



Research Artificial Intelligence—Article

Wasserstein GAN-Based Small-Sample Augmentation for New-Generation Artificial Intelligence: A Case Study of Cancer-Staging Data in Biology



Yufei Liu ^{a,b,d}, Yuan Zhou ^{b,*}, Xin Liu ^a, Fang Dong ^b, Chang Wang ^a, Zihong Wang ^c

^a College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China

^b School of Public Policy and Management, Tsinghua University, Beijing 100084, China

^c School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

^d Center for Strategic Studies, Chinese Academy of Engineering, Beijing 100088, China

ARTICLE INFO

Article history:

Received 5 March 2018

Revised 2 June 2018

Accepted 7 November 2018

Available online 11 January 2019

Keywords:

Artificial intelligence

Generative adversarial network

Deep neural network

Small sample size

Cancer

ABSTRACT

It is essential to utilize deep-learning algorithms based on big data for the implementation of the new generation of artificial intelligence. Effective utilization of deep learning relies considerably on the number of labeled samples, which restricts the application of deep learning in an environment with a small sample size. In this paper, we propose an approach based on a generative adversarial network (GAN) combined with a deep neural network (DNN). First, the original samples were divided into a training set and a test set. The GAN was trained with the training set to generate synthetic sample data, which enlarged the training set. Next, the DNN classifier was trained with the synthetic samples. Finally, the classifier was tested with the test set, and the effectiveness of the approach for multi-classification with a small sample size was validated by the indicators. As an empirical case, the approach was then applied to identify the stages of cancers with a small labeled sample size. The experimental results verified that the proposed approach achieved a greater accuracy than traditional methods. This research was an attempt to transform the classical statistical machine-learning classification method based on original samples into a deep-learning classification method based on data augmentation. The use of this approach will contribute to an expansion of application scenarios for the new generation of artificial intelligence based on deep learning, and to an increase in application effectiveness. This research is also expected to contribute to the comprehensive promotion of new-generation artificial intelligence.

© 2019 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction and background

The concept of artificial intelligence was first proposed in 1956 [1]; since then, profound changes have taken place in the development of artificial intelligence based on information technology and an increased data scale [2]. These changes are particularly outstanding in certain fields, such as the mobile Internet, big data, supercomputing, sensor networks, and brain science. The Development Plan for a Next-Generation Artificial Intelligence [3], which was issued by the State Council of China in July 2017, describes artificial intelligence as moving into a new stage. The plan outlines how the new generation of artificial intelligence will be

characterized by deep learning, cross-border fusion, human-machine collaboration, crowd intelligence, and autonomous intelligence. The foundation of these technologies is big-data-driven methodology [4]. Pan [2] has also described big data intelligence as the basic method and important development direction of the new generation of artificial intelligence.

Deep learning, which was developed by Hinton and Salakhutdinov [5], has become the key technology of big data intelligence [6] and has led to major breakthroughs, such as intelligent driving [7], smart cities [8], voice recognition [9], and information retrieval [10]. Compared with classical statistical machine-learning methodologies, deep learning, as the core of the big data intelligence method, has a relatively complex model structure. The size and quality of a dataset can significantly affect a deep-learning classifier. Large-scale annotated sample data are required in order

* Corresponding author.

E-mail address: zhou_yuan@tsinghua.edu.cn (Y. Zhou).

to fully optimize the model parameters and obtain superior performance [11]. In other words, under the existing framework, the performance of a deep-learning model is determined by the scale and quality of the annotated data; this situation also influences the development of the new generation of artificial intelligence. Nevertheless, it is both difficult and expensive to obtain labeled sample data in many real-world applications. For example, a series of long-term and expensive experiments [12] is required to generate a training sample in biology, which can then be used to train the classifiers with high accuracy. In the field of computerized numerical control (CNC) machine tools, it takes decades to accumulate annotation datasets of a sufficient size, while data on certain specific cases of CNC are rare [13]. Meanwhile, the implementation of big data methods in CNC can be even more difficult in China, where CNC is still in the developmental stage. In strategic intelligence analysis, the labeling of samples requires close and seamless cooperation among outstanding experts from multiple fields [14–20]; thus, it is extremely expensive to obtain sufficiently sized datasets. In addition to the high cost, the data features are complicated and high dimensionality exists. In this situation, the dimensionality of the original feature space is approximately equal to or greater than the number of samples, which is known as “the small sample size problem” [21]. With a small sample size, deep learning is restricted to good generalization performance. Furthermore, the development of a new generation of artificial intelligence is limited because there are considerably more fields with a small sample size than fields with a big data environment [22–26].

In previous studies, several oversampling methods have been proposed in order to address the insufficiency of the data scale. The main advantage of these methods is that they are self-sufficient. In the early stage, the training set can be enlarged by duplicating the training examples of the minority class if the examples of different classes are imbalanced, or by creating a new dataset by adding artificial noises to the existing ones [27]. In 2002, Chawla et al. [28] proposed a classic oversampling method, called the synthetic minority oversampling technique (SMOTE), which involves the creation of a synthetic minority class dataset. On the basis of the SMOTE method, Han et al. [29] proposed two novel minority oversampling techniques, which consider neighboring instances and only the minority instances near the borderline, respectively. In 2008, He et al. [30] proposed the adaptive synthetic sampling approach, which utilizes a weighted distribution for minority class instances according to the level of difficulty in learning. The majority weighted minority oversampling technique was then proposed in 2014 by Barua et al. [31]. This method aims to generate useful synthetic minority class instances by identifying hard-to-learn minority class samples and assigning weights based on their Euclidian distance from the nearest majority class instance. It then generates synthetic instances using a clustering method. Many more methods have been developed to meet dataset demands. In 2015, Xie et al. [32] suggested a minority oversampling technique based on local densities in a low-dimensional space in order to address the problem of dimensionality that affected earlier methods; this technique involved mapping the trained sample into a low-dimensional space and assigning weights. In 2017, Douzas and Bacao [33] proposed a self-organizing map oversampling method, in which artificial data are used for certain classifiers. Most of the abovementioned methods focus on imbalance learning, in which better performance can be achieved by adding oversampling instances to the minority class dataset. However, datasets in many realms remain insufficient, rather than imbalanced, in every class.

In order to address the small sample size problem, a few traditional approaches other than SMOTE have also been raised. In 1995, Bishop [34] proposed that training with noise could lead to a good result; this concept is equivalent to Tikhonov regulation. In 2004, the neural-ensemble-based C4.5, a decision tree method,

was introduced by Zhou and Jiang [35]; this method first trains a neural network, and then employs it to generate a new training set. A virtual sample generation method based on an internalized kernel density estimate was introduced by Li and Lin [36] in 2006; this method involves determining the probability density function of the samples, which is then used to generate training samples. In 2009, Li and Fang [37] again proposed a non-linear virtual sample generation technique using group discovery and parametric equations of the hypersphere. Nevertheless, these methods are unable to make use of the inherent features of the samples, resulting in a limitation of the training models.

In recent years, with the rapid development of the new generation of artificial intelligence and big data, the use of a generative adversarial network (GAN) based on a deep neural network (DNN) has provided opportunities to create new approaches to solve the data problem. It has also made the application of deep learning possible in the case of a small sample size. A GAN is a powerful type of generative model [38] that was introduced in 2014 by Goodfellow et al. [39]; it can be utilized to generate synthetic samples with the same distribution as the real data in order to solve the insufficiency problem of annotated data [40]. A GAN consists of two deep architecture functions for the generator and the discriminator, which can simultaneously learn from the trained data in an adversarial fashion [41]. In the learning process, the generator captures the potential distribution of the real data and generates synthetic samples, while the discriminator discriminates between the real samples and the synthetic samples as accurately as possible.

Recent work has shown that a GAN can successfully be applied to image generation, language processing, and supervised learning with insufficient training data. From the perspective of image generation, Santana and Hotz [42] proposed an approach to generate images with the same distribution as real driving scenarios. Gou et al. [43] utilized a GAN to learn from both real and synthetic images in order to improve the accuracy of eye detection. From the perspective of language processing, Li et al. [44] utilized a GAN to capture the relevance of dialog and to generate corresponding synthetic text, while Pascual et al. [45] proposed a speech-enhancement framework based on a GAN. These studies demonstrate that the synthetic samples generated by a GAN conform to the distribution of the original samples. Moreover, the success of GANs in various fields indicates that this generative model is independent of precise domain knowledge, which is propitious for the application of this approach in other fields. From the perspective of supervised learning with insufficient training data, most of the existing studies deal with the problem of class imbalance, and a GAN is often used as an oversampling method. Fiore et al. [46] utilized a GAN to generate synthetic illicit transaction records, and merged these synthetic records into an augmented training set to improve the effectiveness of credit card fraud detection. Douzas and Bacao [47] utilized a GAN to generate synthetic samples for the minority class of various imbalanced datasets; the results showed that the GAN performs better than the other oversampling methods. These studies demonstrate that data augmentation with a GAN is more effective than traditional oversampling methods in improving the quality of data. Furthermore, most standard augmentation methods have been integrated into Augmentor, a well-accepted data augmentation tool with a high-level application programming interface (API) [48]. Despite the remarkable success of GANs in balancing datasets, their application in multi-classification with a small sample size has, surprisingly, not yet been studied, to the best of our knowledge. Enlargement of the data scale by means of data augmentation makes it possible to improve the performance of supervised machine-learning based on DNNs in various fields. To sum up, a GAN provides an opportunity to solve the problem posed by a small sample size and thus improve multi-classification.

To solve the multi-classification problems posed by a small sample size, we propose an approach that combines a GAN with a DNN. In this research, the original samples were first divided into a training set and a test set. The GAN method was utilized as data augmentation in order to generate synthetic sample data to enlarge the training set scale of cancer staging in biology, and to satisfy the conditions of DNN model training. Next, the DNN method was trained with the synthetic samples and tested using the test set. Finally, the effectiveness of the approach for multi-classification with a small sample size was validated by comparing several indicators with a combination of the classical supervised machine-learning approaches—DNN, SMOTE, and GAN. The approach proposed in this paper is an attempt to transform the classical statistical machine-learning classification method based on original samples into a deep-learning classification method based on data augmentation. Furthermore, this research is conducive to exploring the potential of the application range and reliability of the new generation of artificial intelligence, as represented by deep learning. This is the first attempt to utilize an approach combining GAN with DNN to improve the effectiveness of multi-classification for cancer staging.

2. Methodology

2.1. Workflow of the study

To solve the supervised learning problem posed by a small sample size and extend the scope of application of deep learning, this paper proposes an approach combining a GAN with a DNN classifier for multi-classification. This approach can be outlined as follows (Fig. 1):

- (1) Divide the original samples into a training set and a test set.
- (2) Use the training set to train the GAN and tune its hyperparameters.
- (3) Use the trained generator of the GAN to generate synthetic samples, and use the discriminator to filter these samples.
- (4) Use the synthetic samples to train the DNN classifier, and use the test set to test the DNN classifier.

2.2. Generative adversarial network

The Wasserstein generative adversarial network (WGAN) [49] was used to generate the synthetic samples in this study, as the

training process of the original GAN was a minimax game, and the optimization goal was to reach the Nash equilibrium [40], which posed the vanishing gradient problem [50]. Compared with the original GAN, WGAN uses the Wasserstein distance instead of the Jensen–Shannon (JS) divergence to evaluate the distribution distance between the real samples and the generated samples [51]. Using the Wasserstein distance, the training process of WGAN was more stable and faster than that of the original GAN [52].

In the proposed approach, the process of generating synthetic samples using WGAN consisted of two stages. First, the generator began to generate the original synthetic samples when the loss functions of the generator and the discriminator converged after being trained tens of thousands of times. Second, according to the GAN's concept of the adversary [53], the generator attempted to generate synthetic samples that could fool the discriminator, while the discriminator attempted to discriminate between the real samples and the synthetic samples. In other words, when a synthetic sample was identified as real by the discriminator, that synthetic sample fooled the discriminator. The original synthetic samples that fooled the discriminator were taken as the final synthetic samples.

2.3. Deep neural network

This study used a DNN, which is a deep architecture classifier based on deep learning, as a classifier. A DNN classifier can utilize several models of computation to learn representations of data with multiple layers of abstraction; the models are composed of multiple processing layers. The DNN classifier was trained with a large number of synthetic samples generated by WGAN in order to avoid overfitting. In the proposed approach, the DNN classifier was tested with the test set, which could inspect the generalization performance of the classifier. To test the performance of the DNN classifier, we used three multi-classification metrics based on a confusion matrix (Fig. 2): accuracy, the F -measure, and the G -mean. Accuracy denoted the proportion of predictions that were correct, the F -measure represented the harmonic mean of precision and recall [54], and the G -mean indicated the geometric mean of recall [55]. Accuracy, F -measure, and G -mean are defined in Eqs. (1–3).

$$\text{Accuracy} = \frac{\sum_{i=1}^L n_{ii}}{\sum_{i=1}^L \sum_{j=1}^L n_{ij}} \quad (1)$$

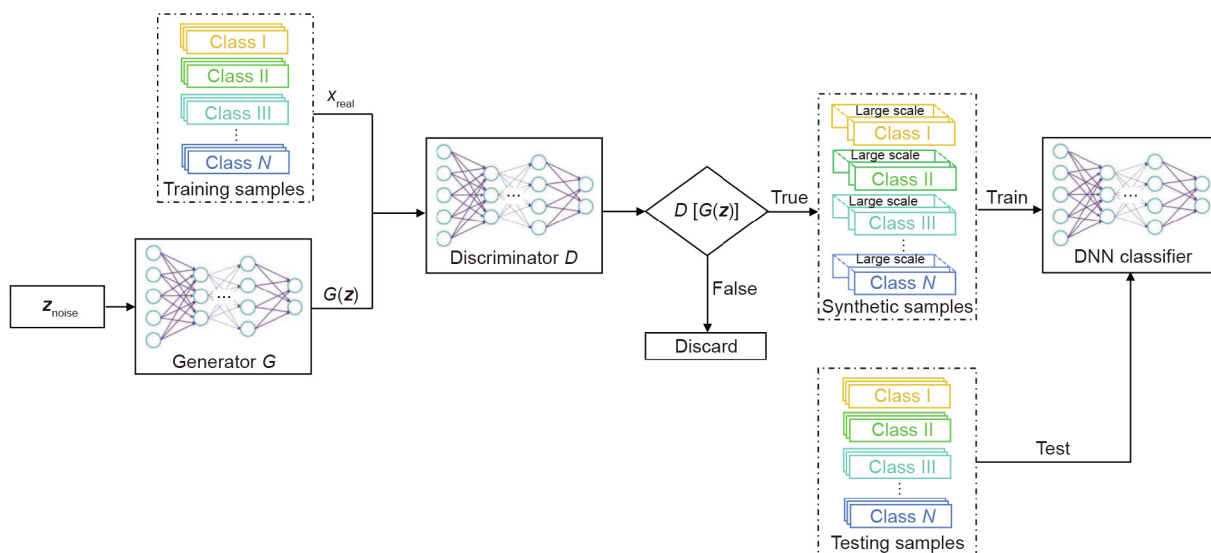


Fig. 1. Workflow of the small sample size multi-classification approach.

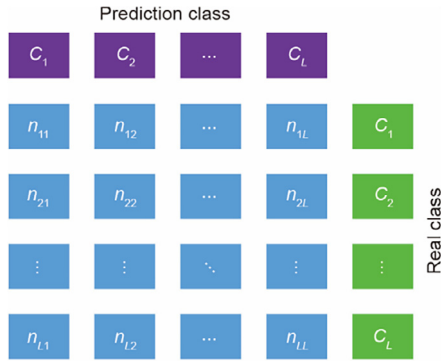


Fig. 2. Diagram of the multi-classification confusion matrix.

$$F\text{-measure} = \frac{2}{L} \frac{\sum_{i=1}^L R_i \sum_{i=1}^L P_i}{\sum_{i=1}^L R_i + \sum_{i=1}^L P_i} \quad (2)$$

$$G\text{-mean} = \left(\prod_{i=1}^L R_i \right)^{1/L} \quad (3)$$

In these equations, L represents the class number; n_{ii} and n_{ij} denote the number of class C_i samples that are correctly predicted as class C_i and incorrectly predicted as class C_j , respectively; and R_i and P_i indicate the recall and the precision of class C_i , respectively, which are defined as follows:

$$R_i = \frac{n_{ii}}{\sum_{j=1}^L n_{ji}} \quad (4)$$

$$P_i = \frac{n_{ii}}{\sum_{j=1}^L n_{ij}} \quad (5)$$

3. Empirical analysis and discussion

Due to the need to protect patient privacy, pathological data are expensive to acquire and the corresponding data annotation is difficult. As a result, pathological studies often encounter the problem of a small sample size. Therefore, the application of data augmentation in the field of pathology is typical. Hepatocellular carcinoma (HCC) is a common malignancy with five-year relative survival

rates of less than 15% [56,57]. The five-year relative survival rates of HCC can be improved effectively by early treatment. Nevertheless, research on the identification of early-stage HCC is limited by the lack of samples with staging information. Glycosylation is one of the most widespread post-translational modifications, and plays crucial roles in various biological processes [58–60]. Numerous cancer-related processes, including oncogenic transformation [59,61], tumor progression [62], and antitumor immunity [63], are associated with the aberrant glycosylation of proteins. Furthermore, various tumor markers are glycoproteins with alterations in serum glycomics [64–67]. Therefore, glycosylation data are an effective means for the prediction of cancer staging. In this section, we discuss the use of WGAN combined with a DNN to identify the stage of HCC, which is significant for the diagnosis and treatment of HCC.

3.1. Data collection

In this study, serum samples donated by Tongji Hospital (Tongji Medical College, Huazhong University of Science and Technology) were used as the experimental data. N -glycans, which are features of data augmentation, were first released from the human serum samples by means of PNGase F prior to a solid-phase permethylation protocol [68]. Next, the mass spectrometry (MS) peak distribution and its relative intensities of permethylated N -glycans (Fig. 3) were detected using 4800 Plus MALDI (AB SCIEX, Concord, Canada). The obtained MS data were processed further using Data Explorer 4.5, and .txt files listing the m/z values and MS intensities were generated (from ASCII Spectrum). The stages of cancer were divided according to the tumor node metastasis (TNM). Through the abovementioned biological processes, 60 HCC cases (TNM stage I, 21 cases; TNM stage II, 24 cases; and TNM stage III, 15 cases) were obtained, each containing 42 features, and 18 healthy samples served as the control group. Each sample was represented as a 42 dimensional feature vector, according to its peak distribution order and relative intensity, as shown in Fig. 3. The HCC cases were divided into a training set (60%) and a test set (40%), as shown in Table 1.

3.2. Result analysis

According to the proposed approach, we first used TNM stage I, TNM stage II, TNM stage III, and the control group's training set to

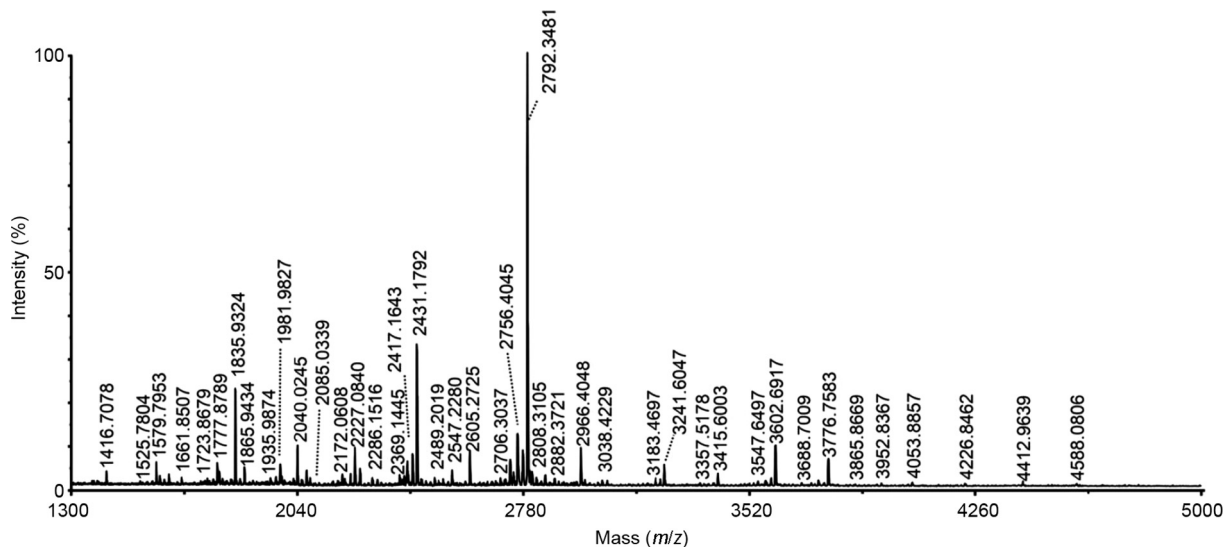


Fig. 3. N -glycan spectra of the HCC sample using MALDI mass spectrometry.

Table 1

Overview of the dataset used in this paper.

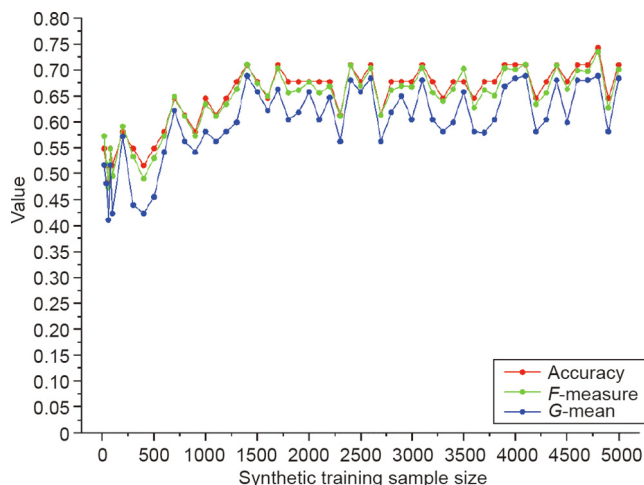
	HCC original samples		Total
	Training set	Test set	
Healthy	11	7	18
TNM stage I	13	8	21
TNM stage II	14	10	24
TNM stage III	9	6	15
Total	47	31	78

train WGAN, and then used the trained WGAN to generate the corresponding synthetic samples. The hyperparameters for the GAN were determined through a series of experiments. The generator had one hidden layer containing 32 rectified linear units (ReLU), and 42 sigmoid units were used as the output layer. The dimension of the noise vector z was set to 15. The discriminator also had one hidden layer, which contained 64 ReLUs; one unit without the activation function was used as the output layer. The hyperparameters of WGAN for each class' training sample were the same. The WGAN's development environment was TensorFlow1.1 and it was trained through the graphics processing unit (GPU). The WGAN training process contained 300 000 iterations. In each iteration of the WGAN training, the discriminator first iterated 100 times, and then the generator iterated one time.

After the generation of the synthetic samples of the HCC cases, these samples were used to train the DNN classifier. The DNN classifier, which was a multi-layer perceptron (MLP), was then validated with the HCC test set. After a series of experiments on the DNN, the hyperparameters were determined. The dimension of the classifier's input was 42, which was equal to the number of features in the HCC samples. The classifier had three hidden layers, each containing 32 ReLUs; the softmax function was used as the output layer and cross-entropy was used as the loss function. TensorFlow1.1 and GPU were used to train the DNN classifier as well, and the number of iterations was set to 3000.

To evaluate the effect of the synthetic training sample size on the performance of the DNN classifier, we used a different number of synthetic samples to train the DNN classifier, and then used real samples to test for the three indicators of accuracy, F -measure, and G -mean. From 20 synthetic training samples of each class, more than 20 synthetic samples were generated each time to train the DNN classifier. In the case of more than 100 synthetic training samples, more than 100 synthetic samples were generated each time to train the DNN classifier. The changes in the accuracy, F -measure, and G -mean are shown in Fig. 4.

With an increase in the synthetic training sample size, the accuracy gradually increased. With 100 synthetic training samples, the accuracy reached 51.61%. When the synthetic training sample size was 1000, the accuracy was greater than 0.6 (64.52%), and when the synthetic training sample size reached 2000, the accuracy was 67.74%. However, after this, increasing the synthetic training sample size did not lead to an improvement in accuracy, and the accuracy continued to fluctuate around 67%. When the synthetic training sample size reached 4000, the accuracy remained stable at approximately 70%. In addition, when the synthetic training sample size was increased, the tendency of the F -measure was basically consistent with that of the accuracy, which showed that the prediction accuracy of each class' real samples was consistent with the overall situation. Furthermore, each stage of HCC could be predicted effectively, and the misdiagnosis rate was very low. With an increase in the number of synthetic samples, the G -mean remained slightly lower than the accuracy but was consistent overall. This finding indicated that the misdiagnosis rate remained low. According to the accuracy, F -measure, and G -mean values, when the synthetic training sample size was 4000, an effective

**Fig. 4.** DNN classifier performance with increase in synthetic training sample size.

DNN classifier for the identification of the HCC stage in a small sample size could be obtained.

In particular, when the number of synthetic samples generated by WGAN was 4000, the test accuracy of the real samples was 70.97% (of the 31 original samples, 22 were predicted correctly), the F -measure was 70.07%, and the G -mean was 68.39%. Table 2 presents the DNN prediction results for all of the stages. All of the healthy samples in the control group were correctly predicted. Of the eight TNM stage I test samples, five samples were correctly predicted, two were predicted to be TNM stage II, and one was predicted to be stage III. Of the ten TNM stage II test samples, seven samples were predicted correctly; two were predicted to be healthy, indicating a risk of misdiagnosis; and one was predicted to be TNM stage III. Of the six TNM stage III real cases, only three were correctly predicted; the rest were predicted to be TNM stage I, for an accuracy of only 50%. This decreased the overall DNN model performance. Therefore, the number of original TNM stage III samples should be increased in a future study, in order to improve the specificity of the DNN model for TNM stage III. TNM stages I and II have similar clinical features and can be classified into a single category called "early-stage cancer." Thus, according to the results of the early-stage HCC (TNM stages I and II) identification, the accuracy of the proposed method reached 77.78%. This level of accuracy has great significance for the early identification and treatment of HCC, because a current study [69] has found that early treatment significantly increases the survival rate of patients with HCC. A recent study by Holzinger et al. [70] indicates that digital pathology will dramatically change medical workflows if pathologists are augmented by machine-learning methods. Thus, our integrated approach with accurate prediction holds potential to promote further research into the pathogenesis of HCC.

Table 2

Confusion matrix of HCC stage identification.

	Predicted				Total
	Healthy	TNM stage I	TNM stage II	TNM stage III	
Real healthy	7	0	0	0	7
Real TNM stage I	0	5	2	1	8
Real TNM stage II	2	0	7	1	10
Real TNM stage III	0	3	0	3	6
Total	9	8	9	5	31

3.3. Evaluation of WGAN combined with a DNN

To test the effectiveness of the proposed approach in classifying the TNM stage of HCC with a limited number of samples, a classical statistical machine-learning classifier and a data oversampling method were used for a comparison. A random forest (RF) is an ensemble learning method with higher accuracy and a better generalization capability than other machine-learning models [71], while a naïve Bayes (NB) classifier has a simple principle and a stable classification performance [72]. These two algorithms were chosen as the representatives of classical statistical machine-learning classifiers. The deep-learning method cannot be applied effectively if there is a limited number of real samples of HCC due to the lack of sufficient data on HCC; however, the classical statistical machine-learning method has a relatively low demand for datasets, so this method can be used. In the classical statistical machine-learning experiments with original samples, the classifiers were trained using the training set and then tested with the test set. In the oversampling experiments, SMOTE was adapted to generate the oversampling samples by using the oversampling HCC training set of all of the stages. The RF, NB, and DNN classifiers were then trained with the oversampling samples. Next, the trained classifiers were validated with the HCC test set. In the proposed framework based on WGAN, a large number of synthetic samples generated by WGAN were used to train the RF, NB, and DNN classifiers. The classifiers were then validated using real samples.

As shown in Table 3, the RF and NB models trained with the training set and tested with the test set resulted in the relatively low accuracies of 54.84% and 32.26%, *F*-measure values of 56.73% and 12.20%, and *G*-mean values of 45.50% and 0, respectively. The NB indexes were considerably lower than those of the RF, which indicated that the NB was more sensitive than the RF in the case of the considered dataset. These results indicate that the misdiagnosis rate was high and that a number of HCC cases were not discriminated. Thus, given a limited training set, classical machine-learning models cannot be effectively trained.

In comparison, the RF and NB models trained with 4000 oversampling samples generated by SMOTE performed better in terms of all the indexes. These results indicate that SMOTE improved the classical machine-learning model performance and reduced the misdiagnosis rate to a certain extent. The DNN model trained with 4000 oversampling samples generated by SMOTE showed significant improvement in all of the indexes: The accuracy increased to 64.52%, the *F*-measure increased to 66.05%, and the *G*-mean increased to 63.32%. These results show that the deep-learning model resulted in better performance than the classical machine-learning model.

The RF and NB models trained with 4000 synthetic samples generated by WGAN exhibited different performance changes. The RF model had worse indexes than the model trained with the oversampling samples, while the NB model's indexes increased considerably. These results imply that with the synthetic samples generated by WGAN, the classical machine-learning models did

not provide good results. In the proposed framework, the deep-learning model DNN with synthetic samples from WGAN performed best among all the considered classifiers. Compared with the oversampling samples generated by SMOTE, this method further increased the accuracy from 64.52% to 70.97%, the *F*-measure value from 66.05% to 70.07%, and the *G*-mean value from 63.32% to 68.39%, which demonstrates that WGAN with a DNN effectively solved the HCC stage recognition problem. Above all, these findings indicate that the deep-learning method can successfully be applied to a multi-classification problem with a limited number of samples.

3.4. Discussion

According to the experimental results given above, WGAN combined with a DNN can be applied to the identification of HCC stages, and results in excellent performance compared with traditional methods. This finding is of significance to cancer research. Research into most cancers is hindered by the small sample size problem; samples with accurate staging information are particularly rare. This problem has led to slow progress in the early diagnosis and treatment of cancers; furthermore, it affects the exploration of the pathogenesis of cancer. Our data augmentation method based on WGAN may well provide a solution for these issues. The proposed approach was designed not only to solve the problem of HCC staging, but also to solve the small sample size problem using supervised learning. Therefore, cancer-staging data based on serum samples were selected, as such data result in unsatisfactory performance with traditional statistical machine learning due to the small sample size problem. The proposed framework does not rely on a precise domain knowledge of cancer, due to the characteristics of deep learning. Therefore, the proposed method has a low barrier to successful application in other biological research domains, and may even be applied in more far-ranging fields once the performance has been optimized. Furthermore, the combination of WGAN with a DNN holds enormous potential for bringing domains that lack samples into the intelligence era [26,73–75].

4. Conclusion

In this paper, a WGAN approach combined with a DNN was presented for cancer stage identification on the basis of a small sample size. Using the indicators of precision, *F*-measure, and *G*-mean, we demonstrated that in comparison with classical machine-learning methods and oversampling methods, the proposed approach substantially improved the classification effectiveness with an increase in the number of synthetic samples. Early recognition of cancer is particularly significant for the diagnosis and treatment of cancer. Because the feature selection did not rely on precise domain knowledge from experts, the proposed supervised deep-learning approach based on a small sample size holds potential to provide effective solutions to other problems involving a small sample size in various fields. Thus, this approach could easily be used to promote intelligent development in other fields. This new approach strongly promotes the new stage of artificial intelligence. In future, the proposed approach will be applied to more datasets from various fields in order to continuously improve our research [76,78].

Acknowledgements

This work was supported by the National Natural Science Foundation of China (91646102, L1724034, L16240452, L1524015, and 20905027), the MOE (Ministry of Education in China) Project of

Table 3
Performance of different classification strategies.

	Accuracy	<i>F</i> -measure	<i>G</i> -mean
RF	0.5484	0.5673	0.4550
NB	0.3226	0.1220	0
RF with SMOTE	0.5806	0.6254	0.4811
NB with SMOTE	0.5484	0.5600	0.5233
DNN with SMOTE	0.6452	0.6605	0.6332
RF with WGAN	0.2903	0.2931	0.2627
NB with WGAN	0.6129	0.6260	0.6043
DNN with WGAN	0.7097	0.7007	0.6839

Humanities and Social Sciences (16JDGC011), the Chinese Academy of Engineering's China Knowledge Center for Engineering Sciences and Technology Project (CKCEST-2018-1-13), the UK-China Industry Academia Partnership Programme (UK-CIAPP \260), Volvo-Supported Green Economy and Sustainable Development at Tsinghua University (20153000181), and the Tsinghua Initiative Research Project (2016THZW).

Compliance with ethics guidelines

Yufei Liu, Yuan Zhou, Xin Liu, Fang Dong, Chang Wang, and Zihong Wang declare that they have no conflict of interest or financial conflicts to disclose.

References

- [1] Crevier D. AI: the tumultuous history of the search for artificial intelligence. New York: Basic Books, Inc.; 1993.
- [2] Pan Y. Heading toward Artificial Intelligence 2.0. *Engineering* 2016;2(4):409–13.
- [3] State Council of the People's Republic of China. Development Plan for a Next-Generation Artificial Intelligence [Internet]. Beijing: www.gov.cn. [cited 2018 Mar 5]. Available from: http://english.gov.cn/policies/latest_releases/2017/07/20/content_281475742458322.htm.
- [4] State Council Information Office of the People's Republic of China. The policy interpretation of Development Planning for a Next-Generation Artificial Intelligence [Internet]. Beijing: www.scio.gov.cn. [cited 2018 Mar 5]. Available from: <http://www.scio.gov.cn/34473/34515/Document/1559231/1559231.htm>. Chinese.
- [5] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science* 2006;313(5786):504–7.
- [6] Zhuang Y, Chen C, Pan Y. Challenges and opportunities: from big data to knowledge in AI 2.0. *Front Inf Technol Electronic Eng* 2017;18(1):3–14.
- [7] Al-Qizwini M, Barjasteh I, Al-Qassab H, Radha H. Deep learning algorithm for autonomous driving using GoogLeNet. In: *Proceedings of the 2017 IEEE Intelligent Vehicles Symposium*; 2017 Jun 11–14; Los Angeles, CA, USA. New York: IEEE; 2017. p. 89–96.
- [8] Wang L, Sng D. Deep learning algorithms with applications to video analytics for a smart city: a survey. 2016: arXiv:1511.06434.
- [9] Mohamed A, Dahl G, Hinton G. Acoustic modeling using deep belief networks. *IEEE Trans Audio Speech Lang Process* 2012;20(1):14–22.
- [10] Jones N. Artificial-intelligence institute launches free science search engine [Internet]. Heidelberg: Springer Nature. c2018 [cited 2018 Mar 5]. Available from: <https://www.nature.com/news/artificial-intelligence-institute-launches-free-science-search-engine-1.18703>.
- [11] Goodfellow I, Bengio Y, Courville A. *Deep learning*. Cambridge: The MIT Press; 2016.
- [12] Zhuang F, Luo P, Qing H, Shi Z. Survey on transfer learning research. *J Software* 2015;26:26–39. Chinese.
- [13] Chen J, Yang J, Zhou H, Xiang H, Zhu Z, Li Y, et al. CPS modeling of CNC machine tool work processes using an instruction-domain based approach. *Engineering* 2015;1(2):247–60.
- [14] Urban F, Zhou Y, Nordensvard J, Narain A. Firm-level technology transfer and technology cooperation for wind energy between Europe, China and India: from north–south to south–north cooperation? *Energy Sustainable Dev* 2015;28:29–40.
- [15] Zhou Y, Zhang H, Ding M, Su J. How public demonstration project affects the emergence of a new industry: an empirical study on electric vehicle demonstration project in China. In: *Proceedings of the 2013 Suzhou Silicon Valley–Beijing International Innovation Conference*; 2013 Jul 8–9. New York: IEEE; 2013. p. 234–9.
- [16] Zhou Y, Minshall T. Building global products and competing in innovation: the role of Chinese university spin-outs and required innovation capabilities. *Int J Technol Manage* 2014;64(2):180–209.
- [17] Xu G, Wu Y, Minshall T, Zhou Y. Exploring innovation ecosystems across science, technology, and business: a case of 3D printing in China. *Technol Forecast Social Change* 2017;136:180–221.
- [18] Li X, Zhou Y, Xue L, Huang L. Roadmapping for industrial emergence and innovation gaps to catch-up: a patent analysis of OLED industry in China. *Int J Technol Manage* 2016;7(1–3):105–43.
- [19] Li X, Zhou Y, Xue L, Huang L. Integrating bibliometrics and roadmapping methods: a case of dye-sensitized solar cell technology-based industry in China. *Technol Forecast Social Change* 2015;97:205–22.
- [20] Zhou Y, Pan M, Urban F. Comparing the international knowledge flow of China's wind and solar photovoltaic (PV) industries: patent analysis and implications for sustainable development. *Sustainability* 2018;10(6):1883.
- [21] Theodoridis S, Koutroumbas K. *Pattern recognition*. 3rd ed. Orlando: Academic Press; 2006.
- [22] Nordensvard J, Zhou Y, Zhang X. Innovation core, innovation semi-periphery and technology transfer: the case of wind energy patents. *Energy Policy* 2018;120:213–27.
- [23] Pan M, Zhou Y, Zhou DK. Comparing the innovation strategies of Chinese and European wind turbine firms through a patent lens. *Environ Innovation Societal Transitions*. Epub 2017 Dec 27.
- [24] Zhou Y, Pan M, Zhou DK, Xue L. Stakeholder risk and trust perceptions in the diffusion of green manufacturing technologies: evidence from China. *J Environ Dev* 2017;27(1):46–73.
- [25] Zhou Y, Li X, Lema R, Urban F. Comparing the knowledge bases of wind turbine firms in Asia and Europe: patent trajectories, networks, and globalisation. *Sci Public Policy* 2016;43(2):476–91.
- [26] Chen L, Xu J, Zhou Y. Regulating the environmental behavior of manufacturing SMEs: interfirm alliance as a facilitator. *J Cleaner Prod* 2017;165:393–404.
- [27] DeRouin E, Brown J. Neural network training on unequally represented classes. New York: ASME Press; 1991.
- [28] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 2002;16(1):321–57.
- [29] Han H, Wang WY, Mao BH. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: Huang DS, Zhang XP, Huang GB, editors. *Advances in intelligent computing*. Berlin: Springer; 2005. p. 878–87.
- [30] He H, Bai Y, Garcia EA, Shuao L. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: *Proceedings of the 2008 IEEE International Joint Conference on Neural Networks*; 2008 Jun 1–8; Hong Kong, China. New York: IEEE; 2008. p. 1322–8.
- [31] Barua S, Islam MM, Yao X, Kazuyuki M. MWMOTE-majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Trans Knowl Data Eng* 2014;26(2):405–25.
- [32] Xie Z, Jiang L, Ye T, Li X. A synthetic minority oversampling method based on local densities in low-dimensional space for imbalanced learning. In: Renz M, Shahabi C, Zhou X, Cheema M, editors. *Database systems for advanced applications*. Cham: Springer; 2015. p. 3–18.
- [33] Douzas G, Bacao F. Self-Organizing Map Oversampling (SOMO) for imbalanced data set learning. *Expert Syst Appl* 2017;82:40–52.
- [34] Bishop CM. Training with noise is equivalent to Tikhonov regularization. *Neural Comput* 1995;7(1):108–16.
- [35] Zhou Z, Jiang Y. Nc4.5: neural ensemble based C4.5. *IEEE Trans Knowl Data Eng* 2004;16(6):770–3.
- [36] Li DC, Lin YS. Using virtual sample generation to build up management knowledge in the early manufacturing stages. *Eur J Operat Res* 2006;175(1):413–34.
- [37] Li D, Fang Y. A non-linearly virtual sample generation technique using group discovery and parametric equations of hypersphere. *Exp Syst Appl* 2009;36(1):844–51.
- [38] Wang K, Gou C, Duan Y, Lin Y, Zheng X, Wang F. Generative adversarial networks: introduction and outlook. *IEEE/CAA J Autom Sin* 2017;4(4):588–98.
- [39] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KG, editors. *Advances in neural information processing systems*. La Jolla: Neural Information Processing Systems Foundation, Inc.; 2014. p. 2672–80.
- [40] Creswell A, White T, Dumoulin V, Arulkumaran K, Sengupta B, Bharath AA. Generative adversarial networks: an overview. *IEEE Signal Process Mag* 2018;35(1):53–65.
- [41] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. 2015:arXiv:1512.03131.
- [42] Santana E, Hotz G. Learning a driving simulator. 2016:arXiv:1608.01230.
- [43] Gou C, Wu Y, Wang K, Wang F, Ji Q. Learning-by-synthesis for accurate eye detection. In: *Proceedings of the 23rd International Conference on Pattern Recognition*; 2016 Dec 4–8; Cancun, Mexico. New York: IEEE; 2017. p. 3362–7.
- [44] Li J, Monroe W, Shi T, Jean S, Ritter A, Jurafsky D. Adversarial learning for neural dialogue generation. 2017:arXiv:1701.06547.
- [45] Pascual S, Bonafonte A, Serrà J. SEGAN: speech enhancement generative adversarial network. 2017:arXiv:1703.09452.
- [46] Fiore U, Santis AD, Perla F, Zanetti P, Palmieri F. Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Inf Sci* 2019;479:448–55.
- [47] Douzas G, Bacao F. Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Syst Appl* 2017;91:464–71.
- [48] Bloice MD, Stocker C, Holzinger A. Augmentor: an image augmentation library for machine learning. 2017:arXiv:1708.04680.
- [49] Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. 2017:arXiv:1701.07875.
- [50] Ratliff LJ, Burden SA, Sastry SS. Characterization and computation of local Nash equilibria in continuous games. In: *Proceedings of the 51st Annual Allerton Conference on Communication, Control, and Computing*; 2013 Oct 2–4; Monticello, IL, USA. New York: IEEE; 2006. p. 917–24.
- [51] Danihelka I, Lakshminarayanan B, Uribe B, Wierstra D, Dayan P. Comparison of maximum likelihood and GAN-based training of real NVPs. 2017:arXiv:1705.05263.
- [52] Yang Q, Yan P, Zhang Y, Yu H, Shi Y, Mou X, et al. Low dose CT Image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. *IEEE Trans Med Imaging* 2017;37(6):1348–57.
- [53] McDaniel P, Papernot N, Celik ZB. Machine learning in adversarial settings. *IEEE Secur Privacy* 2016;14(3):68–72.
- [54] Sousa LR, Miranda T, Sousa RL, Tinoco J. The use of data mining techniques in rockburst risk assessment. *Engineering* 2017;3(4):552–8.
- [55] Sun Y, Kamel MS, Wong AKC, Wang Y. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognit* 2007;40(12):3358–78.

- [56] Farazi AP, DePinho RA. Hepatocellular carcinoma pathogenesis: from genes to environment. *Nat Rev Cancer* 2006;6(9):674–87.
- [57] Arzumanyan A, Reis HMG, Feitelson MA. Pathogenic mechanisms in HBV- and HCV-associated hepatocellular carcinoma. *Nat Rev Cancer* 2013;13(2):123–35.
- [58] Mechref Y, Hu Y, Garcia A, Zhou S, Desantos-Garcia JL, Hussein A. Defining putative glycan cancer biomarkers by MS. *Bioanalysis* 2012;4(20):2457–69.
- [59] Tang Z, Varghese RS, Bekesova S, Loffredo CA, Hamid MA, Kyselova Z, et al. Identification of N-glycan serum markers associated with hepatocellular carcinoma from mass spectrometry data. *J Proteome Res* 2010;9(1):104–12.
- [60] Kronewitter SR, De Leoz MLA, Strum JS, An HJ, Dimapasoc LM, Guerrero A, et al. The glycozyler: automated glycan annotation software for high performance mass spectrometry and its application to ovarian cancer glycan biomarker discovery. *Proteomics* 2012;12(15–16):2523–38.
- [61] Pierce M, Buckhaults P, Chen L, Fregien N. Regulation of N-acetylglucosaminyltransferase V and Asn-linked oligosaccharide $\beta(1,6)$ branching by a growth factor signaling pathway and effects on cell adhesion and metastatic potential. *Glycoconjugate J* 1997;14(5):623–30.
- [62] Lau KS, Dennis JW. N-Glycans in cancer progression. *Glycobiology* 2008;18(10):750–60.
- [63] Saldova R, Royle L, Radcliffe CM, Abd Hamid UM, Evans R, Arnold JN, et al. Ovarian cancer is associated with changes in glycosylation in both acute-phase proteins and IgG. *Glycobiology* 2007;17(12):1344–56.
- [64] Noda K, Miyoshi E, Gu J, Gao CX, Nakahara S, Kitada T, et al. Relationship between elevated FX expression and increased production of GDP-L-fucose, a common donor substrate for fucosylation in human hepatocellular carcinoma and hepatoma cell lines. *Cancer Res* 2003;63(19):6282–9.
- [65] Basu PS, Majhi R, Batabyal SK. Lectin and serum-PSA interaction as a screening test for prostate cancer. *Clin Biochem* 2003;36(5):373–6.
- [66] Arnold JN, Saldova R, Hamid UMA, Rudd PM. Evaluation of the serum N-linked glycome for the diagnosis of cancer and chronic inflammation. *Proteomics* 2008;8(16):3284–93.
- [67] Adamczyk B, Tharmalingam T, Rudd PM. Glycans as cancer biomarkers. *Biochim Biophys Acta Gen Subj* 2012;1820(9):1347–53.
- [68] Deguchi K, Keira T, Yamada K, Ito H, Takegawa Y, Nakagawa H, et al. Two-dimensional hydrophilic interaction chromatography coupling anion-exchange and hydrophilic interaction columns for separation of 2-pyridylamino derivatives of neutral and sialylated N-glycans. *J Chromatography A* 2008;1189(1–2):169–74.
- [69] Siemerink E, Mulder NH, Brouwers AH, Hospers GA. Early prediction of response to sorafenib treatment in patients with hepatocellular carcinoma (HCC) with ^{18}F -fluorodeoxyglucose-positron emission tomography (^{18}F -FDG-PET). *J Clin Oncol* 2008;26(21):1–15.
- [70] Holzinger A, Malle B, Kieseberg P, Roth PM, Müller H, Reihs R. Machine learning and knowledge extraction in digital pathology needs an integrative approach lecture notes in computer science. In: Holzinger A, Goebel R, Ferri M, Palade V, editors. *Towards integrative machine learning and knowledge extraction*. Cham: Springer; 2017.
- [71] Breiman L. Random forests. *Mach Learn* 2001;45(1):5–32.
- [72] Mitchell T. *Machine learning*. Zeng H, Zhang Y, translator. Beijing: China Machine Press; 2003. Chinese.
- [73] Liu P, Zhou Y, Zhou DK, Xue L. Energy Performance Contract models for the diffusion of green-manufacturing technologies in China: a stakeholder analysis from SMEs' perspective. *Energy Policy* 2017;106:59–67.
- [74] Kong D, Feng Q, Zhou Y, Xue L. Local implementation for green-manufacturing technology diffusion policy in China: from the user firms' perspectives. *J Cleaner Prod* 2016;129:113–24.
- [75] Zhou Y, Xu G, Minshall T, Liu P. How do public demonstration projects promote green-manufacturing technologies? A case study from China. *Sustainable Dev* 2015;23(4):217–31.
- [76] Kong D, Zhou Y, Liu Y, Xue L. Using the data mining method to assess the innovation gap: a case of industrial robotics in a catching-up country. *Technol Forecasting Social Change* 2017;119:80–97.
- [77] Li M, Zhou Y. Visualizing the knowledge profile on self-powered technology. *Nano Energy* 2018;51:250–9.
- [78] Wang B, Liu Y, Zhou Y, Wen Z. Emerging nanogenerator technology in China: a review and forecast using integrating bibliometrics, patent analysis and technology roadmap methods. *Nano Energy* 2018;46:322–30.