Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Automated Detection of Poor-Quality Data: Case Studies in Healthcare

M.A. Dakka (milad@presagen.com)
Presagen

T. Nguyen
Presagen

J.M.M. Hall
Presagen

S.M. Diakiw
Presagen

M. VerMilyea
Ovation Fertility

R. Linke
Department of Medical Imaging - SAMI, Women's and Children's Hospital Campus

M. Perugini
Presagen

D. Perugini
Presagen

Research Article

Keywords: Poor-Quality Data, Healthcare, Case Studies, AI models, support a diagnosis

Posted Date: April 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-440365/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Automated Detection of Poor-Quality Data: Case Studies in Healthcare

M.A. Dakka[1,2,*], T. Nguyen[1,3], J.M.M. Hall[1,4], S.M. Diakiw[1], M. VerMilyea[5,6], R. Linke[7], M. Perugini[1,8], and D. Perugini[1,*]

[1]Presagen, Adelaide, SA 5000, Australia
[2]School of Mathematical Sciences, The University of Adelaide, Adelaide, SA 5000, Australia
[3]School of Computing and Information Technology, University of Wollongong, Wollongong, NSW 2522, Australia
[4]Australian Research Council Centre of Excellence for Nanoscale BioPhotonics, Adelaide, SA 5000, Australia
[5]Ovation Fertility, Austin, TX 78731, USA
[6]Texas Fertility Center, Austin, TX 78731, USA
[7]Department of Medical Imaging – SAMI, Women's and Children's Hospital Campus, Adelaide, SA, 5000, Australia
[8]Adelaide Medical School, The University of Adelaide, Adelaide, SA 5000, Australia
[*]Corresponding authors: milad@presagen.com, don@presagen.com

ABSTRACT

The detection and removal of poor-quality data in a training set is crucial to achieve high-performing AI models. In healthcare, data can be inherently poor-quality due to uncertainty or subjectivity, but as is often the case, the requirement for data privacy restricts AI practitioners from accessing raw training data, meaning manual visual verification of private patient data is not possible. Here we describe a novel method for automated identification of poor-quality data, called Untrainable Data Cleansing. This method is shown to have numerous benefits including protection of private patient data; improvement in AI generalizability; reduction in time, cost, and data needed for training; all while offering a truer reporting of AI performance itself. Additionally, results show that Untrainable Data Cleansing could be useful as a triage tool to identify difficult clinical cases that may warrant in-depth evaluation or additional testing to support a diagnosis.

Advances in deep learning using artificial neural networks (ANN)[1,2] have resulted in the increased use of AI for healthcare applications[3–5]. One of the most successful examples of deep learning has been the application of convolutional neural network (CNN) algorithms for medical image analysis to support clinical assessment[6]. AI models are trained with labeled or annotated data (medical images) and learn complex features of the images that relate to a clinical outcome, which can then be applied to classify new unseen medical images. Applications of this technology in healthcare span a wide range of domains including but not limited to dermatology[7,8], radiology[9,10], ophthalmology[11–13], pathology[14–16], and embryo quality assessment in IVF[17].

Despite the enormous potential of AI to improve healthcare outcomes, AI performance can often be sub-optimal as it is crucially dependent on the quality of the data. While AI practitioners often focus on the quantity of data as the driver of performance, even fractional amounts of poor-quality data can substantially hamper AI performance. Good-quality data is therefore needed to train models that are both accurate and generalizable, and which can be relied upon by clinics and patients globally. Furthermore, because measuring AI performance on poor-quality test data can mislead or obfuscate the true performance of the AI, good-quality data is also important for benchmark test sets used in performance reporting, which clinics and patients rely on for clinical decisioning.

We define two types of poor-quality data:

- **Incorrect data**: Mislabeled data, for example an image of a dog incorrectly labeled as a cat. This also includes adversarial attacks by intentionally inserting errors in data labels (especially detrimental to online machine learning methods[18]).
- **Noisy data**: Data itself is of poor quality (e.g. out-of-focus image), making it ambiguous or uninformative, with insufficient information or distinguishing features to correlate with any label.

In healthcare, clinical data can be inherently poor quality due to subjectivity and clinical uncertainty. An example of this is pneumonia detection from chest x-ray images. The labeling of a portion of the image can be somewhat subjective in terms of clinical assessment, often without a known ground truth outcome, and is highly dependent on the quality of the x-ray image taken. In some cases, the ground truth outcome might also involve clinical data that is not present in the dataset used for analysis, such as undiagnosed conditions in a patient, or effects that cannot be seen from images and records provided for

the assessment. This kind of clinical uncertainty can contribute to both the incorrect and noisy data categories. Therefore, poor-quality data cannot always be reliably detected, even by clinical experts. Furthermore, due to data privacy, manual visual verification of private patient data is not always possible.

Several methods exist to account for these sources of reader variability and bias. One method[19] uses a so-called Direct Uncertainty Prediction to provide an unbiased estimate of label uncertainty for medical images, which can be used to draw attention to images requiring a second medical opinion. This technique relies on training a model to identify cases with high potential for expert disagreement. Other methods model uncertainty in poor-quality data through Bayesian techniques[20]. However, such techniques require significant amounts of annotated data from multiple experts, which is often not readily available. Some methods assume that erroneous label distribution is conditionally independent of the data instance given the true label[21], which is an assumption that does not hold true in the settings considered in this article. Other techniques[22] relax this assumption by using domain-adapted generative models to explain the process that generates poor-quality data, though these techniques typically require additional clean data to generate good priors for learning. This is an issue in medical domains such as the assessment of embryo viability, where reader variability is significant[17] and ground truth labels may be impossible to ascertain, so there is no way of identifying data as poor quality *a priori*. There is a need for better approaches to cleanse poor-quality data automatically and effectively, in and beyond healthcare.

Results

In this paper a novel technique is presented for automated data cleansing which can identify poor data quality without requiring a cleansed dataset with known ground truth labels. The technique is called Untrainable Data Cleansing (UDC) and is described in the Methods section. UDC essentially identifies and removes a subset of the data (i.e. cleanses the data) that AI models are unable to correctly label (classify) during the AI training process. From a machine learning perspective, the two types of poor-quality data described above are realized through: (1) identifying data that are highly correlated to the opposite label of what would reasonably be expected, based on the classification of most of the data in the dataset (incorrect data); or (2) identifying data that have no distinguishing features that correlate with any label (noisy data). Results show that UDC can consistently and accurately identify poor-quality data, and that removal of UDC-identified poor-quality data, and thus "cleansing" of the data, ultimately leads to higher performing and more reliable AI models for healthcare applications.

Validation of UDC

UDC was first validated using two types of datasets, cats and dogs for binary classification, and vehicles for multi-classification. These datasets were used because the ground truth can be manually confirmed, and therefore incorrect labels could be synthetically injected.

Binary Classification Using Cats and Dogs

A benchmark (Kaggle) dataset of 37,500 cat and dog images were used to validate UDC for binary classification. This dataset was chosen because the ground truth outcomes (labels) could be manually determined with certainty, and synthetic incorrect labels could be readily introduced by flipping the correct label to an incorrect label. Synthetic incorrect labels were added to this dataset to test UDC under various amounts of poor-quality data.

A total of 24,916 images (12,453 cats, 12,463 dogs) were used for training and 12,349 images (6,143 cats, 6,206 dogs) used as a separate blind test set. Synthetic errors (incorrect labels) were added to the training dataset (but not the test set), which was split 80/20 into training and validation sets. A single round of UDC was applied to the training dataset, poor-quality data identified by UDC was removed, and a new AI model was trained on the UDC-cleansed dataset. The highest balanced AI accuracy achievable on the blind test dataset was reported. Results are shown in **Table** 1.

Incorrect Label Levels (Cats %, Dogs %)	**(0%, 0%)**	**(35%, 5%)**	**(50%, 5%)**	**(30%, 30%)**
Average Level of Incorrect Labels (%)	0.0	20.0	27.5	30.0
Original Balanced Accuracy (%)	99.2	77.9	72.6	63.1
UDC Balanced Accuracy (%)	–	98.3	98.7	94.7
Accuracy Improvement (%)	–	**20.4**	**26.1**	**31.6**
Images Removed by UDC (Total)	–	4,699	6,757	7,578
Images Removed by UDC (%)	–	18.9	27.1	30.4

Table 1. Results from a single round of UDC applied to various incorrect label levels. Cats % is the percentage of cat images incorrectly labeled as dogs, vice versa. Baseline model is assumed to contain 0% error in both classes, denoted (0%, 0%).

Results show that UDC is resilient to even extreme levels of errors, functioning in datasets with up to 50% incorrect labels in one class (while the other class remained relatively clean), and with significant symmetric errors of up to 30% incorrect labels in both classes. Visual assessment verified removal of both incorrect data and a minor proportion of noisy data, where, for example, dogs looked like cats. Significant improvements of greater than 20% were achieved in all cases. Compared to cases with uniform levels of incorrect labels, the asymmetric incorrect label cases, i.e. (35%, 5%) and (50%, 5%), achieved a higher balanced accuracy of over 98% after just one round of UDC. This is expected, since in the asymmetric cases, one class remains as a true correct class, allowing UDC to be more confident in identifying incorrectly labeled samples.

In the uniform cases a slightly lower balanced accuracy of 94.7% was achieved after one round of UDC, which was found to identify and remove 88% of the intentionally mislabeled images. A second round of UDC improved upon the results of a single application of UDC, successively increasing the predictive power of the dataset by removing the remaining incorrect data. In fact, the accuracy achieved after a second round of UDC (99.7%) to the symmetric case (30%, 30%) showed an improvement even when compared to the baseline accuracy (99.2%) on datasets with 0% synthetic error. We posit that this occurs because UDC filters out noisy data that was likely present in the original clean dataset (since the baseline set itself is not guaranteed to be free of poor-quality data), therefore helping to not only recover but surpass the accuracy of models trained on the baseline (0% poor-quality) datasets after a second round of UDC.

For the symmetrical (50%, 50%) case (not shown), UDC simply learns to treat one entire class as incorrect and the other as correct, thereby discarding all samples from the opposite class as data errors. Therefore, as might be expected, UDC fails when data error levels in both classes approach 50%, because there is no longer a majority of good-quality data to allow UDC to confidently identify the minority of poor-quality data. In this case, the dataset is deemed unsuitable for AI training.

Multi-Classification Using Vehicles

An open image database of 27,363 vehicles was used to validate UDC for multi-classification. The dataset comprised four classes of vehicles: airplanes, boats, motorcycles, and trucks. A total of 18,991 images (7,244 airplanes, 5,018 boats, 3,107 motorcycles, 3,622 trucks) were used for training, and 8,372 images (3,101 airplanes, 2,194 boats, 1,424 motorcycles, 1,653 trucks) used as a separate blind test set. As in the previous section, this dataset was chosen because the ground truth outcomes (labels) could be manually ascertained, and synthetic incorrect labels could be readily introduced. Synthetic incorrect labels were uniformly added to each class in the training dataset in increments of 10% to test UDC under various amounts of poor-quality data.

Results are summarized in **Table** 2, which shows the percentage improvement for all cases after only a single round of UDC, removing both noisy and incorrect labels. Errors are calculated as the standard error on the mean for results obtained from eight models overall to reduce bias on a particular validation set (four model architectures based on Residual Convolutional Neural Network (ResNet)[23] and Dense Convolutional Network (DenseNet)[24], each trained on two cross-validation phases).

Error Rate (%)	Before UDC Accuracy (%)	After UDC Accuracy (%)	Enhancement (%)
0	$\mathbf{98.8 \pm 0.1}$	$\mathbf{98.8 \pm 0.1}$	0.0 ± 0.1
10	97.5 ± 0.2	$\mathbf{98.6 \pm 0.1}$	$\mathbf{1.1 \pm 0.2}$
20	96.7 ± 0.3	$\mathbf{98.5 \pm 0.1}$	$\mathbf{1.8 \pm 0.3}$
30	94.7 ± 0.8	$\mathbf{98.2 \pm 0.1}$	$\mathbf{3.5 \pm 0.8}$
40	92.8 ± 0.5	$\mathbf{98.3 \pm 0.1}$	$\mathbf{5.5 \pm 0.5}$
50	87.7 ± 1.9	$\mathbf{98.1 \pm 0.1}$	$\mathbf{10.3 \pm 1.9}$
60	74.7 ± 2.6	$\mathbf{97.0 \pm 0.2}$	$\mathbf{22.3 \pm 2.6}$
70	47.4 ± 2.5	$\mathbf{92.2 \pm 0.6}$	$\mathbf{44.8 \pm 2.6}$
80	$\mathbf{15.3 \pm 2.0}$	3.7 ± 0.5	-11.7 ± 2.0
90	$\mathbf{2.9 \pm 0.6}$	1.2 ± 0.2	-1.7 ± 0.7

Table 2. Results from a single round of UDC applied to various incorrect label levels. Incorrect labels are distributed evenly across all four classes: airplanes, boats, motorcycles, and trucks. UDC is robust to label error rates up to 70%, where a 44.8% gain in AI performance is achieved after a single round of UDC.

For the multi-class case, results show that UDC is resilient and can identify poor-quality data and improve AI accuracy even at more extreme levels of errors (i.e. 30% -70%) compared with the binary case. UDC fails when the percentage of incorrect labels in each class approaches 80%. This is because when 80% or more of a class' labels are distributed into three other classes, this results in fewer correct labels than incorrect labels for that class, making model training impossible as the model is pulled away from convergence by a larger number of incorrectly vs. correctly labeled data, and making such data *uncleansable*.

Taken together, these results suggest that UDC creates cleansed datasets that can be used to develop high performing AI models that are both accurate and generalizable using fewer training data, and reduced training time and cost. Near baseline-level performance was achieved using datasets containing up to 70% fewer training data. We showed that 97% accuracy could be achieved on datasets with up to 60% fractional incorrect labels for all classes, using less than 30% of the original amount of training data. In an even more extreme case with 70% incorrect labels, models trained on less than 5% of the original training set still achieved greater than 92% accuracy on a blind test set. Finally, application of UDC gave greater stability and accuracy during the training process (across epochs), which means that AI model selection for deployment can be automated because the selection is not hyper-dependent on a training epoch once a threshold of accuracy is achieved.

Application of UDC

The UDC technique was then applied to two healthcare problems, pediatric chest x-ray images for identification of pneumonia, and embryo images to identify likelihood of pregnancy (viability) for in vitro fertilization (IVF). Finally, UDC was also shown to be able to cleanse benchmark test datasets themselves to enable a truer and more realistic representation of AI performance.

Assessment of Pediatric Chest X-Rays for Pneumonia Detection

A publicly available dataset of pediatric chest x-ray images with associated labels of "Pneumonia" or "Normal" from Kaggle[25] was used. The labels were determined by multiple expert physicians. There were 5,232 images in the training set and 624 images in the test set. UDC was applied to all 5,856 images. Approximately 200 images were identified as noisy, while no labels were identified as incorrect. This suggests there were no suspected labeling errors in the dataset, but the images identified by UDC were considered poor-quality or uninformative. Poor-quality images in this dataset mean that labels of "normal" or "pneumonia" were not easily identifiable with certainty from the x-ray.

To verify the result, an independent expert radiologist assessed 200 x-ray images from this dataset, including 100 that were identified as noisy by UDC, and 100 that were identified as correct. The radiologist was only provided the image, and not the image label nor the UDC assessment. Images were assessed in random order, and the radiologist's assessment of the label for each image recorded. Results showed that the reader consensus between the radiologist's label and the original label was significantly higher for the correct images compared with the noisy images. Applying Cohen's kappa test[26] on the results gives levels of agreement for noisy ($\kappa \approx 0.05$) and correct ($\kappa \approx 0.65$) labels (refer to **Figure** 1). This confirms that for noisy images detected by UDC, there is insufficient information in the image alone to conclusively (or easily) make an assessment of pneumonia by either the radiologist or the AI. UDC could therefore prove beneficial as a screening tool for radiologists that could help triage difficult to read or suspicious (noisy) images that warrant further in-depth evaluation or additional tests to support a definitive diagnosis.

We then compared AI performance when trained using the original uncleansed x-ray training dataset versus UDC-cleansed x-ray training dataset with noisy images removed. Results are shown in **Figure** 2. The blue bar in the figure represents a theoretical maximum accuracy possible on the test dataset. It is obtained by testing every trained AI model on the test dataset to find the maximum accuracy that can be achieved. The orange bar is the actual (generalized) accuracy of the AI obtained using standard practice for training and selecting a final AI model using the validation set, then testing AI performance on the test set. The difference between the blue bar and orange bar indicates the generalizability of the AI, i.e. the ability of the AI to reliably apply to unseen data. Results show that training the AI on a UDC-cleansed dataset increases both the accuracy and generalizability of the AI. Additionally, the AI trained using a UDC-cleansed dataset achieved 95% generalized accuracy. This exceeds the 92% accuracy reported for other models in the literature using this same chest x-ray dataset[27].

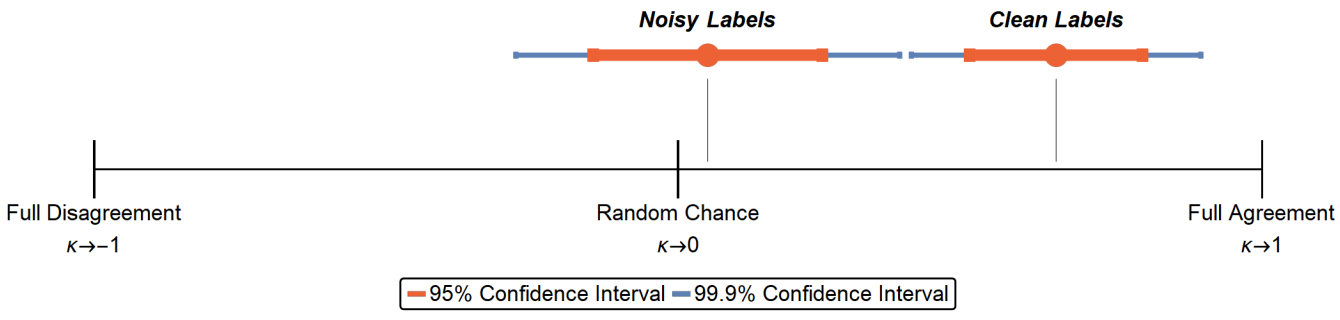

Figure 1. Cohen's kappa test for noisy and Correct labels shows that images with Correct labels lead to a significantly higher level of agreement than random chance, and significantly higher than those with noisy labels.

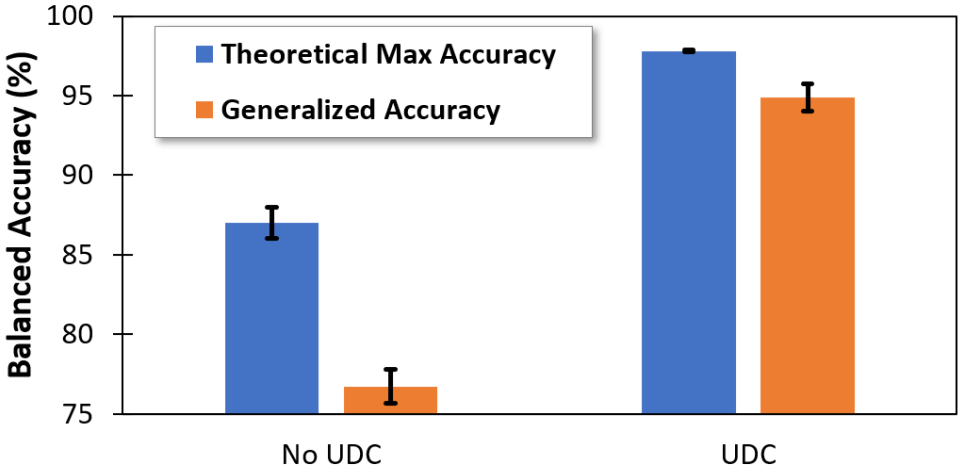

Figure 2. Balanced accuracy before and after UDC. The orange bar represents the AI accuracy on the test dataset using the standard AI training practice. The blue bar represents the theoretical maximum AI accuracy possible on the test dataset. The discrepancy between these two values is indicative of the generalizability of the model.

Lastly, we investigated application of UDC on the test dataset of x-ray images to assess its quality. This is vital because the test dataset is used by AI practitioners to assess and report on AI performance. Too much poor-quality data in a test set means the AI accuracy is not a true representation of AI performance. To evaluate this, we injected the uncleansed test dataset into the training dataset used to train the AI to determine the maximum accuracy that could be obtained on the validation dataset. **Figure** 3 shows reduced performance of AI trained using the aggregated dataset (training dataset plus the noisy test dataset) compared with the AI trained only using the cleansed training set. This suggests that the level of poor-quality data in the test dataset is significant, and thus care should be taken when AI performance is measured using this particular test dataset.

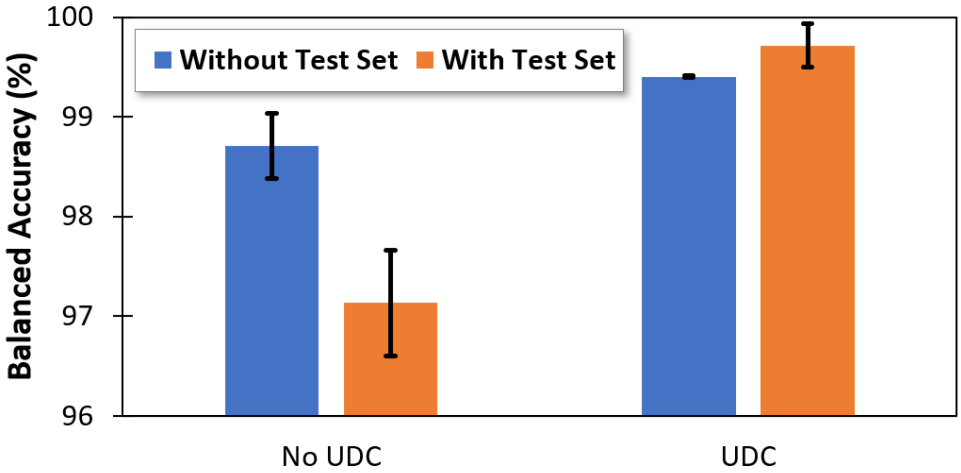

Figure 3. The colors of the bars represent the performance of the model on the validation set, with (orange) and without (blue) the test set included in the training set. AI performance drops when the uncleansed blind test set is included in the training set, indicating a considerable level of poor-quality data in the test set.

Assessment of Embryo Quality for IVF

Finally, UDC was successfully applied to the problem of assessing embryo viability in IVF. UDC was a core technique in developing a commercial AI healthcare product, which is currently being used in IVF clinics globally[17]. The AI model analyzes images of embryos at Day 5 of development to identify which ones are viable and likely to lead to a clinical pregnancy.

Clinical pregnancy is measured by the presence of a fetal heartbeat at the first ultrasound scan approximately six weeks after the embryo is transferred to an IVF patient. An embryo is labeled viable if it led to pregnancy, and non-viable if a pregnancy did

not occur. Although there is certainty in the outcome (pregnancy or no pregnancy), there is uncertainty in the labels, because there may be patient medical conditions or other factors beyond embryo quality that prevent pregnancy (e.g. endometriosis)[17]. Therefore, an embryo that is viable may be incorrectly labeled as non-viable. These incorrect labels impact the performance of the AI if not identified and removed.

UDC was applied to images of embryos to identify incorrect labels. These were predominantly in the training dataset's non-viable class, as expected, as they included embryos that appeared viable but were labeled as unviable. Performance results are shown in **Figure** 4. AI models trained using a UDC-cleansed training dataset achieved an increase in accuracy, from 59.7% to 61.1%, when reported on the standard uncleansed test dataset. This small increase in accuracy can be misleading, though, as the uncleansed test set itself may comprise a significant portion of incorrectly labeled non-viable embryo images, thus reducing specificity as the AI model improves. For the predominantly clean viable class, sensitivity increased significantly from 76.8% to 80.6%. When a UDC-cleansed test set is utilized, AI models trained using a UDC-cleansed training dataset achieved an increase in accuracy from 73.5% to 77.1%. This larger increase is a truer representation of the AI performance.

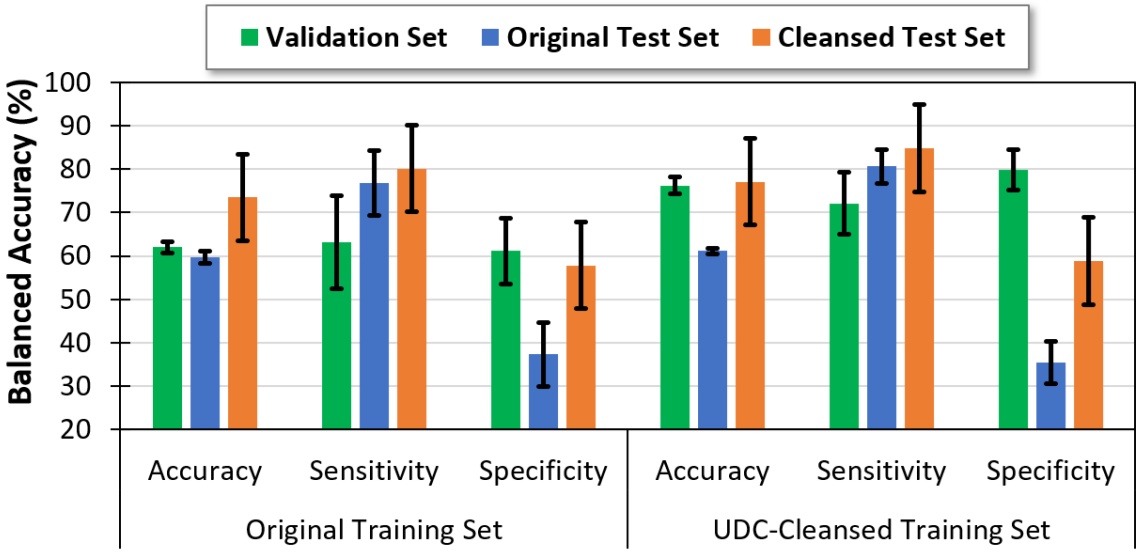

Figure 4. Performance metrics of AI model predicting clinical pregnancy, trained on original (left section) and UDC-cleansed (right section) training data. Both graphs show results on the validation set (green), and corresponding original test set (blue) and UDC-cleansed test set (orange).

Effect sizes before and after UDC are represented using Cohen's d, as shown in **Table** 3, along with p-values. Effect sizes larger than 0.6 are considered "large", meaning that for all test sets (including validation), UDC has a large effect on training and inference (test) performance, except for specificity results for both cleansed and uncleansed (expected due to the large proportion of incorrectly labeled non-viable embryos) test sets. This can be interpreted as there being a significant pair-wise uplift in sensitivity without much cost to specificity. In all cases, there is very large ($d > 1.4$) effect on overall performance. Taken together these results suggest that using UDC to cleanse training datasets can improve the accuracy of the AI even in clinical datasets with a high level of mislabeled, poor-quality data.

Dataset	**Validation Set**			**Original Test Set**			**Cleansed Test Set**		
Test Statistic	**Acc.**	**Sens.**	**Spec.**	**Acc.**	**Sens.**	**Spec.**	**Acc.**	**Sens.**	**Spec.**
d	8.656	2.439	3.234	1.414	0.664	-0.329	1.611	0.788	0.144
p	<0.001	0.199	0.049	0.002	0.401	0.449	0.062	0.385	0.482

Table 3. Results from a single round of UDC applied to various incorrect label levels. Incorrect labels are distributed evenly across all four classes: airplanes, boats, motorcycles, and trucks. UDC is robust to label error rates up to 70%, where a 44.8% gain in AI performance is achieved after a single round of UDC.

Discussion

This study characterizes a novel technique, Untrainable Data Cleansing (UDC), that serves to automatically identify, and thus allow removal of, poor-quality data to improve AI performance and reporting. In the clinical setting, accurate AI could mean the difference between life and death, or early diagnosis versus missed diagnosis. Thus it is critical that poor-quality data are identified and removed so as not to confuse the AI training process and impact clinical outcomes. It can be difficult for even the most experienced clinicians to identify poor-quality data, particularly when the clinical outcome is uncertain, or the quality and integrity of the data does not allow for a definitive labeling of the image. Furthermore, due to data privacy laws, it may not even be possible to manually assess data quality of private medical datasets. Because UDC can be "shipped" to the secure location of private data, it offers an automated way of addressing data quality concerns while respecting data privacy laws.

UDC was validated across two problem sets, (1) cats vs. dogs, and (2) vehicles, or binary and multi-classification problems, respectively, because image labels could be manually verified. In both cases UDC was effective at identifying synthetically introduced incorrect labels. Training AI models following removal of poor-quality data significantly improved the AI performance, in terms of both accuracy and generalizability. UDC was applied to two medical problem sets, one for pneumonia detection in chest x-rays, and the other for embryo selection in IVF. Both are challenging clinical assessment areas due to varying degrees of noisy or incorrectly labeled data. In both studies UDC was effective (as measured on double blind datasets) at identifying poor quality data, and yielded significant improvements in accuracy and generalizability.

In the same way that poor-quality training data can impact AI training, the reporting (or testing) of AI performance done so on a poor-quality test dataset (that contains noisy or incorrectly labeled data) can lead to inaccurate performance reporting. Inaccurate reporting can mislead clinicians and patients on the true performance and reliability of the AI, with potential real-world consequences for those that may rely on AI results. We assessed the utility of UDC for cleansing test datasets and showed that the accuracy of the AI reported on test datasets cleansed with UDC was different to that reported on uncleansed test datasets. The reporting of AI accuracy on a UDC-cleansed test set was shown to be a truer representation of the AI performance based on independent assessment.

Finally, we showed that UDC was able to identify noisy data, which in the case of the pneumonia x-rays neither the AI nor the radiologist could consistently classify. The ability for UDC to identify these cases suggests it can be used as a triage tool and direct clinicians to those cases that warrant new tests or additional in-depth clinical assessment. This study demonstrates that the performance of AI for clinical applications is highly dependent on the quality of clinical data, and the utility of a method like UDC that can automatically cleanse otherwise poor-quality clinical data cannot be overstated.

Methods

UDC Algorithm

Untrainable Data Cleansing (UDC) can identify three categories of image-label pairs:

- **Correct** – strongly expected to be correct (i.e. label matches the ground-truth).
- **Incorrect** – strongly expected to be incorrect (i.e. label does not match ground-truth).
- **Noisy** – data is ambiguous or uninformative to classify with certainty (i.e. label may or may not match ground-truth).

UDC delineates between correct, incorrect, and noisy images using the method described in Algorithm 1, which utilizes a mixture between sampling different (n) model architectures and sampling across data using k-fold cross validation (KFXV). These $n \times k$ models vote on each image-label in this manner to reduce bias and to increase robustness to outliers.

We describe UDC as "*turning AI onto itself*", as it uses the AI training process to identify poor-quality data. Multiple AI models using different architectures and parameters are trained using the data (to be cleansed), then the AI models are applied back on the same training data to infer their labels. Data which cannot be consistently classified correctly are identified as poor-quality (i.e. incorrect or noisy).

The idea behind UDC is that if data cannot be consistently correctly classified within the AI training process itself, which is where AI models are likely to find the strongest correlations and correct classifications on the dataset used to train/create it, then the data is likely to be poor-quality.

The intuition behind using UDC to subdivide image-label pairs into three types of labels is based on probability theory. A correct label can be thought of as a positively weighted coin ($p \gg 0.5$), where p is the probability of being correctly predicted by a certain model. In contrast, an incorrect label can be thought of as a negatively weighted coin ($p \ll 0.5$) – very likely to be incorrectly predicted. A noisy label can be thought of as a fair coin ($p \approx 0.5$) – equally likely to be correctly or incorrectly predicted. To illustrate how this intuition applies to UDC, we consider a hypothetical dataset of N image-label pairs. Algorithm 1 is applied to this dataset to produce a number of successful predictions (s_j) for each image j. A histogram of s_j values (with increasing s on the x-axis) then shows how correct, incorrect, and noisy labels tend to cluster at high, low, and medium values of s, respectively.

UDC Applied to Cats and Dogs

Synthetic errors (incorrect labels) were added to the training dataset of 24,916 images, which is split 80/20 into training and validation sets with the following parameters used for each study t, where $\boldsymbol{n}^{(t)} = \left(\boldsymbol{n}_{cat}^{(t)}, \boldsymbol{n}_{dog}^{(t)}\right)$, where $0 \leq \boldsymbol{n}(\%) \leq 100$, contains the fractional level of incorrect labels for cat and dog classes (see Table 4), respectively, and where RN stands for ResNet[23], and DN stands for DenseNet[24].

k	n	$\mathscr{M}^{(1)}$	$\mathscr{M}^{(2)}$	$\mathscr{M}^{(3)}$	$\boldsymbol{n}^{(1)}$	$\boldsymbol{n}^{(2)}$	$\boldsymbol{n}^{(3)}$	$\boldsymbol{n}^{(4)}$	$\boldsymbol{n}^{(5)}$
5	3	DN-121	RN-50	RN-18	(35,5)	(50,5)	(30,30)	(70,70)	(50,50)

Table 4. Parameters used to test Untrainable Data Cleansing (Algorithm 1). 'RN' stands for ResNet[23] and 'DN' stands for DenseNet[24] neural network architectures, respectively.

UDC Applied to Vehicles

Synthetic errors (incorrect labels) were added to the training dataset of 18,991 images, which is split 80/20 ($k = 5$) into training and validation sets and where for each study t, $\boldsymbol{n}^{(t)} = \left(\boldsymbol{n}_A^{(t)}, \boldsymbol{n}_B^{(t)}, \boldsymbol{n}_M^{(t)}, \boldsymbol{n}_T^{(t)}\right)$ represents fractional levels of incorrect labels for airplane (A), boat (B), motorcycle (M), and truck (T) classes (see Table 5), respectively, where $0 \leq \boldsymbol{n}(\%) \leq 100$, and where R stands for ResNet[23], and D stands for DenseNet[24]. Note, in Table 5, the fractional level of incorrect labels was kept constant across classes in each study, so only one value is shown.

k	n	$\mathscr{M}^{(1)}$	$\mathscr{M}^{(2)}$	$\mathscr{M}^{(3)}$	$\mathscr{M}^{(4)}$	$\boldsymbol{n}^{(0)}$	$\boldsymbol{n}^{(1)}$	$\boldsymbol{n}^{(2)}$	$\boldsymbol{n}^{(3)}$	$\boldsymbol{n}^{(4)}$	$\boldsymbol{n}^{(5)}$	$\boldsymbol{n}^{(6)}$	$\boldsymbol{n}^{(7)}$	$\boldsymbol{n}^{(8)}$	$\boldsymbol{n}^{(9)}$
5	4	D-121	R-101	R-50	R-18	0	10	20	30	40	50	60	70	80	90

Table 5. Parameters used to test Untrainable Data Cleansing (Algorithm 1) for a multi-classification problem.

Reader Consensus Between Radiologists for Correct vs. Noisy Labels

Images identified by UDC to have noisy labels are suspected to have inconsistencies rendering their annotation (or labeling) more difficult. As such, we expect the reader consensus of Pneumonia/Normal assessments between different radiologists to be lower for images with noisy labels than for those with correct labels that are easily identified by the AI model and for which we expect a relatively high reader consensus between radiologists. The following two hypotheses are formulated and can be directly tested using the (Cohen's) kappa test:

- $\mathscr{H}_0^{(1)}$: The level of agreement between radiologists for noisy labels is different from random chance.
- $\mathscr{H}_a^{(1)}$: The level of agreement between radiologists for noisy labels is no different from random chance.
- $\mathscr{H}_0^{(2)}$: The level of agreement between radiologists for correct labels is no greater than random chance.
- $\mathscr{H}_a^{(2)}$: The level of agreement between radiologists for correct labels is greater than random chance.

We prepare an experimental dataset by splitting the data into correct and noisy labels as follows, where the two subsets are used in a clinical study to test the above hypotheses and validate UDC:

1. A dataset $\mathscr{D}$ with 200 elements $\boldsymbol{\varkappa}_j = (\mathbf{x}_j, \hat{y}_j)$ has images $\mathbf{x}_j$ and (noisy) annotated labels $\hat{y}_j$. This dataset is split into two equal subsets of 100 images each:
   (a) $\mathscr{D}_{clean}$ – labels identified as correct by UDC, with the following breakdown:
      i. 48 Normal
      ii. 52 Pneumonia (39 Bacterial / 13 Viral)
   (b) $\mathscr{D}_{noisy}$ – labels identified as noisy by UDC, with the following breakdown:
      i. 51 Normal
      ii. 49 Pneumonia (14 Bacterial / 35 Viral)

2. The dataset $\mathscr{D}$ is randomized to create a new dataset $\hat{\mathscr{D}}$ for an expert radiologist to label, and to indicate confidence or certainty in those labels (Low, Medium, High). This randomization is to address fatigue bias and any bias related to the ordering of the images.

3. The *reader consensus* between the expert radiologist and the original labels is calculated using Cohen's kappa test26 and is compared between datasets $\mathscr{D}_{clean}$ vs. $\mathscr{D}_{noisy}$.

Figure 1 provides visual evidence showing that both null hypotheses, $\mathscr{H}_{\mathbf{0}}^{(\mathbf{1})}$ and $\mathscr{H}_{\mathbf{0}}^{(\mathbf{2})}$, are rejected with very high confidence ($> 99.9\%$) and effect size (> 0.85). Therefore, both alternate hypotheses are accepted: $\mathscr{H}_{\mathbf{a}}^{(\mathbf{1})}$, stating that labels identified as noisy have levels of agreement no different from random chance, and $\mathscr{H}_{\mathbf{a}}^{(\mathbf{2})}$, stating that labels identified by UDC as correct have levels of agreement greater than random chance.

Algorithm 1 A given dataset $\mathscr{D}$ contains a set of N elements $\mathscr{z}_j = (\mathbf{x}_j, \hat{y}_j)$ with each image $\mathbf{x}_j$ paired with its corresponding (noisy) target label $\hat{y}_j$ (e.g. $y_j \in 0,1$ for binary classification problem), where $j \in 1..N$. $\mathscr{D}$ (excluding a subset used as a blind test set) is split into k mutually exclusive validation datasets $\mathscr{D}^{(i)}$ with training sets $\mathscr{D} \setminus \mathscr{D}^{(i)}$, where $i \in 1..k$ is the cross-validation phase index. A set of model architectures $\mathscr{M}$ with n elements $M^{(m)} \in \mathscr{M}$, where $m \in 1..n$, is trained on each dataset $\mathscr{D} \setminus \mathscr{D}^{(i)}$. Algorithm A is used to produce a set of learned mappings $\mathscr{f} \ni f^{(i,m)} \leftarrow A\left(M^{(m)}, \mathscr{D} \setminus \mathscr{D}^{(i)}\right)$, which are chosen using confidence metrics described above. Phase index i is included in the learned mapping since each model is trained on a different dataset $\mathscr{D} \setminus \mathscr{D}^{(i)}$, and because dropout is used during training. These learned mappings can then be tested against the entire dataset $\mathscr{D}$ to generate predicted outcomes $\bar{y}^{(i,m)} \leftarrow f^{(i,m)}(\mathbf{x}_j)$, which can be used to find a per-element successful prediction count $\mathscr{s}_j = \sum_{i=1}^{k} \sum_{m=1}^{n} \mathscr{s}_j^{(i,m)}$, where $\mathscr{s}_j^{(i,m)}$ equals unity if the model prediction equals the noisy target label, or $\hat{y}_j = \bar{y}^{(i,m)} \leftarrow f^{(i,m)}(\mathbf{x}_j)$, and equals zero otherwise. The vector $\mathscr{s}$ containing all elements $\mathscr{s}_j \in 0..n \times k$ is returned.

Define: $UDC\,(\mathscr{D}, k, \mathscr{M}, A)$

Require: $\mathscr{D}$, the given dataset with N elements $\mathscr{z}_j = (\mathbf{x}_j, \hat{y}_j)$; images $\mathbf{x}_j$, target labels $\hat{y}_j$

Require: k, the number of folds used to split dataset $\mathscr{D}$ in k mutually exclusive subsets $\mathscr{D}^{(i)}$

Require: $\mathscr{M}$, a set of n model architectures with elements $M^{(m)} \in \mathscr{M}$, where $m \in 1..n$

Require: A, the learning algorithm, maps a dataset and a model into a learned function $f^{(i,m)}$

Initialize: $\mathscr{s} \leftarrow \mathbf{0}_N$

Split $\mathscr{D}$ into k mutually exclusive validation subsets $\mathscr{D}^{(i)}$, whose union is $\mathscr{D}$

for m from 1 to n **do**

  for i from 1 to k **do**

    $f^{(i,m)} \leftarrow A\left(M^{(m)}, \mathscr{D} \setminus \mathscr{D}^{(i)}\right)$

    for $\mathscr{z}^{(j)}$ in $\mathscr{D}$ **do**

      if $\hat{y}_j = \bar{y}^{(i,m)} \leftarrow f^{(i,m)}(\mathbf{x}_j)$ **then**

        $\mathscr{s}_j \leftarrow \mathscr{s}_j + 1$

      end if

    end for

  end for

end for

Return $\mathscr{s}$

References

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *nature* **521**, 436–444 (2015). Publisher: Nature Publishing Group.
2. Goodfellow, I., Bengio, Y. & Courville, A. *Deep learning* (MIT press, 2016).
3. Esteva, A. *et al.* A guide to deep learning in healthcare. *Nat. medicine* **25**, 24–29 (2019). Publisher: Nature Publishing Group.
4. Fourcade, A. & Khonsari, R. H. Deep learning in medical image analysis: A third eye for doctors. *J. Stomatol. Oral Maxillofac. Surg.* **120**, 279 – 288, DOI: https://doi.org/10.1016/j.jormas.2019.06.002 (2019).
5. Lundervold, A. S. & Lundervold, A. An overview of deep learning in medical imaging focusing on mri. *Zeitschrift fur Medizinische Physik* **29**, 102 – 127, DOI: https://doi.org/10.1016/j.zemedi.2018.11.002 (2019).
6. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. image analysis* **42**, 60–88 (2017). Publisher: Elsevier.
7. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017). Publisher: Nature Publishing Group.
8. Haenssle, H. A. *et al.* Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals Oncol.* **29**, 1836–1842 (2018). Publisher: Oxford University Press.
9. Cheng, J.-Z. *et al.* Computer-aided diagnosis with deep learning architecture: applications to breast lesions in us images and pulmonary nodules in ct scans. *Sci. reports* **6**, 1–13 (2016). Publisher: Nature Publishing Group.
10. Kooi, T. *et al.* Large scale deep learning for computer aided detection of mammographic lesions. *Med. image analysis* **35**, 303–312 (2017). Publisher: Elsevier.
11. Gulshan, V. *et al.* Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama* **316**, 2402–2410 (2016). Publisher: American Medical Association.
12. Poplin, R. *et al.* Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* **2**, 158 (2018). Publisher: Nature Publishing Group.
13. De Fauw, J. *et al.* Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. medicine* **24**, 1342–1350 (2018). Publisher: Nature Publishing Group.
14. Ciresan, D. C., Giusti, A., Gambardella, L. M. & Schmidhuber, J. Mitosis detection in breast cancer histology images with deep neural networks. 411–418 (Springer, 2013).
15. Charoentong, P. *et al.* Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. *Cell reports* **18**, 248–262 (2017). Publisher: Elsevier.
16. Beck, A. H. *et al.* Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci. translational medicine* **3**, 108ra113–108ra113 (2011). Publisher: American Association for the Advancement of Science.
17. VerMilyea, M. *et al.* Development of an artificial intelligence-based assessment model for prediction of embryo viability using static images captured by optical light microscopy during ivf. *Hum. Reproduction* (2020).
18. Zhang, X. & Lessard, L. Online data poisoning attacks (2020).
19. Raghu, M. *et al.* Direct uncertainty prediction for medical second opinions. *arXiv:1807.01771 [cs, stat]* (2019). ArXiv: 1807.01771.
20. Kendall, A. & Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? 5574–5584 (2017).
21. Natarajan, N., Dhillon, I. S., Ravikumar, P. K. & Tewari, A. Learning with noisy labels. 1196–1204 (Curran Associates, Inc., 2013).
22. Xiao, T., Xia, T., Yang, Y., Huang, C. & Wang, X. Learning from massive noisy labeled data for image classification. 2691–2699 (2015).
23. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. 770–778 (2016).
24. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. 4700–4708 (2017).
25. Mooney, P. Chest x-ray images (pneumonia).

26. Sim, J. & Wright, C. C. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys. therapy* **85**, 257–268 (2005). Publisher: Oxford University Press.

27. Kermany, D. S. *et al.* Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122–1131 (2018). Publisher: Elsevier.

Author contributions statement

D.P. invented the concept, M.A.D. designed the algorithm, M.A.D. and J.M.M.H and T.N. and D.P. conceived the experiments, M.A.D. and J.M.M.H. and T.N. conducted the experiments, M.V. and R.L. provided clinical data and clinical review, D.P. and M.A.D. and J.M.M.H and T.N. and S.M.D. and M.P. drafted the manuscript and provided critical review of the results.

Additional Information

Provisional Patent

Applicant: Presagen Pty Ltd
Filing Date: April 3 2020
Title: Artificial Intelligence (AI) Method for Cleaning Data for Training AI Models
Number: 2020901043
Status: Pending

Competing Interests

J.M.M.H., D.P., and M.P. are co-owners of Presagen. S.M.D., T.N., and M.A.D. are employees of Presagen.

Figures

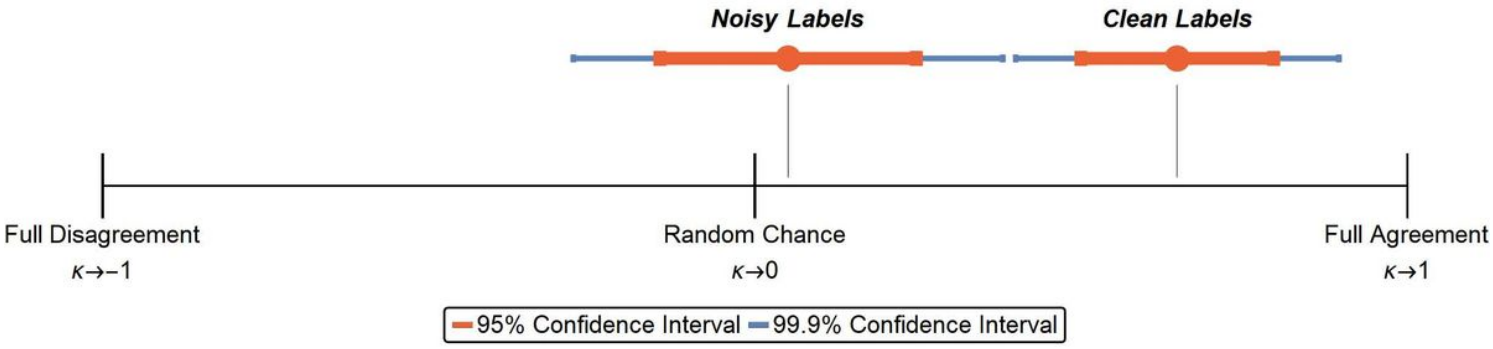

Figure 1

Cohen's kappa test for noisy and Correct labels shows that images with Correct labels lead to a significantly higher level of agreement than random chance, and significantly higher than those with noisy labels.

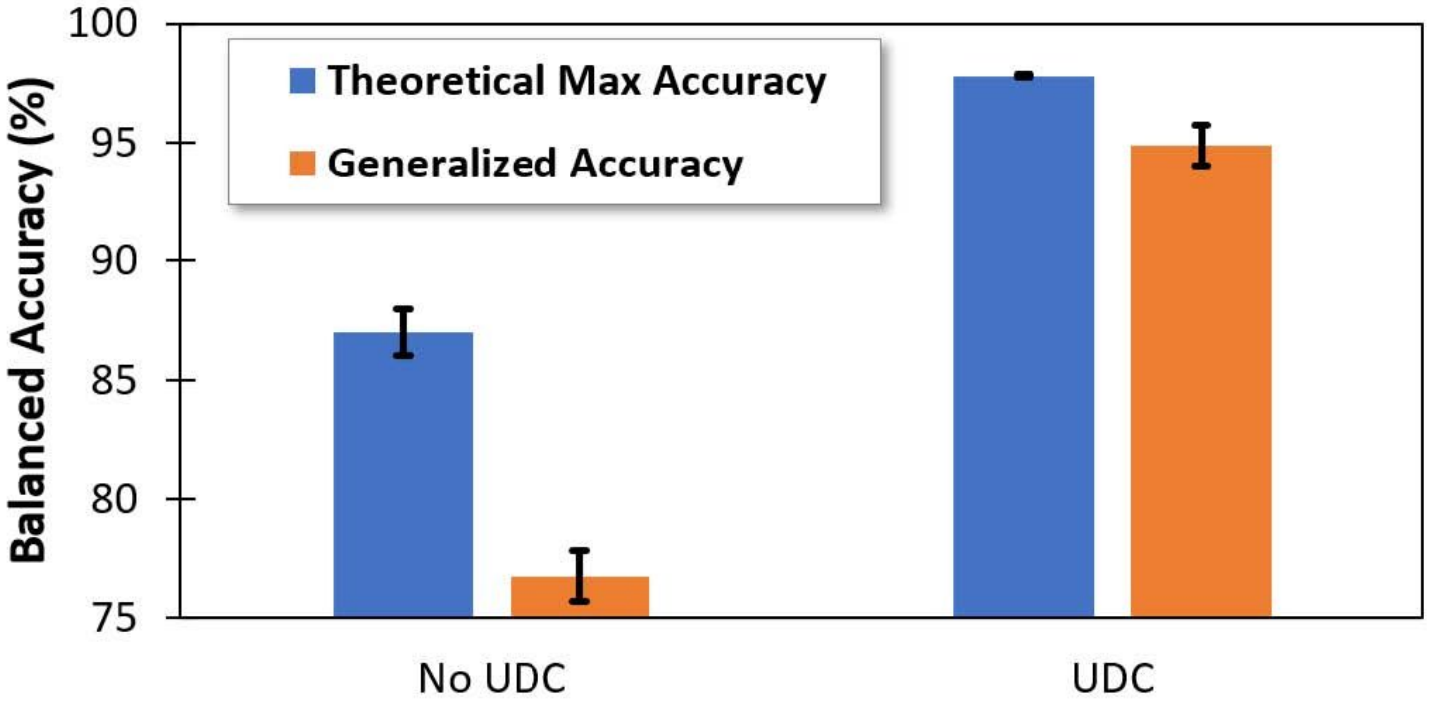

Figure 2

Balanced accuracy before and after UDC. The orange bar represents the AI accuracy on the test dataset using the standard AI training practice. The blue bar represents the theoretical maximum AI accuracy possible on the test dataset. The discrepancy between these two values is indicative of the generalizability of the model.

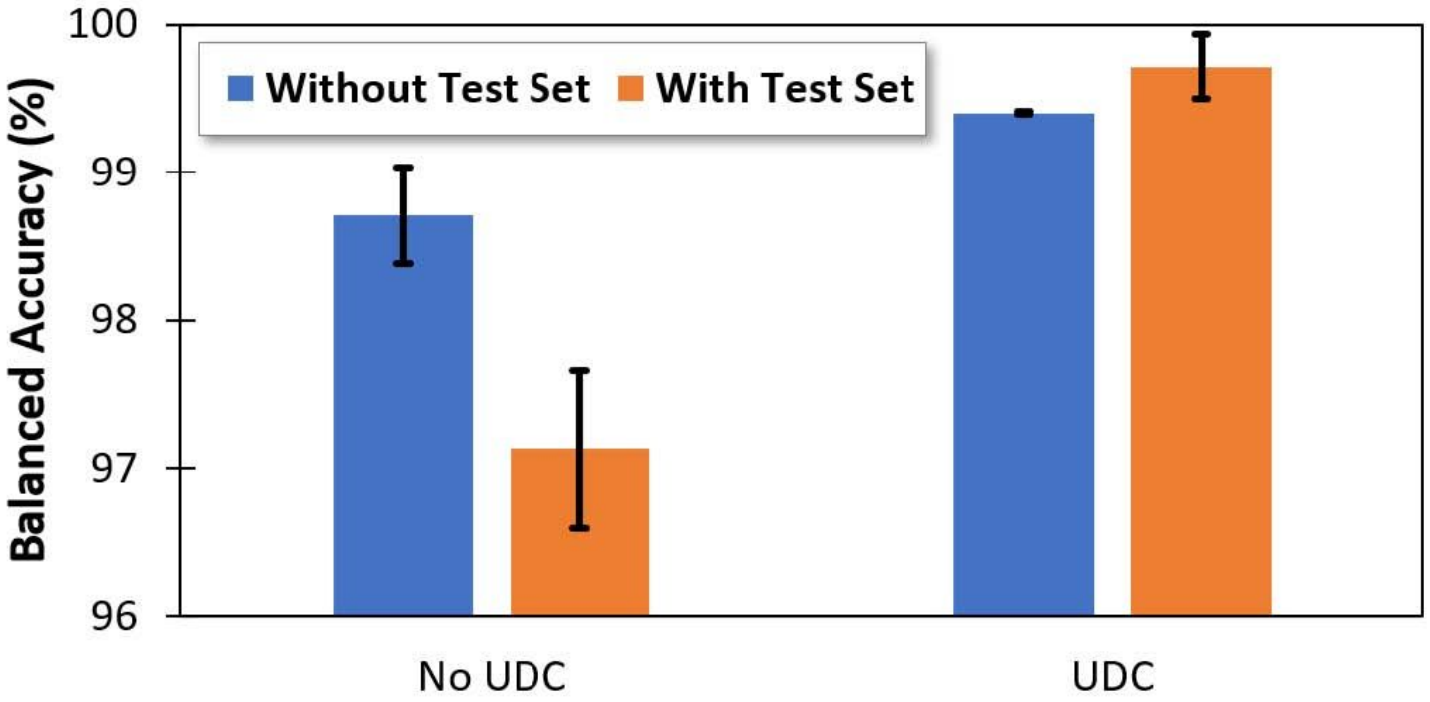

Figure 3

The colors of the bars represent the performance of the model on the validation set, with (orange) and without (blue) the test set included in the training set. AI performance drops when the uncleansed blind test set is included in the training set, indicating a considerable level of poor-quality data in the test set.

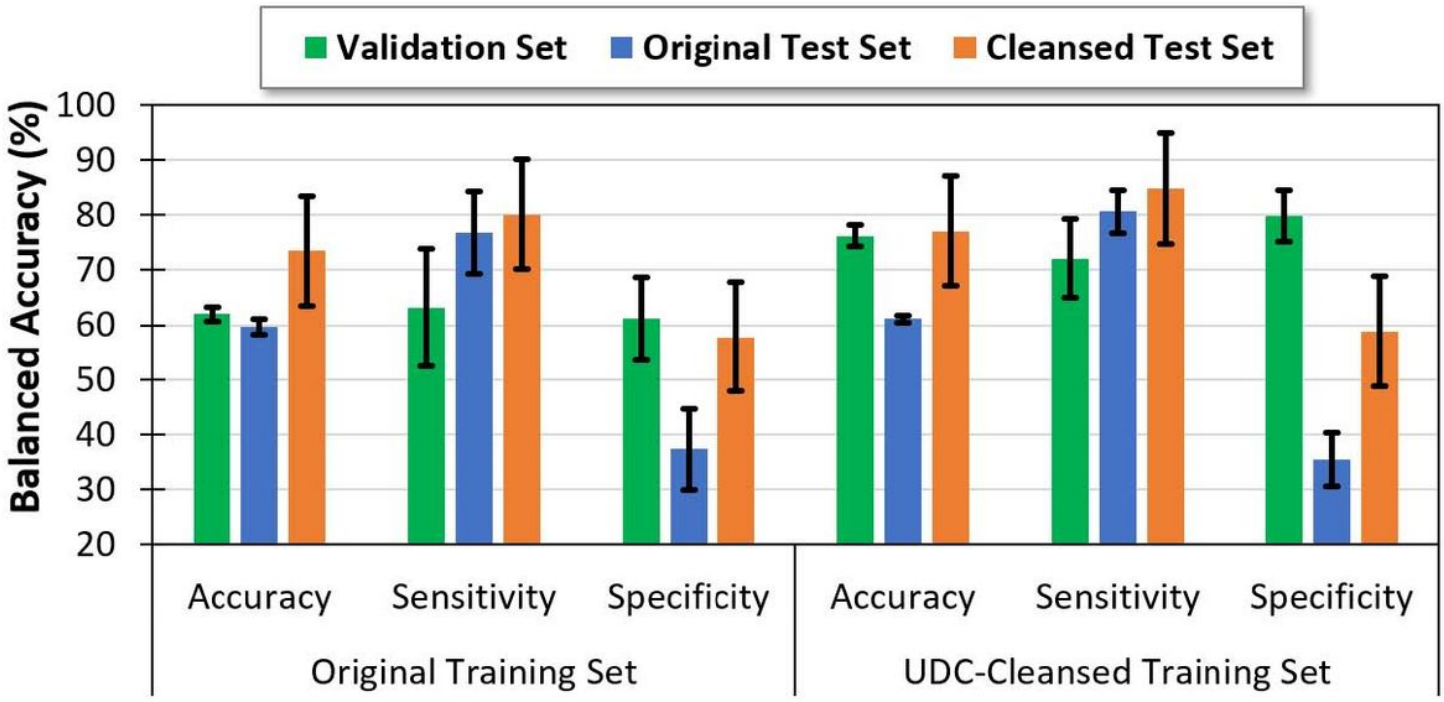

Figure 4

Performance metrics of AI model predicting clinical pregnancy, trained on original (left section) and UDC-cleansed (right section) training data. Both graphs show results on the validation set (green), and corresponding original test set (blue) and UDC-cleansed test set (orange).