Deep Unsupervised Clustering for Conditional Identification of Subgroups Within a Digital Pathology Image Set

Mariia Sidulova[1,3], Xudong Sun[2], and Alexej Gossmann[1(✉)]

[1] U.S. Food and Drug Administration, Center for Devices and Radiological Health, Silver Spring, MD, USA
alexej.gossmann@fda.hhs.gov
[2] Institute of AI for Health, Helmholtz Munich, Munich, Germany
[3] Department of Biomedical Engineering, George Washington University, Washington, D.C., USA

Abstract. Consideration of subgroups or domains within medical image datasets is crucial for the development and evaluation of robust and generalizable machine learning systems. To tackle the domain identification problem, we examine deep unsupervised generative clustering approaches for representation learning and clustering. The Variational Deep Embedding (VaDE) model is trained to learn lower-dimensional representations of images based on a Mixture-of-Gaussians latent space prior distribution while optimizing cluster assignments. We propose the Conditionally Decoded Variational Deep Embedding (CDVaDE) model which incorporates additional variables of choice, such as the class labels, as conditioning factors to guide the clustering towards subgroup structures in the data which have not been known or recognized previously. We analyze the behavior of CDVaDE on multiple datasets and compare it to other deep clustering algorithms. Our experimental results demonstrate that the considered models are capable of separating digital pathology images into meaningful subgroups. We provide a general-purpose implementation of all considered deep clustering methods as part of the open source Python package DomId (https://github.com/DIDSR/DomId).

Keywords: Domain Identification · Deep Clustering · Subgroup Identification · Variational Autoencoder · Generative Model

1 Introduction

Machine learning (ML), specifically deep learning (DL), algorithms have shown exceptional performance on numerous medical image analysis tasks [2]. Never-

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43993-3_64.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14227, pp. 666–675, 2023.
https://doi.org/10.1007/978-3-031-43993-3_64

theless, comprehensive reviews highlight major issues of generalizability, robustness, and reproducibility in medical imaging AI/ML [9,15]. For a generalizability assessment, reporting only aggregate performance measures is not sufficient. Due to model complexity and limited training data, ML performance often varies across data subgroups or domains, such as different patient subpopulations or varied data acquisition scenarios. Aggregate performance measures (e.g., sensitivity, specificity, ROC AUC) can be dominated by the larger subgroups, masking the poor ML model performance on smaller but clinically important subgroups [11]. Thus, achieving (through training) and demonstrating (as part of testing) satisfactory ML model performance across relevant subgroups is crucial before the real-world clinical deployment of a medical ML system [13].

However, a challenging situation arises when relevant subgroups are unrecognized. One solution to this issue is to apply a clustering algorithm to the data, with the goal of identifying the unannotated subgroups. The main objective of unsupervised clustering is to group data points into distinct classes of similar traits. However, due to the complexity and high dimensionality of the medical imaging data and the resulting difficulty in establishing a concrete notion of similarity, extracting low-dimensional characteristics becomes the key to establishing the best criteria for grouping. Unsupervised generative clustering aims to simultaneously address both domain identification and dimensionality reduction. Deep unsupervised clustering algorithms could map the medical imaging data back to their causal factors or underlying domains, such as image acquisition equipment, patient subpopulations, or other meaningful data subgroups. However, there is a practical need to be able to guide the deep clustering model towards the identification of grouping structures in a given dataset that have not been already annotated. To that end, we propose a mechanism that is intended to constrain the model towards identifying clusters in the data that are not associated with given variables of choice (already known class labels or subgroup structures). The resulting algorithmic cluster assignments could then be used to improve ML algorithm training, or for generalizability and robustness evaluation.

2 Methods

We provide a PyTorch-based implementation of all deep clustering algorithms described below (VaDE, CDVaDE, and DEC) in the open source Python package DomId that is publicly available under https://github.com/DIDSR/DomId.

2.1 Variational Deep Embedding (VaDE)

Variational Deep Embedding (VaDE) [6] is an unsupervised generative clustering approach based on Variational Autoencoders [10]. In our study, VaDE is deployed as a deep clustering model using Convolutional Neural Network (CNN) architectures for the encoder $g(\mathbf{x};\boldsymbol{\phi})$ and the decoder $f(\mathbf{z};\boldsymbol{\theta})$. The encoder learns to

compress the high-dimensional input images $\mathbf{x}$ into lower-dimensional latent representations $\mathbf{z}$. Using a Mixture-of-Gaussians (MOG) prior distribution for the latent representations $\mathbf{z}$, we examine subgroups or domains within the dataset, revealed by the individual Gaussians within the learned latent space, and how $\mathbf{z}$ affects the generation of $\mathbf{x}$. The model can be used to perform inference, where observed images $\mathbf{x}$ are mapped to corresponding latent variables $\mathbf{z}$ and their cluster/domain assignments c. We denote the latent space dimensionality by d (i.e., $\mathbf{z} \in \mathbb{R}^d$), and the number of clusters by D (i.e., $c \in \{1, 2, \ldots, D\}$). The trained decoder CNN can also be used to generate synthetic images from the algorithmically identified subgroups.

VaDE is optimized using Stochastic Gradient Variational Bayes [10] to maximize a statistical measure called the Evidence Lower Bound (ELBO). We denote the true data distribution by $p(\mathbf{z}, \mathbf{x}, c)$ and the variational posterior distribution by $q(\mathbf{z}, c|\mathbf{x})$. The ELBO of VaDE can be written as

$$\begin{aligned} \mathcal{L}_{ELBO}(\mathbf{x}) &= E_{q(\mathbf{z},c|\mathbf{x})}\left[\log \frac{p(\mathbf{z}, \mathbf{x}, c)}{q(\mathbf{z}, c|\mathbf{x})}\right] \\ &= E_{q(\mathbf{z},c|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z}) + \log p(\mathbf{z}|c) \\ &\quad + \log p(c) - \log q(\mathbf{z}|\mathbf{x}) - \log q(c|\mathbf{x})], \end{aligned} \tag{1}$$

where $p(\mathbf{x}|\mathbf{z})$ is modeled by the decoder CNN, and $q(\mathbf{z}|\mathbf{x})$ is modeled by the encoder CNN $g(\mathbf{x}; \boldsymbol{\phi})$ as

$$q(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\mathbf{z}; \tilde{\boldsymbol{\mu}}, \mathrm{diag}\left(\tilde{\boldsymbol{\sigma}}^2\right)\right), \quad \left(\tilde{\boldsymbol{\mu}}, \log \tilde{\boldsymbol{\sigma}}^2\right) = g(\mathbf{x}; \boldsymbol{\phi}).$$

Finally, the cluster assignments can be determined via

$$q(c|\mathbf{x}) \approx p(c|\mathbf{z}) = \frac{p(c)p(\mathbf{z}|c)}{\sum_{c'=1}^{D} p(c')p(\mathbf{z}|c')}, \tag{2}$$

$$p(c) = \mathrm{Cat}(\boldsymbol{\pi}), \quad p(\mathbf{z}|c) = \mathcal{N}\left(\mathbf{z}; \boldsymbol{\mu}_c, \mathrm{diag}\left(\boldsymbol{\sigma}_c^2\right)\right), \tag{3}$$

where the probability distributions $p(c)$ and $p(\mathbf{z}|c)$ come from the MOG prior of the latent space, with the respective distributional parameters $\boldsymbol{\pi}$, $\boldsymbol{\mu}_c$, $\boldsymbol{\sigma}_c^2$ (for $c \in \{1, 2, \ldots, D\}$) optimized by maximizing the ELBO of Eq. (1). Note that Eq. (2) follows from the observation that in order to maximize the ELBO in Eq. 1, the KL Divergence between $q(c|\mathbf{x})$ and $p(c|\mathbf{z})$ needs to be equal to 0. We refer to [6] for details.

In all our experiments, we apply VaDE with CNN architectures for the encoder and decoder. The CNN encoder consists of convolution layers with 32, 64, 128 filters, respectively, followed by a fully-connected layer. Respectively, the CNN decoder consists of a fully-connected layer followed by transposed convolution layers with the number of input/output channels decreasing as 128, 64, 32, 3. Batch normalization and the leaky ReLU activation functions are used.

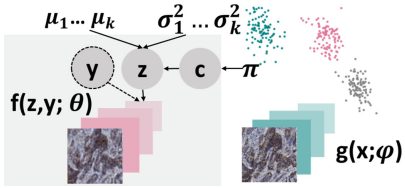

Fig. 1. Diagram of CDVaDE: $\mathbf{z}$ distribution is driven by Gaussian mean and covariance parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\sigma}_k^2$, prior cluster probabilities π_k, and conditioning variables $\mathbf{y}$.

2.2 Conditionally Decoded Variational Deep Embedding (CDVaDE)

We propose the Conditionally Decoded Variational Deep Embedding (CDVaDE) model as an extension to VaDE as shown in Fig. 1. The generative process of CDVaDE differs from VaDE in that it concatenates additional variables $\mathbf{y}$ to the latent representation $\mathbf{z}$. For example, $\mathbf{y}$ may contain the available class labels or already known subgroup structures, which do not need to be discovered. It is assumed that these additional variables $\mathbf{y}$ are available at training and test time. Specifically, the generative process of CDVaDE takes the form

$$p(c) = \text{Cat}(\boldsymbol{\pi}) \tag{4}$$

$$p(\mathbf{z}|c) = \mathcal{N}\left(\mathbf{z}; \boldsymbol{\mu}_c, \text{diag}\left(\boldsymbol{\sigma}_c^2\right)\right), \tag{5}$$

$$\left(\boldsymbol{\mu}_{xy}, \log \boldsymbol{\sigma}_{xy}^2\right) = f(\mathbf{z}, \mathbf{y}; \boldsymbol{\phi}), \tag{6}$$

$$p(\mathbf{x}|\mathbf{z}, \mathbf{y}) = \mathcal{N}\left(\mathbf{x}; \boldsymbol{\mu}_{xy}, \text{diag}\left(\boldsymbol{\sigma}_{xy}^2\right)\right) \tag{7}$$

Since our goal is to find clusters c that are unassociated with the available variables $\mathbf{y}$ of choice and to learn latent representations $\mathbf{z}$ that do not contain information about $\mathbf{y}$, the generative process of CDVaDE assumes that $\mathbf{z}, c$ are jointly independent of $\mathbf{y}$.

The changes compared to the generative process of VaDE can also be regarded as imposing a structure on the model, where the encoder learns hidden representations of the image $\mathbf{x}$ conditioned to the additional variables $\mathbf{y}$ (i.e., $q(\mathbf{z}|\mathbf{x}, \mathbf{y})$), but acts as an identity function with respect to $\mathbf{y}$ (i.e., $\mathbf{y}$ can be regarded as being simply concatenated to the latent space representations $\mathbf{z}$). The decoder then translates this data representation in the form of $(\mathbf{z}, \mathbf{y})$ to the input space (i.e., $p(\mathbf{x}|\mathbf{z}, \mathbf{y})$). Given that the underlying VAE architecture seeks to efficiently compress the input data $\mathbf{x}$ into a learned representation, this incentivizes the model to exclude information about $\mathbf{y}$ from the learned variables $\mathbf{z}$ and c.

The ELBO of CDVaDE can be derived as follows,

$$\begin{aligned} \mathcal{L}_{ELBO}(\mathbf{x}|\mathbf{y}) &= E_{q(\mathbf{z},c|\mathbf{x})}\left[\log \frac{p(\mathbf{z}, \mathbf{x}, c|\mathbf{y})}{q(\mathbf{z}, c|\mathbf{x})}\right] \\ &= E_{q(\mathbf{z},c|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z}, \mathbf{y}) + \log p(\mathbf{z}|c) \\ &\quad + \log p(c) - \log q(\mathbf{z}|\mathbf{x}) - \log q(c|\mathbf{x})], \end{aligned} \tag{8}$$

where we use the fact that by the generative process of CDVaDE it holds that

$$p(\mathbf{x}, \mathbf{z}, c|\mathbf{y}) = p(\mathbf{x}|\mathbf{z}, \mathbf{y})p(\mathbf{z}|c, \mathbf{y})p(c|\mathbf{y}) = p(\mathbf{x}|\mathbf{z}, \mathbf{y})p(\mathbf{z}|c)p(c), \tag{9}$$

and we adopt from VaDE the assumption that $q(\mathbf{z}, c|\mathbf{x}) = q(\mathbf{z}|\mathbf{x})q(c|\mathbf{x})$ holds. Hence, once the base VaDE decoder CNN is replaced by its modified version $f(\mathbf{z}, \mathbf{y}; \boldsymbol{\theta})$ in CDVaDE, there are no further differences between the ELBO loss function of Eq. (8) compared to Eq. (1).

While in this work we present our conditioning mechanism as an extension to VaDE, it can be combined with any deep clustering algorithm that follows an encoder-decoder architecture. In all our experiments, we use the same CNN architectures for the encoder and decoder as in VaDE (see Sect. 2.1).

2.3 Deep Embedding Clustering (DEC)

Deep Embedding Clustering (DEC) [14] is a popular state-of-the-art clustering approach that combines a deep embedding model with k-means clustering. In this study, we include comparisons of VaDE and the proposed CDVaDE to DEC, because it is a model that belongs to a different family of deep clustering algorithms which are not based on variational inference. In our DEC experiments, we use the same autoencoder architecture and the same initialization as for the VaDE.

2.4 Related Works in Medical Imaging

A number of studies have been conducted with several approaches of deep clustering for medical imaging data. Typically, clustering is performed on top of features extracted with the use of an encoder neural network, and the cluster assignments are determined by using conventional clustering algorithms, such as k-means, on top of the learned latent representations [1,5,7,12]. In contrast, this work investigates models which enforce a clustering structure in the latent space through the use of a MOG prior distribution, as well as guidance of the clustering model via the proposed conditioning mechanism.

3 Experiments

3.1 Colored MNIST

The Colored MNIST is an extension to the classic MNIST dataset [3], which contains binary images of handwritten digits. The Colored MNIST includes colored images of the same digits, where each number and background have a color assignment. We present results of the experiments with five distinct colors and five digits of MNIST (0–4). To enhance computational efficiency and expedite experiments, we utilized only 1% of the MNIST images, which were sampled at random. This simple dataset can be used to investigate whether a given clustering algorithm will categorize the images by color or by the digit label

and whether the proposed conditioning mechanism of CDVaDE can successfully guide the clustering away from the categorization we want to avoid (e.g., condition the model to avoid clustering by color, in order to distinguish the digits in an unsupervised fashion). We compare CDVaDE to the deep clustering models VaDE and DEC that do not incorporate such conditioning. We use latent space dimensionality $d = 20$ for all models.

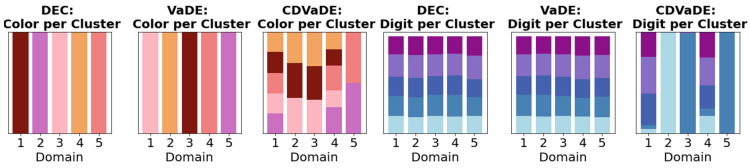

Fig. 2. Both VaDE and DEC cluster the Colored MNIST digits by the color, while CDVaDE clusters are associated with the digit label. Bar graphs labeled "Color" – each color represents a specific color of Colored MNIST digits. In the "Digit" plots, colors correspond to digits labels.

In Fig. 2 a summary of the results for the experiments on the colored MNIST dataset is presented. The results demonstrate that by allowing for the incorporation of additional information, particularly color labels, the proposed CDVaDE model is more sensitized to learning other underlying features, which allows for distinguishing between the different digits in this particular example. Notably, both VaDE and DEC end up clustering the data by color, as it is the most striking distinguishing characteristic of these images. On the other hand, the predicted domains of CDVaDE have no association with color, and the data are separated by the shapes in the images, distinguishing some of the digit labels (albeit imperfectly). This example serves as a proof of concept for the proposed conditioning mechanism of CDVaDE.

3.2 Application to a Digital Pathology Dataset

HER2 Dataset. Human epidermal growth factor receptor 2 (HER2 or HER2/neu) is a protein involved in normal cell growth, which plays an important role in the diagnosis and treatment of breast cancer [8]. The dataset consists of 241 patches extracted from 64 digitized slides of breast cancer tissue which were stained with HER2 antibody. Each tissue slide has been digitized at three different sites using three different whole slide imaging systems, evaluated by 7 pathologists on a 0–100 scale, and following clinical practice labeled as HER2 Class 1, 2, or 3 (based on mean pathologists' scores with cut-points at 33 and 66). We use a subset of this dataset consisting of 672 images (the remainder is held out for future research). Because the intended purpose is finding subgroups in the given dataset only, a separate test set is not used. The dimensions of the images vary from 600 to 826 pixels, and we scale all data to a uniform size of 128×128 pixels before further processing. We refer to [4,8] for more details about this dataset.

This retrospective human subject dataset has been made available to us by the authors of the prior studies [4,8], who are not associated with this paper. Appropriate ethical approval for the use of this material in research has been obtained.

Deep Clustering Models Applied to the HER2 Dataset. We evaluate the performance and behavior of the DEC, VaDE, and CDVaDE models on the HER2 dataset. We investigate whether the models will learn to distinguish the HER2 class labels, the scanner labels, or other potentially meaningful data subgroups in a fully unsupervised fashion. To investigate the clustering abilities of CDVaDE on the HER2 dataset, we inject the HER2 class labels into the latent embedding space. We hypothesize that this will disincentivize the encoder network from including information related to the HER2 class labels in the latent representations z. Thus, with CDVaDE we aim to guide the clustering towards identifying subgroup structures that are not associated with the HER2 classes, and potentially were not previously recognized. The dimensionality of the latent embedding space was set to $d = 500$ for all three models.

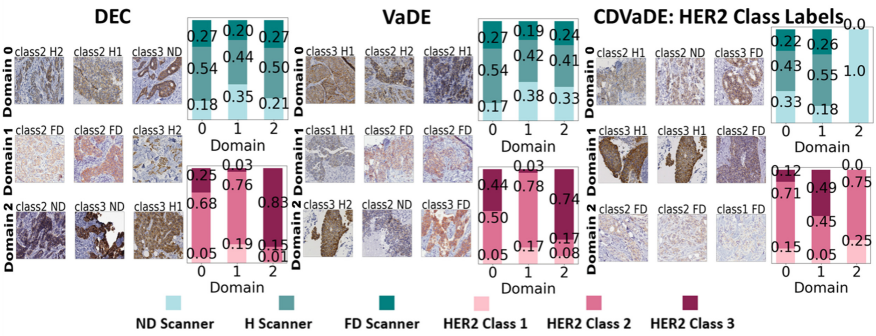

Fig. 3. Results summary for VaDE, CDVaDE, and DEC, all with $D = 3$. Example images from the identified clusters are visualized for each method. Distributions of HER2 class and scanner labels are shown per cluster (i.e., predicted domain).

Figure 3 demonstrates that even without scrutinizing, one can observe a strong visual separation between the algorithmically identified image domains for both VaDE and DEC experiments. For example, in the first predicted domain by VaDE in Fig. 3 images tend to have slightly visible boundaries but a comparatively uniform light appearance overall. In the second predicted domain, images have less visible boundaries and more pail staining. In the third predicted domain, images have more visible staining and sharper edges compared to the other domains.

As illustrated by the bar graphs in Fig. 3, there is an association between HER2 class 2 and predicted domain 2, as well as between HER2 class 3 and

predicted domain 3. Similarly to the VaDE model, the DEC model has also shown the ability to separate between HER2 class 2 and HER2 class 3. To investigate these observations further, we look at the distribution of the ground truth HER2/neu scores within each of the predicted domains. The boxplots in Fig. 4 show that both the VaDE and DEC models tend to separate high HER2/neu scores from the lower ones. The Pearson's correlation coefficient between the clustering assignments c of VaDE and the HER2/neu scores is 0.46. The correlation coefficient between the DEC clusters and the HER2/neu scores is 0.71. However, neither VaDE nor DEC clusters are associated to the scanner labels.

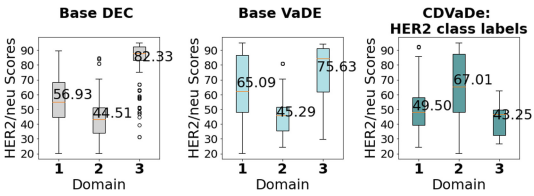

Fig. 4. Boxplots of HER2/neu scores per predicted domain for all experiments.

We investigate the proposed CDVaDE model with the goal of identifying meaningful data subgroups which are not associated with the already known HER2 class labels. As visualized in Fig. 3, the predicted domains are again clearly visually disparate. However, as intended, there is a weaker association with the HER2 class labels and a stronger association with the scanner labels, compared to the results of VaDE and DEC. In Fig. 4, HER2/neu median scores of the three clusters move closer together, illustrating the decrease of association with HER2 class labels, as intended by the formulation of the CDVaDE model. The correlation coefficient between the CDVaDE cluster assignments and the HER2/neu scores is 0.39. While the CDVaDE model does not achieve full independence between the identified clusters and the HER2 labels, it decreases this association compared to VaDE and DEC. Moreover, the clusters identified by CDVaDE are distinctly different from those of VaDE, with a 0.43 proportion of agreement between the two algorithms (after matching the two sets of cluster assignments using the Hungarian algorithm).

4 Conclusion

We investigated deep clustering models for the identification of meaningful subgroups within medical image datasets. The proposed CDVaDE model incorporates a conditioning mechanism that is capable of guiding the clustering model away from subgroup structures that have already been annotated and towards the identification of yet unrecognized image subgroups/domains. Our experimental findings on the HER2 digital pathology dataset surmise that VaDE and DEC are capable of finding, in an unsupervised fashion, image subgroups related to the

HER2 class labels, while CDVaDE (conditioned on the HER2 labels) identifies visually distinct subgroups that have a weaker association to the HER2 labels. Because the CDVaDE clusters do not clearly correspond to the scanner labels either, future work involves a review by a pathologist to see whether these subgroups capture meaningful but unannotated characteristics in the images. While CDVaDE can be used as an exploratory tool to unveil unknown subgroups in a given dataset, developing specialized quantitative evaluation metrics for this unsupervised task is inherently difficult and will also be a focus in our future work.

Acknowledgments. The authors would like to thank Dr. Marios Gavrielides for providing access to the HER2 dataset and for helpful discussion. This project was supported in part by an appointment to the Research Participation Program at the U.S. Food and Drug Administration administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. Food and Drug Administration. XS acknowledges support from the Hightech Agenda Bayern.

References

1. Ahn, E., Kumar, A., Feng, D., Fulham, M., Kim, J.: Unsupervised feature learning with k-means and an ensemble of deep convolutional neural networks for medical image classification. arXiv preprint arXiv:1906.03359 (2019)
2. Barragán-Montero, A., et al.: Artificial intelligence and machine learning for medical imaging: a technology review. Physica Med. **83**, 242–256 (2021)
3. Deng, L.: The MNIST database of handwritten digit images for machine learning research. IEEE Sig. Process. Mag. **29**(6), 141–142 (2012)
4. Gavrielides, M.A., Gallas, B.D., Lenz, P., Badano, A., Hewitt, S.M.: Observer variability in the interpretation of HER2/*neu* immunohistochemical expression with unaided and computer-aided digital microscopy. Arch. Pathol. Lab. Med. **135**(2), 233–242 (2011). https://doi.org/10.5858/135.2.233
5. Gossmann, A., Cha, K.H., Sun, X.: Performance deterioration of deep neural networks for lesion classification in mammography due to distribution shift: an analysis based on artificially created distribution shift. In: Medical Imaging 2020: Computer-Aided Diagnosis, vol. 11314, p. 1131404. SPIE (2020). https://doi.org/10.1117/12.2551346
6. Jiang, Z., Zheng, Y., Tan, H., Tang, B., Zhou, H.: Variational deep embedding: an unsupervised and generative approach to clustering. In: IJCAI (2017)
7. Kart, T., Bai, W., Glocker, B., Rueckert, D.: DeepMCAT: large-scale deep clustering for medical image categorization. In: Engelhardt, S., et al. (eds.) DGM4MICCAI/DALI -2021. LNCS, vol. 13003, pp. 259–267. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-88210-5_26
8. Keay, T., Conway, C.M., O'Flaherty, N., Hewitt, S.M., Shea, K., Gavrielides, M.A.: Reproducibility in the automated quantitative assessment of HER2/neu for breast cancer. J. Pathol. Inform. **4**(1), 19 (2013)
9. Kim, D.W., Jang, H.Y., Kim, K.W., Shin, Y., Park, S.H.: Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. Korean J. Radiol. **20**(3), 405–410 (2019). https://doi.org/10.3348/kjr.2019.0025

10. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: ICLR (2013). arxiv.org/abs/1312.6114v10
11. Oakden-Rayner, L., Dunnmon, J., Carneiro, G., Re, C.: Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: CHIL 2020, pp. 151–159. ACM (2020). https://doi.org/10.1145/3368555.3384468
12. Perkonigg, M., Sobotka, D., Ba-Ssalamah, A., Langs, G.: Unsupervised deep clustering for predictive texture pattern discovery in medical images. arXiv preprint arXiv:2002.03721 (2020)
13. Vokinger, K.N., Feuerriegel, S., Kesselheim, A.S.: Mitigating bias in machine learning for medicine. Commun. Med. **1**(1), 25 (2021). https://doi.org/10.1038/s43856-021-00028-w
14. Xie, J., Girshick, R., Farhadi, A.: Unsupervised deep embedding for clustering analysis. In: Balcan, M.F., Weinberger, K.Q. (eds.) Proceedings of the 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, New York, USA, vol. 48, pp. 478–487. PMLR (2016). https://proceedings.mlr.press/v48/xieb16.html
15. Yu, A.C., Mohajer, B., Eng, J.: External validation of deep learning algorithms for radiologic diagnosis: a systematic review. Radiol. Artif. Intell. **4**(3), e210064 (2022). https://doi.org/10.1148/ryai.210064