Comprehensive Analysis of Different Techniques for Data Augmentation and Proposal of New Variants of BOSME and GAN

Asier Garmendia-Orbegozo[1(✉)], Jose David Nuñez-Gonzalez[1], Miguel Angel Anton Gonzalez[2], and Manuel Graña[3]

[1] Department of Applied Mathematics, University of the Basque Country UPV/EHU, 20600 Eibar, Spain
asier.garmendiao@ehu.eus

[2] TECNALIA, Basque Research and Technology Alliance (BRTA), 20009 San Sebastian, Spain

[3] Computation Intelligence Group, University of the Basque Country UPV/EHU, 20018 San Sebastian, Spain

Abstract. In many environments in which detection of minority class instances is critical, the available data intended for training Machine Learning models is poorly distributed. The data imbalance usually produces deterioration of the trained model by generalisation of instances belonging to minority class predicting as majority class instances. To avoid these, different techniques have been adopted in the literature and expand the original database such as Generative Adversarial Networks (GANs) or Bayesian network-based over-sampling method (BOSME). Starting from these two methods, in this work we propose three new variants of data augmentation to address data imbalance issue. We use traffic event data from three different areas of California divided in two subgroups attending their severity. Experiments show that top performance cases where reached after using our variants. The importance of data augmentation techniques as preprocessing tool has been proved as well, as a consequence of performance drop of systems in which original databases with imbalanced data where used.

Keywords: Data augmentation · Data imbalance · GANs

1 Introduction

Machine Learning (ML) and specially Deep Learning (DL) models have become one of the most effective and useful tool for prediction and inference in different environments such as biomedicine or smart cities. Although data availability is not a matter to be concerned, data imbalance could cause deterioration of performance of the models. No matter the size of the database if there are few

P. García Bringas et al. (Eds.): HAIS 2023, LNAI 14001, pp. 145–155, 2023.
https://doi.org/10.1007/978-3-031-40725-3_13

instances of one of the possibles classes to be determined, the algorithm might generalise by classifying almost all instances as part of the majority class.

In many cases it is of special interest the correct classification of the minority class. For instance, in a cancer diagnosis problem the cost of predicting wrongly a patient with cancer as a cancer-free case is critical. Generally, in databases the minority cases become from patient that suffer cancer, and if this imbalance is extreme, models should generalize by classifying almost all instances as part of majority class, obtaining a high accuracy yet. Other examples of such imbalances could be found in fraud commitment detection where fraudulent cases are less frequent by far.

In traffic event prediction different factors are responsible for causing traffic delays or accidents, and identification of them in real time is crucial for avoiding uncomfortable situations. In this area the instances belonging to traffic misfortunes are a minority in comparison to usual traffic sensor readings too.

To avoid these situations in which minority class instances could not be detected, training the models with as much instances from minority class as majority class instances should be the solution. Over-sampling is a suitable methodology to modify the class variable distribution at a data-level stage (preprocessing), before the learning process. By this way, the model obtains enough information from the minority class to detect these exceptions while performing in real scenarios.

In this work, starting from two different alternatives for expanding original data we proposed three novel ways for generating new instances. Each of them are evaluated in a large, real-world dataset consisting of traffic sensor observations and from the different metropolitan areas from the state of California over a period of three months.

The rest of the paper is organised as follows. Section 2 reviews some of the most representatives works published in the literature. Section 3 specifies the new alternatives proposed by this work. In Sect. 4 the materials and methodology applied in this work are presented. In Sect. 5 we conduct different experiments of classification tasks using data generated by all the alternatives proposed and the results are presented. In Sect. 6 discussion of these results and conclusions are made.

2 State of the Art

Different methodology has been applied to expand original databases to obtain a more generalised source of knowledge resulting on an optimized inference model. In [7] they propose the primitive GAN algorithm in which a generator and discriminator play an adversarial process in which they simultaneously train two models: a generative model that captures the data distribution, and a discriminative model that estimates the probability that a sample came from the training data rather than from the generator. The training procedure for the generator is to maximize the probability of the discriminator making a mistake. By this way, new instances are created with similar characteristics of the original data.

Cibersecurity systems usually face the problem of data imbalance. In [9] they proposed a Multi-task learning model with hybrid deep features (MEMBER) to address different challenges like class imbalance or attack sophistication. Based on a Convolutional Neural Network (CNN) with embedded spatial and channel attention mechanisms, MEMBER introduces two auxiliary tasks (i.e., an auto-encoder (AE) enhanced with a memory module and a distance-based prototype network) to improve the model generalization capability and reduce the performance degradation suffered in imbalanced databases. Continuing with intrusion detection area, a tabular data sampling method to solve the imbalanced learning task of intrusion detection, which balances the normal samples and attack samples was proposed in [6]. In [14] TGAN was presented, as a method for creating tabular data creating discrete and continuous variables like medical or educational records. In [15] they developed CTAB-GAN, a novel conditional table GAN architecture with the ability to model diverse data types, including a mix of continuous and categorical variables, solving data imbalance and long tail issues, i.e., certain variables having drastic frequency differences across large values. In [4] they proposed a method to train generative adversarial networks on multivariate feature vectors representing multiple categorical values.

Bayesian network-based over-sampling method (BOSME) was introduced in [12], which is a new over-sampling methodology based on Bayesian networks. What makes BOSME different is that it relies on a new approach, generating artificial instances of the minority class following the probability distribution of a Bayesian network that is learned from the original minority classes by likelihood maximization.

Some other researchers opted for treating multi-modal data in order to optimize the trained network's inference accuracy. In [13] they proposed an end-to-end framework named Event Adversarial Neural Network (EANN), which is able to obtain event- invariant features and thus benefit the detection of fake news on newly arrived events. In [8] they proposed an audio-visual Deep CNNs (AVDCNN) SE model, which incorporates audio and visual streams into an unified network model. For traffic event detection were also used other approaches that include data from multiple type and sources. In [2] they annotated social streams such as microblogs as a sequence of labelling problem. They presented a novel training data creation process for training sequence labelling models. This data creation process utilizes instance level domain knowledge. In [3] they proposed Restricted Switching Linear Dynamical System (RSLDS) to model normal speed and travel time dynamics and thereby characterize anomalous dynamics. They used the city traffic events extracted from text to explain those anomalous dynamics. In [10] they used human mobility and social media data. A detected anomaly was represented by a sub-graph of a road network where people's routing behaviors significantly differ from their original patterns. They then try to describe a detected anomaly by mining representative terms from the social media that people posted when the anomaly happened. In [5] they used Twitter posts and sensor data observations to detect traffic events using semi-supervised deep learning models such as Generative Adversarial Networks. They extend the

multi-modal Generative adversarial Network model to a semi-supervised architecture to characterise traffic events.

3 Proposed Approach

As we mentioned in the introductory part in classification environments in which data imbalances can cause the performance deterioration of the machine learning model, it is of special interest to have a balanced class distribution. For this purpose, BOSME was proposed tackling this issue by generating synthetic data following the probability distribution of a Bayesian network. Moreover, in the majority of the cases GANs become the first option at extending databases and address imbalance learning tasks. In this work, we assess both option and propose three new variants that raise from both methodologies.

3.1 Variant 1: Feeding the Discriminator of GAN with Data Proceeding from BOSME

The idea of GAN is to maximize the capability of the generator of creating instances as equal as possible as original ones by trying to confuse the discriminator and this last trying to distinguish real data from synthetic data. Originally, discriminator is fed by data generated by the generator raised from normal distribution. If we substitute these data by data generated by BOSME which expand databases including synthetic data from the minority class, the capability of distinguishing real data of the discriminator might be enhanced. Following this idea, we propose this variant, in which first BOSME is applied to the original database and next, a modified version of GAN is applied, where the discriminator is fed by the synthetic data proceeding from BOSME.

3.2 Variant 2: Feeding the Discriminator of GAN with Data Proceeding from BOSME+data Proceeding from the Generator

As the continuation of the variant proposed above, we expand the data with which the discriminator of GAN is fed. We mixed two types of data, the data proceeding from the generator, which is raised from noise, and the synthetic data proceeding from BOSME. By this way, a more general vision of the synthetic data could obtain the discriminator improving its ability to distinguish fake data from real data.

In the previous variant, the data proceeding from BOSME only feeds the discriminator with data from the minority class which could cause some problems in certain environments. In contrast, with this last variant this issue is tackled. A simplistic graphic description is given in Fig. 1

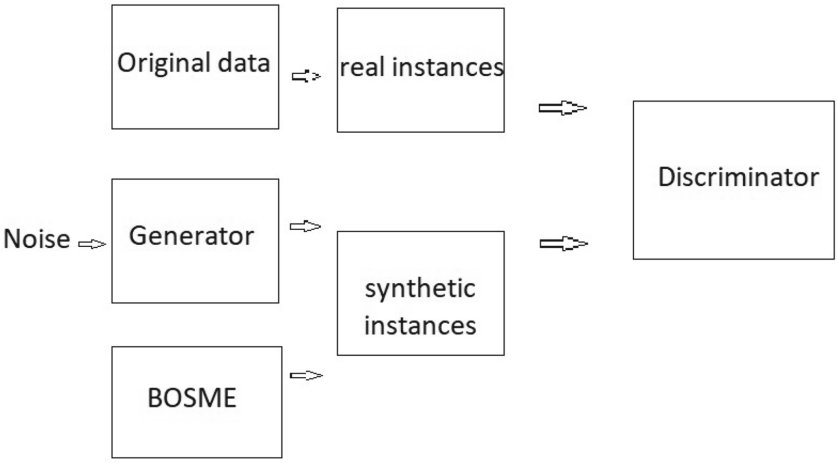

Fig. 1. Graphic diagram of Variant 2

3.3 Variant 3: Application of GAN with Minority Class Data

Finally, we opted for dividing the original data based on its class. The data that belong to minority class is used to feed GAN, and synthetic data is created following the GAN architecture. By this way, data imbalance issue is addressed and the resulting classification task should be enhanced.

4 Materials and Methods

4.1 Material and Environment

In this work we tested each of the variants proposed in the above section as well as the original GAN and BOSME methodology by expanding original data from a large, real-world dataset consisting of traffic sensor observations and from the different metropolitan areas from the state of California over a period of three months.

The Caltrans Performance Measurement System (PeMS) [1] provides large amount of traffic sensor data that has been widely used by the research communities. We collected traffic events within a three months period from 31st July 2013 to 31st October 2013, for three different metropolitan area of the state of California, i.e., Bay Area, North Central and Central Coast. We divided each traffic event depending their level of risk, i.e., hazard and control. In each case we identified the minority class to proceed with each variant proposed in this work.

The environment in which all testing and training procedure took place is the following. We used Machine Learning oriented sklearn [11] library of Python in a 64 bit Windows operating system running on Intel Core i5-2010U CPU at 1,6 GHz $\times$ 4.

4.2 Methodology

First of all, we applied each of the three variants proposed in Sect. 3 as well as the original BOSME and GAN methodologies. By this way, we had 5 ways of generating synthetic data starting from the original databases. Next, each of the expanded databases were used to feed 7 well-known ML classifiers that are listed below. Each of them has been used in the default configuration of sklearn except for the attributes mentioned.

- DT: Decision Tree (criterion = entropy)
- RF: Random Forest (number of estimators = 150, criterion = entropy)
- knn: k-Nearest Neighbors(number of neighbors = 3, weights = distance)
- GNB: Gaussian Naive Bayes
- AB: Adaboost (Base Classifier: Decision Tree)
- MLP: Multilayer Perceptron
- SVM: Supported Vector Machine

Different performance metrics were used for determining which of the aforementioned techniques for extending the original data fits best with traffic event prediction task. These are accuracy, recall, precision, F1 score and AUC (Area Under the ROC Curve). AUC is the area below the ROC curve, i.e., a graph showing the performance of a classification model at all classification thresholds. What is plotted in the curve is the FPR and TPR in the x and y axes, respectively, whose definitions are given in Eq. 4 and 5. The definitions of the rest of the metrics mentioned above are given in Eqs. 1 ,3 and 6, where TP, TN, FP, and FP stand for True Positives, True Negatives, False Positives, and False Negatives, respectively.

$$Accuracy(Acc) = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Recall(Re) = \frac{TP}{TP + FN} \tag{2}$$

$$Precision(Pr) = \frac{TP}{TP + FP} \tag{3}$$

$$TruePositiveRate(TPR) = \frac{TP}{TP + FN} \tag{4}$$

$$FalsePositiveRate(FPR) = \frac{FP}{FP + TN} \tag{5}$$

$$F1 - score(F1) = \frac{2 * Pr * Re}{Pr + Re} \tag{6}$$

5 Experimental and Results

Each of the methodologies cited in this work were tested for data augmentation of original traffic event databases. For 10 different seeds a 10-fold cross validation were developed in each case to obtain all performance metrics. In each table, in the first column the metropolitan area of event detection and the data augmentation methodology applied are given, where BA, NC and CC stand for Bay Area, North Central and Central Coast respectively. Vx stand for x variant we proposed in Sect. 3, and Original means that the evaluation was done using the original database. The abbreviation of each classifier is given in Sect. 4.

As it could be observed in Table 1 in the majority of the cases the application of GAN or BOSME independently outperforms the variants we proposed in terms of accuracy. This metric is not very representative, because the original database also is useful. In fact, due to the class imbalance in these databases, the accuracy does not degrade, i.e., the few minority class instances could be wrongly classified even offering a good accuracy overall. Other metrics are needed to obtain more general conclusion, so we opted for attending Precision, Recall, F1-Score and AUC. As the most determining action is the correct classification of instances from the minority class, Recall is the most representative metric, since it determines how good is a classifier predicting a positive instance as positive, i.e., it defines a ratio between instances classified as positive ones and all positive

Table 1. Accuracies of different classifiers for different data augmentation techniques for different areas.

Area-Method	RF (%)	DT (%)	knn (%)	GNB (%)	AB (%)	MLP (%)	SVM (%)
BA-GAN	99.3225	99.3225	99.3225	88.1021	99.3225	97.3432	96.6745
BA-V1	97.1225	97.1572	96.8251	95.2354	45.4789	96.8561	97.0731
BA-V2	97.0996	97.1572	96.8605	95.2354	16.0777	97.0553	97.0731
BA-V3	99.3225	99.3225	99.2871	88.2349	99.3225	96.9801	96.7631
BA-BOSME	99.3225	99.3225	99.2915	94.6687	99.3225	96.0369	95.1868
BA-Original	97.3282	96.4740	97.2193	96.8827	96.9784	97.2423	97.4849
NC-GAN	98.9076	98.9076	98.9076	92.1753	98.9076	94.8792	94.8929
NC-V1	93.6228	93.6774	92.7625	89.526	50.2666	92.7489	93.2815
NC.V2	93.6091	93.6774	92.8444	89.526	92.4212	92.7489	93.2815
NC-V3	98.9076	98.9076	98.8529	91.4788	98.9076	94.729	94.8929
NC-BOSME	98.9076	98.9076	98.9076	92.4212	98.9076	93.1177	93.2407
NC-Original	94.4505	93.0577	94.5378	94.4424	93.6777	94.8301	94.8661
CC-GAN	99.6858	99.6858	99.6858	94.7013	99.686	96.1925	96.1535
CC-V1	99.6466	99.6858	96.7422	4.9455	99.6863	2.9438	96.1533
CC-V2	95.0855	95.1241	94.6209	90.8156	91.3716	94.5051	94.5051
CC-V3	99.686	99.6858	99.6858	94.7013	99.686	95.918	96.1535
CC-BOSME	99.6858	99.6858	99.6858	94.0726	99.6858	92.9348	93.5239
CC-Original	95.3254	92.7939	95.4468	95.6437	94.2579	95.6121	95.8008

instances. As shown in Table 2 in the Original database cases the performance metric drops significantly. Thus, it is necessary more instances from the minority lass for the proper training of each of the classifiers.

Other interesting metric to be observed is the AUC. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example. In this case, the original database offers the worst performance given the moderate percentage of random negative instances due to the generalization as a consequence of data imbalance. Attending the rest of the variants mentioned and proposed within this work, we can not deduce which is the best given that in each area for some method some variants perform better than others and vice-versa for other methods. For instance, variant 2 suits best for Random Forest classifier, whereas for knn classifier is the worst option. The application of GAN and BOSME independently offers a regular performance between different classifiers. However, the highest percentage is obtained by the combination V3-AB in Bay Area, V1-AB in North Central and V2-AB in Central Coast, which means that our variants are the most adequate applying the best classifier. Table 3 shows all these measurements of the aforementioned metric.

Table 2. Recall of different classifiers for different data augmentation techniques for different areas.

Area-Method	RF (%)	DT (%)	knn (%)	GNB (%)	AB (%)	MLP (%)	SVM (%)
BA-GAN	99.3225	99.3225	99.3225	88.1021	99.3225	97.3432	96.6745
BA-V1	97.1225	97.1572	96.8251	95.2354	45.4789	96.8561	97.0731
BA-V2	97.0996	97.1572	96.8605	95.2354	16.0777	97.0553	97.0731
BA-V3	99.3225	99.3225	99.2871	88.2349	99.3225	96.9801	96.7631
BA-BOSME	99.3225	99.3225	99.2915	94.6687	99.3225	96.0369	95.1868
BA-Original	66.129	65.2011	63.9704	64.8414	66.21	56.6379	58.0993
NC-GAN	98.9076	98.9076	98.9076	92.1753	98.9076	94.8792	94.8929
NC-V1	93.6228	93.6774	92.7625	89.526	50.2666	92.7489	93.2815
NC.V2	93.6091	93.6774	92.8444	89.526	92.4212	92.7489	93.2815
NC-V3	98.9076	98.9076	98.8529	91.4788	98.9076	94.729	94.8929
NC-BOSME	98.9076	98.9076	98.9076	92.4212	98.9076	93.1177	93.2407
NC-Original	69.3718	68.9104	69.0945	66.1001	68.3082	64.2413	61.3296
CC-GAN	99.6858	99.6858	99.6858	94.7013	99.686	96.1925	96.1535
CC-V1	99.6466	99.6858	96.7422	4.9455	99.6863	2.9438	96.1533
CC-V2	95.0855	95.1241	94.6209	90.8156	91.3716	94.5051	94.5051
CC-V3	99.686	99.6858	99.6858	94.7013	99.686	95.918	96.1535
CC-BOSME	99.6858	99.6858	99.6858	94.0726	99.6858	92.9348	93.5239
CC-Original	67.337	65.3427	65.278	71.0973	66.0623	63.9602	65.1639

Table 3. AUC of different classifiers for different data augmentation techniques for different areas.

Area-Method	RF (%)	DT (%)	knn (%)	GNB (%)	AB (%)	MLP (%)	SVM (%)
BA-GAN	90.2775	89.871	90.5662	49.1252	91.5185	63.6251	58.9844
BA-V1	90.8908	89.871	71.726	65.9578	56.018	57.1239	58.1556
BA-V2	90.4126	89.871	72.8234	65.9578	71.1632	59.0527	58.1556
BA-V3	90.3776	89.871	85.553	49.1935	92.5633	62.1791	58.9629
BA-BOSME	90.6981	89.871	90.1002	66.1345	90.3663	65.4085	66.256
BA-Original	82.6257	68.4206	74.1477	71.2723	75.8556	80.7843	71.7135
NC-GAN	91.8614	91.4632	91.4632	68.8858	91.4632	60.176	62.4061
NC-V1	91.9706	91.4632	76.2103	68.8858	97.6291	67.3004	62.4061
NC-V2	91.8896	91.4632	76.568	68.8858	74.5844	67.3004	62.4061
NC-V3	91.9108	91.4632	91.9939	68.5141	92.5419	60.0394	62.4061
NC-BOSME	91.835	91.4632	91.4632	69.1838	93.9506	69.3278	68.0071
NC-Original	83.5906	70.2484	76.7554	74.5938	74.9902	81.4973	76.1547
CC-GAN	97.0206	97.0206	97.0206	72.4319	98.4706	66.0791	67.1447
CC-V1	97.1699	96.8562	71.374	71.8002	97.7436	63.1151	66.9084
CC-V2	96.5661	97.0206	71.6273	0.5	99.5428	69.3871	67.0409
CC-V3	97.2947	97.0206	97,0206	72.4319	97.8178	70.6532	67.1447
CC-BOSME	97.0206	97.0206	97.0206	75.007	97.0206	76.2153	74.9746
CC-Original	79.7884	65.2088	72.5906	82.3193	75.6649	77.5265	75.4132

6 Discussion and Conclusion

In this work we realised the importance of having a balanced data in classification tasks in order to avoid generalisation of the resulting training of the classifiers. In different environments an incorrect classification of an instance belonging to minority class could have a critical impact. Thus, a data preprocessing is needed to extend minority class instances and address this issue.

First, we looked through the accuracies of different classifiers after applying every data augmentation methodology described in previous sections. We saw that there was no evident difference between the application of different method for balancing data or starting from the original database training the classifiers. The low number belonging to the minority class was causing this, their incorrect classification not degrading severely.

However, if we look other metrics such Recall or AUC, we can figure out the importance of these data augmentation techniques. By this way, the classifiers have enough instances from both classes for training phase, and the problem of generalisation is tackled. In each metropolitan area used analysed in the experimental process the original data augmentation techniques perform more regularly than the variants proposed in this work. In fact, if we determine their goodness attending their overall performance metric within all the classifiers we can deduce that they outperform the variants we proposed. Nevertheless, the

highest AUC values where obtained by one of the variants proposed in Sect. 3 for each metropolitan area.

For Bay Area, the highest AUC value was obtained after applying our third variant, i.e., the application of GAN for the minority class instances, and posterior use of Adaboost (Decision Tree as base classifier) as classifier.

For North Central area, our first approach gives the highest AUC value, i.e., the use of the new instances proceeding from BOSME and the instances proceeding from the generator as the entry for the discriminator, and the posterior use of GAN for the creation of new instances. Finally, Adaboost (Decision Tree as base classifier) was used as classifier.

In case of Central Coast, the second approach gives the best AUC value, i.e., the use of the new instances proceeding from BOSME as the entry for the discriminator, and the posterior use of GAN for the creation of new instances. Finally, Adaboost (Decision Tree as base classifier) was used as classifier.

To summarize, the power of the data augmentation techniques as preprocessing tool in data imbalance environments has been exceedingly demonstrated in this work. The adequateness of each variant proposed in this work depends on the characteristics and distribution of the original database, and the posterior machine learning model to be adopted for the classification task. For more complex models such as Adaboost or Random Forest where more than a single classifier are evaluated our variants outperform the original GAN and BOSME. Although, the highest values of AUCs are obtained by one of these variants in each metropolitan area the overall performance in more simple models is better for the simplest data augmentation methodologies. Depending the application or the limitations of the hardware to be deployed all the system, some options would be more adequate than others. For instance, if we have to adjust the models size or the training time is critic lighter models should be used and the original GAN and BOSME would be the option to adopt in these cases. In contrast, if there are no such restrictions, the possibility of finding the best classifier and the best variant for addressing data imbalance issue would be the best alternative. Following this line, finding an automatic way of finding the best combination of data augmentation and classification model would alleviate big part of finding the best alternative, improving system's time efficiency.

Acknowledgments. Authors received research funds from 59 the Basque Government as the head of the Grupo de Inteligencia Computacional, Universidad del Pais Vasco, UPV/EHU, from 2007 until 2025. The current code for the grant is IT1689-22. Additionally, authors participate in Elkartek projects KK-2022/00051 and KK-2021/00070. The Spanish MCIN 5has also granted the authors a research project under code PID2020-116346GB-I00.

References

1. Caltrans. performance measurement system (pems). Accessed 07 Mar 2023, http://pems.dot.ca.gov,

2. Anantharam, P., Barnaghi, P., Thirunarayan, K., Sheth, A.: Extracting city traffic events from social streams. ACM Trans. Intell. Syst. Technol. **6**, 1–27 (2015). https://doi.org/10.1145/2717317
3. Anantharam, P., Thirunarayan, K., Marupudi, S., Sheth, A., Banerjee, T.: Understanding city traffic dynamics utilizing sensor and textual observations, vol. 30 (2016)
4. Camino, R.D., Hammerschmidt, C.A., State, R.: Generating multi-categorical samples with generative adversarial networks. ArXiv abs/1807.01202 (2018)
5. Chen, Q., Wang, W., Huang, K., De, S., Coenen, F.: Multi-modal generative adversarial networks for traffic event detection in smart cities. Expert Syst. Appl. **177**, 114939 (2021). https://doi.org/10.1016/j.eswa.2021.114939, https://www.sciencedirect.com/science/article/pii/S0957417421003808
6. Ding, H., Chen, L., Dong, L., Fu, Z., Cui, X.: Imbalanced data classification: A KNN and generative adversarial networks-based hybrid approach for intrusion detection. Future Gener. Comput. Syst. **131**, 240–254 (2022). https://doi.org/10.1016/j.future.2022.01.026, https://www.sciencedirect.com/science/article/pii/S0167739X22000346
7. Goodfellow, I., et al.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. Curran Associates Inc
8. Hou, J.C., Wang, S.S., Lai, Y.H., Tsao, Y., Chang, H.W., Wang, H.M.: Audio-visual speech enhancement using multimodal deep convolutional neural networks. IEEE Trans. Emerg. Top. Comput. Intell. **2**(2), 117–128 (2018). https://doi.org/10.1109/TETCI.2017.2784878
9. Lan, J., Liu, X., Li, B., Sun, J., Li, B., Zhao, J.: Member: a multi-task learning model with hybrid deep features for network intrusion detection. Comput. Secur. **123**, 102919 (2022). https://doi.org/10.1016/j.cose.2022.102919, https://www.sciencedirect.com/science/article/pii/S016740482200311X
10. Pan, B., Zheng, Y., Wilkie, D., Shahabi, C.: Crowd sensing of traffic anomalies based on human mobility and social media. IN: Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (2013)
11. Pedregosa, F., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2012)
12. Rosario, D., Nuñez-Gonzalez, J.D.: Bayesian network-based over-sampling method (bosme) with application to indirect cost-sensitive learning. Sci. Rep. 12 (2022). https://doi.org/10.1038/s41598-022-12682-8, https://www.nature.com/articles/s41598-022-12682-8
13. Wang, Y., et al.: Eann: event adversarial neural networks for multi-modal fake news detection. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 849–857. KDD 2018, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3219819.3219903
14. Xu, L.: Synthesizing tabular data using generative adversarial networks (2018)
15. Zhao, Z., Kunar, A., van der Scheer, H., Birke, R., Chen, L.Y.: Ctab-gan: Effective table data synthesizing. ArXiv abs/2102.08369 (2021)