. Alternatively, a neural network toolkit must obtain sufficient data for training and validating the model. Moreover, generalization properties should be especially emphasized. Only by meeting all these requirements will the selected tool operate adequately and synthesize the required amount of synthetic data of the required quality. Thus this paper aims to investigate neural network methods for enlarging tabular datasets to improve the accuracy of classification based on them.

2. Materials and methods

This section includes a description of two neural-network-based approaches for numerical data augmentation used in this paper. The main objective is to improve the classification accuracy in Clinical Medicine based on expanded datasets.

2.1.1. ANN-based numerical data augmentation methods

The first approach selected is a new method for generating an artificial dataset based on a Generative adversarial network (GAN) [4]. To this end, the author of the technique modified neural networks to deal specifically with numerical datasets. The modification was as follows. The authors proposed to use Conditional GANs as a generator of numerical data. This approach is explained by:

- efficient performance in the event of an unbalanced dataset;
- independence of the type of variables: discrete and continuous, with the possibility of modeling them both at the same time;
- a flexible approach to modeling the distribution of probabilities within the dataset;
- the possibility of synthesizing high-quality synthetic samples that are very similar to the observations from the initial dataset.

A peculiarity of this method is that the authors use a special normalization method and a set of state-of-the-art model learning methods, and a post-annotated network. In other respects, the method works like a conventional GAN.

Another interesting method is data augmentation based on a variational autoencoder [5]. It is referred to as generative models. Learning methods of this family consist of mapping objects into a given latent space and reproducing them back. The task related to the autoencoder is to find the functions that will allow mapping the latent variable area to another one, an understandable and simple space. A customarily distributed space is a case in point.

While designing methods based on variational autoencoder, one should define the number of neurons in the first and the second latent layers and set the number of latent factors. It will contain all applicable information and serve as a decoder to recover all initial inputs. After all the necessary settings have been made, the learning procedure can be performed.

If the hidden dependencies between variables are linear, the variational autoencoder works as a PCA method [6]. In this case, to each element according to the method [5] some random noise will be added to get the best autoencoder performance. As investigated by the author of the method, this approach provides the possibility of obtaining an artificial set closer to the real dataset compared to the method without noise. That is why this method was taken for comparison. Details of its implementation are given in [7]

The synthesis of the new data is as follows. Beforehand we know the variance and the mean of our latent variables, which are determined by the autoencoder. The next step is the use of a normal distribution with the variance and the mean for each of the latent variables. It is needed for the selection of the value for all latent variables. It is these that serve as starting points from which all attributes of the initial data set can be reproduced.

In case the latent dependencies between variables are linear, the variational autoencoder works like the PCA method [6], which suggests that some random noise will be added to each item according to the method of [5] to obtain better results from the autoencoder. As the reported by author of the method investigated in [5], this approach allows producing an artificial set more similar to the real dataset compared with the method without using noise. That is why this method was taken for comparison. Details of its implementation are presented in [7].

2.1.2. Dataset description

Diagnostics tasks are widespread in the medical industry. The majority of them are reduced to a classification task, and we can apply machine learning techniques. For example, in [8], a dataset is submitted, and the task of predicting the development of diabetes task is formulated. The original variable is represented as 0 or 1. Thus, it is a binary classification problem. Fig. 1 shows a few distributions for some variable pairs.

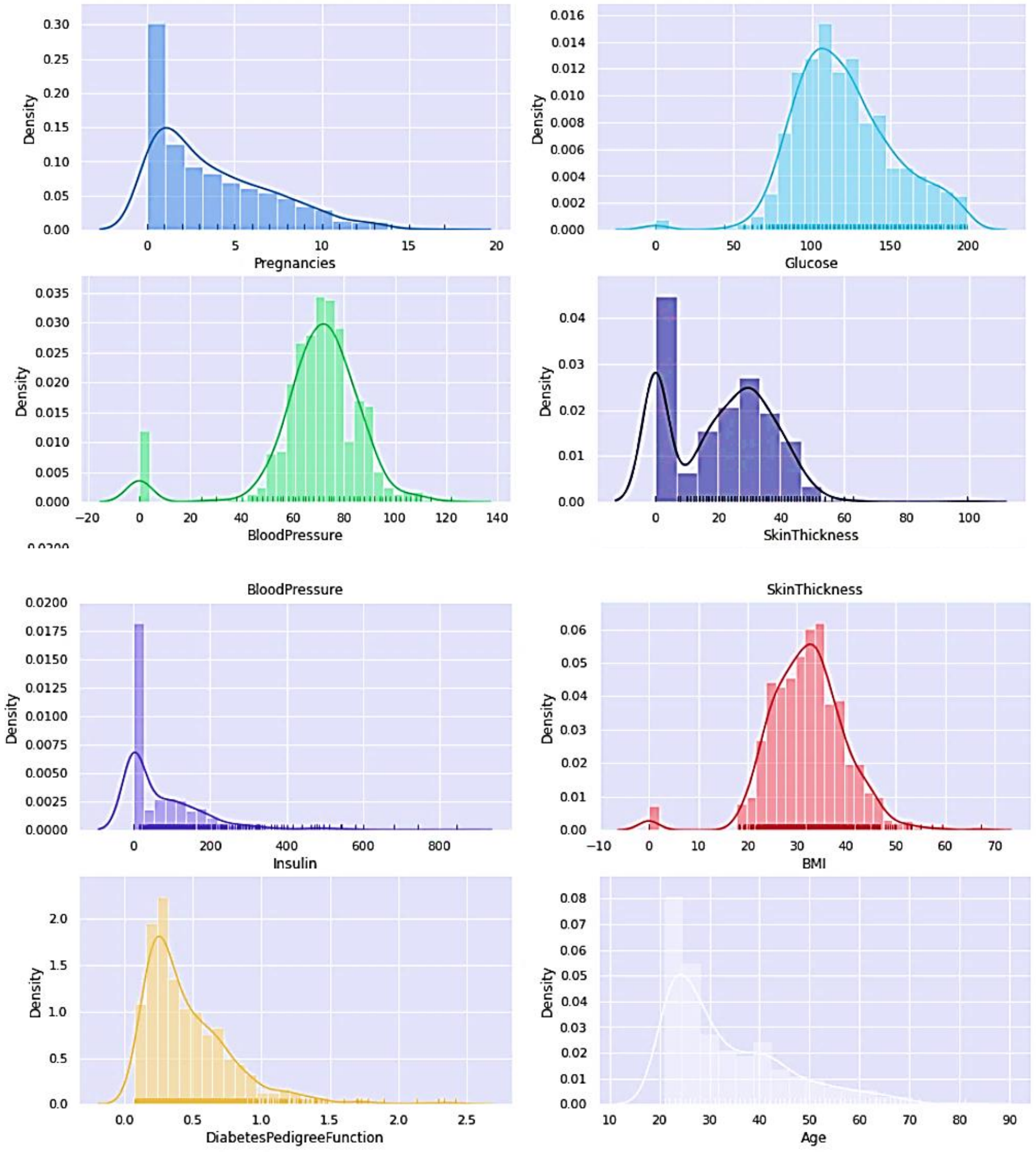


Figure 1: Distribution of features

The dataset includes 9 independent variables:

- Number of times pregnant;
- Plasma glucose concentration a 2 hours in an oral glucose tolerance test;
- Diastolic blood pressure (mm Hg);
- Triceps skin fold thickness (mm);
- 2-Hour serum insulin (muU/ml);
- Body mass index (weight in kg/(height in m)^2);
- Diabetes pedigree function;
- Age (years)

The original variable is represented by 268 cases of no diabetes development and 768 positive cases (diabetes development is diagnosed).

3. Modeling and results

Simulation of the classifiers investigated based on extended datasets using neural network methods. Experimental studies were carried out by dividing the dataset randomly into two parts at a ratio of 80% to 20%. Cross-validation was then applied (5 times). In this way, the reliability of the results was ensured.

The paper presents two neural network approaches for the artificial expansion of short datasets. Let us consider the outcomes and evaluations of the synthesized data for each of them.

3.1.1. Classification using augmented data via autoencoder

Autoencoder-based modeling was carried out in order to synthesize a new dataset whose size would match the size of the original dataset. A comparison between the synthesized dataset with the original one based on several indicators from [9], has revealed the following results: the mean correlations between fake and real columns are 0.97, MAPE estimator results are 0.84, and a similarity score is 0.93.

In addition, Figure 2 presents a comparison of the feature distributions for the initial and synthesized datasets.

It should be noted that the autoencoder generated the same number of instances of each class as the number of instances in the initial dataset.

The application results of different classifiers on the extended dataset are summarised in Table 1.

Table 1
Classification accuracy for investigated ML-based methods using extended dataset obtained by autoencoder

Machine learning algorithm	Total accuracy	Recall	Precision
Random forest classifier	0.8249	0.7139	0.7846
AdaBoost classifier	0.8301	0.6990	0.8230
Logistic regression classifier	0.8295	0.6773	0.8334
SVM classifier	0.7985	0.6205	0.8016

Table 1 reveals that all methods demonstrate high classification accuracy.

Figure 3 show the dependence of the accuracy of the machine learning methods on the amout of the artificially generated vectors added to the initial set.

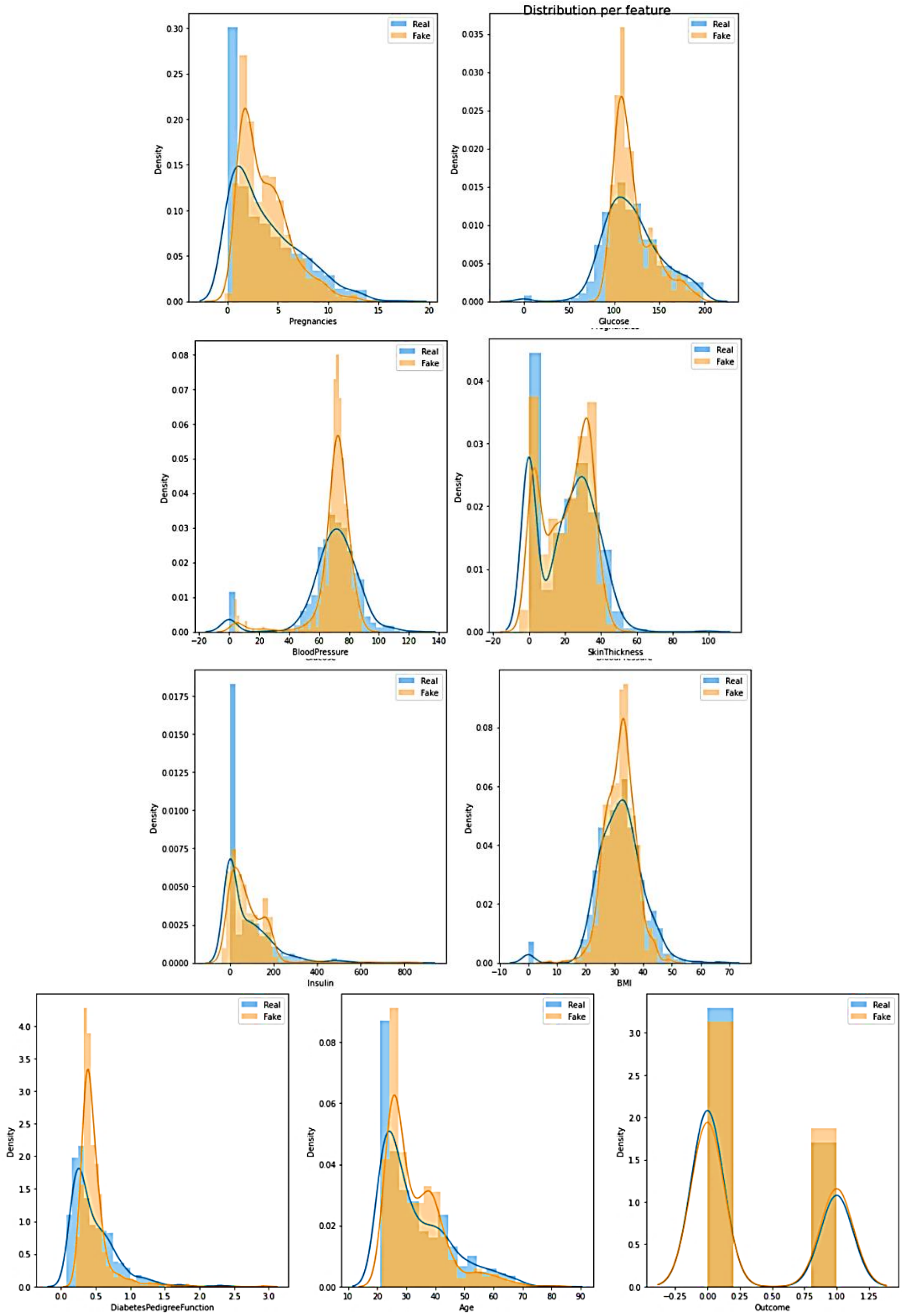
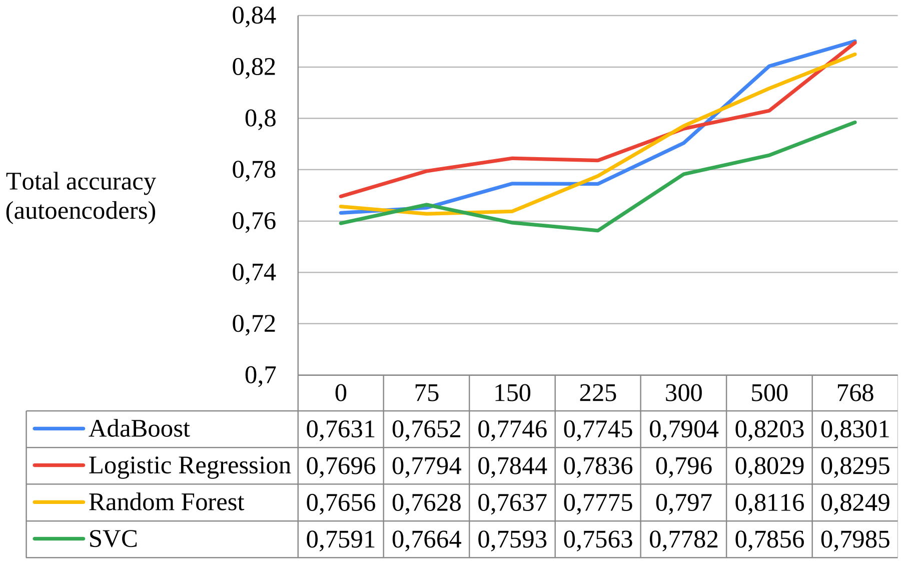


Figure 2: Distribution per features for initial and synthetic datasets



	0	75	150	225	300	500	768
AdaBoost	0,7631	0,7652	0,7746	0,7745	0,7904	0,8203	0,8301
Logistic Regression	0,7696	0,7794	0,7844	0,7836	0,796	0,8029	0,8295
Random Forest	0,7656	0,7628	0,7637	0,7775	0,797	0,8116	0,8249
SVC	0,7591	0,7664	0,7593	0,7563	0,7782	0,7856	0,7985

Figure 3: Dependence of the classification accuracy on the number of generated additional vectors by the autoencoder (0 - initial sample, 768 - 100% of additional vectors)

As can be seen from Figure 3 the increase in the number of additionally added, artificially generated vectors to the initial data set led to an increase in the accuracy of all classifiers. The highest accuracy was obtained by added to the initial set the same dimension of the artificial sample (768 additional vectors).

3.1.2. Classification using augmented data via GAN

The authors adopted the method from [4] for numerical data augmentation. As in the previous cases, the size of the new dataset is equal to the size of the initial one. A comparison between the simulated dataset and the initial one based on several indicators from [9] has shown the following results: mean correlations between fake and real columns are 0.92, MAPE estimator results are 0.67, and a similarity score is 0.59.

Figure 4 shows a comparison of the feature distributions for the initial and synthesized datasets.

It is worth remarking that the GAN-based method attempted to balance the dataset. It generated significantly more instances of the smaller class compared to the initial dataset.

The application results of different classifiers on the extended dataset are summarised in Table 2.

Table 2
Classification accuracy for investigated ML-based methods using extended dataset obtained by GAN

Machine learning algorithm	Total accuracy	Recall	Precision
Random forest classifier	0.7946	0.5853	0.7603
AdaBoost classifier	0.8015	0.6991	0.7734
Logistic regression classifier	0.7352	0.5358	0.6907
SVM classifier	0.7326	0.5101	0,6714

Table 2 suggests that the ensemble-based classifiers exhibit significantly higher accuracy than the other two methods.

Figure 5 show the dependence of the accuracy of the machine learning methods on the amout of the artificially generated vectors added to the initial set.

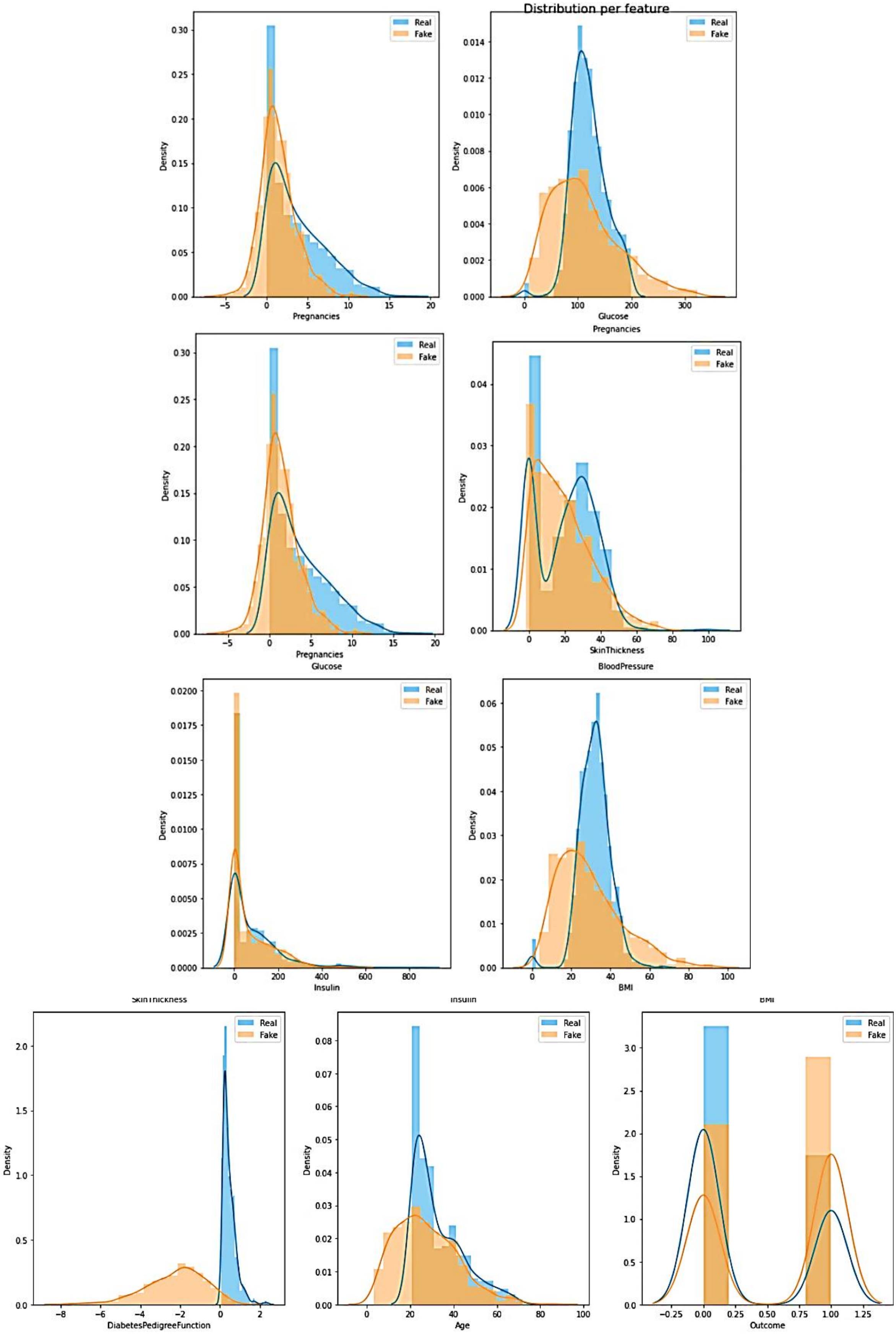
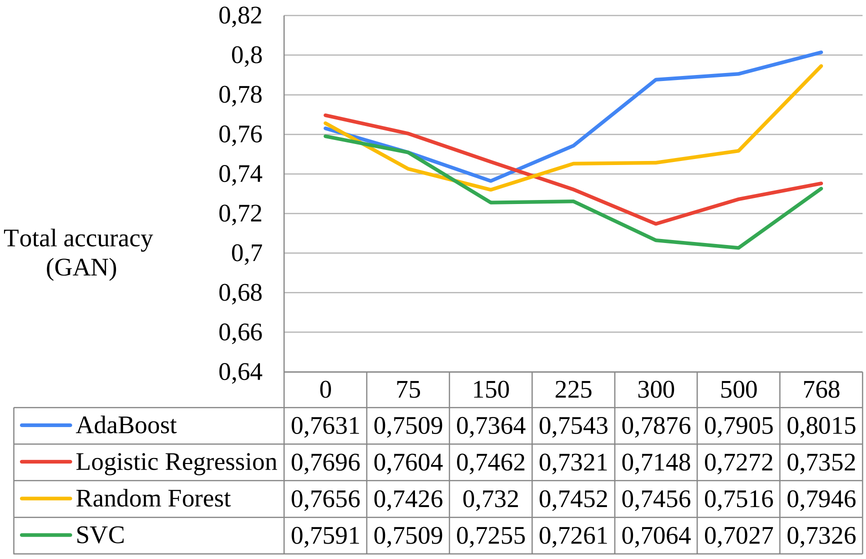


Figure 4: Distribution per features for initial and synthetic datasets



	0	75	150	225	300	500	768
AdaBoost	0,7631	0,7509	0,7364	0,7543	0,7876	0,7905	0,8015
Logistic Regression	0,7696	0,7604	0,7462	0,7321	0,7148	0,7272	0,7352
Random Forest	0,7656	0,7426	0,732	0,7452	0,7456	0,7516	0,7946
SVC	0,7591	0,7509	0,7255	0,7261	0,7064	0,7027	0,7326

Figure 5: Dependence of the classification accuracy on the number of generated additional vectors by GAN (0 - initial sample, 768 - 100% of additional vectors)

As can be seen from Figure 5 the increase in the number of additionally added, artificially generated vectors to the initial data set did not always lead to an increase in the accuracy of the classifiers. Only extension of the initial set by more than 60% of new, artificially synthesized data vectors helped to reduce the errors of classifiers. The highest accuracy, as in the previous case, was obtained by added to the initial dataset the same dimension of the artificial sample (768 additional vectors).

3.2. Comparison and discusion

This section compares both the new datasets generated by both methods under investigation and the classification results based on their application.

3.2.1. Numerical evaluation of the synthetic datasets

In this paper, a comparison of the results obtained by every method investigated was performed on the basis of some indicators from [9]. The results of the comparison between the real dataset and one synthesized by GAN or autoencoder are summarized in Table 3.

Table 3
An evaluation of the synthetic datasets

Indicator	obtained by Autoencoder	obtained by GAN
Mean correlations between fake and real columns	0,97	0,92
Mape estimator results	0,84	0,67
Similarity score	0,93	0,59

As shown in Table 1, the data augmentation method based on the autoencoder provides significantly higher results in comparison with the dataset obtained by GAN. This can significantly affect the performance of classifiers with these data.

However, such dataset decorrelation enables the construction of ensemble models based on a single classifier to process different datasets. This approach can significantly improve the accuracy of classification methods in medicine.

3.2.2. Comparison of the classification accuracy of the different classifiers

The performance of both neural network approaches was compared by determining the accuracy of a few known classifiers: Random forest classifier; AdaBoost classifier; Logistic regression classifier; SVM classifier. They were employed for classification based on the initial and new datasets. It should be noted that the dimensionality of the new datasets was doubled, thus the synthesized data were added to the original one. The outcomes that are based on Total accuracy, Precision, and Recall are shown in Fig. 6. Since the problem is not balanced, F-measure was not taken into account.

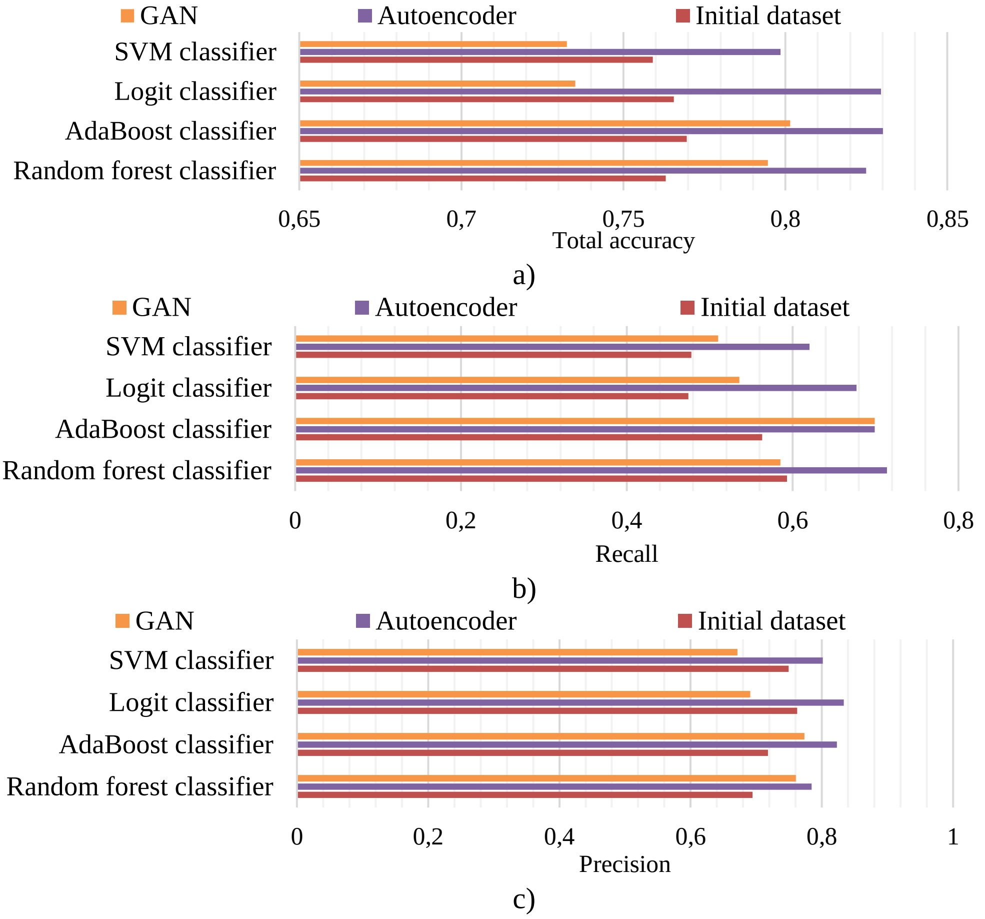


Figure 6: The outcomes of different classifiers based on the initial datasets, and datasets generated using GAN and using autoencoders data augmentation methods: a) Total Accuracy; b) Recall; c) Precision

From the graphs in Fig. 6, it follows that the highest accuracy based on all performance indicators is achieved by using a synthesized dataset with an autoencoder. The application of GAN for data augmentation shows a much lower performance of the known classifiers compared to processing an initial dataset (based on the total accuracy). However, AdaBoost and Random Forest algorithms provide more accurate results in this case. This can be explained by the insufficient amount of training data for effective GAN performance, which affected the one of SVM and Logistic Regression classifiers.

4. Conclusion

This paper deals with the numerical data augmentation task in Clinical Medicine. The authors have experimentally evaluated the performance of modern neural network methods to solve the problem: autoencoders and a GAN. Such an approach helps to reduce risks of overfitting or underfitting when using machine learning models in case of small data processing.

The modeling of the performance of these methods has been carried out using the dataset for solving a classification task. In this case, we tried to predict the possibility of diabetes development. The dataset is not balanced. Experiments have shown that autoencoders generate the most similar data according to the Similiarity score. In addition, the accuracy of classifiers based on these data is significantly higher. Compared with the initial dataset, we have improved the target resolution accuracy by about 10%.

Given the different results of the similarity evaluation of the synthesized datasets concerning the initial one and the different accuracy of the classifiers based on such data, the ensemble learning approach can be used in further research to improve the accuracy of various diagnostics tasks. In particular, the approach of constructing a stacking ensemble of homotypic classifiers that will process different systematically studied datasets seems promising. This very approach can provide a significant increase in the accuracy of classifiers when solving applied tasks of diagnostics in various fields [10-15] when processing average datasets.

The National Research Foundation of Ukraine funds this study from the state budget of Ukraine within the project “Decision support system for modeling the spread of viral infections” (№ 2020.01 / 0025).

5. References

[1] D. Snow, DeltaPy: A Framework for Tabular Data Augmentation in Python, Social Science Research Network, Rochester, NY, 2020. https://doi.org/10.2139/ssrn.3582219.

[2] O. Berezsky, G. Melnyk, T. Datsko, S. Verbovy, An intelligent system for cytological and histological image analysis, in: The Experience of Designing and Application of CAD Systems in Microelectronics, IEEE, Lviv - Polyana, Ukraine, 2015: pp. 28–31. https://doi.org/10.1109/CADSM.2015.7230787.

[3] N. Boyko, M. Kuba, L. Mochurad, S. Montenegro, Fractal Distribution of Medical Data in Neural Network, CEUR-WS.Org. 2488 (2019) 307–318.

[4] L. Xu, M. Skoularidou, A. Cuesta-Infante, K. Veeramachaneni, Modeling Tabular data using Conditional GAN, ArXiv:1907.00503 [Cs, Stat]. (2019). http://arxiv.org/abs/1907.00503 (accessed December 26, 2020).

[5] Deep Learning for tabular data augmentation, Data Science Blog von Lschmiddey. (2021). https://lschmiddey.github.io/fastpages_/2021/04/10/DeepLearning_TabularDataAugmentation.html (accessed May 16, 2021).

[6] V. Kotsovsky, A. Batyuk, M. Yurchenko, New Approaches in the Learning of Complex-Valued Neural Networks, in: 2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP), 2020: pp. 50-54, doi: 10.1109/DSMP47368.2020.9204332.

[7] lschmiddey, lschmiddey/deep_tabular_augmentation, 2021. https://github.com/lschmiddey/deep_tabular_augmentation (accessed May 16, 2021).

[8] Pima Indians Diabetes Database, (n.d.). https://kaggle.com/uciml/pima-indians-diabetes-database (accessed May 16, 2021).

[9] TableEvaluator — table evaluator 15-08-2019 documentation, (n.d.). https://baukebrenninkmeijer.github.io/table-evaluator/table_evaluator.html (accessed May 16, 2021).

[12] D. Chumachenko, O. Sokolov, S. Yakovlev, Fuzzy recurrent mappings in multiagent simulation of population dynamics systems, IJC. (2020) 290–297. https://doi.org/10.47839/ijc.19.2.1773.

[13] M. Zharikova, V. Sherstjuk, Situation diagnosis based on the spatially-distributed dynamic disaster risk assessment, in: 2019 IEEE 14th Intern. Conf. CSIT, 2019: pp. 205-209.

[14] Sergii Babichev, Jiří Škvor, Jiří Fišer, Volodymyr Lytvynenko, Technology of Gene Expression Profiles Filtering Based on Wavelet Analysis, *International Journal of Intelligent Systems and Applications(IJISA),* vol.10, no.4, 2018: pp. 1-7.

[15] S. A. Babichev, V. I. Lytvynenko, M. A. Taif, Estimation of the inductive model of objects clustering stability based on the k-means algorithm for different levels of data noise, *Radio Electronics, Computer Science, Control,* no. 4, 2016: 54-60