Original Article

Combining autoencoder with clustering analysis for anomaly detection in radiotherapy plans

Peng Huang[#]**, Hui Yan**[#]**, Zhiyue Song, Yingjie Xu, Zhihui Hu, Jianrong Dai**

Department of Radiation Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

Contributions: (I) Conception and design: P Huang, H Yan; (II) Administrative support: J Dai; (III) Provision of study materials or patients: Y Xu; (IV) Collection and assembly of data: Z Hu; (V) Data analysis and interpretation: P Huang, Z Song; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work and should be considered as co-first authors.

Correspondence to: Jianrong Dai. Department of Radiation Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China. Email: dai_jianrong@cicams.ac.cn.

Background: To develop an unsupervised anomaly detection method to identify suspicious error-prone treatment plans in radiotherapy.

Methods: A total of 577 treatment plans of breast cancer patients were used in this study. They were labeled as either normal or abnormal plans by experienced clinicians. Multiple features of each plan were extracted and selected by the learning algorithms. The training set consisted of feature samples from 400 normal plans and the testing set consisted of feature samples from 158 normal plans and 19 abnormal plans. Using the k-means clustering algorithm in the training stage, 4 normal plan clusters were formed. The distance between the samples in the testing set and the cluster centers were then determined. To evaluate the effect of dimensionality reduction (DR) on detection accuracy, principal component analysis (PCA) and autoencoder (AE) methods were compared.

Results: The sensitivity of the anomaly detection model based on PCA and AE methods were 84.2% (16/19) and 94.7% (18/19), respectively. The specificity of the anomaly detection model based on PCA and AE methods were 64.6% (102/158) and 69.0% (109/158), respectively. The areas under the receiver operating characteristic (ROC) curve (AUCs) based on PCA and AE methods were 0.81 and 0.90, respectively.

Conclusions: The unsupervised learning method was effective for detecting anomalies from the feature samples. Accuracy could be improved with the introduction of AE-based DR technique. The combination of AE and k-means clustering methods provides an automated way to identify abnormal plans among clinical treatment plans in radiotherapy.

Keywords: Anomaly detection; treatment plan; autoencoder (AE); clustering analysis

Submitted Aug 04, 2022. Accepted for publication Jan 29, 2023. Published online Feb 24, 2023.
doi: 10.21037/qims-22-825
View this article at: https://dx.doi.org/10.21037/qims-22-825

Introduction

Alongside recent technical advances, radiotherapy has evolved into a complex process which requires high precision (1,2). Radiotherapy can achieve treatment goals by delivering a high dose to the target volume while protecting the surrounding normal tissues. However, any small deviation could lead to higher costs and risks to patient health. Therefore, extreme caution and enhanced quality control are necessary during radiotherapy planning. Nevertheless, incidents that compromise treatment output in radiotherapy have continued to occur worldwide. In the worst case, patients may die due to unintentional delivery of an extremely high dose. An example is the Panama incident, where 16 patients were severely overexposed (approximately twice the prescribed dose) in late 2000 and early 2001, resulting in 8 treatment-related deaths (3).

Although physics plan and chart check are the most effective quality control methods for reducing human error in radiotherapy (2,4,5), they rely heavily on manual reviewing, which is inefficient and can lack accuracy (6). A thorough plan check requires a review of the diagnosis, prescription, planning, and physician approval, in addition to approximately 20 field-specific parameters (5,7). Each of these parameters could affect the treatment outcome and patient safety. For manual reviewing, each reviewer could take an hour to thoroughly check hundreds of parameters on different charts; even an experienced human reviewer could unintentionally miss an error. A study by Gopan *et al.* showed that only 38% of errors that were potentially detectable on treatment plans were identified in the review procedure (8). Therefore, automated plan checking tools are important to assist current manual reviewing processes.

Automated plan checks have been widely adopted in clinical practice to partially or completely replace the manual reviewing process (4-7,9-11). Early research focused on developing a system which compared data in the planning system against data in the recording and verification system before treatment delivery. Lack *et al.* developed a software tool to detect two categories of errors in data transferring and planning using root cause analysis (4). Furhang *et al.* developed software to automatically carry out intraplan and interplan reviews (7). Yang and Moore conducted a study to use dynamic scripting to automatically verify treatment plan integrity (6,11). Dewhurst developed a semiautomated system named AutoLock for radiotherapy treatment plan quality control (QC). This system integrates automated treatment plan QC, an electronic checklist, and plan finalization (9). Siochi developed an electronic quality assurance (QA) software that can read data and compare it against parameters in the recording and verification system's database. Compared with the original paperless system, error detection was improved significantly (10).

Azmandian *et al.* (1) first used machine learning (ML) to help detect human errors in radiotherapy treatment plan checking in 2007. They used 1,000 plans to build the clusters, and another 650 plans were used to test the proposed method. A total of 8 distinct features were extracted from each patient's plan. This preliminary work has been promising for automatic plan checking, but more efforts are needed to meet strict clinical requirements. Deep learning-based anomaly detection methods have been successfully applied in industries such as traffic control and bank fraud detection, among others (12-18). With proper modification, these models could be used in radiotherapy plan QA and checking. Recently, ML for patient-specific QA has been increasingly investigated in patient-specific QA. Wang *et al.* developed a novel multitask autoencoder (AE)-based classification-regression model for patient-specific volumetric modulated arc therapy (VMAT) QA (19). The model was compared with other popular models and showed significant improvement in prediction accuracy. For plan checking, there are limited ML applications due to the low number of available data sets.

In this study, 577 hybrid intensity-modulated radiotherapy treatment (IMRT) plans for patients undergoing breast cancer radiotherapy at our institute were collected, and 35 features were extracted. Dimensionality reduction (DR) techniques were introduced to enhance model learning. The k-means clustering algorithm was used to partition samples into clusters in which normal plans remained together and outliers were isolated. The rest of this paper is organized as follows. In Methods, patient data, workflow, DR algorithms, and the unsupervised ML model are explained. In Results, the performance of both AE and principal component analysis (PCA) methods are compared and analyzed quantitatively. In Discussion, the advantages and disadvantages of the proposed anomaly detection model are discussed. We present the following article in accordance with the Standards for Reporting Diagnostic accuracy studies (STARD) reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-22-825/rc).

Methods

The workflow of the proposed anomaly detection method is shown in *Figure 1*. There were four steps: (I) data preprocessing; (II) DR; (III) model learning; and (IV) anomaly detection. All programs were developed on MATLAB_R2018b (MathWorks, Natick, MA, USA). The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The Institutional Ethics Committee of the Cancer Hospital, Chinese Academy of Medical Sciences, and Peking Union Medical College

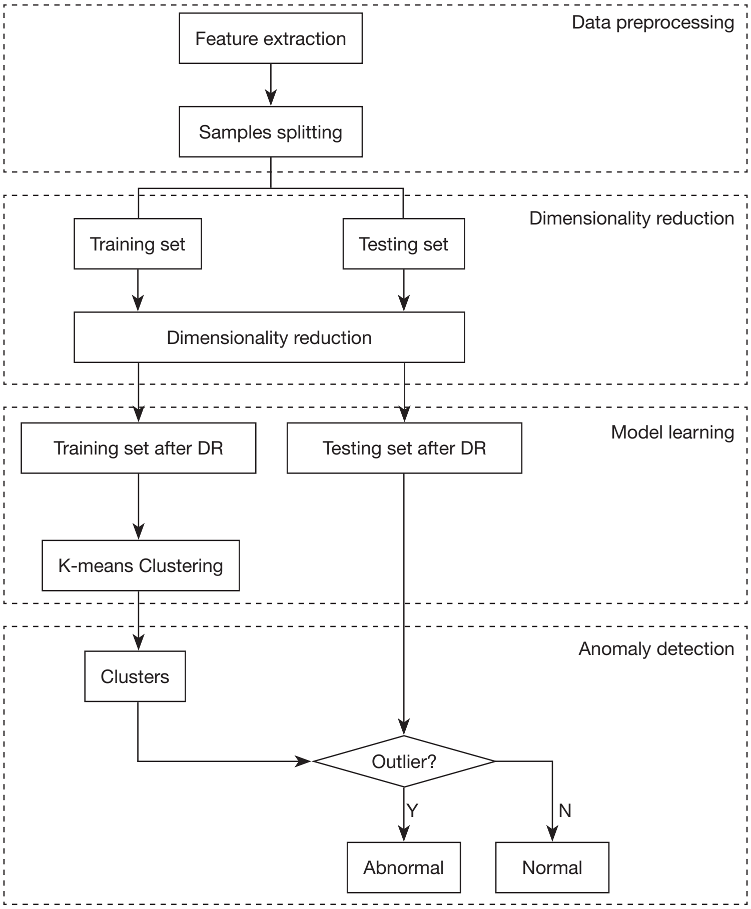

Figure 1 Workflow of the proposed anomaly detection process for plan checking in radiotherapy. DR, dimensionality reduction.

approved this study. Informed consent was waived due to the nature of the retrospective study.

Data preprocessing

Treatment plans for 577 breast cancer patients treated at our hospital between 2010 and 2020 were extracted from MOSAIQ (Elekta, Stockholm, Sweden). The treatment plans were labelled as either “normal” (558/577) or “abnormal” (19/577) by 3 experienced medical physicists. All plans were deliverable and clinically approved for treatment. Normal plans were those which fully matched our treatment planning protocol. Abnormal plans were those which met the goal of radiotherapy treatment but had low quality or were less rational in some respects. For example, beam orientations may not have been optimized to deliver high doses to organs at risk, or additional radiation fields or higher monitor units (MUs) were used, extending the treatment time significantly. These abnormal plans were not erroneous but were considered to have low quality in some aspects. In routine practice, treatment plans are first checked by a rule-based computer program and the error plans are identified when certain thresholds are not meet or rules are not followed. These clinically feasible plans are then manually reviewed by an experienced clinical physicist to check the quality and rationality. If the values of certain plan parameters deviate from their normal distribution, a warning is issued to the planners to request they verify their plan parameters and make the necessary modifications.

In this study, all treatment plans were hybrid plans consisting of 2 conventional tangent fields and 2 IMRT fields (20). To characterize the plans, the most important 35 features were extracted from each plan (*Table 1*). The segment in tangent fields was always 1 and was considered to have no impact on anomaly detection. The number of segments of IMRT field was variable and each plan had 2

Table 1 Summary of the original features obtained from IMRT plans

Feature	Description	Fields	Number of features	Type	Unit
Segment	Number of segments of the field	IMRT	2	Integer	number
SSD	Source to skin distance	IMRT/tangent	4	Float	cm
Coll_{x1}	Collimator position in x1 direction	IMRT/tangent	4	Float	cm
Coll_{x2}	Collimator position in x2 direction	IMRT/tangent	4	Float	cm
Coll_{y1}	Collimator position in y1 direction	IMRT/tangent	4	Float	cm
Coll_{y2}	Collimator position in y2 direction	IMRT/tangent	4	Float	cm
C_θ	Angle of collimator	IMRT/tangent	4	Integer	degree
G_θ	Angle of gantry	IMRT/tangent	4	Integer	degree
Meter set	MU per field	IMRT/tangent	4	Float	MU
Dose	Dose per fraction	IMRT/tangent	1	Float	cGy

IMRT, intensity-modulated radiation therapy; SSD, source to skin distance; MU, monitor unit.

IMRT fields. The dose was 290 or 200 MU for an IMRT field or tangent field, respectively. For the other features, including collimator position, collimator angle, gantry angle, and MU per field, there were 4 features for the 4 fields of each plan. To provide a common scale for the variables, all feature values were normalized using z-score normalization (21). About 70% (400/558) of the normal plans were randomly selected and used for generating the training samples, and the remaining 30% (158/558) of the normal plans and the 19 abnormal plans were used for generating the testing samples.

DR

To improve the efficiency and robustness of model learning, these 35 features were further reduced to fewer features. For this purpose, DR methods were needed. The standard method is PCA, which represents high-dimensional data by fewer principal components without losing major information (22,23). PCA aims to find a linear mapping from high-dimension space to lower dimension space. However, in practice, high-dimensional data often contain highly nonlinear structures which violate the basic assumption of the linear DR models.

Recently, AE was successfully used in a deep-learning network for the feature extraction layers. AE can compress high-dimensional data with nonlinear structures (12-18). As shown in *Figure 2*, the AE model mainly consisted of 3 components: the encoder, a bottleneck layer, and the decoder (12-14). The encoder learned the mapping to transfer the input data into latent representation (or latent variables), which was the bottleneck layer. The decoder then learned the mapping to reconstruct the input data from the latent representation in the bottleneck layer. The mean square error (MSE) between input data and reconstructed data was used to train the AE model. Through this training process, the latent representation of the input data could be achieved and used to represent the input data with fewer dimensions.

The network structure of the AE model is shown in *Figure 2* and was constructed using the trainAutoencoder toolbox provided in MATLAB (https://ww2.mathworks.cn/help/deeplearning/ref/trainautoencoder.html). The network consisted of 7 layers of neurons, including 1 input layer, 1 output layer, 2 encoding layers, 2 decoding layers, and 1 bottleneck layer. The number of neurons in the 7 layers was 35, 24, 16, 9, 16, 24, and 35, respectively (24). The structure could also be predetermined by neuroevolution methods for the optimal hyperparameters (25-27). The transfer function used for the second decoding layer was pure linear transfer function (purelin). The transfer function for the other encoding and decoding layers was saturating linear transfer function (satlin). Early stopping was used to find a sufficient number of training epochs and prevent overtraining the model. The main training steps of the AE model were as follows:

(I) Load and scale training dataset for network;

(II) Define the number of features and the encoder dimensions;

(III) Build up the encoding and decoding layers to form

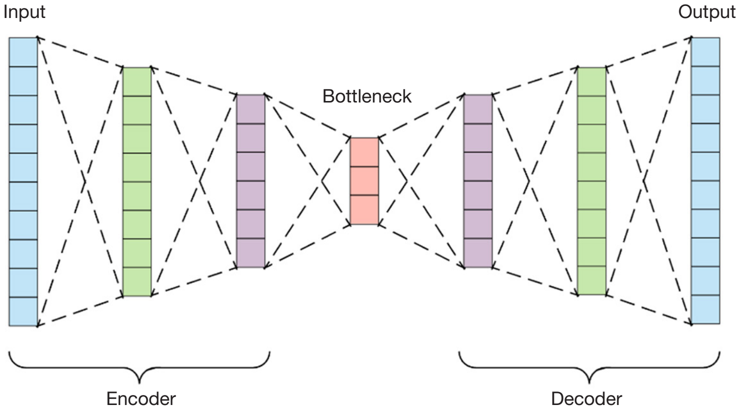

Figure 2 The structure of the AE model for dimensionality reduction. AE, autoencoder.

the entire model;

(IV) Predict the new training data and calculate the cost function;

(V) Adjust the network weights to minimize the cost function;

(VI) Repeat steps (IV) and (V) until the training algorithm converges.

Unsupervised learning

Unsupervised learning is an important branch of ML. Users do not need to label samples and teach the model how to learn the mapping relationship. Instead, it allows the model to work on its own to discover patterns and information from a collection of unlabeled data. Clustering is the most important unsupervised learning method for partitioning data into groups (or clusters). As a result, samples in the same cluster are similar to each other and dissimilar to samples in other clusters. Among the different types of clustering algorithms, k-means clustering is popular and easy to implement (28). The main steps of the k-means clustering algorithm are as follows:

(I) Choose k initial cluster centers;

(II) Estimate the center of each cluster by calculating the mean and the standard deviation (σ) of the data points within each cluster. The mean of each cluster is used as the cluster's center. The standard deviation of each cluster is used as the cluster's boundary;

(III) Assign each data point to its closest cluster based on its Euclidean distance to the cluster's center;

(IV) Repeat steps (II) and (III) until the assignments do not change.

It should be noted that the number k, the initial positions of cluster centroids (or seeds), and the distance metric must be predetermined. Improper setting of these parameters will cause the algorithm to fall into the local optimum (3). In this study, the parameters and methods were investigated and examined by 4 evaluation criteria: silhouette coefficient (Sil), Davies-Bouldin index (DB), Calinski-Harabasz index (CH), and Dunn validity index (DVI) (29-34). Details of candidate distance metrics and centroid initialization methods are described in Appendix 1.

Anomaly detection

With the clusters established based on the training samples by the k-means clustering algorithm, the center and boundary of each cluster were formed and used for anomaly detection. In this study, the standard deviation of each cluster established by the training set was defined as the boundary of each cluster and was used to select the preset threshold. The testing samples were judged to be normal or abnormal plans based on their Euclidean distance to the center of the closest cluster. If this distance was greater than a preset threshold, the plan was assumed to be abnormal; otherwise, it was normal. An optimal threshold was selected to assure a low false positive rate (FPR; i.e., identifying a point as an anomaly when in fact it is normal) and high true positive rate (TPR; i.e., identifying a point as an anomaly when it is, in fact, an anomaly).

Evaluation

The results of clustering based on the samples generated by the 2 DR algorithms were compared using the silhouette

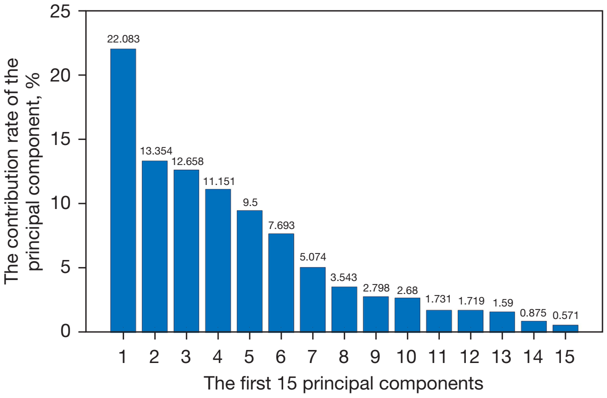

Figure 3 The contribution of the top 15 principal components in the PCA method. PCA, principal component analysis.

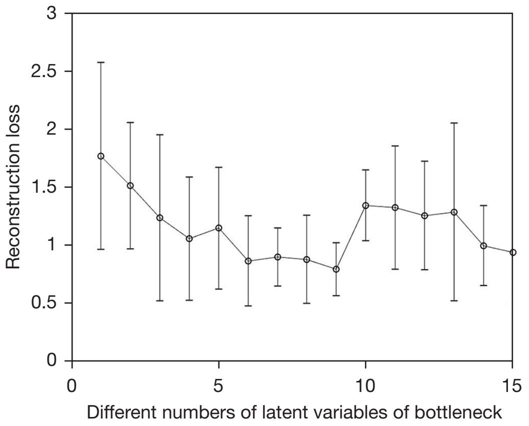

Figure 4 The reconstruction loss of the top 15 latent variables in the AE method. AE, autoencoder.

coefficient, which is calculated as below:

$$S = \frac{\left(b(i) - a(i)\right)}{\max\left(a(i), b(i)\right)} \quad [1]$$

Where *a* and *b* represent the mean distance between sample *i* and all other samples in the same cluster and in the other clusters, respectively. *S* represents the silhouette coefficient for 1 sample. The silhouette coefficient used here was the coefficient for a whole sample set, which was the mean of the coefficient for all samples. Silhouette coefficient measures how well a sample is matched to its identified cluster compared to other clusters. The larger the silhouette coefficient is, the more correct the cluster distribution is (35).

The effect of the 2 DR algorithms, AE and PCA, on the accuracy of anomaly detection was evaluated by 10 repeat tests (36,37). In each test, the samples were shuffled and split into a training set and testing set. The training set was then used to generate clusters of samples compressed by the AE method and PCA method. Performance of anomaly detection was evaluated over the testing set based on the area under the receiver operating characteristic curve (AUC) (38). In addition, model performance was also quantified in terms of sensitivity/true posititve rate (TPR) = TP/(TP + FN), specificity/true negative rate (TNR) = TN/(TN + FP), false posititve rate (FPR) = FP/(FP + TN), and false negative rate (FNR) = FN/(FN + TP), where TP is true positive, TN is true negative, FP is false positive, FN is false negative, and R is rate. In clinical practice, the FPR indicates the ratio of normal plans being wrongly detected as an abnormal plan, whereas FNR indicates the ratio of abnormal plans being wrongly detected as a normal plan.

In addition to the k-means clustering algorithm, 2 other popular clustering algorithms, fuzzy c-means (FCM) and partitioning around medoids (PAM), were also used for comparison. FCM is a popular unsupervised clustering algorithm that groups data into n clusters with every data point related to every cluster. Data points close to a cluster center are assigned a high degree of belonging (connection) to that cluster. Data points far away from the cluster center are assigned a low degree of belonging to that cluster (39). PAM is another typical clustering algorithm. K data points are initially selected and moved towards the best cluster repeatedly. All combinations of data points are then analyzed and the clustering quality for all pairs of data points are derived. If a point is found with the best-improved value of distortion function, the new data point will replace the current best data point. These newly generated best data points form the new medoids (40-42).

Results

DR

The contribution of the top 15 features was calculated as shown in *Figure 3* and *Figure 4* for the PCA and AE methods, respectively. For the PCA method, the contribution of the top 10 principal components was 90.5%. The contribution of the 11th principal component was 1.7%, which was less. Thus, the top 10 principal components were adopted in the PCA method. For AE, the average reconstruction loss was decreased as the number of latent variables of the bottleneck layer increased. The

Table 2 Average value (mean ± standard deviation) of clustering indexes for different number k

K	Sil	DB	CH	DVI
2	0.68±0.11	0.87±0.25	828.81±198.35	0.49±0.07
3	0.79±0.00	0.52±0.01	908.73±34.38	0.48±0.13
4	0.82±0.09	0.46±0.00	1,804.52±423.92	0.51±0.16
5	0.76±0.06	0.80±0.05	1,479.15±343.41	0.07±0.03
6	0.71±0.06	0.99±0.13	1,349.24±64.63	0.05±0.02
7	0.62±0.08	1.15±0.08	1,257.43±29.37	0.05±0.01
8	0.59±0.08	1.22±0.09	1,142.37±40.95	0.05±0.01
9	0.52±0.05	1.28±0.05	1,043.51±30.52	0.04±0.01
10	0.53±0.09	1.31±0.11	1,010.32±41.61	0.05±0.01

Sil, silhouette coefficient; DB, Davies-Bouldin index; CH, Calinski-Harabasz index; DVI, dunn validity index.

average reconstruction loss increased quickly as this number was more than 9. As a compromise, 9 latent variables were adopted in the AE method as their contribution had the smallest mean and standard deviation of reconstruction errors.

The quality of the clusters using the 2 DR algorithms was evaluated using the silhouette coefficient. The silhouette coefficients for clusters from the PCA- and AE-based models were 0.18±0.02 and 0.47±0.01, respectively. The quality of clusters of the AE-based method was better than that of the PCA-based method. Multicollinearity tests were also performed to evaluate the linear correlation among latent variables resulting from the PCA and AE methods. Variance inflating factor (VIF) was used to test the presence of multicollinearity. For 35 features in the raw data set, the mean value of VIF was 89, whereas their values were 1 and 3.5 using the lower number of features resulting from the PCA and AE methods, respectively. Note that VIF values between 1–5 indicate low multicollinearity and values higher than 5 indicate high multicollinearity.

Clustering parameters

To determine the best value of the number k and the optimal distance metrics and centroid initialization methods, different settings were evaluated using 4 indexes (Sil, CH, DVI, and DB). Larger values of Sil, CH, and DVI and a smaller value of DB indicated better clustering performance. As shown in *Table 2*, the optimal cluster number with the best values among the 4 indexes was 4. For distance metrics, sqeuclidean (squared Euclidian distance) showed the best performance among the 4 indexes (*Figure 5*). For centroid initialization methods, the plus method showed the best performance among the 4 indexes (*Figure 6*). Note that in our study, these optimal parameters and methods were adopted in the learning model.

Anomaly detection

The receiver operating characteristic (ROC) curve of the 3 anomaly detection models with and without DR are shown in *Figure* 7. The ROC showed the performance of classifiers based on the PCA and AE methods for different threshold values. The performance of the AE-based k-means clustering method was better than that of the PCA-based k-means clustering method, and both were better than that of k-means clustering without DR. For the competitor models, the performance of the AE/PCA-based k-means clustering method was better than that of the AE/PCA-based FCM/PAM clustering methods. However, without DR, the performance of the k-means clustering method was inferior to that of the FCM/PAM clustering methods. Performance measures of the different methods are presented in *Table 3*. The AUC scores were 0.83, 0.81, and 0.90 for the original, PCA-based, and AE-based k-means clustering methods, respectively. The AUC scores were 0.83, 0.81, and 0.81 for the original, PCA-based, and AE-based FCM clustering methods, respectively. The AUC scores were 0.82, 0.76, and 0.85 for the original, PCA-based, and AE-based PAM clustering methods, respectively. Among the 9 unsupervised learning models, the AE-based k-means clustering method demonstrated the best performance in anomaly detection.

When the threshold was 1σ, the mean sensitivity of the PCA-based and AE-based k-means clustering models were 100% (19/19) and 100% (19/19), respectively. The specificity of the PCA-based and AE-based k-means clustering models were 5.7% (9/158) and 6.3% (10/158), respectively. When the threshold was 2σ, the sensitivity of the PCA-based and AE-based clustering models were 84.2% (16/19) and 94.7% (18/19), respectively. The specificity of the PCA-based and AE-based k-means clustering models were 64.6% (102/158) and 69.0% (109/158), respectively. When the threshold was set at 3σ, the sensitivity of the PCA-based and AE-based k-means clustering models were 42.1% (8/19) and 57.9% (11/19), respectively. The specificity of the PCA-based and AE-based clustering

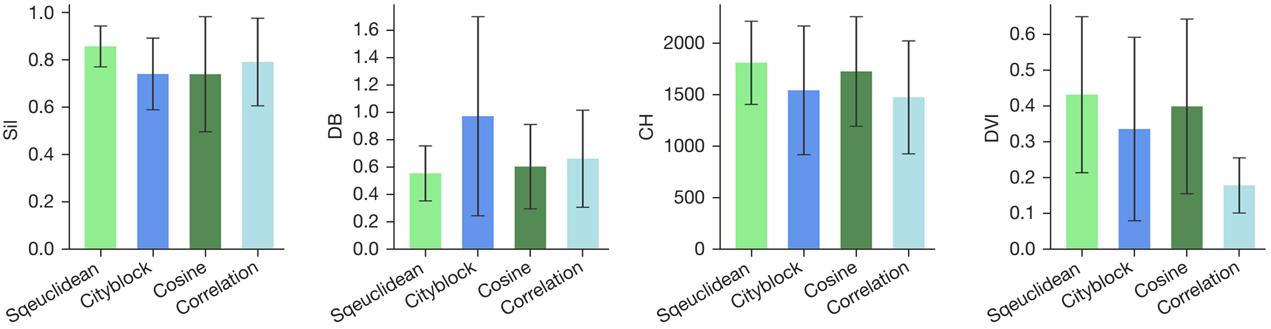

Figure 5 The indexes evaluated for 4 distance metrics. Sil, silhouette coefficient; DB, Davies-Bouldin index; CH, Calinski-Harabasz index; DVI, dunn validity index.

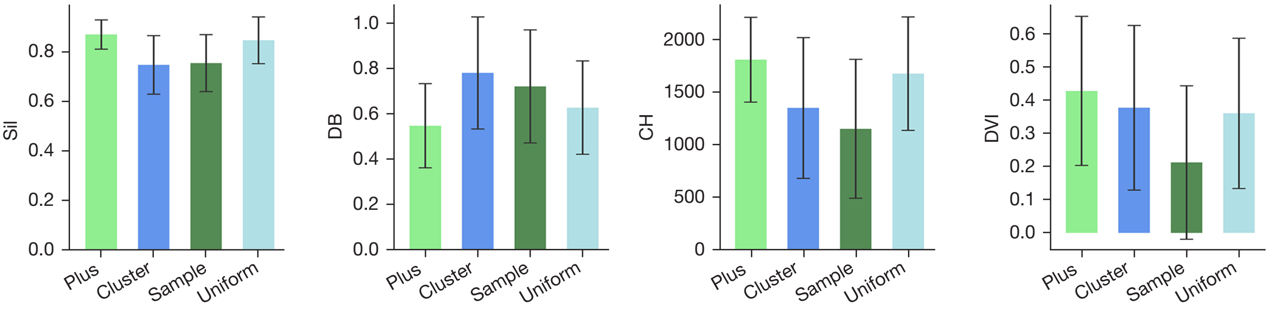

Figure 6 The indexes evaluated for 4 centroid initialization methods. Sil, silhouette coefficient; DB, Davies-Bouldin index; CH, Calinski-Harabasz index; DVI, dunn validity index.

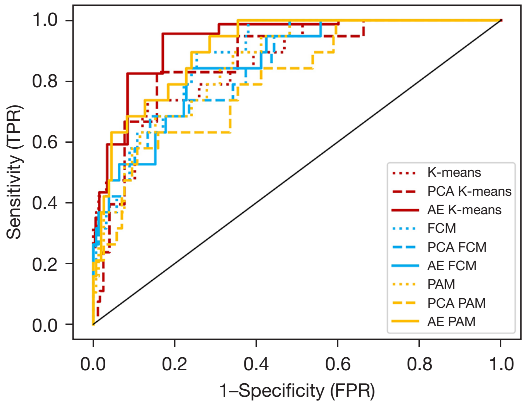

Figure 7 ROC curves for 3 clustering models with and without PCA/AE-based dimensionality reduction. PCA, principal component analysis; AE, autoencoder; FCM, fuzzy c-means; PAM, partitioning around medoids; TPR, true positive rate; FPR, false positive rate; ROC, receiver operating characteristic.

models were 92.4% (146/158) and 91.8% (145/158), respectively. For balancing sensitivity and specificity, 2σ was the optimal threshold.

Discussion

Both DR methods, PCA and AE, were quantitatively evaluated in this study. The results showed that the AE-based clustering model outperformed the PCA-based clustering model in several aspects. First, the AE method employed nonlinear transfer function in encoding/decoding layers. This allowed the network to learn more complex mapping relationships between high-dimension space and low-dimension space, which may have caused less information loss during the DR process. Second, the AE-based clustering model could result in better clusters by fewer parameters, which were 9 latent variables (*Figure 4*). The clustering results were quantified by silhouette coefficient, which also showed that the quality of AE-based clusters was better than PCA-based clusters. Third, the AE-based clustering model showed higher sensitivity and specificity in

Table 3 The performance of 3 clustering models with and without PCA/AE-based dimensionality reduction

Models	Sensitivity	Specificity	FPR	FNR	AUC
K-means	0.37	0.98	0.02	0.63	0.83
PCA k-means	0.84	0.65	0.35	0.16	0.81
AE k-means	0.95	0.69	0.31	0.05	0.90
FCM	0.32	0.96	0.04	0.68	0.83
PCA FCM	0.26	0.97	0.03	0.74	0.81
AE FCM	0.32	0.96	0.04	0.68	0.81
PAM	0.11	0.99	0.01	0.89	0.82
PCA PAM	0.11	1.00	0.00	0.89	0.76
AE PAM	0.11	0.99	0.01	0.89	0.85

PCA, principal component analysis; AE, autoencoder; FCM, fuzzy c-means; PAM, partitioning around medoids; FPR, false positive rate; FNR, false positive rate; AUC, area under curve.

anomaly detection. The ROC and AUC scores also showed that the AE-based clustering model outperformed the PCA-based clustering model.

The application of AE-based clustering in anomaly detection of radiotherapy treatment plans was promising. However, the method could be improved in several aspects. First, the original features extracted from plans were mainly based on expert experience and less quantitative. Several features that may have been sensitive to abnormal plans were not included in our study due to inconsistencies among plans. Second, the current AE-based clustering model should be improved to meet more complex clinical scenarios. For example, treatment plans become more complex and error-prone for new treatment technologies such as VMAT, adaptive radiotherapy, and magnetic resonance (MR)-guided radiotherapy. Third, the AE model used in this study was a simple model, and most model parameters were not optimized for the specific task. With better models and more optimized parameters, the performance of AE could be further improved. In the future, neuroevolution could be used to learn the optimal architecture of AE and its parameters. The benefit of neuroevolution is that it is a scalable and non-gradient method.

The specificity of the AE-based clustering model was 69.0% when the threshold was set to 2σ. The high FPR of the AE model caused about 31.0% of the normal plans to be identified as abnormal plans. In clinical use, plans that are identified as abnormal plans need further verification by clinical physicists, and this high FPR may increase clinical workload. When the threshold was set to 2σ, the TPR dropped to 84.2% and 94.7% for the PCA-based and AE-based clustering models, respectively. Alternatively, the FNR for both models increased to 15.8% and 5.3%, respectively. This means that fewer abnormal plans were wrongly identified as normal plans for the AE-based model than that for the PCA-based model. In addition, only 19 plans in this study were labeled as abnormal by the experienced physicists. It is possible that some abnormal plans may have not been identified by the experienced physicists, which could have contributed to the high FNR. The verification of abnormal plans is important but challenging for clinical physicists.

Conclusions

The proposed AE-based k-means clustering method is feasible for detecting abnormal plans from normal plans for patients undergoing IMRT treatment. The k-means clustering model with AE-based DR is more effective than the other models and has potential to replace the current manual procedure for automatically identifying low-quality plans from regular plans in radiotherapy plan checking.

Acknowledgments

Funding: This work was partially supported by the National Natural Science Foundation of China (Nos. 11875320, 11975312).

Footnote

Reporting Checklist: The authors have completed the STARD reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-22-825/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-22-825/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The Institutional Ethics Committee of the Cancer Hospital, Chinese Academy of Medical Sciences, and Peking Union Medical College approved this study. Informed consent was waived due to the retrospective nature of the study.

References

1. Azmandian F, Kaeli D, Dy JG, Hutchinson E, Ancukiewicz M, Niemierko A, Jiang SB. Towards the development of an error checker for radiotherapy treatment plans: a preliminary study. Phys Med Biol 2007;52:6511-24.
2. Ford E, Conroy L, Dong L, de Los Santos LF, Greener A, Gwe-Ya Kim G, Johnson J, Johnson P, Mechalakos JG, Napolitano B, Parker S, Schofield D, Smith K, Yorke E, Wells M. Strategies for effective physics plan and chart review in radiation therapy: Report of AAPM Task Group 275. Med Phys 2020;47:e236-72.
3. Vatnitsky S, Ortiz Lopez P, Izewska J, Meghzifene A, Levin V. The radiation overexposure of radiotherapy patients in Panama 15 June 2001. Radiother Oncol 2001;60:237-8.
4. Lack D, Liang J, Benedetti L, Knill C, Yan D. Early detection of potential errors during patient treatment planning. J Appl Clin Med Phys 2018;19:724-32.
5. Covington EL, Chen X, Younge KC, Lee C, Matuszak MM, Kessler ML, Keranen W, Acosta E, Dougherty AM, Filpansick SE, Moran JM. Improving treatment plan evaluation with automation. J Appl Clin Med Phys 2016;17:16-31.
6. Yang D, Wu Y, Brame RS, Yaddanapudi S, Rangaraj D, Li HH, Goddu SM, Mutic S. Technical note: electronic chart checks in a paperless radiation therapy clinic. Med Phys 2012;39:4726-32.
7. Furhang EE, Dolan J, Sillanpaa JK, Harrison LB. Automating the initial physics chart checking process. J Appl Clin Med Phys 2009;10:129-35.
8. Gopan O, Zeng J, Novak A, Nyflot M, Ford E. The effectiveness of pretreatment physics plan review for detecting errors in radiation therapy. Med Phys 2016;43:5181.
9. Dewhurst JM, Lowe M, Hardy MJ, Boylan CJ, Whitehurst P, Rowbottom CG. AutoLock: a semiautomated system for radiotherapy treatment plan quality control. J Appl Clin Med Phys 2015;16:5396.
10. Siochi RA, Pennington EC, Waldron TJ, Bayouth JE. Radiation therapy plan checks in a paperless clinic. J Appl Clin Med Phys 2009;10:43-62.
11. Yang D, Moore KL. Automated radiotherapy treatment plan integrity verification. Med Phys 2012;39:1542-51.
12. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science 2006;313:504-7.
13. Jiang X, Gao J, Xia H, Cai Z. Gaussian Processes Autoencoder for Dimensionality Reduction. Pacific-Asia Conference on Knowledge Discovery & Data Mining 2014; 62-73.
14. Pang G, Shen C, Cao L, Hengel AVD. Deep learning for anomaly detection: A review. ACM Computing Surveys (CSUR) 2021;54:1-38.
15. Schreyer M, Sattarov T, Borth D, Dengel A, Reimer B. Detection of anomalies in large scale accounting data using deep autoencoder networks. CoRR 2017. Available online: https://doi.org/10.48550/arXiv.1709.05254
16. Schreyer M, Sattarov T, Schulze C, Reimer B, Borth D. Detection of accounting anomalies in the latent space using adversarial autoencoder neural networks. CoRR 2019. Available online: https://doi.org/10.48550/arXiv.1908.00734
17. An J, Cho S. Variational autoencoder based anomaly detection using reconstruction probability. Special Lecture on IE 2015;2:1-18.
18. Ji T, Vuppala S T, Chowdhary G, Driggs-Campbell K. Multi-modal anomaly detection for unstructured and

uncertain environments. CoRR 2020. Available online: https://doi.org/10.48550/arXiv.2012.08637

19. Wang L, Li J, Zhang S, Zhang X, Zhang Q, Chan MF, Yang R, Sui J. Multi-task autoencoder based classification-regression model for patient-specific VMAT QA. Phys Med Biol 2020;65:235023.
20. Mayo CS, Urie MM, Fitzgerald TJ. Hybrid IMRT plans--concurrently treating conventional and IMRT beams for improved breast irradiation and reduced planning time. Int J Radiat Oncol Biol Phys 2005;61:922-32.
21. Pinaya WHL, Mechelli A, Sato JR. Using deep autoencoders to identify abnormal brain structural patterns in neuropsychiatric disorders: A large-scale multi-sample study. Hum Brain Mapp 2019;40:944-54.
22. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans A Math Phys Eng Sci 2016;374:20150202.
23. Hotelling H. Analysis of a complex of statistical variables into principal components. J. Educ. Psychol 1933;24:417-41, 498-520.
24. Alkhayrat M, Aljnidi M, Aljoumaa K. A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA. Journal of Big Data 2020;7:1-23.
25. Galván E, Mooney P. Neuroevolution in deep neural networks: Current trends and future challenges. IEEE Transactions on Artificial Intelligence 2021;2:476-93.
26. Okada H. Neuroevolution of autoencoders by genetic algorithm. International Journal of Science and Engineering Investigations 2017;6:127-31.
27. Pietroń M, Żurek D, Faber K, Corizzo R. Fast and scalable neuroevolution deep learning architecture search for multivariate anomaly detection. CoRR 2021. Available online: https://doi.org/10.48550/arXiv.2112.05640
28. Steinley D. K-means clustering: a half-century synthesis. Br J Math Stat Psychol 2006;59:1-34.
29. Kaufman L, Rousseeuw P J. Finding groups in data: an introduction to cluster analysis. John Wiley & Sons 2009.
30. Davies DL, Bouldin DW. A cluster separation measure. IEEE Trans Pattern Anal Mach Intell 1979;1:224-7.
31. Caliński T, Harabasz J. A dendrite method for cluster analysis. Communications in Statistics-theory and Methods 1974;3:1-27.
32. Dudoit S, Fridlyand J. A prediction-based resampling method for estimating the number of clusters in a dataset. Genome Biol 2002;3:RESEARCH0036.
33. Zhou S, Xu Z, Liu F. Method for Determining the Optimal Number of Clusters Based on Agglomerative Hierarchical Clustering. IEEE Trans Neural Netw Learn Syst 2017;28:3007-17.
34. Rendón E, Abundez I, Arizmendi A, Quiroz E. Internal versus external cluster validation indexes. International Journal of Computers and Communications 2011;5:27-34.
35. Franco EF, Rana P, Cruz A, Calderón VV, Azevedo V, Ramos RTJ, Ghosh P. Performance Comparison of Deep Learning Autoencoders for Cancer Subtype Detection Using Multi-Omics Data. Cancers (Basel) 2021;13:2013.
36. Wang H, Wang J, Dong C, Lian Y, Liu D, Yan Z. A Novel Approach for Drug-Target Interactions Prediction Based on Multimodal Deep Autoencoder. Front Pharmacol 2020;10:1592.
37. Zhang YD, Hou XX, Lv YD, Chen H, Wang SH. Sparse Autoencoder based deep neural network for voxelwise detection of cerebral microbleed. 2016 IEEE 22nd International Conference on Parallel and Distributed Systems (ICPADS) 2016:1229-32.
38. Nakao T, Hanaoka S, Nomura Y, Murata M, Takenaga T, Miki S, Watadani T, Yoshikawa T, Hayashi N, Abe O. Unsupervised Deep Anomaly Detection in Chest Radiographs. J Digit Imaging 2021;34:418-27.
39. Ghosh S, Dubey S K. Comparative analysis of k-means and fuzzy c-means algorithms. International Journal of Advanced Computer Science and Applications 2013;4:35-9.
40. Sureja N, Chawda B, Vasant A. An improved K-medoids clustering approach based on the crow search algorithm. Journal of Computational Mathematics and Data Science 2022;3:100034.
41. Salmanpour MR, Shamsaei M, Hajianfar G, Soltanian-Zadeh H, Rahmim A. Longitudinal clustering analysis and prediction of Parkinson's disease progression using radiomics and hybrid machine learning. Quant Imaging Med Surg 2022;12:906-19.
42. Huang P, Yan H, Hu Z, Liu Z, Tian Y, Dai J. Predicting radiation pneumonitis with fuzzy clustering neural network using 4DCT ventilation image based dosimetric parameters. Quant Imaging Med Surg 2021;11:4731-41.

Cite this article as: Huang P, Yan H, Song Z, Xu Y, Hu Z, Dai J. Combining autoencoder with clustering analysis for anomaly detection in radiotherapy plans. Quant Imaging Med Surg 2023;13(4):2328-2338. doi: 10.21037/qims-22-825

Appendix 1

Distance metrics for k-means clustering

The choice of distance measures is a critical step in clustering. It defines how the similarity of 2elements (x, y) is calculated and will influence the shape of the clusters. In this study, the impact of 4 different distance metrics on clustering results was evaluated. The following table contains the description of different distance metrics and their formulas.

Distance metric	Description	Formula
Sqeuclidean	Squared Euclidean distance. Each centroid is the mean of the points in that cluster.	$\mathrm{d}(\mathrm{x},\mathrm{y})=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$ (1)
Cityblock	Sum of absolute differences. Each centroid is the component-wise median of the points in that cluster.	$\mathrm{d}(\mathrm{x},\mathrm{y})=\sum_{i=1}^{n}\lvert(x_i-y_i)\rvert$ (2)
Cosine	1 minus the cosine of the included angle between points. Each centroid is the mean of the points in that cluster, after normalizing those points to unit Euclidean length.	$\mathrm{d}(\mathrm{x},\mathrm{y})=1-\frac{\lvert\sum_{i=1}^{n}x_iy_i\rvert}{\sqrt{\sum_{i=1}^{n}x_i^2\sum_{i=1}^{n}y_i^2}}$ (3)
Correlation	1 minus the sample correlation between points (treated as sequences of values). Each centroid is the component-wise mean of the points in that cluster, after centering and normalizing those points to 0 mean and unit standard deviation.	$\mathrm{d}(\mathrm{x},\mathrm{y})=1-\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$ (4)

Centroid initialization methods for k-means clustering

K-means clustering aims to converge on an optimal set of cluster centers (centroids) and cluster membership based on distance from these centroids via successive iterations. It is intuitive in that the more optimal the positioning of these initial centroids, the fewer iterations of the k-means clustering algorithms will be required for convergence. This suggests that some strategic consideration to the initialization of these initial centroids could prove useful. Four methods for centroid initialization were tested in this study, as follows:

Method	Description
Plus	First, a data point is randomly selected from the input data set, and its distance from the nearest cluster center is then calculated. A new data point with larger distance is then selected as the new cluster center. This procedure is repeated until k cluster centers are selected. The idea is that the initial seeds should be as far away from each other as possible.
Sample	The k cluster centers are selected from the data set randomly.
Cluster	Perform a preliminary clustering on a subset of 10% data points. This preliminary clustering is initialized using the 'sample' method. If the number of data points in the subset is less than k, then k data points are selected from the data set randomly.
Uniform	Select k data points uniformly and randomly from the range of data set.