Biomedical Signal Processing and Control 92 (2024) 106048

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

A cost focused framework for optimizing collection and annotation of ultrasound datasets

Alistair Lawley [a,b,c,*], Rory Hampson [b], Kevin Worrall [a,c], Gordon Dobie [a,b]

[a] Future Ultrasound CDT (FUSE), University of Strathclyde, 204 George St., Glasgow G1 1XW, UK
[b] Centre for Ultrasonic Engineering (CUE), University of Strathclyde, 204 George St., Glasgow G1 1XW, UK
[c] James Watt School of Engineering, University of Glasgow, Glasgow G12 8QQ, UK

ARTICLE INFO

Keywords:
Active learning
Medical imaging
Ultrasound
Cost effectiveness
Deep learning

ABSTRACT

Machine learning for medical ultrasound imaging encounters a major challenge: the prohibitive costs of producing and annotating clinical data. The issue of cost vs size is well understood in the context of clinical trials. These same methods can be applied to optimize the data collection and annotation process, ultimately reducing machine learning project cost and times in feasibility studies. This paper presents a two-phase framework for quantifying the cost of data collection using iterative accuracy/sample size predictions and active learning to guide/optimize full human annotation in medical ultrasound imaging for machine learning purposes. The paper demonstrated potential cost reductions using public breast, fetal, and lung ultrasound datasets and a practical case study on Breast Ultrasound. The results show that just as with clinical trials, the relationship between dataset size and final accuracy can be predicted, with the majority of accuracy improvements occurring using only 40–50% of the data dependent on tolerance measure. Manual annotation can be reduced further using active learning, resulting in a representative cost reduction of 66% with a tolerance measure of around 4% accuracy drop from theoretical maximums. The significance of this work lies in its ability to quantify how much additional data and annotation will be required to achieve a specific research objective. These methods are already well understood by clinical funders and so provide a valuable and effective framework for feasibility and pilot studies where machine learning will be applied within a fixed budget to maximize predictive gains, informing resourcing and further clinical study.

1. Introduction

1.1. Motivation for cost analysis and 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]. Despite this there is extremely limited availability of annotated data for machine learning (ML). There are very few large-scale public ultrasound datasets available and where clinical data does exist there is often no useful annotation to produce an effective ground truth. This is not a problem unique to ultrasound, the inherent cost of producing high quality data and subsequent complex clinical annotation required to inform the neural network means that generating appropriate datasets for diagnostic quality deep learning is a major investment [2]. Therefore, when designing or commissioning a research project applying machine learning to ultrasound, it is important to factor in the financial and clinical cost of producing and annotating the data as well as the machine learning itself. There are many methods aimed at optimizing neural network response to training, such as transfer learning [3,4] and few-shot learning [5,6], as well as methods for reducing the burden of annotation [7] by reducing human-model supervision [8,9] and self-supervision [10,11] such as masked autoencoders [12], and clustering [13]. These methods while effective are not designed to consider the real-world barriers [14,15] to machine learning research, factors such as the cost and time of data collection that could completely prevent a study from being performed [16]. Fortunately, there is already a tried and tested methodologies within medical research for performing this type of analysis that is well known to clinical funding bodies and commissioners: those used for designing clinical and random control trials [17], where is often not clinically or financially viable to sample a large population, therefore a smaller feasibility study is first performed, and the results analyzed to calculate the size of subsequent trials.

* Corresponding author at: Centre for Ultrasonic Engineering (CUE), University of Strathclyde, 204 George St., Glasgow G1 1XW, UK.
E-mail address: alistair.lawley@strath.ac.uk (A. Lawley).

https://doi.org/10.1016/j.bspc.2024.106048
Received 9 June 2023; Received in revised form 9 November 2023; Accepted 29 January 2024
Available online 7 February 2024

The process of developing a dataset for machine learning has many similarities to designing random control trials, the addition of more data has diminishing returns on how much it improves the accuracy of the result some tradeoffs can be made to maintain the validity of the study while making it cost effective [18]. Where funding and clinical resources are finite, it is important to weigh the value of additional data and annotation against the time and cost of producing it. This paper seeks to apply a framework similar to that of clinical trials [19], such as using a statistical power curve function during sampling to quantify diminishing returns in training result [20]. Cost will then be further optimized by applying active learning [21] to reduce the cost of data annotation after collection by targeting manual annotation to those parts of the data that most require additional attention by an expert clinician. Time and cost are critical metrics to decision makers attempting to balance the risks and priorities of medical device and imaging research involving machine learning but has seen limited focus in the current literature.

1.2. Cost / time optimization methods and applications

Statistical power analysis is the concept of estimating the effect of a result within a given sample and to what extent this can be generalized to a larger sample size based upon its statistical significance [22,23]. A power curve represents every arrangement of power and difference for each sample size when the significance level and the standard deviation are held constant. Factors that may effect statistical power are as follows: the statistical significance criterion used in the test, the magnitude of the effect of interest in the population and the sample size used to detect the effect [24,25]. Statistical power analysis has been previously used to predict classification performance [26,27], and used to determine sample size in cases such as retinal optical coherence tomography (OCT) [28] and in magnetic resonance imaging (MRI) [29] and small phantom studies [30] but has yet to be explored for ultrasound.

Cost is often the limiting factor when proposing a machine learning study, especially in medical imaging where sample size can play a major role in study cost. One common method of sample selection is to use a derivation of the Widrow-Hoff learning rule [31] that suggests the use of an number (such as 10, 100 or 1000) sets of data for every imaging feature that will be used in the model. This method is somewhat arbitrary and may come up with sample sizes that are too small or large for actual training purposes depending on the feature-set being examined with limited possibility for cost to performance comparison. Model-based sampling based on the algorithmic characteristics such as generalization [32,33], or convergence [34] can provide good baseline for sample size selection based on threshold criteria but can be more difficult to directly relate to costs. The proposed method uses empirical curve fitting for sample size determination [35], allowing for accurate prediction of time and cost of producing the data similar to that used for control trials [36].

Where ultrasound data is available without annotation, there is an opportunity to apply a targeted approach to sample labelling. There are many ways to reduce the cost of annotation in the early stages of data analysis such as using unsupervised clustering methods [37], in this case active learning used to target manual clinical annotation time more effectively. While more expensive than self-supervised and automated methods, full human annotation is already recognized as appropriate by regulatory bodies for medical device research and as such was used as the exampled method in this study. Use of a *meta*-heuristic optimization algorithms for thresholding and feature selection should be considered as a potential method for increasing network accuracy when defining the dataset annotation [38].

Active learning is a technique where a neural network is used to analyze a dataset and the resulting predictions used to target future training effort on a selected portion of the dataset [7]. There are many common methods of active learning within machine learning [7], such as using an unlabeled pool where a network chooses the best examples of a classifier known as diversity sampling [39]. In this case selective uncertainty sampling [40] is used to identify where the neural network has the lowest confidence in its prediction and target those images for annotation. This has been used previously as a method of dataset selection criteria for omni-supervised learning [41]. This forms an active learning loop, allowing for the consistent querying of the learning network to better inform the annotation process (Fig. 1). Active learning has already been successfully applied to breast ultrasound using a weakly supervised approach, as well as in the detection of breast masses [42,43], in the multi-model detection of liver fibrosis for ultrasound elastography [44], and in semi-supervised covid lung disease classification [45].

1.3. Structure and scope of the paper

This paper proposes a 2-phase prescriptive framework for optimizing data capture and data annotation. The datasets, machine learning algorithms, and data control measures of each phase is shown in section 2. The efficacy of each phase is shown independently in section 3.1 and 3.2. A case study is presented in section 3.3 using publicly available data, demonstrating the framework for reducing the cost of data capture and annotation compared to the common approach of using fully annotated arbitrary data sets.

This paper examines how:

- Ultrasound dataset size effects neural network accuracy performance for three publicly available datasets to determine optimal sampling size.
- Uses predictive curve from a small sample to inform further data collection for machine learning based on a cost benefit analysis.
- Compares that sample size prediction compared to the real result from the dataset.
- Tests the effectiveness of uncertainty sampled active learning for ultrasound data for reducing the cost of annotation.
- Combines these methods to determine optimal sample size and annotation level for maximizing accuracy whilst minimizing cost.

The use of curve fitting for determination of sample size may not be as efficient as formulaic or model-based sample size selection methods, but this empirical testing provides a simple robust basis for predictive modeling of dataset cost. While other semi-supervised, fully automated, or clustering methods may provide inexpensive labelling, where manual

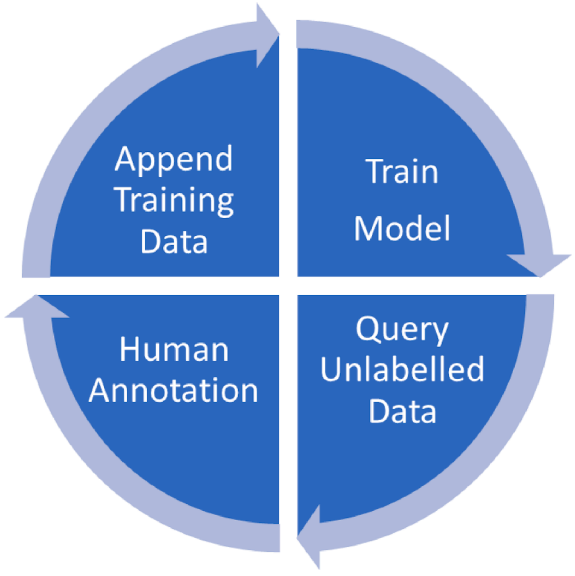

Fig. 1. Active Learning Cycle. This shows the cyclical iterative nature of active learning within machine learning.

annotation must be maintained as the primary form of annotation, such as cases where regulatory approval is a future consideration, uncertainty sampled active learning allows for manual annotation to be targeted at those classifications with the weakest predictions while stronger classifiers are off loaded to a semi-automated labelling process. When combined these methods form a novel 2 phase process to reduce the cost of producing a dataset for machine learning for ultrasound by optimizing collection and labelling of data.

Despite increased demand for scans and a shortage of sonographers [46–48], funding for ultrasound-based machine-learning research is limited. Researchers must therefore ensure that studies are efficient and cost effective, leveraging as much value as possible from collection studies and clinical time.

2. Materials and methods

2.1. Proposed method for optimizing sampling / annotation

Phase 1 uses power curves, a method common in determining size of clinical trials based off factors such as population size and available resources. Applying this technique allows these same factors can be considered during the collection of data for machine learning and used to determine a rough performance estimate from the size of dataset. 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 like those described in section 2.2. Phase 2 uses semi-supervised active learning to automatically annotate a proportion of the dataset, where a reduction in performance can be accepted, a error tolerance threshold can be applied, leading to significant cost and time savings. In this work, a convolutional neural network (CNN) is used as described in section 2.3, Alexnet is used a well-known and understood benchmark, however other neural network architectures can be easily substituted, and would likely provide a more optimal result. Hyperparameters are also exemplar and not intended as recommendation of optimal settings, but merely to demonstrate the framework in action.

Phase 1 – Optimized data set Capture

Phase 1: Estimate the power curve and predict required dataset size (Fig. 2). Below are explanatory notes for the flow chart:

1. A neural network is trained on a small sample of annotated data (dependent on experimental constraints e.g. 10–100 samples). The dataset should be split into training and test sets. Validation metrics (such as accuracy) are saved.
2. An additional subset of annotated samples (ideally in equal chunks) is added to the dataset (re-randomising training and validation sets is advised).
3. Neural network is retrained and tested. The chosen validation metrics are saved.
4. The validation metrics are plotted against dataset size and a power curve is fitted to the data.
5. Repeat steps 2–4 until curve fit is 'stable' at desired statistical power and accuracy. Stability is when subsequent sample groups predict end accuracies within your desired tolerance (such as within 2 %).
6. Plot the power curve (e.g. accuracy vs sample size) can then be used to determine the required dataset size for desired/acceptable validation metric.

Phase 2 – Optimizing annotation

In cases where excess samples have been captured, particularly in large unannotated datasets or where data is being repurposed, active learning, detailed in section 2.5, can be used to selectively target samples that the CNN has the most difficulty identifying for manual annotation, thus optimizing the annotated sample set as seen in Fig. 3. This process uses selective uncertainty sampling to minimize manual annotation of remaining data. Below are explanatory notes for the flow chart:

1. Train and validate a neural network on available annotated data (such as the sample set produced in phase one).
2. Identify least certain samples on unannotated data, where the CNN has least certainty detecting particular classifiers (e.g. bottom 50 samples).
3. Manually annotate next batch of data with additional focus on identified weak classifiers.
4. Combine new and old batches and reshuffle the dataset.
5. Train new neural network and evaluate result using the validation set.

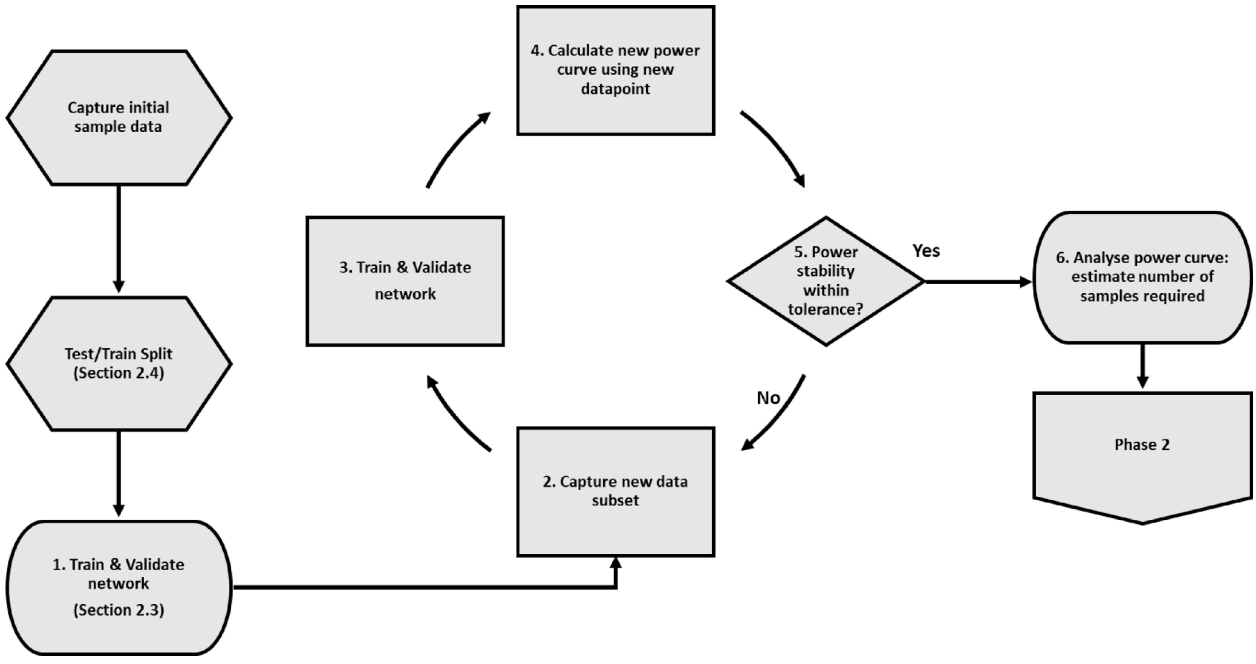

Fig. 2. Phase 1 flow chart showing collection cycle and subsequent power curve analysis leading to the determination of dataset size based on curve fit.

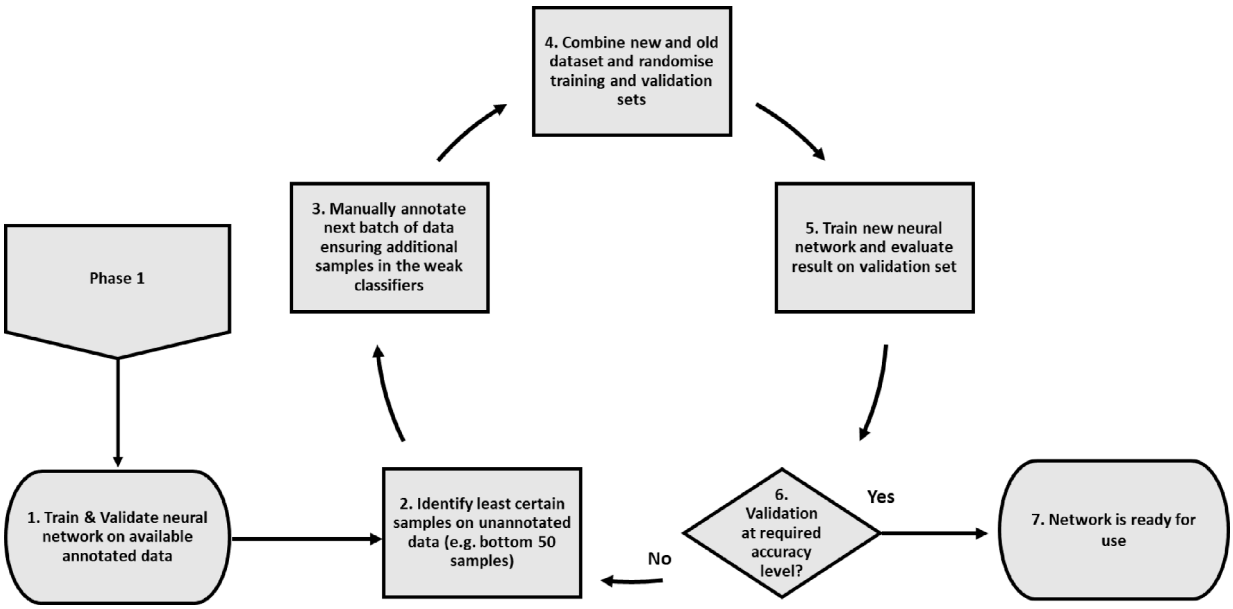

Fig. 3. Phase 2 Active learning cycle for annotation.

6. Use the validation metrics (such as accuracy level) to decide if additional samples are required. Consideration should be given to the effect of dataset balance on classification as well as comparisons made to predicted values from Phase 1 to ensure training validity.
7. Cease annotation when (cost and accuracy) metrics are within acceptable parameters.

This iterative process allows this technique to be applied naturally during the data collection and annotation process, such as during a pilot study. A new round of training can be performed upon receipt of a new batch of data, adding an additional datapoint for the power curve. If data is in a single large batch it, like those introduced in subsection 2.2, can be divided into percentages such as in this study, to produce the required increments. This is shown in the case study, Section 3.3.

2.2. Datasets

2.2.1. Breast lesion

The BUSI breast lesion ultrasound dataset (available at [49]) (Fig. 4) consists of breast ultrasound images of 600 women between the ages of 25 and 75. The ground truth images were presented with original images. The images were categorized into three classifiers, which are normal, benign, and malignant, while a segmentation mask is available in this dataset, a simple classification ground truth is substituted to enable comparison with other datasets studied.

2.2.2. Covid lung

The lung ultrasound dataset [50,51] (available at [52])(Fig. 5), consists of 179 videos (64 COVID, 49 bacterial pneumonia, 66 healthy), 53 images (18x COVID, 20x bacterial pneumonia, 15x healthy) from convex probes and 17 videos (6 COVID, 2 bacterial pneumonia, 9 healthy) and 6 images (4 COVID, 2 bacterial pneumonia) from linear probes. Cases of viral pneumonia in the dataset were excluded as it consisted of only 6 cases and there is evidence to suggest ultrasound can differentiate between viral and bacterial pneumonia [53,54] meaning including it in a single pneumonia classifier would be counter intuitive.

2.2.3. Fetal planes

The fetal ultrasound dataset [55] (available: [56]) consists of around 12,000 images from 1792 patients and is split into 6 classifiers: fetal abdomen, brain, femur, thorax, maternal cervix, and a generic 'other' classifier as exampled in Fig. 6.

2.3. Deep learning

The experimentation was performed using the Pytorch framework [57], on a computer with an Intel CPU with a clock speed of 2.4Ghz and a Nvidia 3060 GPU. A standard Alexnet neural network that had been pretrained using the ImageNet Challenge dataset [58] was used with the final layer output reduced to fit the classification requirements of the dataset. Alexnet [59] was selected to provide a baseline to study dataset

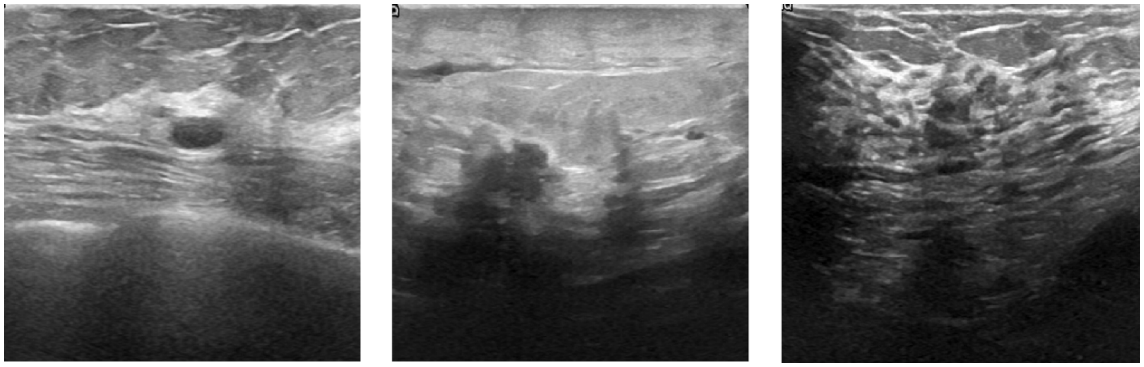

Fig. 4. Examples of breast lesion ultrasound classifiers from the BUSI dataset [23], left benign, center: malignant, right: normal.

Fig. 5. Examples of Covid Lung Ultrasound Dataset [24]. left: Covid, center: bacterial pneumonia, right: normal.

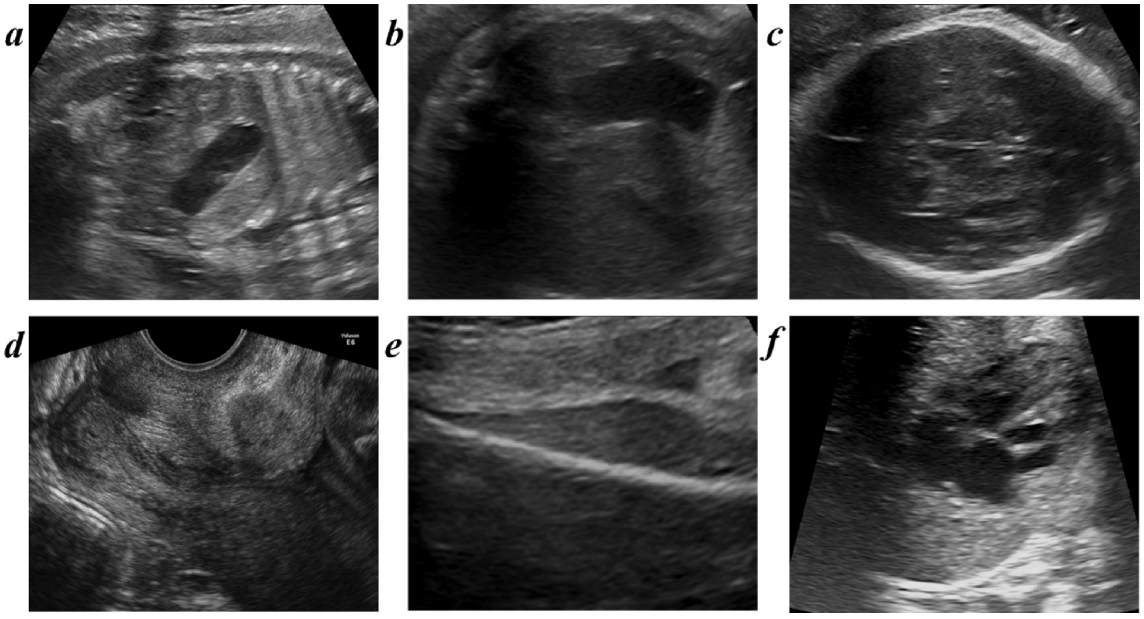

Fig. 6. Examples of fetal Plane Ultrasound Dataset [28]. a: other, b: abdomen, c: brain, d: maternal cervix, e: femur, f: thorax.

selection size, using the training parameters in Table 1. Network selection and hyperparameter settings are simple and designed to example the framework in action but not optimal for high precision machine learning tasks. The complexity of the data and requisite predictors and feature sets should be considered when designing machine learning studies using techniques as seen in these recent studies [60–63]. The images were contrast normalized and compressed into tensors sized 299 × 299. Test/train split was performed on a per image basis as there is no known repeated data within the set. No additional data augmentation was performed as this would potentially confound results.

Table 1
Sample Hyperparameters used in example network in this paper.

Hyperparameters	Value
Activation Functions	SoftMax
Learning Rate	0.001
Training Iterations	80
Epochs	20
Optimizer	ADAM
Momentum	0.9
Dropout	0.5

2.4. Training set size

The dataset was initially split randomly 80/20 into test and training sets as documented above, an additional split was performed on the training set taking a percentage of the data for training using stratified random sampling [64], to ensure a representative sampling of each classifier. The percentage of data is increased with each iteration in order to determine the relations between sample size and accuracy. The percentage of the breast and fetal datasets were between 1 % and 5 % then in increments of 10 % thereafter. The Lung dataset has a smaller sample size and so the smallest split tested was 5 %, then in increments of 10 %. This was done to simulate the collection of data over time. Each experimental training run was performed 80 times over 20 epoch each. The use of an unseen test set means that any overfitting that has occurred will already be reflected in the test result that is used in the calculation of the power curve, with the accuracy of an overfitted network substantially lower when validated on the unseen test data. Each subsequent round of collection increases the size of both training and test sets and testing adds a new datapoint to the power curve further refining the curve fit. The exampled approach is overly simplified, performing a stratified random split with each new experimental run and looking at only very simple classification tasks, more complex datasets should consider folded-based cross validation and the potential improvements that could be made through careful feature selection and

ensemble learning models as methods to reduce overfitting as exampled in recent studies [65–67].

2.5. Active learning

At its core, active learning is a technique whereby the learner plays a role in specifying the content they learn [7]. An uncertainty sampling [39,41] method is used whereby the images with the lowest confidence was selected for annotation. This was performed for each percentage of the dataset from 10 to 90 % with the active learning performed on the remaining percentage of the dataset also using a threshold percentage to specify an additional proportion of the dataset for annotation (as can be seen in Fig. 7). Each active learning threshold was tested 20 times over 20 epochs each.

3. Results

3.1. Size to accuracy of dataset

Examining the breast dataset mean accuracy results for 80 neural networks per threshold percentage (Fig. 8), the highest mean accuracy of 85.42 % was achieved using 90 % of the data, contrast to 79.6 % using 40 % of the data and 75.29 % at 20 %, a difference of 5.82 % and 10.13 % respectively. Increasing dataset size reduces the variation as seen in the standard deviation between neural networks with an average of 6.17 at 1 % of the data, down to 2.42 at 90 %. Selecting the neural network with the highest accuracy for each percentile shows that the highest accuracy network with 91.72 % was produced with only 60 % of the dataset, in comparison to 82.8 % using 20 % of the dataset a difference of just 8.92 %, there were significant diminishing returns on data investment after 30–40 % of the data is used.

Using the mean accuracy data, it is possible to extrapolate a close approximation of data to fit classification accuracy, a fitted curve from just 10 % of the data can be used to approximate the amount of additional data required to reach a certain level of accuracy similar to that used in clinical studies, with each additional data point improving the fit further.

In the lung dataset (Fig. 9) the difference between the highest mean accuracy of 83.66 % and 80.87 % using 40 % of the dataset was just 2.79 %. The trend of reducing standard deviation as dataset size increases is less obvious, while the initial deviation is 12.34 at only 5 % of the dataset it is reduced to 7.62 by around 10 % of the dataset but remains unstable but does achieve the lowest standard deviation at 5.25 at around 90 % of the dataset. When the highest accuracy neural networks are considered, an accuracy of 89.93 % is achieved at just 30 % of the data, with diminishing returns until 70–80 % where a significant improvement is achieved with results of 94.36 % at 70 % and 95.96 % at 80 % of the data. As previously seen in Fig. 6 the same data trend is possible and is visible in the mean accuracy data for the lung dataset. A statistical curve is used to predict CNN accuracy for sample sizes.

The fetal plane dataset contains over 12,000 samples from over 1700 patients making it the largest dataset assessed, using only 20 % (around 2400 samples) the mean accuracy reached over 90 %, with additional data providing diminishing returns for the additional data added. The 1SD trends downwards from 3.40 to 0.54 using 80 % of the dataset. The fetal data also exhibits the same trend from the power curve (Fig. 10) despite containing substantially more data and classifiers than the previous two datasets, the difference between mean and highest result after 50 % of the dataset (6000 samples) is 0.66 % of that achieved with 90 % of the dataset, it is also within 0.56 % of the highest achieved result of 94.97 %.

The use of curve fitting as a method of sample selection is empirical has clearly shown to be effective at determining sample size in all three datasets, a clear followable trend that can be seen in the network response and can be used to extrapolate data requirements based off this trend. The high initial standard deviation in accuracy result in training result seen in all three datasets is due to factors such as overfitting, sample randomisation and training performance. As sample size increases the so does the stability of the training process, due the test set being unseen the results from the networks form a clear accuracy trend regardless of these factors. Where sample size will be consistently small harmonic mean (F-1 Score in Table 2) should be factored into result metrics to ensure network response is truly representative of learnt classification.

3.2. Active learning

Comparing the results of using active learning to target the lowest predicted accuracy using a threshold to that of annotating the same percentage with no targeting shows a small consistent improvement. The highest accuracy of 92.99 % is achieved at 60 % of the dataset, as can be seen in Table 2, the neural network is already performing consistently with a weighted average precision of 90 %, recall of 92 % leading to an F-1 Score, the combination of precision and recall of 0.91.

Comparing the default annotation values to those using active learning shown in Table 3 and Fig. 9, shows that the majority of learning can be achieved using between 40 and 50 % of the data (in the region of 300–400 sample sets). For example, when 10 % of the data is used for the teacher network and an additional 30 % is annotated using active learning then a mean result of 82.99 was achieved that is only 4.3 % less than when trained with the complete dataset where a mean result of 87.29 % was achieved. Where 20 % of the dataset is used to train the teacher network then this difference drops to just 3.28 %. The variation of accuracy after 60 % of the dataset is likely due to the probabilistic nature of neural network training rather than the dataset itself. The statistical maximum result of 92.99 % was achieved at all subsequent dataset proportions above 50 % when trained exhaustively, but this may not be feasible to achieve in practice. This supports the hypothesis that additional data provides limited, to no, return on investment after this point.

When comparing the data for the default annotation technique with active learning for the breast dataset (Fig. 11), the mean active learning

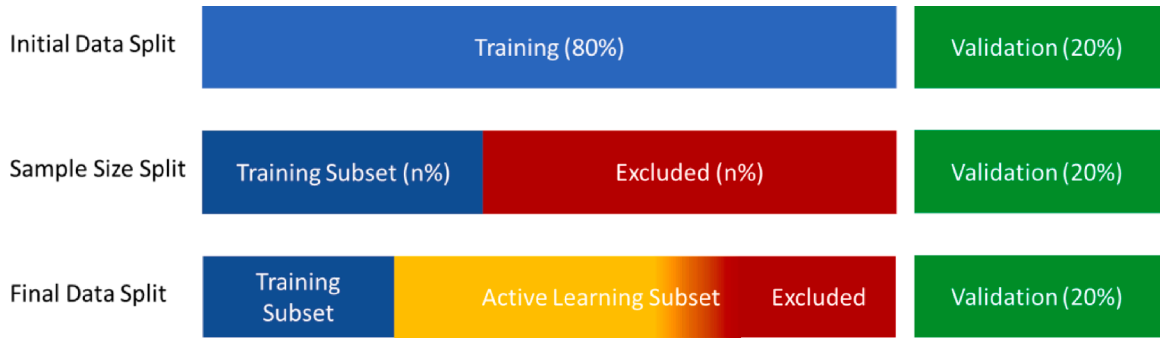

Fig. 7. Diagram of active learning dataset split method showing proportion of data used for training and threshold for additional annotation.

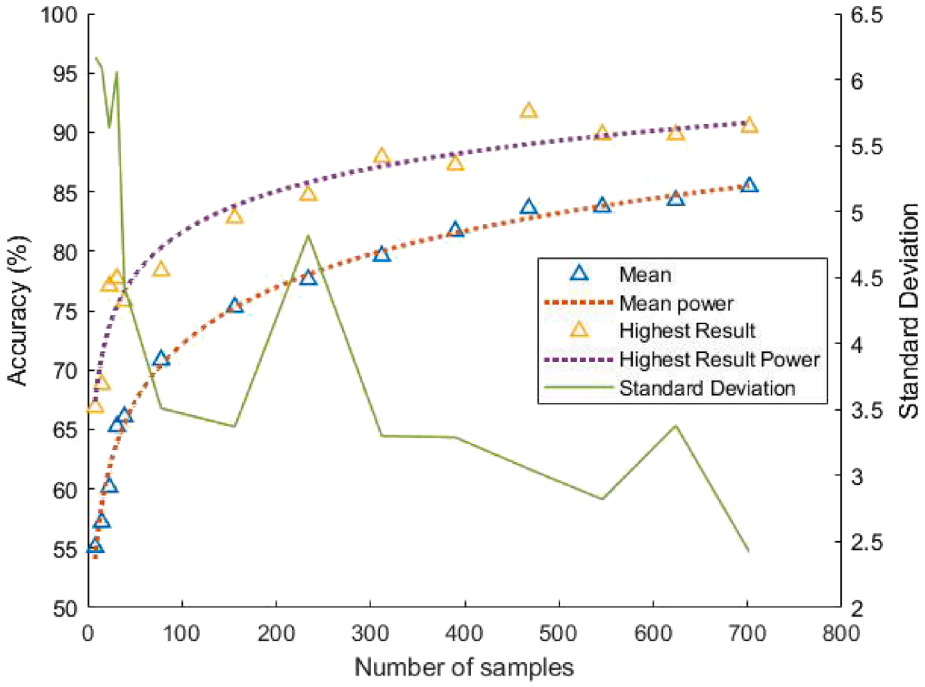

Fig. 8. Accuracy of mean and highest result with associated power curves for neural network response for breast dataset.

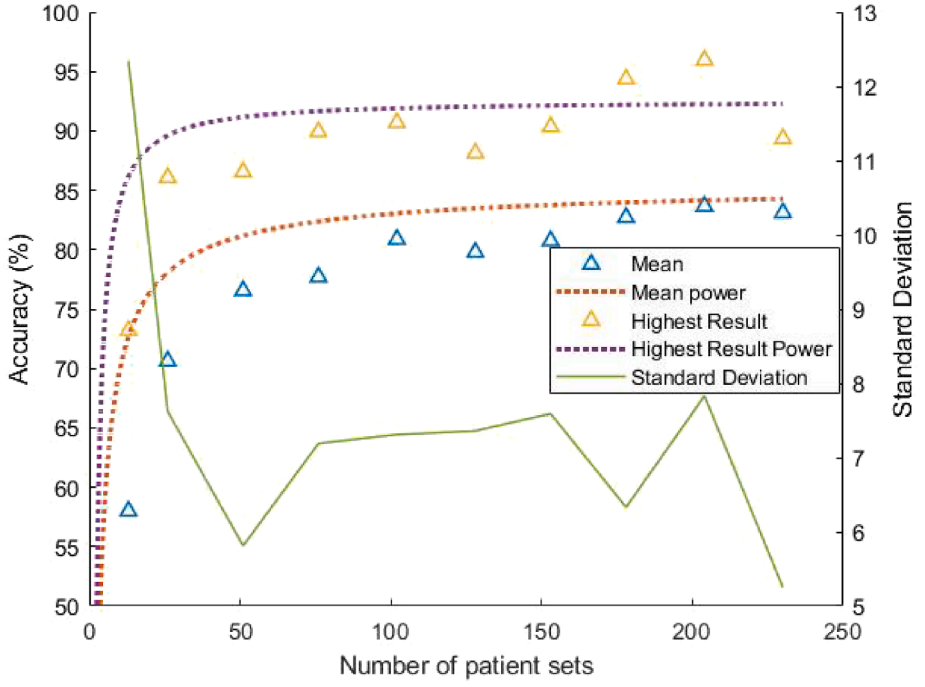

Fig. 9. Accuracy of mean and highest result with associated power curves for neural network response for lung dataset.

consistently is above that of the default annotation technique, although this improvement suffers from diminishing returns after 50 % of the dataset is used. While there is significant improvement in the highest accuracy results achieved even with only 10 % active learning there is significant training variance at lower dataset sizes that would need to be accounted for in the training methodology.

Lung ultrasound also performed well with active learning (Table 4), with a network trained on 20 % and an additional 10 % active learning was able to achieve a mean accuracy of 82.35 %, just 4.3 % less than the highest achieved mean accuracy of 86.65 % from a network trained on 30 % of the data and an additional 50 % targeted through active learning.

When comparing active learning to default annotation methods (Fig. 12), the active learning does improve mean accuracy results but does not significantly improve training of high accuracy models after 60 % of the data is in use due to the wide variation in training accuracy achieved.

The variation in training accuracy is highest for this dataset, out of the three, which is attributed to the low base dataset size, meaning that the CNN training is more susceptible to statistical anomalies in the data,

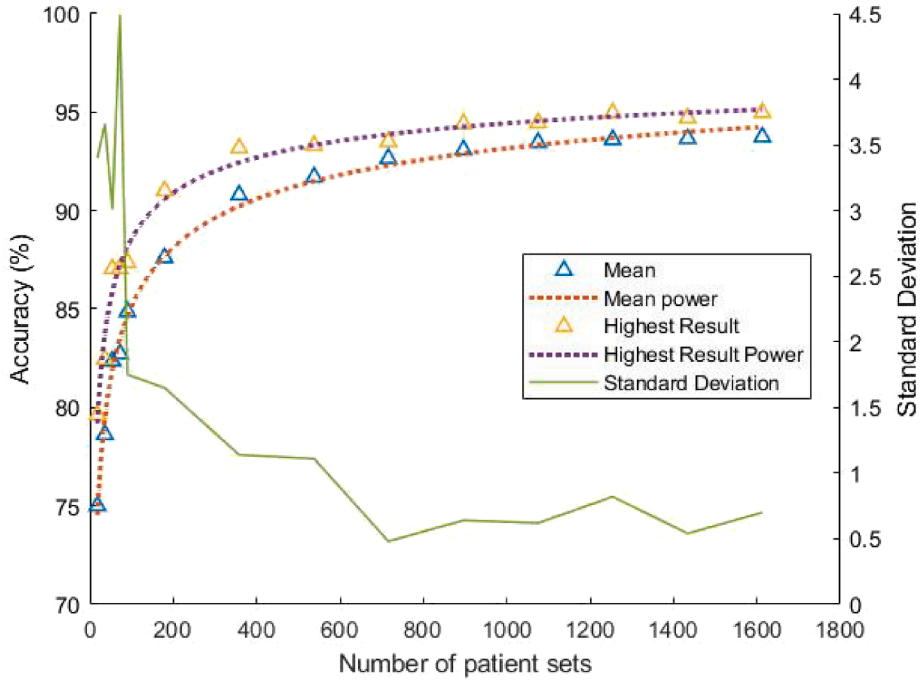

Fig. 10. Accuracy of mean and highest result with associated power curves for neural network response for fetal plane.

Table 2
Precision, Recall and F-1 score for top performing network (BUSI (breast) dataset based off a network trained on 10% of the dataset with an additional 40% annotated using active learning.

Classifier	Precision	Recall	F1-Score
0 – Benign	90	88	0.89
1 – Malignant	94	92	0.93
2 – Normal	87	96	0.91
Average	90	92	0.91

a well-known phenomenon. Despite this, the trend improvement is still evident, although with smaller returns initially than larger datasets.

The fetal ultrasound data has a significantly larger sample size and double the number of classifiers than the previous two datasets. As seen in Table 5, there was a 2.22 % improvement when active learning was used to annotate 10 % of the dataset but is subject to diminishing returns as the highest accuracy result achieved was 94.40 % using 80 % of the dataset where the active learning had been trained using a dataset with 60 % of the data an improvement of only 1.39 %.

When comparing active learning to default annotation methods in Fig. 13, an initial accuracy boost, after using above 60 % data, training variance becomes a significant factor with active learning achieving only limited improvements.

The use of uncertainty sampled active learning is shown to boost classification accuracy performance of all three datasets with most improvement seen prior to 50 % human annotation of the dataset with substantial diminishing returns after this point. Targeting human annotation to the most essential data while allow semi-automated annotation to potentially fill in for classifiers where the network already has high confidence. Active learning where 50 % of the dataset has already been annotated shows no performance drop over full human annotation, suggesting this data could annotated using active learning with almost no loss of reliability despite the potential of labelling error.

3.3. Case study

Having shown in section 3.1 that accuracy vs. sample size follows a power law and therefore has significantly diminishing returns after a certain point. As shown in section 3.2, active learning produces improvements in accuracy with low amounts of initially annotated data, again with diminishing returns as annotation proportion increases, we can now consider what this means in terms of costs for collection and annotation.

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 CNNs respectively [68].

Table 3
Active Learning improvement in BUSI (breast) dataset based on the percentage of data used – Mean Results.

	Percentage of dataset used for training										
Percentage of training data actively learned	-	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %	100 %
	0 %	70.84	75.29	77.60	79.60	81.69	83.61	83.73	84.29	85.42	85.42
	10 %		77.39	80.51	82.99	85.25	86.56	86.62	85.67	85.86	86.85
	20 %			80.10	84.01	84.78	85.73	85.57	85.35	86.69	87.20
	30 %				83.03	84.94	86.59	85.70	86.08	87.17	86.78
	40 %					84.32	85.49	85.46	85.56	85.73	85.99
	50 %						84.28	85.16	86.56	86.37	86.07
	60 %							85.29	85.64	86.88	87.29
	70 %								85.42	85.42	85.42
	80 %									85.92	86.14
	90 %										86.02

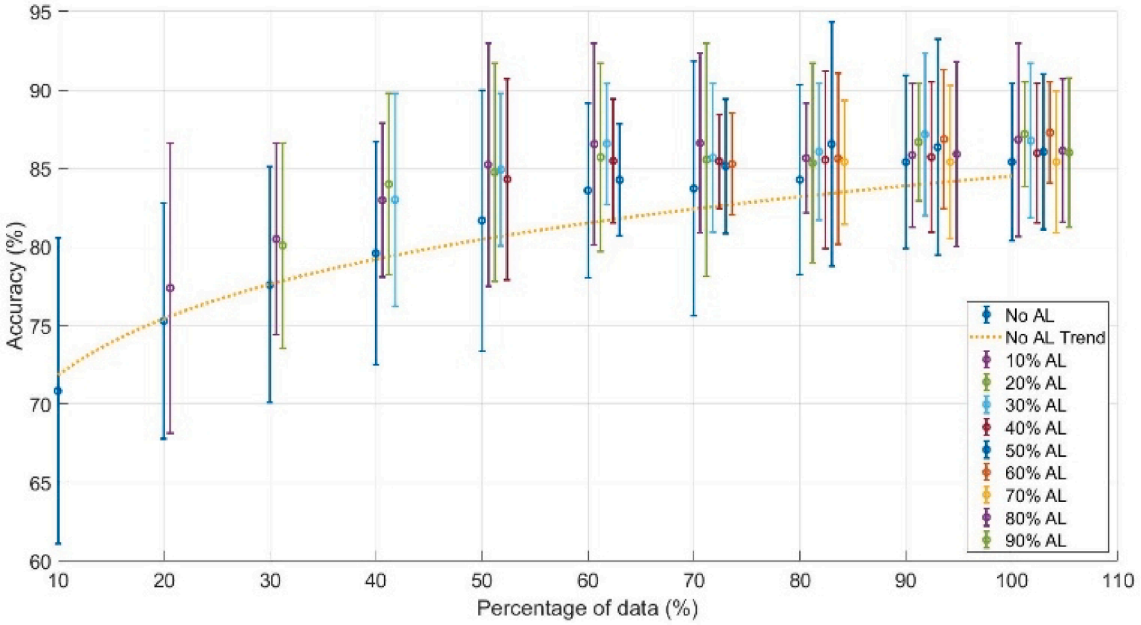

Fig. 11. Comparison of mean Active Learning (AL) to default annotation for breast dataset.

Table 4
Active Learning improvement in lung dataset based on the percentage of data used – Mean Results.

	Percentage of dataset used for training										
Percentage of training data actively learned	**-**	**10 %**	**20 %**	**30 %**	**40 %**	**50 %**	**60 %**	**70 %**	**80 %**	**90 %**	**100 %**
	0 %	**70.62**	**76.55**	**77.71**	**80.87**	**79.79**	**80.75**	**82.72**	**83.66**	**83.11**	**84.43**
	10 %		**78.41**	**78.66**	**81.09**	**80.86**	**84.33**	**84.24**	**83.37**	**84.59**	**84.59**
	20 %			**82.35**	**83.77**	**80.09**	**83.22**	**85.69**	**83.77**	**85.09**	**84.43**
	30 %				**81.46**	**81.91**	**83.51**	**82.97**	**86.65**	**85.23**	**84.81**
	40 %					**80.86**	**83.33**	**83.90**	**84.36**	**84.51**	**84.57**
	50 %						**84.33**	**84.81**	**85.23**	**83.11**	**84.59**
	60 %							**84.24**	**84.36**	**84.81**	**85.23**
	70 %								**83.50**	**85.23**	**84.43**
	80 %									**84.84**	**84.81**
	90 %										**84.59**

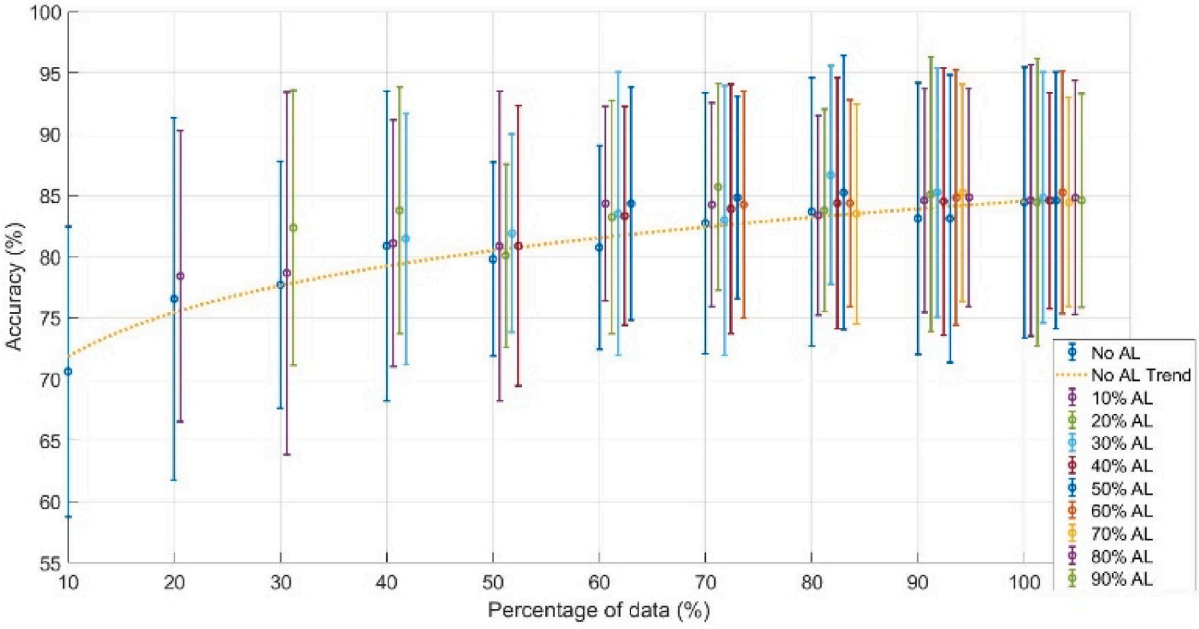

Fig. 12. Comparison of mean active learning to default annotation for lung dataset.

Table 5
Active Learning for Fetal ultrasound based on the percentage of data used – Mean Results.

	Percentage of dataset used for training										
Percentage of training data actively learned	-	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %	100 %
	0 %	87.61	90.79	91.70	92.65	93.08	93.45	93.60	93.65	93.74	94.13
	10 %		93.01	93.83	93.95	93.89	93.89	94.19	93.78	94.15	93.92
	20 %			93.85	94.10	93.92	93.78	94.02	93.97	93.98	94.09
	30 %				94.15	93.91	93.83	94.03	94.00	93.84	94.06
	40 %					94.01	93.80	93.77	94.13	94.18	94.08
	50 %						93.89	93.92	93.94	94.00	93.98
	60 %							93.31	94.40	93.68	94.34
	70 %								93.92	93.91	93.95
	80 %									93.75	94.23
	90 %										94.34

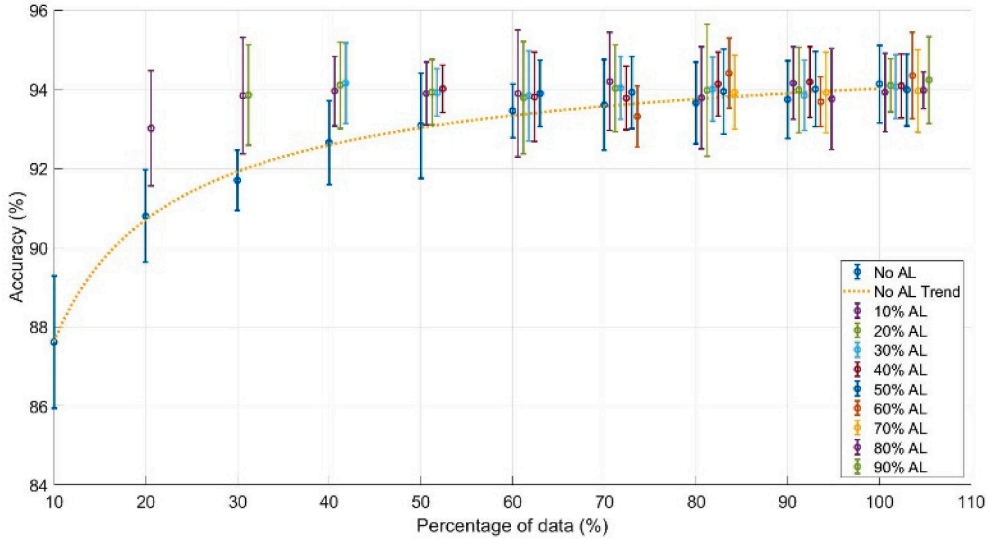

Fig. 13. Comparison of mean active learning to default annotation for lung dataset fetal dataset.

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. 14a. 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 used for validation.

Phase 2 was then applied with an initial CNN 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 CNN 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. 14b for illustrative purposes, but the process would stop once the acceptable tolerance is reached.

All networks were trained for 100 times for a cycle of 20 epoch, using

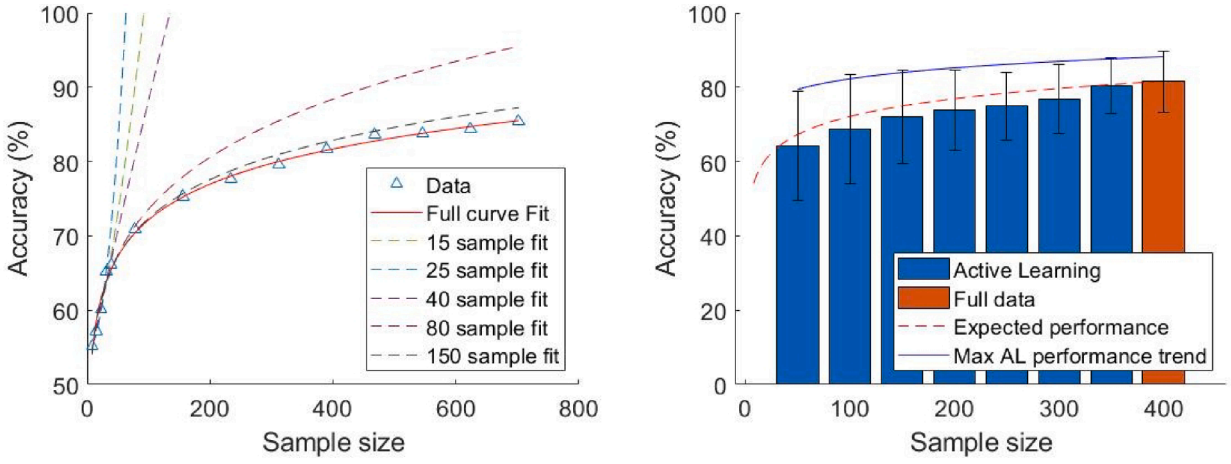

Fig. 14. A.) comparing the power curve fit for sampling steps of the breast dataset from phase1b.) Performance of active learning in comparison to full annotation of 400 samples from phase 2.

an Alexnet CNN 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.

From the combined method of phase 1 and phase 2, considering the maximum response form the CNNs, an accuracy of 85 % was achievable using only 400 samples compared with the theoretical maximum of 88 % at the full BUSI dataset size (from Fig. 8). 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 costs.

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. 15, calculated using (1), where P is the price of collection or annotation and N is the numbers predicted by phase 1 and 2 respectively.

$$Cost = (P_{Collect} \times N_{Collect}) + (P_{Annotate} \times N_{Annotate}) \quad (1)$$

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. 15. A similar analysis has been performed for differing overall acceptable tolerances from the maximum prediction from Fig. 8. This allows for further optimization of costs when accuracy can be acceptably traded. 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 CNN type used, and the costing model, but this method is always expected to return a cost reduction for minimal accuracy loss.

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. Similarly, if theoretical maximum accuracy is required, the method allows for a 40–50 % cost reduction with negligible loss in accuracy depending on the robustness criterion used; demonstrating the power of active learning in boosting accuracy at low sample numbers. Even when using just phase 2, a cost reduction of ~ 25 % is feasible for no accuracy drop using active learning to take some of the annotation burden. If the case study were a 'quick pass' feasibility study, then a massive 90 % cost saving can be made for an accuracy tradeoff of 10 %. Although cost is important, this would be most significant in terms of time as it allows proof of concepts to be demonstrated quickly and efficiently. This method is a powerful tool for planners to maximize gains and productivity under a fixed budget or time frame.

4. Conclusions

This paper demonstrated a 2-phase method that can be used to perform a cost analysis for the collection and annotation of data. Three publicly available ultrasound datasets were investigated using the prescribed method:

- Using an empirical curve-fit model of sample size determination was shown to provide an indicative method for determining cost and providing a method for scaling research studies at the cost of stability.
- Uncertainty sampling with active learning provides a cost-effective method of augmenting manual annotation by targeting samples with the lowest confidence for human annotation while those with high confidence can be annotated using active learning.

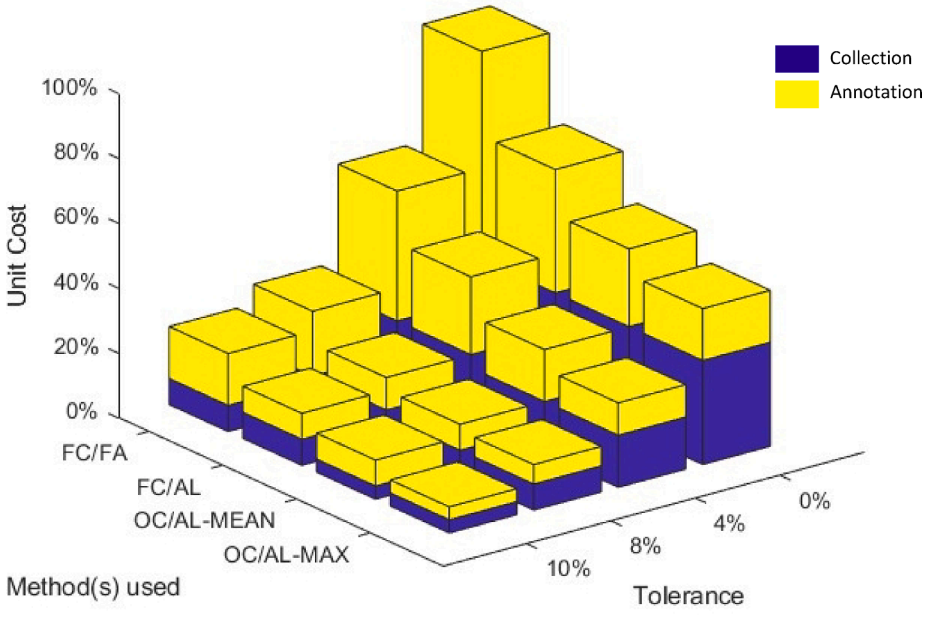

Fig. 15. Cost saving of capture and annotation for methods: Full capture/Full annotate (FC/FA), Full capture/Active learning (FC/AL), optimized capture/Active learning from mean accuracy (OC/Mean-AL), optimized capture/Active learning from max accuracy (OC/Max).

- Defining an error tolerance on required performance metrics can be used with this framework to substantially lower cost. With a 50 % cost reduction possible in the case study with an error tolerance of 4 %.
- The point of diminishing returns can be clearly defined using this method, potentially reducing over collection for that specific use case.
- Using this framework aligns machine learning research with other clinical trials that already use these and similar sampling techniques.

This framework provides ultrasound researchers with an empirical method, using power theory to identify the most effective sample size and therefore cost of future collection but and using methods such as uncertainty sampling to provide a robust, targeted method of augmenting manual labelling, allowing the clinician to focus on targets with low predictive certainty, with semi-supervised active learning labelling those with high confidence. This will allow for better scaling of studies to encourage additional ultrasound machine learning studies to be funded in future by providing a clear empiric indicator of expected performance that is easily converted to cost metrics.

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 by scaling feasibility studies and using targeted sampling methods such as the one presented in this work. Future works will explore additional cost reduction methodologies for both collection, labelling and machine learning. It will further examine small scale collection for machine learning such as the use of decentralized collection and federated learning. As well as exploring additional methods to reduce labelling cost using automation such as masked autoencoders and clustering.

The implications of lower cost studies with clear empirical indicators of results that can be expected in future studies, is that machine learning research into ultrasound will become a less risky endeavor allowing for more prospective studies to be conducted and more ultrasound data suitable for machine learning to become available to the academic community.

CRediT authorship contribution statement

Alistair Lawley: Conceptualization, Data curation, Writing – original draft, Investigation, Validation, Software. **Rory Hampson:** Writing – review & editing, Validation. **Kevin Worrall:** Writing – review & editing, Supervision, Validation. **Gordon Dobie:** Funding acquisition, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All data already publicly available

Acknowledgments

This work was supported by a UK Engineering and Physical Sciences Research Council (EPSRC) Future Ultrasonic Engineering Center for Doctoral Training (FUSE CDT) under grant EP/S023879/1 and 2296317.

References

[1] B. Luijten, et al., Adaptive ultrasound beamforming using deep learning, IEEE Trans. Med. Imaging 39 (12) (2020) 3967–3978.
[2] L. H. Lee, Y. Gao, 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] J.-Y.-L. Chan, K.T. Bea, S.M.H. Leow, S.W. Phoong, W.K. Cheng, State of the art: a review of sentiment analysis based on sequential transfer learning, Artif. Intell. Rev. 56 (1) (2023) 749–780.
[4] M.A. Morid, A. Borjali, G. Del Fiol, A scoping review of transfer learning research on medical image analysis using ImageNet, Comput. Biol. Med. 128 (2021) 104115.
[5] M. Karnes, S. Perera, S. Adhikari, A. Yilmaz, Adaptive Few-Shot Learning PoC Ultrasound COVID-19 Diagnostic System, in: 2021 IEEE Biomedical Circuits and Systems Conference (BioCAS), 2021, pp. 1-6: IEEE.
[6] A. Patra, J.A. Noble, Hierarchical class incremental learning of anatomical structures in fetal echocardiography videos, IEEE J. Biomed. Health Inform. 24 (4) (2020) 1046–1058.
[7] S. Budd, E.C. Robinson, B. Kainz, A survey on active learning and human-in-the-loop deep learning for medical image analysis, Med. Image Anal. 71 (2021) 102062.
[8] L. Schmarje, M. Santarossa, S.-M. Schröder, R. Koch, A survey on semi-, self-and unsupervised learning for image classification, IEEE Access 9 (2021) 82146–82168.
[9] J. Peng, Y. Wang, Medical image segmentation with limited supervision: a review of deep network models, IEEE Access 9 (2021) 36827–36851.
[10] R. Huang, J.A. Noble, 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.
[11] V. Rani, S.T. Nabi, M. Kumar, A. Mittal, K. Kumar, Self-supervised learning: a succinct review, Archi. Comput. Methods Eng., pp. 1-15, 2023.
[12] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009.
[13] S. Gorgin, M. Gholamrezaei, D. Javaheri, J.-A. Lee, An energy-efficient K-means clustering FPGA accelerator via most-significant digit first arithmetic, in: 2022 International Conference on Field-Programmable Technology (ICFPT), 2022, pp. 1-4: IEEE.
[14] C. Balsano, et al., Artificial intelligence and liver: opportunities and barriers, Dig. Liver Dis. (2023).
[15] D. Ben-Israel, et al., The impact of machine learning on patient care: a systematic review, Artif. Intell. Med. 103 (2020) 101785.
[16] S.M. Eldridge, et al., Defining feasibility and pilot studies in preparation for randomised controlled trials: development of a conceptual framework, PLoS One 11 (3) (2016) e0150205.
[17] J. Kendall, Designing a research project: randomised controlled trials and their principles, Emerg. Med. J. 20 (2) (2003) 164–168.
[18] K. Hemming, S. Eldridge, G. Forbes, C. Weijer, M. Taljaard, How to design efficient cluster randomised trials, bmj, vol. 358, 2017.
[19] D.J. Biau, S. Kernéis, R. Porcher, Statistics in brief: the importance of sample size in the planning and interpretation of medical research, Clin. Orthop. Relat. Res. 466 (9) (2008) 2282–2288.
[20] N.C. Thompson, K. Greenewald, K. Lee, G.F. Manso, Deep learning's diminishing returns: The cost of improvement is becoming unsustainable, IEEE Spectr. 58 (10) (2021) 50–55.
[21] P. Ren, et al., A survey of deep active learning, ACM Comput. Surveys (CSUR) 54 (9) (2021) 1–40.
[22] M.G. Arend, T. Schäfer, Statistical power in two-level models: A tutorial based on Monte Carlo simulation, Psychol. Methods 24 (1) (2019) 1.
[23] A. Reito, L. Raittio, O. Helminen, Revisiting the sample size and statistical power of randomized controlled trials in orthopaedics after 2 decades, JBJS Rev. 8 (2) (2020) e0079.
[24] J. Cohen, Statistical power analysis for the behavioral sciences, Academic press, 2013.
[25] J. Correll, C. Mellinger, G.H. McClelland, C.M. Judd, Avoid Cohen's 'small', 'medium', and 'large'for power analysis, Trends Cogn. Sci. 24 (3) (2020) 200–207.
[26] R.L. Figueroa, Q. Zeng-Treitler, S. Kandula, L.H. Ngo, Predicting sample size required for classification performance, BMC Med. Inf. Decis. Making 12 (2012) 1–10.
[27] J. Uttley, Power analysis, sample size, and assessment of statistical assumptions—Improving the evidential value of lighting research, Leukos 15 (2–3) (2019) 143–162.
[28] A. Rokem, Y. Wu, 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.
[29] J. Cho, K. Lee, E. Shin, G. Choy, S. Do, How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?, arXiv preprint arXiv:1511.06348, 2015.
[30] R. Hampson, A. Lawley, G. Dobie, Phantom study of arterial localization using tactile sensor array and a normal vs. shear pulse pressure propagation method, in: In 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2023), 2023, pp. 1–4.
[31] N. Saha, A. Swetapadma, M. Mondal, A Brief Review on Artificial Neural Network: Network Structures and Applications, in: 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS), 2023, vol. 1, pp. 1974-1979: IEEE.
[32] M.M. Bejani, M. Ghatee, A systematic review on overfitting control in shallow and deep neural networks, Artif. Intell. Rev. (2021) 1–48.

[33] E. Ragusa, E. Cambria, R. Zunino, P. Gastaldo, A survey on deep learning in image polarity detection: Balancing generalization performances and computational costs, Electronics 8 (7) (2019) 783.
[34] V.N. Vapnik, A.Y. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, in: in Measures of complexity: festschrift for alexey chervonenkis: Springer, 2015, pp. 11–30.
[35] I. Balki, et al., Sample-size determination methodologies for machine learning in medical imaging research: a systematic review, Can. Assoc. Radiol. J. 70 (4) (2019) 344–353.
[36] H.A. Glick, J.A. Doshi, S.S. Sonnad, D. Polsky, Economic evaluation in clinical trials. OUP Oxford, 2014.
[37] T. Liu, H. Yu, R.H. Blair, Stability estimation for unsupervised clustering: A review, Wiley Interdiscip. Rev. Comput. Stat. 14 (6) (2022) e1575.
[38] W. Ren, A.H. Bashkandi, J.A. Jahanshahi, A.Q.M. AlHamad, D. Javaheri, M. Mohammadi, Brain tumor diagnosis using a step-by-step methodology based on courtship learning-based water strider algorithm, Biomed. Signal Process. Control 83 (2023) 104614.
[39] Y. Yang, Z. Ma, F. Nie, X. Chang, A.G. Hauptmann, Multi-class active learning by uncertainty sampling with diversity maximization, Int. J. Comput. Vis. 113 (2) (2015) 113–127.
[40] V.-L. Nguyen, M.H. Shaker, E. Hüllermeier, How to measure uncertainty in uncertainty sampling for active learning, Mach. Learn. 111 (1) (2022) 89–122.
[41] L. Venturini, A.T. Papageorghiou, J. A. Noble, A.I. Namburete, Uncertainty estimates as data selection criteria to boost omni-supervised learning, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23, 2020, pp. 689-698: Springer.
[42] J. Yun, J. Oh, I. Yun, Gradually applying weakly supervised and active learning for mass detection in breast ultrasound images, Appl. Sci. 10 (13) (2020) 4519.
[43] G. Liu, et al., Breast ultrasound tumor detection based on active learning and deep learning, EasyChair2516-2314 (2021).
[44] 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.
[45] L. Liu, W. Lei, X. Wan, L. Liu, Y. Luo, C. Feng, 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.
[46] C. Cohen, J. Childs, S. Maranna, Behind closed doors: are sonographers coping? A literature review of sonographer burnout, Sonography 8 (1) (2021) 3–11.
[47] D. Zhang, M. Yan, H. Lin, G. Xu, H. Yan, Z. He, Evaluation of work-related musculoskeletal disorders among sonographers in general hospitals in Guangdong province, China, Int. J. Occup. Saf. Ergon. 26 (4) (2020) 802–810.
[48] J. Nightingale, M. Burton, R. Appleyard, T. Sevens, S. Campbell, Retention of radiographers: a qualitative exploration of factors influencing decisions to leave or remain within the NHS, Radiography 27 (3) (2021) 795–802.
[49] W. Al-Dhabyani, M. Gomaa, H. Khaled, A. Fahmy. Dataset of breast ultrasound images, 2020. Available: https://scholar.cu.edu.eg/?q=afahmy/pages/dataset.
[50] J. Born, et al., Accelerating detection of lung pathologies with explainable ultrasound image analysis, Appl. Sci. 11 (2) (2021) 672.
[51] J. Born et al., L2 Accelerating COVID-19 differential diagnosis with explainable ultrasound image analysis: an AI tool, ed: BMJ Publishing Group Ltd, 2021.
[52] Available: https://github.com/jannisborn/covid19_ultrasound.
[53] J.W. Tsung, D.O. Kessler, V.P. Shah, Prospective application of clinician-performed lung ultrasonography during the 2009 H1N1 influenza A pandemic: distinguishing viral from bacterial pneumonia, Crit. Ultrasound J. 4 (1) (2012) 1–10.
[54] D. Malla, V. Rathi, S. Gomber, L. Upreti, Can lung ultrasound differentiate between bacterial and viral pneumonia in children? J. Clin. Ultrasound 49 (2) (2021) 91–100.
[55] X.P. Burgos-Artizzu, et al., Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes, Sci. Rep. 10 (1) (2020) 1–12.
[56] Available: https://zenodo.org/record/3904280.
[57] A. Paszke et al., "Automatic differentiation in pytorch," 2017.
[58] O. Russakovsky, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.
[59] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: In Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[60] A. Ahmad, et al., Deep-AntiFP: Prediction of antifungal peptides using distanct multi-informative features incorporating with deep neural networks, Chemom. Intel. Lab. Syst. 208 (2021) 104214.
[61] S. Akbar, M. Hayat, M. Tahir, S. Khan, F.K. Alarfaj, cACP-DeepGram: classification of anticancer peptides via deep neural network and skip-gram-based word embedding model, Artif. Intell. Med. 131 (2022) 102349.
[62] S. Akbar, S. Khan, F. Ali, M. Hayat, M. Qasim, S. Gul, iHBP-DeepPSSM: Identifying hormone binding proteins using PsePSSM based evolutionary features and deep learning approach, Chemom. Intel. Lab. Syst. 204 (2020) 104103.
[63] S. Akbar, M. Hayat, M. Iqbal, M.A. Jan, iACP-GAEnsC: Evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space, Artif. Intell. Med. 79 (2017) 62–70.
[64] V.L. Parsons, Stratified sampling, Wiley StatsRef: Statistics Reference Online, pp. 1-11, 2014.
[65] F. Ali, S. Ahmed, Z.N.K. Swati, S. Akbar, DP-BINDER: machine learning model for prediction of DNA-binding proteins by fusing evolutionary and physicochemical information, J. Comput. Aided Mol. Des. 33 (2019) 645–658.
[66] A. Ahmad, S. Akbar, M. Tahir, M. Hayat, F. Ali, iAFPs-EnC-GA: identifying antifungal peptides using sequential and evolutionary descriptors based multi-information fusion and ensemble learning approach, Chemom. Intel. Lab. Syst. 222 (2022) 104516.
[67] S. Akbar, A. Ahmad, M. Hayat, A.U. Rehman, S. Khan, F. Ali, iAtbP-Hyb-EnC: Prediction of antitubercular peptides via heterogeneous feature representation and genetic algorithm based ensemble learning model, Comput. Biol. Med. 137 (2021) 104778.
[68] A. Lawley, R. Hampson, K. Worrall, G. Dobie, Prescriptive method for optimizing cost of data collection and annotation in machine learning of clinical ultrasound. In 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2023), 2023.