Prescriptive Method for Optimizing Cost of Data Collection and Annotation in Machine Learning of Clinical Ultrasound

Alistair Lawley*, Rory Hampson, *Member, IEEE*, Kevin Worrall, Gordon Dobie

***Abstract*— Machine learning in medical ultrasound faces a major challenge: the prohibitive costs of producing and annotating clinical data. Optimizing the data collection and annotation will improve model training efficiency, reducing project cost and times. This paper prescribes a 2-phase method for cost optimization based on iterative accuracy/sample size predictions, and active learning for annotation optimization. *Methods:* Using public breast, fetal, and lung ultrasound datasets we can: Optimize data collection by statistically predicting accuracy for a desired dataset size; and optimize labeling efficiency using Active Learning, where predictions with lowest certainty were labelled manually using feedback. A practical case study on BUSI data was used to demonstrate the method prescribed in this work. *Results:* With small data subsets, ~10%, dataset size vs. final accuracy relations can be predicted with diminishing results after 50% usage. Manual annotation was reduced by ~10% using active learning to focus the annotation. *Conclusion:* This led to cost reductions of 50%-66%, depending on requirements and initial cost model, on BUSI dataset with a negligible accuracy drop of 3.75% from theoretical maximums.**

***Clinical Relevance*— This work provides methodology to optimize dataset size and manual data labelling, this allows generation of cost-effective datasets, of interest to all, but particularly for financially limited trials and feasibility studies, Reducing the time burden on annotating clinicians.**

I. INTRODUCTION

A. Motivation for Cost Optimization

Ultrasound is one of the most commonly used diagnostic modalities in the world today due to its low cost and minimally invasive approach [1]. There is very limited availability of public annotated data for machine learning (ML) in clinical ultrasound. This is not a problem unique to ultrasound, the inherent cost of producing high quality data and subsequent accurate clinical annotation often means that generating appropriate datasets for diagnostic quality deep learning is a major investment [2]. There are many techniques that have been explored to reduce the cost of image annotation, such as self-supervision [3, 4], amongst others [5].

Datasets for machine learning are similar to cluster random control trials, where the addition of more data has diminishing returns on how much it improves the accuracy of the result [6]. When training neural networks (NNs), it is important to weigh the value of additional data and annotation against the time and cost of producing it. This optimization, and analysis focusing on time and cost is critical to decision makers attempting to balance priorities on medical device and imaging projects. This can be partially automated, reducing researcher and clinical burdens using ML and applied clinical trial protocols adapted to ML [7]. There is a lack of detailed prescriptive methods for optimal data capture and annotation in the clinical literature, that this work aims to address.

B. Cost / Time Optimization Methods and Applications

Statistical power curves have so far been used with retinal optical coherence tomography (OCT) [8], and magnetic resonance imaging (MRI) [9] but has yet to be explored for ultrasound. The use of sample size determination is a common technique in clinical trials, but it is possible to adapt the technique for machine learning by augmenting the sample selection process. This method is known to have significant difficulty predicting situations where there are only subtle variances between classifiers and reduces overall generalizability [10]. Machine learning has shown potential to significantly improve control trials such as in reducing sample size requirements in MRI studies of cognitive impairment [11], in augmenting the selection process to better identify and retain sampled population [12-14], and for improving statistical analysis [15, 16].

Where ultrasound data is available without annotation, there is an opportunity to apply a targeted approach to sample selection. Active learning is a technique where a neural network is used to analyze the dataset allowing for targeted training on a selected portion of the dataset [17]. There are many common methods of active learning within machine learning, such as diversity sampling [18]. In this case selective uncertainty sampling [19] is used to identify where the neural network has the lowest confidence in its prediction and target those images for annotation. This forms an active learning loop, allowing for the consistent querying of the learning network to better inform the annotation process. Active learning has already been successfully applied to breast ultrasound using a weakly supervised approach, as well as in the detection of breast masses [20, 21], in the multi-model detection of liver fibrosis for ultrasound elastography [22], and in semi-supervised covid lung disease classification [23].

This paper examines how optimal ultrasound dataset size can be determined and also investigates effectiveness of uncertainty sampled active learning for ultrasound data in reducing the cost of dataset production for clinical ultrasound A case study is explored on the BUSI dataset [24], with discussion applied to Covid Lung [25] and Fetal datasets [26]. This shall result in an understanding of the optimal sample size for maximizing accuracy whilst minimizing cost.

*Research supported by Canon Medical Systems (EDI/JP).

Alistair Lawley is with the Future Ultrasound CDT (FUSE), University of Strathclyde, 204 George St., Glasgow, G1 1XW, GB (correspondence e-mail: alistair.lawley@strath.ac.uk). Rory Hampson and Gordon Dobie are with the Centre for Ultrasonic Engineering (CUE), University of Strathclyde, GB. Kevin Worral is with the University of Glasgow, Glasgow, GB.

This is a peer reviewed, accepted author manuscript of the following research contribution: Lawley, A., Hampson, R., Worrall, K., & Dobie, G. (2023). Prescriptive method for optimizing cost of data collection and annotation in machine learning of clinical ultrasound. Paper presented at *45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2023),* Sydney, Australia.

II. Method for Optimized Sampling/Annotation

Ensuring a representative sampling within the training set assists in the subsequent extrapolation of the statistical power curve. This also allows the simulation of the data collection process with each subsequent iteration representing an additional round of data collection of datasets. In this work, an AlexNet [27] NN is used as a well-known benchmark, however other NNs can be easily substituted.

A. Phase 1 – Optimized Data Set Capture:

To predict required data set size from iterating power curves:

1. A NN is trained on a small sample of annotated data (10-50 samples, dependent on experimental constraints). Overfitting should be taken into account when designing your initial model training, with a portion of annotated data set aside to validate accuracy. The validation accuracy of this NN, determined as described in Subsection D is then stored.
2. Add an additional 10-50 annotated samples (in even chunks as before) and retrain NN on cumulative sample set. Record the validation accuracy.
3. Plot NN accuracy vs. sample size and fit a power curve to the data. Continue adding data in chosen chunk sizes until curve fit is 'stable' at desired accuracy. Stability is when subsequent sample groups predict end accuracies within your desired tolerance, say 2%.
4. Use the power curve to determine the required sample size for desired/acceptable accuracy.
5. Capture remainder of the predicted data set, without immediate annotation + 25% (based on conventional 80/20 split) again of annotated data for validation set.

B. Phase 2 – Optimizing annotation:

In cases where excess samples have been captured, particularly in large unannotated datasets, active learning can be used to selectively annotate samples that the NN has the most difficulty identifying, thus optimizing the annotated sample set. This can also be used to optimize annotation of small pre-captured datasets.

1. Using NN trained on annotated data from Phase 1, evaluate the probability outcome (NN classification certainty) on the remaining unannotated data.
2. Identify least certain samples on unannotated data, where NN has least certainty, say bottom 50 samples.
3. Manually annotate this sample group.
4. Retrain NN on phase 1 annotated samples + phase 2 manual annotation group.
5. Evaluate result on validation set.
6. Compare accuracy to predicted value from Phase 1.
7. If accuracy is outside of acceptable tolerance for theoretical accuracy from phase 1, say 5%, use step 8-9, if within tolerance, terminate the process.
8. Reshuffle all annotated data (phase 1 + phase 2 group + validation set) into a new 80/20 split. (Transfer learning could also be used but this is out of this work).
9. Iterate back to step 2 using remaining unannotated sample pool until accuracy is within an acceptable tolerance of theoretical accuracy from phase 1, say 5%.

III. Results from BUSI Case Study

A. Dataset Size Optimization

In order to demonstrate the potential saving incrementally, phase 1 and phase 2 of the prescribed method were applied independently to the BUSI data set, and then as a combined method considering mean response and max response of the NNs respectively. A standard Alexnet NN that had been pretrained using the ImageNet Challenge dataset [28] was used with the final layer output reduced to fit the classification requirements of the dataset.

Phase one was applied to the BUSI breast dataset, with an initial sample size of 15. The process was iterated until power curve stability was achieved at 150 samples as shown in Fig. 1a. This allowed a prediction that 400 samples were required to be within 4% of the theoretical maximum accuracy achievable with the full dataset. These remaining samples (250) were then 'collected' by randomly sampling the BUSI dataset. All remaining BUSI data was for validation.

Phase 2 was then applied with an initial NN trained on a sub sample of 50, and then predicted the annotation for the remaining 350 unannotated datasets with 50 chosen for annotation using uncertainty sampling and added to the training set. A new NN was then trained, validated, and then used to select an additional 50 samples from the remaining unannotated patient sets. This was repeated until all 400 samples were selected as shown in Fig. 1b for illustrative purposes, but the process would stop at the desired tolerance.

All networks were trained for 100 times for a cycle of 20 epoch, using an Alexnet NN and ADAM optimization method. Depending on the experimental robustness requirements, the best result of the training epochs or the mean result can be considered with differing conclusions.

For the BUSI dataset with 400 samples from Phase 1, with an acceptable tolerance of 2%, 350 samples of the 400 need annotation for the mean response to be within tolerance but only 200 samples of the 400 need be annotated for the maximum result to be within acceptable tolerance.

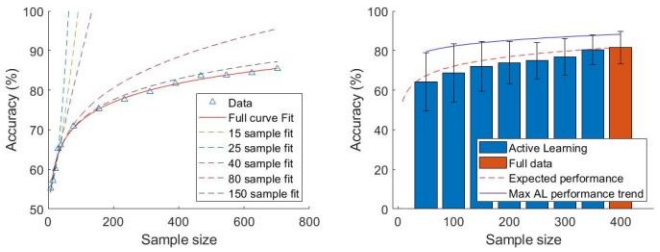

Figure 1 – a.) Comparing the power curve fit for sampling steps of the breast dataset from phase 1. b.) Performance of active learning in comparison to full annotation of 400 samples from phase 2.

From the combined method of phase 1 and phase 2, considering the maximum response form the NNs, an accuracy of 85% was achievable using only 400 samples compared with the theoretical maximum of 88% at the full BUSI dataset size. Additionally with only 200 of the 400 samples manually annotated, accuracy only drops to 84.7% for a 50% reduction in annotation burden, directly translatable into both time and costs.

B. *Cost Saving Vs. Accuracy*

For completeness, the cases of simply performing phase 1 alone (with 400 captured and annotated samples) and performing phase 2 alone on the full BUSI dataset, yielding 50% annotation, were also considered to illustrated cost differences. Using an initial representative costing model of 1:2 for data collection and annotation the relative costs of each method and phase can be seen in Fig. 2, calculated using (1), where P is the price of collection or annotation and N is the numbers predicted by phase 1 and 2 respectively.

$$\text{Cost} = (P_{\text{Collect}} \text{ X } N_{\text{Collection}}) + (P_{\text{Annotate}} \text{ X } N_{\text{Annotate}}) \quad (1)$$

Dependent on accuracy and robustness requirements, significant cost savings can be made by optimizing collection using a statistical power curve, and by targeting annotation by applying active learning as described in our method. Combining the methods shows the potential to reduce costs even further, up to 66% where the best performing network is taken into account as shown in (Fig. 2). The shape of this graph shows that regardless of the initial costing model used, the prescribed method will always yield a cost reduction in comparison to capturing arbitrary amounts of data and annotating it all, which is an important result allowing decision makers to optimize their clinical applications of machine learning. The scale of the cost saving is related to the complexity of the data, the NN type used, and the costing model, but this method is always expected to return a cost reduction for minimal accuracy loss.

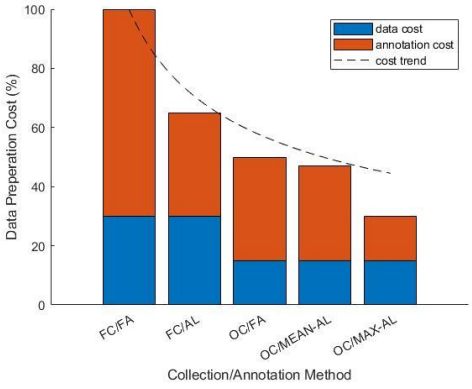

Figure 2 - Cost saving of capture and annotation for methods: Full capture/Full annotate (FC/FA), Full capture/Active learning (FC/AL), Optimized capture/Full annotate (OC/FA), optimized capture/Active learning from mean accuracy (OC/Mean-AL), optimized capture/Active learning from max accuracy (OC/Max-AL).

This case study has shown statistical power curves and active learning allow for significant optimization in both sample and annotation set size. This reduction in sample size represents a direct cost reduction in producing a viable dataset. Through the example case study on the BUSI dataset, this gave a 50% cost reduction for an accuracy loss of 4% when considering mean response or a 66% cost reduction for an accuracy loss of 3.75% from theoretical maximums at full dataset size using Alexnet as a performance benchmark.

IV. Discussion

Estimating clinical trial sample size is a standardized practice allowing clinical researchers to fit the size of studies so they are feasible clinically and financially within the timeframe available. This same approach has be used to predict the effectiveness of increasing the dataset for machine learning and inform researchers as to the usefulness of further annotation. Examining the results from the three datasets in this simple classifier per image study, the NN response to sample size trend is clear and can be exploited to cap data collection and annotation costs. The breast and lung datasets both showed diminished returns after 40-50% of the dataset with the much larger fetal dataset reaching diminishing returns between 10-20%. This means that data collection and annotation can be reduced without significant accuracy loss.

This is an iterative process that can be done throughout the data collection process to plan subsequent data capture and annotation with each successive cycle providing a more accurate indicator of how much additional data is required to achieve the desired results. This is due to the power curve convergence observed (Fig. 1a), where the power curve converges onto a stable value predicting accuracy for arbitrary sample sizes.

This process can also be used retrospectively to determine possible accuracy increases from additional sampling. Considering Fig. 3, for the BUSI dataset, extrapolation of the BUSI power curve suggests that additional accuracy can be achieved but a doubling of sample size will only yield a 4% improvement in accuracy. Similarly, no improvement is expected for the large fetal dataset with increasing sample size, allowing decision makers to plan appropriately.

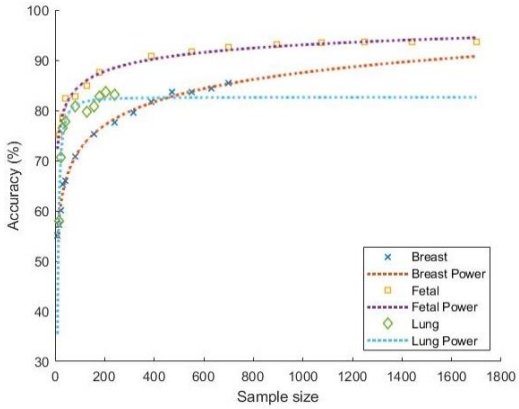

Figure 3 - Power curve extrapolated trends from mean NN accuracy results of all three datasets normalized to a patient set sample size.

When evaluating the effectiveness of active learning across each of the datasets there is a measurable decrease in dataset annotation requirement but only up to ~60% dataset usage, above this point, it provided minimal improvement over untargeted annotation methods. Using a targeted approach such as active learning it is possible to further reduce the sample size by identifying where the neural network is least sure of the result, but this effect also diminishes in value as the training set increases in size.

V. Conclusion

In order to progress machine learning research further in ultrasound, significant investment in data collection and annotation will be required, but this burden can be significantly reduced using targeted sampling methods. This paper proposed a 2 phase method to optimize the collection and annotation processes, demonstrated rigorously and repeatably per phase on public datasets, before being shown as effective together on a practical example. These methods provide ultrasound researchers with a method to not only identify the most effective sample size when collecting data but also a method to maximize annotation effectiveness potentially producing robust algorithms at reduced cost.

Using a simple statistical power curve to predict accuracy results will allow researchers to provide an estimate as to the usefulness of additional data in NN training. This allows the results of limited feasibility studies using relatively small datasets to inform the selection and design of subsequent larger studies just as with clinical trials. Active learning using uncertainty sampling to reduces annotation effort on a smaller subset, useful especially where a data collection is ongoing or a large unlabeled dataset is available. These methods can be used independently, but give best results when used together. This method is to be used to streamline future clinical trials by the authors, and applied to other ultrasound applications like NDT and industrial inspection to standardize methods.

While this study has focused on accuracy as the sole metric, additional validation metrics would allow for these methods to be significantly finetuned to produce the best network response at the lowest cost. Alexnet and per image classification provided an adequate baseline but further research into more complex architectures and ultrasound datasets with a more complex taxonomy would offer further insight into developing additional methods to reduce the cost of producing and annotating ultrasound data, especially where overlapping classifiers exist within the dataset.

References

[1] B. Luijten *et al.*, "Adaptive ultrasound beamforming using deep learning," *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 3967-3978, 2020.

[2] L. H. Lee, Y. Gao, and J. A. Noble, "Principled ultrasound data augmentation for classification of standard planes," in *International Conference on Information Processing in Medical Imaging*, 2021, pp. 729-741: Springer.

[3] P.-H. Yeung, M. Aliasi, A. T. Papageorghiou, M. Haak, W. Xie, and A. I. Namburete, "Learning to map 2D ultrasound images into 3D space with minimal human annotation," *Medical Image Analysis*, vol. 70, p. 101998, 2021.

[4] R. Huang, J. A. Noble, and A. I. Namburete, "Omni-supervised learning: scaling up to large unlabelled medical datasets," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2018, pp. 572-580: Springer.

[5] L. Schmarje, *et al.*, "A survey on semi-, self-and unsupervised learning for image classification," *IEEE Access*, vol. 9, pp. 82146-82168, 2021.

[6] K. Hemming, *et al.* "How to design efficient cluster randomised trials," *bmj*, vol. 358, 2017.

[7] D. J. Biau, S. Kernéis, and R. Porcher, "Statistics in brief: the importance of sample size in the planning and interpretation of medical research," *Clinical orthopaedics and related research*, vol. 466, no. 9, pp. 2282-2288, 2008.

[8] A. Rokem, Y. Wu, and A. Y. Lee, "Assessment of the need for separate test set and number of medical images necessary for deep learning: a sub-sampling study," *bioRxiv*, p. 196659, 2017.

[9] J. Cho, *et al.*, "How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?," *arXiv preprint arXiv:1511.06348*, 2015.

[10] H. G. Schnack and R. S. Kahn, "Detecting neuroimaging biomarkers for psychiatric disorders: sample size matters," *Frontiers in psychiatry*, vol. 7, p. 50, 2016.

[11] J. Escudero, J. P. Zajicek, and E. Ifeachor, "Machine Learning classification of MRI features of Alzheimer's disease and mild cognitive impairment subjects to reduce the sample size in clinical trials," in *2011 Annual international conference of the IEEE engineering in medicine and biology society*, 2011, pp. 7957-7960: IEEE.

[12] D. Calaprice-Whitty, K. Galil, W. Salloum, A. Zariv, and B. Jimenez, "Improving clinical trial participant prescreening with artificial intelligence (AI): a comparison of the results of AI-assisted vs standard methods in 3 oncology trials," *Therapeutic innovation & regulatory science*, vol. 54, no. 1, pp. 69-74, 2020.

[13] D. L. Labovitz, *et al.*, "Using artificial intelligence to reduce the risk of nonadherence in patients on anticoagulation therapy," *Stroke*, vol. 48, no. 5, pp. 1416-1419, 2017.

[14] B. S. Glicksberg *et al.*, "Automated disease cohort selection using word embeddings from electronic health records," in *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2018: Proceedings of the Pacific Symposium*, 2018, pp. 145-156: World Scientific.

[15] J. Romero, S. Chiang, and D. M. Goldenholz, "Can machine learning improve randomized clinical trial analysis?," *Seizure*, vol. 91, pp. 499-502, 2021.

[16] N. Zhou and P. Manser, "Does including machine learning predictions in ALS clinical trial analysis improve statistical power?," *Annals of Clinical and Translational Neurology*, vol. 7, no. 10, pp. 1756-1765, 2020.

[17] M. Hasenjäger and H. Ritter, "Active learning in neural networks," in *New learning paradigms in soft computing*: Springer, 2002, pp. 137-169.

[18] Y. Yang, Z. Ma, F. Nie, X. Chang, and A. G. Hauptmann, "Multi-class active learning by uncertainty sampling with diversity maximization," *International Journal of Computer Vision*, vol. 113, no. 2, pp. 113-127, 2015.

[19] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *SIGIR'94*, 1994, pp. 3-12: Springer.

[20] J. Yun, J. Oh, and I. Yun, "Gradually Applying Weakly Supervised and Active Learning for Mass Detection in Breast Ultrasound Images," *Applied Sciences*, vol. 10, no. 13, 2020.

[21] G. Liu *et al.*, "Breast Ultrasound Tumor Detection Based on Active Learning and Deep Learning," EasyChair 2021.

[22] L. Gao *et al.*, "Multi-modal active learning for automatic liver fibrosis diagnosis based on ultrasound shear wave elastography," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021, pp. 410-414: IEEE.

[23] L. Liu, *et al.*, "Semi-supervised active learning for COVID-19 lung ultrasound multi-symptom classification," in *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, 2020, pp. 1268-1273: IEEE.

[24] W. Al-Dhabyani, *et al.*, "Dataset of breast ultrasound images," *Data in brief*, vol. 28, p. 104863, 2020.

[25] J. Born *et al.*, "Accelerating detection of lung pathologies with explainable ultrasound image analysis," *Applied Sciences*, vol. 11, no. 2, p. 672, 2021.

[26] X. P. Burgos-Artizzu *et al.*, "Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes," *Scientific Reports*, vol. 10, no. 1, pp. 1-12, 2020.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097-1105, 2012.

[28] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211-252, 2015.