Using Cluster Analysis to Assess the Impact of Dataset Heterogeneity on Deep Convolutional Network Accuracy: A First Glance

Conference Paper · October 2019

CITATIONS
0

READS
73

4 authors, including:

Mauro Méndez
Costa Rican Institute of Technology (ITCR)
5 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Pascal N Tyrrell
University of Toronto
111 PUBLICATIONS **3,046** CITATIONS
SEE PROFILE

Saúl Calderón Ramírez
De Montfort University
28 PUBLICATIONS **48** CITATIONS
SEE PROFILE

Using Cluster Analysis to Assess the Impact of Dataset Heterogeneity on Deep Convolutional Network Accuracy: A First Glance

M. Mendez[1], S. Calderon[1], and P. N. Tyrrell[2]

[1] School of Computing, Costa Rica Institute of Technology, Costa Rica
mamendez@ic-itcr.ac.cr, sacalderon@itcr.ac.cr
[2] Departments of Medical Imaging and Statistical Sciences, University of Toronto
pascal.tyrrell@utoronto.ca

Abstract. In this paper we performed cluster analysis using Fuzzy K-means over the image-based features of two models, to assess how dataset heterogeneity impacts model accuracy. A highly heterogeneous dataset is linked with sparse data samples, which usually impacts the overall model generalization and accuracy with test samples. We propose to measure the Coefficient of Variation (CV) in the resulting clusters, to estimate data heterogeneity as a metric for predicting model generalization and test accuracy. We show that highly heterogeneous datasets are common when the number of samples are not enough, thus yielding a high CV. In our experiments with two different models and datasets, higher CV values decreased model test accuracy considerably. We tested ResNet 18, to solve binary classification of x-ray teeth scans, and VGG16, to solve age regression from hand x-ray scans. Results obtained suggest that cluster analysis can be used to identify heterogeneity influence on CNN model testing accuracy. According to our experiments, we consider that a CV $< 5\%$ is recommended to yield a satisfactory model test accuracy.

Keywords: Cluster Analysis · Heterogeneity · Transfer Learning · Small Dataset · Convolutional Neural Network

1 Introduction

The latest advances in medicine and computer science have resulted in the Precision Medicine (PM) field, where several problems in medicine, involving images, patients and medical data are solved by state-of-the-art computer science techniques and can be adequate to individual cases [9]. For instance, physicians and oncologists from all over the world deal with cancer patients every day using imaging techniques due to its non-invasive nature, low risk and cost [9]. X-ray Computed Tomography (CT) is a regularly used imaging technique that measures tissue density at high resolution and exhibition of strong contrasts among different tissue types [34]. In the past few years different initiatives in PM have been solved by making use of Convolutional Neural Networks (CNN). Sampaio et al talks about how segmentation and feature extraction of mammograms are

done by a CNN architecture in order to identify tumorous masses [36]. Liang et al works on a CNN used to automatically classify single cells in thin blood smears on standard microscope slides as either malaria infected or uninfected [30]. Whereas in some cases CNNs can perform in astounding ways, these models have several issues that are far from resolved: hyper-parameter optimization, spatial information loss, cost effective and high data dependency in both quality and quantity [43, 37, 2].

Due to its high data dependency, CNN models are at risk of being overfit or underfit to the training dataset and therefore usually require the process of tweaking hyper-parameters in order to obtain the best fit. However, when a small dataset is used for training, most of Artificial Intelligence (AI)-models have a hard time to overcome a space with limited samples, which translates in to a highly heterogeneous dataset. This usually results in an unsatisfactory testing accuracy [19]. Literature recommends enlarging the dataset, which can be as simple as collecting more samples, involving costs on experts to label the data; or using artificial data augmentation approaches. Han and Le Guennec et al show several techniques such as: spatial and morphological transformations, slicing, zooming, noising and filtering, in order to augment the dataset and improve model accuracy [21, 29]. Recently, new approaches have been applied to generate new samples using generative adversarial networks [16, 4]. These models aim to learn the data distribution and are able to create new samples that were never used on training. However, when we are using a medical imaging dataset it is often not possible to obtain more data given the availability of specific patient types, accurate labeling and high data generation cost. Nevertheless, if it is possible to enlarge the dataset, variation (i.e. heterogeneity) of data points is often ignored at model training. Handling an homogeneous dataset could yield a model that performs poorly on new very atipic data points, whereas having a too heterogeneous but small dataset could limit the learning of the model resulting in a poor testing accuracy for outlier points. [20, 32]

This paper assess CNN model test accuracy with small datasets using cluster analysis metrics such as the Coefficient of Variation (CV), of image-based features. By evaluating patient clusters that differed in accuracy, dataset-oriented decisions can be done to improve model test accuracy by, for example, adding more layers to the CNN model and tweaking hyper-parameters for reducing overfitting on small heterogeneous sample data sets. Subsection 1.1 refers to the implications that entails using a small dataset to train a CNN model and Subsection 1.2 details how cluster analysis can help to identify the problem. Later, on Section 2 the proposed method is explained and is tested on Section 3 of experiments and results, where the correlation between the proposed CV data heterogeneity metric and model test accuracy is done. Finally, main conclusions and future work is addressed in Section 4.

1.1 Importance of data

Researchers on AI and Data Science often face dataset related short-comings. A common issue is small dataset size, leading to overfitting models and non-

optimal solutions. Medical imaging datasets often require complex labels, making labeling a time-consuming and expensive task, which often prevents yielding not enough training samples for the AI model [43]. Several Machine Learning (ML) architectures have been successfully applied on small datasets, where it is shown that supervised fine-tuning with a relatively small dataset on a network pre-trained with a large image dataset of generic objects (e.g., ResNet [22], VGG16 [38]) can lead to significant improvement in performance; an approach known as transfer learning [33, 8, 12, 18, 44]. However, using transfer learning does not guarantee a model that performs well for every sample, due to sample size, the observed data points and outliers, which can be linked to dataset heterogeneity.

Data heterogeneity is often analyzed in medical studies from three perspectives: the clinical perspective, referring to data variation from observed subjects. When data heterogeneity is generated from study procedures, it is referred as methodological heterogeneity. Finally, statistical heterogeneity refers to variation on study measurements [14]. In this paper we focus in two applications with clinical and statistical heterogeneity.

Medical heterogeneous datasets have been investigated extensively [5, 39]. For instance, Altman et al assesses heterogeneity on epidemiological clinical trials, aiming to identify different subgroups of patients, to find whether the observed relationship between an exposure and disease is different among these subgroups [3].

For an AI model, a relatively large heterogeneous dataset is desired to train a model, as it can learn from a lot of samples that introduce variation, making the model to adjust better for the problem to estimate, i.e. improving model generalization, as it is better trained for samples that were not used on the training stage. A relatively small heterogeneous dataset forces the model to adjust for the observed sample set, yielding poor accuracy for samples not used in the training stage, i.e. yielding a reduced generalization. Thus, in order to mitigate the effect of heterogeneity from small sample sizes, statistical approaches are used by Frantziskonis and Wardenaar et al where methods such as: latent class analysis, parametric functions, factor analyses and mixture growth analyses, attempt to change or remove heterogeneous samples to build a more homogeneous dataset [15, 40].

1.2 Cluster analysis of features as a mean to assess the impact of heterogeneity

Cluster analysis is the formal study of methods and algorithms for grouping or clustering objects, according to measured or perceived intrinsic characteristics or similarity [25]. There are several techniques to apply clustering analysis on a dataset, used to identify low-heterogeneous subgroups of data with common features [26]. In the medical field, cluster analysis is often used as a way to understand clusters of patients and improve medical diagnosis [41, 23]. For instance, Fitzpatrick et al uses ward hierarchical clustering to identify asthma phenotypes, applied to school-age children with persistent asthma across a wide range of severities, giving as outcome 5 significant clusters [13].

Whereas literature have focused lately on CNN models and cluster analysis, the powerful mixture between both has not yet been fully addressed in literature, specially in medical imaging analyisis. Xu et al addresses sparseness of text representation with a CNN architecture used to extract features from word embeddings to later cluster them with K-means. The model yields a short text classification pipeline capable of outperform related works [42]. Donahue et al works on semantic clustering on trained CNN's features in order to create a framework for semi-supervised learning [12]. Moreover, no work could be found using both methods and the study of heterogeneity on medical imaging data. Although, similar work has been addressed on the statistical calibration of models as the Hosmer-Lemeshow test [24], where subgroups from a numerical dataset are identified to assess the goodness of fit for logistic regression on small sample sizes.

Sample dimensionality is an issue when clustering is performed [31]. High dimensional samples impairs clustering algorithms performance due to the decreasing significance of the clusters yielded in high dimensional spaces, linked to the curse of dimesionality. In this work the clustered samples are features extracted from a CNN. Moreover, image-based features are often of high dimensionality, making the use of a dimensionality reduction method necessary, like Principal Component Analysis (PCA).

2 Proposed method

In this work, we aim to evaluate how well a CNN model deals with heterogeneity on small and large datasets using clustering analysis as a tool to measure its impact, using a cluster quantitative measure as the CV. Common CNN-transfer learning based model can be divided in two sections: back model and top model. The back model contains all the convolutional calculation and feature extraction filters, whereas the top model implements data transformations to yield the intended predictions, based on the features from previous stage [19].

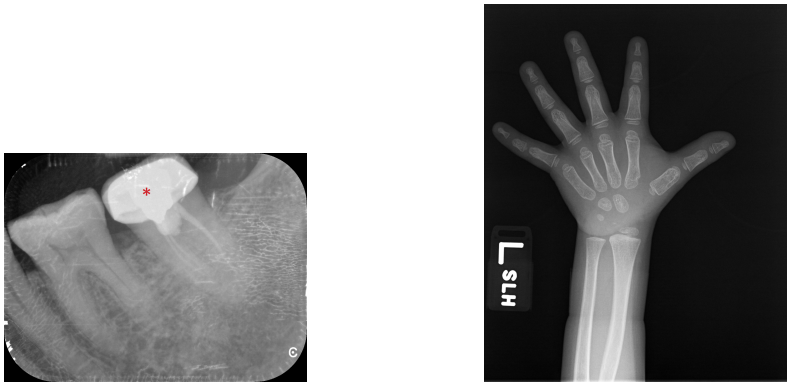
This research proposes to perform cluster analysis on the principal components of the back model outputs. Thus, we propose to reduce the original feature space created by the convolutional layers using PCA. We performed PCA assuming that medical data is often generated from normal distributions [27], creating a new feature space with the highest dataset covariance dimensions.

We propose to evaluate model accuracy in the clustered data, using the predictions of the top model. Both the back and top model need to be trained before applying the proposed method.

For the cluster analysis, an exploration on the number of clusters and the algorithm that perform the best for the extracted data was made. If every cluster performs similarly as if the whole dataset was evaluated, we can infer that the back model was able to generalize well, i.e. the heterogeneity in the dataset did not affect model test accuracy. However, differences between the clusters accuracies and the whole dataset are found, suggesting a back model unable to generalize the distribution for every sample given, i.e. data heterogeneity affected

the model performance. We propose to use CV [1] as a quantitative normalized measure that takes into account the accuracy for each cluster. CV measures the dispersion on data points, in our case, our data points correspond to cluster's accuracy; if dispersion is relatively high, it means that there is a significant difference in model performance. On the other hand, a relatively low dispersion is a sign of well-generalized training.

We aim to measure and find an appropriate CV interval, which can ensure enough data homogeneity and subsequent model generalization. We tested two models with different datasets and objectives, predicting classification and regression. VGG16 and Resnet18 are popular architectures, the models were selected for their transfer-learning affine structure, the ability to extract image-based features, training simplicity and dataset availability. Tested datasets belong to the medical imaging domain, a domain where the lack of data is usual, yielding into more small and possible heterogeneous datasets. The models are detailed in sections 2.1 and 2.2.



(a) PSP plate image (b) Bone Age Image

Fig. 1. Image from PSP plates dataset (a). Image from Bone Age dataset (b).

2.1 PSP plates with Resnet18

This model solves the binary classification of each sample between discard the plate (yes) or keep it (no). Due to the number of samples, high resolution images and prediction complexity this model converged easily, making it a good model to try with different sample sizes. A simple metric as the accuracy, number of right decisions over total amount of decisions, will be used.

The first test model was ResNet 18 [22] back model and a fully connected layer as top model. ResNet introduces skip connections in order to avoid the vanishing gradient problem. The features used on this model were the outputs of the last layer of the back model giving 512 image-features to work with.

The dataset for this model is composed by the superimposition of two images, a Physical Photostimulable Phosphor (PSP) plate and a CMOS teeth scan. The dataset was built at the Faculty of Dentistry at the University of Toronto (Toronto, Canada) using a Carestream CS 7600 for the former, MiPACS and Carestream RVG 6200 for the latter. A selection of 25 PSP plates were mixed with 100 cases of CMOS scans giving a total of 2928 samples. The 25 PSP plates consists of 10 severe damage plates, 10 with intermediate damage, 4 new plates, and one blank mask. As well, 25 dentists with at least 1 year of experience labeled the scans in two categories: keep the plate or discard it. The dataset is composed by teeth x-ray scans of 1152×869, as reference see figure 1.

2.2 Bone Age with VGG16

This model predicts the bone age for each subject, solving a regression problem. The metric used for this model was the 1-Normalized Root Mean Squared Error (1-NRMSE).

The model implemented is based in a VGG16 [38] back model, and a 4-layers fully connected model as top model [6]. VGG16 uses the smallest filter size capable of encoding directional information, i.e. 3×3 filters. The filters are used along the whole network allowing to learn the same information as the larger filters used on other networks. This feature allows VGG16 to train with significantly fewer parameters. As the previous model, 512 image-features are extracted.

The dataset used in this model consists of radiographs from left hands of both male and female subjects with different races and ages ranging from 1 month to 228 months (19 years) for a total of 12600 samples. The Radiological Society of North America made the dataset publicly available, and was acquired from Stanford Children's Hospital and Colorado Children's Hospital [17], as reference see figure 1. For the following experiments because of the computation complexity on training with thousands of images and the number of tests that we had to run, just the female radiographs were used giving a total of 5674 images to work with.

3 Experiments and Results

Four sample sizes on each model were selected to show the increase of accuracy and the expected decrease of heterogeneity; for model 2.1 the sample sizes were 100, 300, half (1464) and the whole dataset (2928), as well, for model 2.2, 300, 700, half (2837) and the whole dataset (5674).

The models were trained using 4-fold validation for every sample size, to minimize data split randomness and keep computational requirements reasonable. For every fold the image-based features (back model outputs) were extracted as well as the model prediction (top model outputs). Cluster analysis was performed on every fold and the metrics were averaged to give an overall representation for

the sample size. The metrics were calculated using the model prediction and were performed separately for every cluster.

The extracted image-based features were reduced from their original amount to 10 principal components using PCA. Later, clustering was performed on principal component based space. Dimensionality reduction was performed to attenuate the curse of dimensionality and minimization of correlated features. For cluster analysis, we picked a clustering algorithm and the number of clusters. The latter was unknown as clustering was performed on the extracted image features with PCA dimensionality reduction, where the number of classes is unknown. We tested different clustering algorithms: K-Means, Fuzzy K-Means and Gaussian Mixture Model using unsupervised clustering performance metrics such as Elbow Method [28], Silhouette Coefficient [35], Calinski-Harabaz Index [7], Davies-Bouldin Index [11] and amount of patients in each cluster. For both models Fuzzy K-Means with 4 clusters were found to give the most demonstrative results for all combinations.

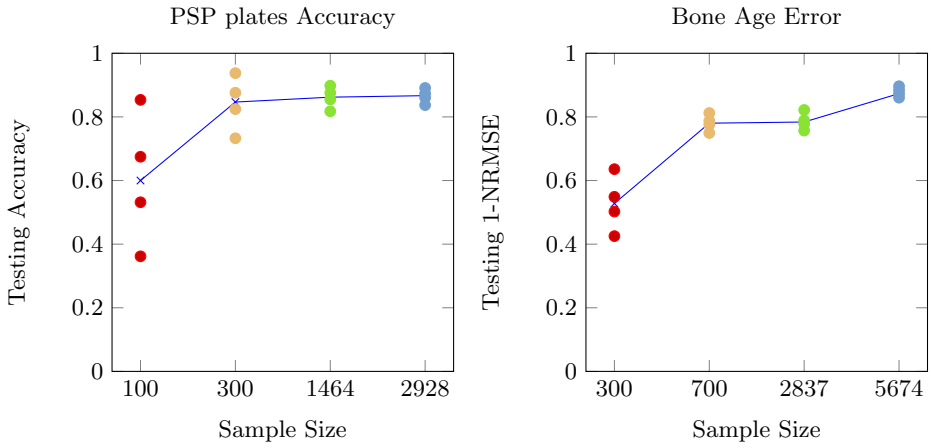


Fig. 2. PSP plates clusters testing accuracy over different sample sizes (left). Bone Age clusters testing error over different sample sizes (right). The blue line measure the whole dataset.

Due to the stochastic nature of clustering, several trainings of the algorithm were made in order to average accurate clustering metrics. The way that a cluster algorithm assigns labels is arbitrarily, meaning that no criteria is used to label them making it random within each training. Thus, the algorithm was trained 10 times and cluster identifying and re-labeling was performed using the cluster patients and accuracy; in that way, each cluster had the same label across trainings. Every metric shown is the average of the several runs.

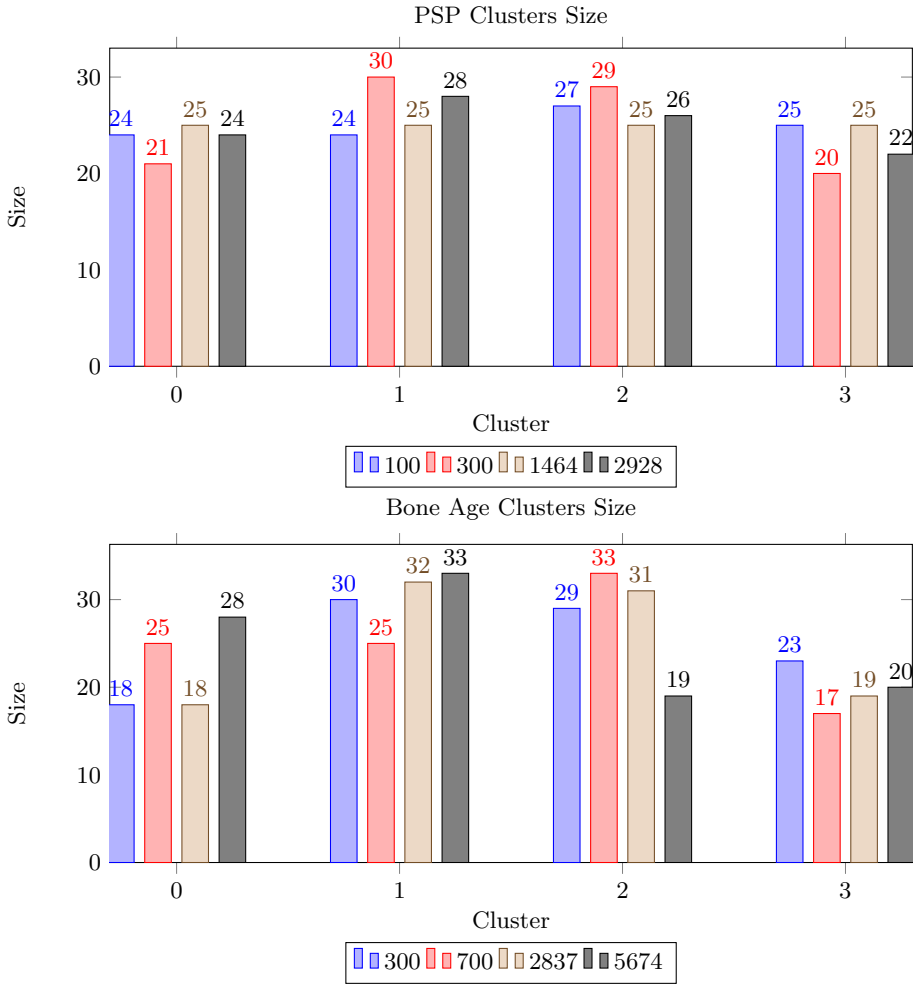


Fig. 3. PSP plates clusters' size over different sample sizes (top). Bone Age clusters' size over different sample sizes (bottom). The amount is shown as a percentage of the whole sample size. The clusters were made using Fuzzy K-Means.

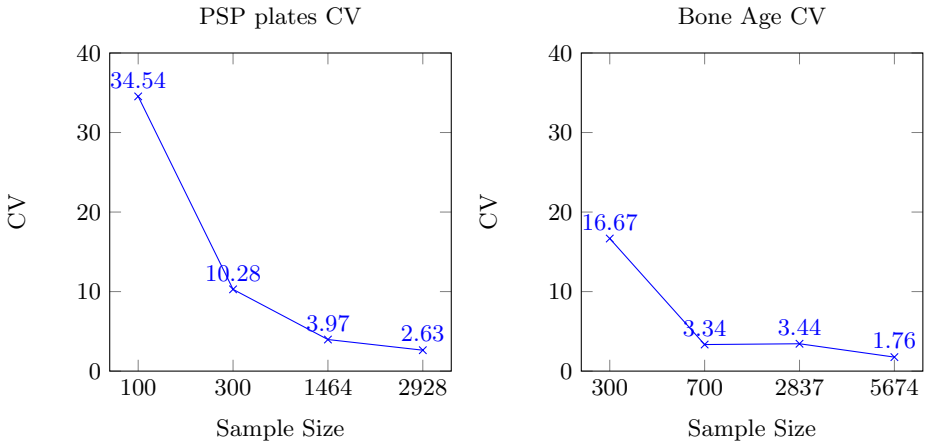


Fig. 4. The Coefficient of Variation for PSP plates clusters over different sample sizes (left). The Coefficient of Variation for Bone Age clusters over different sample sizes (right).

In Figure 2 is shown the evaluated clusters for each model using the selected sample sizes. In this figure is demonstrated that there was a relationship between sample size and heterogeneity that affected model test accuracy. The more data you have the better the model will generalize for every cluster of samples, reducing the negative effect of its heterogeneity.

Figure 3 displays clusters' sizes in percentage for each sample size, showing that Fuzzy K-Means is able to find significant clusters with characteristics in common and similar number of patients between each other. A bad clustering algorithm would give unrealistic clusters with few isolated patients that do not reflect real-world data samples.

Figure 4 shows the level of variation between clusters over the sample sizes. In clinical chemistry literature a CV $< 10\%$ is very good, $10\% - 20\%$ is good, $20\% - 30\%$ is acceptable, and CV $> 30\%$ is not acceptable [10]. The plots show that CV stability is achieved when more samples make relative small change to the dispersion of the clusters, that can translate into a model that performs almost the same for all clusters of samples.

Figure 5 shows the correlation between model accuracy and the CV for both models. In this figure is demonstrated how we are able to estimate the accuracy or error given a CV and viceversa. Both models show negative correlation with a Pearson Coefficient of -0.9871 for the PSP plates model and -0.9832 for the Bone Age model. A negative correlation means that the lower CV we have, the higher testing accuracy our model will get. Thus, if we want our model to perform over a certain accuracy, we have to achieve a generalization that gives a favorable CV. Although, 4 points are not enough for calculating a reliable coefficient, it

shows an expectable negative trend. From the results in Figure 5, we can infer that a CV lower than 5% is desirable in order to keep testing accuracy high.

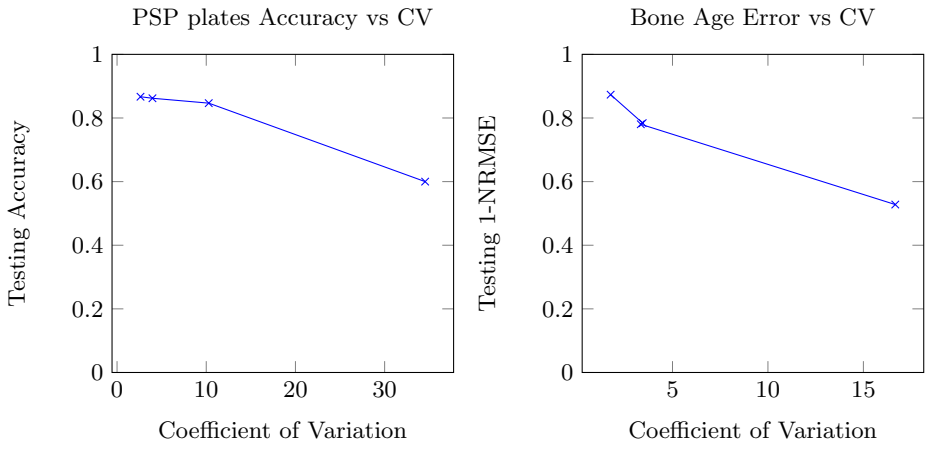


Fig. 5. The PSP plates Testing Accuracy over its Coefficient of Variation (left). The Bone Age Testing Error over its Coefficient of Variation (right).

Table 1. Centroid and Closest-Point euclidean distances between each cluster.

Model		PSP plates with Resnet18						Bone Age with VGG16						
Data Sizes / Clusters		0-1	0-2	0-3	1-2	1-3	2-3		0-1	0-2	0-3	1-2	1-3	2-3
Centroid Distance	100	10.88	10.53	11.06	10.83	10.13	10.05	300	5.12	5.07	6.2	5.22	5.94	5.96
	300	10.91	10.30	10.16	10.57	10.79	10.74	700	13.06	11.92	9.27	9.42	11.26	9.76
	1464	11.26	10.29	9.82	10.59	10.98	10.21	2837	9.53	10.49	10.79	10.25	9.42	10.16
	2928	11.48	11.34	10.83	11.75	11.06	12.38	5674	11.5	10.72	10.52	11.21	11.67	11.65
Closes-Point Distance	100	6.63	6.74	6.7	6.79	6.18	6.98	300	1.44	1.48	1.49	1.13	1.26	1.29
	300	4.89	4.69	4.92	4.28	4.41	4.55	700	1.62	1.26	1.24	1.6	1.63	1.36
	1464	2.64	2.61	2.62	2.67	2.82	2.58	2837	1.47	1.22	1.79	1.20	1.31	1.57
	2928	2.01	1.96	2.02	1.91	2.14	2.14	5674	1.04	1.19	1.52	1.02	1.48	2.34

We calculated the inter-cluster and intra-cluster Euclidean distances to analyze cluster quality and consistency. Table 1 shows how the distances between cluster change over sample sizes. On the other hand, Table 2 shows how clusters change when we increase the sample size. Both tables show the behavior of the clusters with increasing sample sizes. Within a specific sample size, the centroids appear to remain relatively fixed; however, as sample size increases, clusters grow in spread and get closer to each other until it becomes almost one big cluster. In a well-trained model every sample cluster should have almost the

Table 2. Euclidean distance between the cluster centroids and their points.

Model		PSP with Resnet18					Bone Age with VGG16			
Clusters / Distance		Min	Max	Mean	StDv		Min	Max	Mean	StDv
Cluster 0	100	5.65	18.38	11.52	3.02	300	2.49	21.40	6.10	3.80
	300	5.73	21.52	11.97	3.37	700	2.62	43.66	8.11	5.61
	1464	4.70	26.51	12.22	3.68	2837	4.10	56.36	11.36	6.45
	2928	4.42	29.84	11.67	3.48	5674	3.91	76.97	12.73	7.39
Cluster 1	100	6.55	18.85	11.92	3.3	300	2.44	27.51	6.81	4.35
	300	5.72	21.81	11.77	3.24	700	1.89	45.92	7.95	6.27
	1464	4.83	26.24	11.88	3.61	2837	3.01	78.48	9.69	6.08
	2928	4.41	31.11	12.11	4.03	5674	3.26	90.62	11.69	7.61
Cluster 2	100	6.68	21.95	13.38	3.65	300	2.38	27.46	7.08	4.38
	300	5.51	21.03	11.91	3.23	700	1.37	42.81	5.08	4.17
	1464	4.84	25.85	12.04	3.59	2837	3.72	80.60	12.83	7.30
	2928	4.14	30.49	11.49	3.94	5674	3.06	70.42	9.74	5.47
Cluster 3	100	7.55	22.75	13.82	3.58	300	2.44	23.73	6.57	3.93
	300	5.93	21.06	11.93	3.33	700	2.45	48.78	7.85	7.47
	1464	4.70	24.91	11.75	3.48	2837	3.70	53.70	11.47	6.19
	2928	4.40	27.94	11.60	3.78	5674	3.66	63.21	10.15	6.23

same accuracy as others; this is the result of how image-based extracted features improve its representation, and also the dataset increases its homogeneity when we more training samples are avaliable.

4 Conclusions and future work

All figures showed that cluster analysis, using the right algorithm and number of clusters, is an effective way to identify and assess how affected the model testing accuracy is by dataset heterogeneity, in classification and regression problems. Moreover, CV can be used as a predictor of model testing accuracy and viceversa. We consider a dataset with a CV of less than 5%, homogeneous enough to allow the model generalize data properly and avoid overfitting.

Cluster analysis could be used as a tool to identify if the CNN model is affected by heterogeneity from the dataset fed; even if your model has a high accuracy it can still perform poorly for a cluster of samples and this method may serve to identify it. Furthermore, this approach could assist to calibrate a model to be able to perform well for every sample cluster, or discard under-performing data clusters.

As future work, we aim to experiment on more datasets and more models generalizing the application of using cluster analysis to identify heterogeneity affection on model training for a CNN transfer learning problem, obtaining more evidence of our claim.

Another approach proposed as future work is the tracking and identification of the image-based features belonging to a principal component, and identify which features define a good cluster of images. Understanding how features are

related to patient samples can be useful to gain insight on how CNNs extract information from data. As well, knowing which features produce a better patient cluster can be useful to improve those samples that did not perform well.

References

1. Abdi, H.: Coefficient of variation. Encyclopedia of research design **1**, 169–171 (2010)
2. Ahmadvand, P., Ebrahimpour, R., Ahmadvand, P.: How popular cnns perform in real applications of face recognition. In: 2016 24th Telecommunications Forum (TELFOR). pp. 1–4. IEEE (2016)
3. Altman, D.G., Matthews, J.N.: Statistics notes: Interaction 1: heterogeneity of effects. Bmj **313**(7055), 486 (1996)
4. Antoniou, A., Storkey, A., Edwards, H.: Data augmentation generative adversarial networks. stat **1050**, 8 (2018)
5. Bowden, J., Tierney, J.F., Copas, A.J., Burdett, S.: Quantifying, displaying and accounting for heterogeneity in the meta-analysis of rcts using standard and generalised q statistics. BMC medical research methodology **11**(1), 41 (2011)
6. Calderon, S., Fallas, F., Zumbado, M., Tyrrell, P., Stark, H., Emersic, Z., Meden, B., Solis, M.: Assessing the impact of the deceived non local means filter as a preprocessing stage in a convolutional neural network based approach for age estimation using digital hand x-ray images. In: 2018 25th IEEE International Conference on Image Processing (ICIP). pp. 1752–1756. IEEE (2018)
7. Caliński, T., Harabasz, J.: A dendrite method for cluster analysis. Communications in Statistics-theory and Methods **3**(1), 1–27 (1974)
8. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: Delving deep into convolutional nets. CoRR **abs/1405.3531** (2014)
9. Collins, F.S., Varmus, H.: A new initiative on precision medicine. New England Journal of Medicine **372**(9), 793–795 (2015)
10. Cui, Z.: Allowable limit of error in clinical chemistry quality control. Clinical chemistry **35**(4), 630–631 (1989)
11. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence (2), 224–227 (1979)
12. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: International conference on machine learning. pp. 647–655 (2014)
13. Fitzpatrick, A.M., Teague, W.G., Meyers, D.A., Peters, S.P., Li, X., Li, H., Wenzel, S.E., Aujla, S., Castro, M., Bacharier, L.B., et al.: Heterogeneity of severe asthma in childhood: confirmation by cluster analysis of children in the national institutes of health/national heart, lung, and blood institute severe asthma research program. Journal of allergy and clinical immunology **127**(2), 382–389 (2011)
14. Fletcher, J.: What is heterogeneity and is it important? Bmj **334**(7584), 94–96 (2007)
15. Frantziskonis, G.: Heterogeneity and implicated surface effects: statistical, fractal formulation and relevant analytical solution. Acta mechanica **108**(1-4), 157–178 (1995)
16. Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: Synthetic data augmentation using gan for improved liver lesion classification. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 289–293. IEEE (2018)

17. Gertych, A., Zhang, A., Sayre, J., Pospiech-Kurkowska, S., Huang, H.: Bone age assessment of children using a digital hand atlas. Computerized Medical Imaging and Graphics **31**(4-5), 322–331 (2007)
18. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 580–587 (2014)
19. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016)
20. Guibas, J.T., Virdi, T.S., Li, P.S.: Synthetic medical images from dual generative adversarial networks. CoRR **abs/1709.01872** (2017), http://arxiv.org/abs/1709.01872
21. Han, D., Liu, Q., Fan, W.: A new image classification method using cnn transfer learning and web data augmentation. Expert Systems with Applications **95**, 43–56 (2018)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
23. Hervier, B., Devilliers, H., Stanciu, R., Meyer, A., Uzunhan, Y., Masseau, A., Dubucquoi, S., Hatron, P.Y., Musset, L., Wallaert, B., et al.: Hierarchical cluster and survival analyses of antisynthetase syndrome: phenotype and outcome are correlated with anti-trna synthetase antibody specificity. Autoimmunity reviews **12**(2), 210–217 (2012)
24. Hosmer Jr, D.W., Lemeshow, S., Sturdivant, R.X.: Applied logistic regression, vol. 398. John Wiley & Sons (2013)
25. Jain, A.K.: Data clustering: 50 years beyond k-means. Pattern recognition letters **31**(8), 651–666 (2010)
26. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM computing surveys (CSUR) **31**(3), 264–323 (1999)
27. Jolliffe, I.: Principal component analysis. Springer (2011)
28. Kodinariya, T.M., Makwana, P.R.: Review on determining number of cluster in k-means clustering. International Journal **1**(6), 90–95 (2013)
29. Le Guennec, A., Malinowski, S., Tavenard, R.: Data augmentation for time series classification using convolutional neural networks. In: ECML/PKDD workshop on advanced analytics and learning on temporal data (2016)
30. Liang, Z., Powell, A., Ersoy, I., Poostchi, M., Silamut, K., Palaniappan, K., Guo, P., Hossain, M.A., Sameer, A., Maude, R.J., et al.: Cnn-based image analysis for malaria diagnosis. In: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 493–496. IEEE (2016)
31. Liu, Y., Hayes, D.N., Nobel, A., Marron, J.: Statistical significance of clustering for high-dimension, low–sample size data. Journal of the American Statistical Association **103**(483), 1281–1293 (2008)
32. Neff, T., Payer, C., Stern, D., Urschler, M.: Generative adversarial network based synthesis for supervised medical image segmentation. In: Proc. OAGM and ARW Joint Workshop (2017)
33. Ng, H.W., Nguyen, V.D., Vonikakis, V., Winkler, S.: Deep learning for emotion recognition on small datasets using transfer learning. In: Proceedings of the 2015 ACM on international conference on multimodal interaction. pp. 443–449. ACM (2015)
34. Parmar, C., Leijenaar, R.T., Grossmann, P., Velazquez, E.R., Bussink, J., Rietveld, D., Rietbergen, M.M., Haibe-Kains, B., Lambin, P., Aerts, H.J.: Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer. Scientific reports **5**, 11044 (2015)

35. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20**, 53–65 (1987)
36. Sampaio, W.B., Diniz, E.M., Silva, A.C., De Paiva, A.C., Gattass, M.: Detection of masses in mammogram images using cnn, geostatistic functions and svm. Computers in biology and medicine **41**(8), 653–664 (2011)
37. Severyn, A., Moschitti, A.: Twitter sentiment analysis with deep convolutional neural networks. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 959–962. ACM (2015)
38. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
39. Terrin, N., Schmid, C.H., Lau, J., Olkin, I.: Adjusting for publication bias in the presence of heterogeneity. Statistics in medicine **22**(13), 2113–2126 (2003)
40. Wardenaar, K.J., de Jonge, P.: Diagnostic heterogeneity in psychiatry: towards an empirical solution. BMC medicine **11**(1), 201 (2013)
41. Wirapati, P., Sotiriou, C., Kunkel, S., Farmer, P., Pradervand, S., Haibe-Kains, B., Desmedt, C., Ignatiadis, M., Sengstag, T., Schütz, F., et al.: Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. Breast Cancer Research **10**(4), R65 (2008)
42. Xu, J., Peng, W., Guanhua, T., Bo, X., Jun, Z., Fangyuan, W., Hongwei, H., et al.: Short text clustering via convolutional neural networks (2015)
43. Yamashita, R., Nishio, M., Do, R.K.G., Togashi, K.: Convolutional neural networks: an overview and application in radiology. Insights into imaging **9**(4), 611 (2018)
44. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Advances in neural information processing systems. pp. 3320–3328 (2014)