Limits on transfer learning from photographic image data to X-ray threat detection

Matthew Caldwell and Lewis D. Griffin*

Department of Computer Science, University College London, London, UK

September 27, 2019

Abstract

BACKGROUND: X-ray imaging is a crucial and ubiquitous tool for detecting threats to transport security, but interpretation of the images presents a logistical bottleneck. Recent advances in Deep Learning image classification offer hope of improving throughput through automation. However, Deep Learning methods require large quantities of labelled training data. While photographic data is cheap and plentiful, comparable training sets are seldom available for the X-ray domain.
OBJECTIVE: To determine whether and to what extent it is feasible to exploit the availability of photo data to supplement the training of X-ray threat detectors.
METHODS: A new dataset was collected, consisting of 1901 *matched pairs* of photo & X-ray images of 501 common objects. Of these, 258 pairs were of 69 objects considered *threats* in the context of aviation. This data was used to test a variety of transfer learning approaches. A simple model of threat cue availability was developed to understand the limits of this transferability.
RESULTS: Appearance features learned from photos provide a useful basis for training classifiers. Some transfer from the photo to the X-ray domain is possible as $\sim$40% of danger cues are shared between the modalities, but the effectiveness of this transfer is limited since $\sim$60% of cues are not.
CONCLUSIONS: Transfer learning is beneficial when X-ray data is very scarce—of the order of *tens* of training images in our experiments—but provides no significant benefit when hundreds or thousands of X-ray images are available.

Keywords: Automated Threat Detection, Deep Learning, Transfer Learning, Security Imaging.

*Corresponding author: Lewis Griffin, Department of Computer Science, University College London, London, WC1E 6BT, UK. Tel.: +44 20 3108 7107; E-mail: l.griffin@cs.ucl.ac.uk

1 Introduction

X-ray imaging has been an important tool for aviation security since the 1970s. The technology is well understood and widely deployed, and image acquisition is reasonably fast: many thousands of images are acquired every hour at busy airports worldwide. X-rays have the obvious benefit of being able to 'see inside' baggage, parcels and cargo without needing to open them. Imaging simultaneously at two or more X-ray energies allows spectroscopic estimation of the material properties of scanned objects, in particular the effective atomic number.

However, the interpretation of X-ray security images represents a significant bottleneck in the screening process. Scrutinising the images for prohibited or suspicious content requires skill and attention, while also being extremely repetitive and boring [3]. As in many security contexts, truly dangerous target objects are vanishingly rare, and also likely to be obfuscated. Equipment manufacturers, airport management and regulators resort to tricks such as *threat image projection* [9] to help counter operator inattention or desensitisation. This interpretation bottleneck is therefore an attractive target for augmentation or replacement by automated methods.

Recent years have seen dramatic advances in the performance of computational image analysis and recognition, driven by parallel increases in processing power, algorithmic sophistication and—crucially—data availability. The prevalence of digital photography on the world wide web and social media, combined with labelling efforts such as ImageNet [10] and more recently Google Open Images [18], provide a large body of training data and also underpin de facto benchmarks like the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), which involves identification of 1000 distinct object classes in photographic images [25]. *Deep learning* (DL) methods based on multi-layered *convolutional neural networks* (CNNs) surpassed human performance on ILSVRC in 2015 and continue to improve [17, 12, 27, 28, 6, 13]. Practical implementations of difficult computer vision problems such as facial recognition are now routinely available in mass market consumer products such as smartphones and tablets.

While there are some important differences in the image formation process of typical X-ray scanners compared to photographic cameras, the resulting data is structurally very similar. So it makes sense to apply DL approaches to the security X-ray domain. A number of groups, ourselves included, have demonstrated good results from such application [1, 14, 15, 24, 21, 5, 2].

However, methods for training and evaluating deep models are data intensive and the availability of security X-ray data is poor. In contrast to the ubiquity of portable digital cameras, X-ray equipment

is bulky and expensive. Although image data is captured in large quantities during screening at sensitive locations such as airports, that data is not collected and published, let alone annotated. There have been attempts by individual X-ray equipment manufacturers to collect datasets for internal or machine-specific purposes, and also some data collection efforts undertaken by security-related government agencies, but these tend to be of constrained scope and access to the data is often restricted. Publicly-available repositories of labelled X-ray image data for security research are rare and on a much smaller scale than those for photos [20].

This problem is not unique to X-ray security. There are many tasks for which domain-specific data is scarce compared to the abundance of natural photography [8, 19, 11, 7]. In these cases it is sometimes possible to leverage the high availability of photographic image data via *transfer learning* [29, 22, 30]. Typically this involves *representation transfer* of network weights or features learned from photographs as a starting point for learning in the target domain, though in some cases it may also be possible to employ more general photographic data directly as a supplement to the scarce target domain data.

Here we investigate the usefulness of photographic data as a transfer source for X-rays using a newly-acquired dataset consisting of matched pairs of photographic and X-ray images of the same objects in the same poses. This similarity of content should allow for maximal transferability, with limitations arising primarily from differences is the nature of danger cues across the two modalities. We develop a simple model of the overlaps between these cues in order to assess the potential utility of photographic data in training automated threat detection for X-ray images.

2 Dataset

In order to support useful comparative experiments across the X-ray and photographic modalities, we have created a new dataset, COMPASS-XP, consisting of *matched pairs of images, in which the same object in the same pose is captured as both an X-ray and a photo. The dataset includes 1901 such pairs, encompassing 501 objects from 369 object classes, with each object imaged several times in different poses.*

The dataset is intended to provide a sampling of the *variety* of the space of objects potentially present in aviation baggage. This sampling is highly *granular*—e.g. we distinguish many different types of clothing, even different types of *hats*—but *shallow*. Classes are represented by only a small number of instances—in many cases just one—which would not be sufficient to learn to identify them individually. Rather, the data is intended to discover commonalities across objects within much

higher-level semantic groupings—specifically, in this case, the grouping of dangerous vs. benign in the context of an aircraft cabin.

Danger is multivalent: it is not captured by a single trait or object class. An item such as a hatchet is dangerous, but it does not exhaust the possibilities of danger. Given many hatchet instances it would in principle be possible to learn to recognise hatchets perfectly, without learning much more about danger than was apparent from the first one or two. But the things that identify a hatchet as dangerous may also be shared by many other sharp-edged or pointed metallic objects. The same can be true for many benign objects: e.g., fabric items such as clothing, towels, bedding etc, tend to be very poorly distinguished in X-ray images, and it would be difficult to classify them precisely, but they are interchangeably unthreatening. By including instances of many different such objects, we aim to improve discoverability of their shared features. We make the explicit assumption that there is at least some generalisability among different classes. But rather than impose such threat/benign groupings *a priori*, which would prejudice the process of discovery, we allow the classifier to learn pertinent characteristics from the broad spread of the data.

The list of object classes for scanning was based on a subset of the 1000 classes of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [25] identified as plausible to be carried within aviation baggage (e.g. car and dog were excluded), supplemented with additional classes deemed relevant to aviation baggage. In particular, classes were added to increase the range of dangerous items, which are under-represented in ILSVRC. Labels for danger vs. benign were assigned manually to the classes: e.g., *hatchet* and *blowtorch* are dangerous, whereas *running shoe* and *ukulele* are not. The final tally of 369 classes is not claimed to be exhaustive or perfectly sampled. Some potentially important classes were excluded for safety reasons—for example, the set includes no true explosives, although those are certainly relevant threats to aviation. Nevertheless, these classes provide a useful basis for experimentation.

Digital photographs of each object were taken with a Sony DSC-W800 compact digital camera mounted on a custom rostrum. X-ray scans were obtained using a Gilardoni FEP ME 536 mailroom X-ray machine, distributed in the UK by Todd Research under the name TR50. This is a single-view, dual energy conveyor-belt scanner with a tunnel size of 506×360 mm. Objects were photographed and scanned in a weighted crate to maintain pose through the scanner's protective lead curtains. Additional unboxed scans were also obtained for objects that were large and heavy enough to do so.

Each X-ray image is present in five variant forms that represent the scanned data somewhat differently:

Low Raw 8-bit greyscale data from the scanner's low energy X-ray channel.

High Raw 8-bit greyscale data from the scanner's high energy X-ray channel.

Density 8-bit greyscale data representing inferred material density.

Grey RGB PNG image representing a combination of both low and high energy channels in a marginally more visually readable way.

Colour RGB PNG image with false colour palette representing inferred material properties.

The colour, grey and density images are calculated internally by the scanner using unknown proprietary algorithms. Comparison with the raw single channel data indicates that some spatial filtering is performed on the colour and grey images to improve the image sharpness and apparent resolution. All X-ray images are 580 pixels high, with variable width depending on the object size. For objects imaged in our weighted crate, the width is typically ∼1200 pixels, of which some ∼350 is typically empty belt leading in and out. Trimmed scan size is approximately 850×580 pixels. Photos were captured as RGB JPEG at 5152×3864 pixels.

Between 1–6 *instances* of each object class were scanned, and each instance was scanned in 3–8 *poses*. Examples of image types, instances and poses for the class *carving knife* are shown in Fig. 5, while Table 1 summarises the contents of the final collected dataset. All the images and corresponding metadata are released under the Creative Commons Attribution License (CC-BY) in the Zenodo data repository, and can be freely downloaded via doi:10.5281/zenodo.2654887.

3 Danger detection

We consider the problem of detecting threats as a supervised binary classification problem, i.e. where the classification boundary is learned algorithmically from a training set of images that have been explicitly labelled as either dangerous or not. We do not attempt to distinguish between different *degrees* of danger.

Unless otherwise specified, we used a *logistic regression* (LR) classifier with L^2 regularisation (see Section 3.3 for tests using different classifiers). All tests were repeated multiple times (reported as, e.g., $N = 100$) using the full dataset randomly partitioned into training and test sets, with the following constraints:

1. All poses of the same object instance were always in the same set.

2. The proportion of dangerous objects present in each set was maintained, to the extent possible given constraint 1.

Note that since the number of poses of each instance varies, the exact number of images in each set was not always identical. The dataset contains more benign than threat objects, so samples were weighted to compensate for the mismatch. In general, where several different configurations were to be directly compared (e.g., when results are presented in the same plot), the same training/test splits were used for all configurations.

Classifier performance was evaluated using the area under the receiver operating characteristic curve (AUC) [4]. This has an intuitive interpretation as the rate at which a system can correctly distinguish which is which of a danger and a benign example. AUC does not require us to decide what detection rate is needed nor what false alarm rate is acceptable for the aviation security domain. While these considerations are essential for a system nearer to deployability, for earlier research such as this it is more useful to report the overall performance without prejudging the operational demands.

All computational experiments were written in Python (https://python.org) with the SciPy stack of scientific packages (https://scipy.org). Deep learning models were implemented using Keras (https://keras.io) and TensorFlow (https://tensorflow.org), while classic machine learning methods were implemented using SciKit-Learn (https://scikit-learn.org/). Additional statistical analysis and plotting was performed using R (https://www.r-project.org).

3.1 Baseline tests

As a human baseline, we tested the performance of human observers using a *two-alternative forced choice* (2AFC) paradigm [16]. The subjects were presented with random image pairs of which exactly one was of an object from a class that had been labelled as dangerous, and were required to choose which image contained the threat. The test was undertaken by naïve participants who were not familiar with the image data ($N = 5$ for each modality, non-overlapping). They were not shown any training images, but instead were briefed on the general nature of the content and the danger criteria. Subjects achieved a median AUC of 95.9% for colour X-ray images and 97.6% for photos.

As an algorithmic baseline, we tested performance of a machine classifier trained on simple engineered features expressing basic image statistics: pixel histograms; Tukey five number summaries (minimum, first quartile, median, third quartile, maximum) for the three colour channels (abbreviated RGB5); 'busyness', assessed as the sum of squared deviations from the mean value; and the

latter supplemented by the mean itself. Results are shown in Fig. 5. The best performing feature set for colour X-rays was RGB5 (median AUC 82.9%), while for photos it was the histogram (median AUC 74.6%).

Notably, the classifiers on engineered features all performed better on X-rays than photos. This is in contrast to the results for human observers and what is found below using learned features, for which photo performance is always better. We infer that the false colour palette indicating material density provides information that helps to identify some of the 'low-hanging fruit' in the danger classification when using simple RGB summary features—in particular the presence of metal objects—but that this advantage is outweighed by the greater detail and familiarity of photographic content when assessing more difficult cases.

3.2 Deep learning from scratch

Deep learnt image classification models have a very large number of trainable parameters (e.g. roughly 18 million for DenseNet, 22 million for Inception v3), and consequently require a large quantity of data to train. Given the meagre quantities of labelled X-ray data available, training such a model from scratch almost inevitably leads to significant overfitting.

In tests using our X-ray dataset, while the CNNs we trained consistently converged to high accuracy on the training data within 30–40 epochs, their performance only weakly generalised to the test data. Over repeated trial runs of 100 training epochs, typical final median AUCs were 81–83%. This performance is notably no better than was achieved using logistic regression on the engineered RGB5 vectors with just 15 feature dimensions.

3.3 Using pre-trained appearance features

Rather than training a complex deep neural network from scratch, we can exploit knowledge already learned from a large corpus of out of domain photographic images in the form of pre-trained *appearance features*. In a deep model trained on labelled images such as those in ImageNet, the outputs of the pre-classification pooling layers are a *representation* of an image in terms of attributes useful for the learned classification task, which the model performs very well. This feature space is of significantly lower dimension (10^3–10^4) than the original image space (10^5–10^6), but the feature dimensions are more informative. We transfer this representation as a starting point for our own different classification task, in the expectation that the image properties these features distill will also prove useful in our target domain.

We tested this technique with features from five candidate pre-trained networks, details of which are given in Table 2. Summary results from a range of tests on these features are shown in Fig. 5.

We first investigated the relative performance of classifying danger vs benign in different image modalities (Fig. 5(a)). Classifier performance was higher when analysing photographic images rather than X-ray images. This is unsurprising given that the appearance features are learned from and tailored to photo classification. The five X-ray variants did not vary much in their performance, but the false colour version showed a small consistent advantage and we have focussed on that variant in most subsequent tests.

We also compared the performance of a range of different classification algorithms applied to these appearance features, finding that regularised logistic regression (LR) outperformed other approaches, including fully-connected deep networks with up to 3 hidden layers (FC0–FC3). Notably the worst performer was a nearest-neighbour (NN) approach, suggesting that dangerous images are not strongly clustered in appearance space. On the basis of these results, we focus on LR for classification in most of our tests.

Finally, we compared the performance of the different appearance models for both photos and false colour X-rays (Fig. 5(c)). Once again the performance on photos was better than for X-rays, probably reflecting the greater similarity with the image domain on which the features were originally trained. The best performing features for photos were those from the Xception network. However, these features were actually the worst for X-ray images, for which DenseNet features provided the best performance. Since targeting X-rays is our primary concern, we focus on DenseNet features in the remainder.

It is worth noting that all feature models significantly outperformed the engineered features of Fig. 5, confirming the merit of the transferred feature approach. However, they all fall short of the accuracy of human observers.

3.4 Cross-modality training between X-ray and photo

In the previous tests we made use of photo-trained features, but made no attempt to employ photos directly in model training. To see whether there is additional benefit to be gained from such training, we performed a series of tests in which the availability of training data from the target modality was artificially restricted.

For each trial, the full data set of matched pairs was divided into training and test sets, but then only a fraction of the training set was used for training in the target modality. The remainder of

the training data was optionally added as supplementary training data using the 'wrong' image type—simulating transfer of data that is from the correct application domain but a different modality. Performance of the trained classifiers was then evaluated using only the target modality images from the test set. Results are shown in Fig. 5, with the light bars showing results that include the supplementary, cross-modality training data and the dark ones showing results with only the restricted target modality training data. Note the non-uniform sampling of the horizontal axis, with more dense sampling of low target data levels, corresponding to likely scenarios of low data availability.

The leftmost light bar in each plot shows pure cross-training with no data at all from the real target modality. In this situation it would be impossible to train a classifier at all without recourse to transferred data. So, even though the performance is not especially impressive (median AUC 77.3% targeting colour, 74.3% targeting photos), the transfer is clearly beneficial. Comparing both modalities, it appears that photos represent a better training proxy for X-rays than vice versa.

Conversely, the rightmost dark bar in each plot shows the effect of training with all available target modality training data and no transfer data. This clearly provides the best performance of all the configurations: given ideal data availability there would be no point in transferring data from a different modality.

Based on the intermediate fraction results, we can make two observations.

First, even a small amount of true target data can significantly improve the training. For both modalities, training with just 4% of target data alone ($\sim$60 images) achieves better median performance than using *all* the cross-modality data ($\sim$1500 images), although the variance is higher with the small training set—with so few data points, the performance is particularly sensitive to a good or bad selection of training examples.

Second, the benefits accruing from the cross-modality data tail off rapidly as the target modality availability increases. For both modalities, the effect of including the transfer data is close to neutral once the target fraction gets to about 30%.

Together, these results suggest that while there is some applicability of photo data to the training of X-ray threat detectors, that applicability is quite limited in practice. In the next section, we attempt to understand what this implies about how cues to danger overlap between the two image spaces.

4 Estimating the transferability of learned danger cues

Consider a danger classifier operating in some space of image features, **F**. For a given body of labelled training data, some feature combinations in **F** will represent cues to danger class (either *for* or *against*), while others will be neutral. Given any test image, the evidence of the cues for and against will ultimately determine the classification. If the training data is well-sampled from the target domain (and assuming danger and not-danger are in fact discriminable in **F**), the learned cues will provide a good basis for successful classification.

When the training domain differs from the target, there will be a mismatch between the set of cues that are learned and the set that are actually informative in the target domain. In the case of our X-ray and photo domains, we will denote the true set of danger cues for photos as **P** and the true set of danger cues for X-rays as **X**. The transferability of learning from one domain to another depends on the intersection of these cue sets, $\mathbf{PX} = \mathbf{P} \cap \mathbf{X}$. The non-overlapping subsets $\mathbf{P^0} = \mathbf{P} \setminus \mathbf{X}$ and $\mathbf{X^0} = \mathbf{X} \setminus \mathbf{P}$ are neutral with respect to the other domain, and so should have no effect on the transfer (except potentially as noise sources that could reduce the efficacy of learning). The intersection of X-ray and photo cues can be considered a disjoint union $\mathbf{PX} = \mathbf{PX}^+ \cup \mathbf{PX}^-$ of features that *consistently* provide evidence for the classification in both domains ($\mathbf{PX}^+$) and features that provide *contradictory* evidence between the domains ($\mathbf{PX}^-$). This model is illustrated in Fig. 5(a).

The particular feature sets cueing danger may be difficult or impossible to identify in all but the simplest feature spaces, but we can draw conclusions about the 'magnitude' of the sets since they determine the classifier accuracy. For example, when training on photos to target X-rays, classification will be better than random to an extent determined by the informative cues $\mathbf{PX}^+$, and worse than it could be to an extent determined by the misleading cues $\mathbf{PX}^-$.

We can further probe the cue sets by manipulating the labels on the training data: with the labels inverted, the contributions of the $\mathbf{PX}^+$ and $\mathbf{PX}^-$ fractions are reversed. Cues that were contradictory become confirmatory and vice versa. Median AUC results from $N = 100$ trials using such different training configurations with both DenseNet appearance features and the engineered RGB5 features are shown in Table 3.

The raw results are difficult to interpret quantitatively because AUC values are not additive—indeed, they do not combine in any clearly defined way. To get around this, we propose a simple model for the relationship between the size of the cue sets and the observable AUC values.

Assume that each cue is independent and binary—that is, either it is present in a consistent way

in an image or set of images or it is not. The *size* of a cue set, $X = |\mathbf{X}|$, represents the *information value* it contributes to classification. Sizes are additive in the following restricted sense: for disjoint sets the sizes add and subtract; a set and its negation cancel each other; but a set added to itself remains the same size.

We assert that classification accuracy is a function f of the information value. The exact form of this relationship is unknown, but we make the following assumptions about it: that f is *non-decreasing* (i.e., better information never leads to worse accuracy) and *decelerating* (i.e., the performance boost from an additional increment of information gets smaller the larger the baseline of information to which it is added). Our metric of accuracy in this case is how much the AUC exceeds 50% (so a perfectly random classifier corresponds to zero information), though the same analysis should also be applicable to any alternative scoring method. For convenience we normalise both the information and accuracy into $[0, 1]$. The model is illustrated in Fig. 5(b).

Now, consider accuracy results from the different training configurations from the upper half of Table 3, where the target domain is X-ray. Denote the values as follows:

- A_h: measured by human testing. This does not operate in the same feature space as the other tests, so it does not encompass the same cues. However, we assume that humans can in some sense approximate the cues available to a classifier in **F**, together with some set of additional cues **HX** external to **F**, giving an overall cue pool of size $X^0 + PX^+ + PX^- + HX$
- A_x: trained directly on X-rays, accessing a cue pool of size $X^0 + PX^+ + PX^-$
- $A_{x,p}$: trained on both X-rays and photos, accessing a cue pool of size $X^0 + PX^+$
- $A_{x,\neg p}$: trained on a combination of correctly-labelled X-rays and incorrectly-labelled photos, accessing a cue pool of size $X^0 + PX^-$
- A_p: trained on photos only, accessing a cue pool of size $PX^+ - PX^-$

Given these values and the monotonicity assumptions on f, we can infer lower bounds on the relative sizes of the different cue sets:

$$\frac{PX^+}{X^0} \geq \frac{A_x - A_{x,p}}{A_{x,p} + A_{x,\neg p} - A_x} \tag{1}$$

$$\frac{PX^-}{X^0} \geq \frac{A_x - A_{x,\neg p}}{A_{x,p} + A_{x,\neg p} - A_x} \tag{2}$$

$$\frac{HX}{X} \geq \frac{A_h - A_x}{A_x} \tag{3}$$

We cannot immediately place upper bounds on PX^+ and PX^- because the two sets can *jointly* become arbitrarily large. But a larger PX^+ provides no benefit for transfer learning if offset by equal losses from PX^-. By making the additional assumption that PX^- is negligible, we can also infer an upper bound on PX^+:

$$\frac{PX^+}{X^0} \leq \frac{A_p}{A_{x,\neg p}} \tag{4}$$

Bounds can be estimated from results targeting photos in exactly the same way. (In that case we denote the overlap set **XP** for distinguishability, and the human cues as **HP**.)

Estimates for all these are shown in Table 4. The key value of interest is the upper bound on the ratio PX^+ / X^0. This ratio compares the amount of information about danger common and consistent between X-ray images and photos to the amount of information present in X-ray images but not photos. The larger the ratio, the greater the potential usefulness of photo transfer learning to the X-ray domain. Our results suggest that the commonality of danger cues between the two modalities in the DenseNet appearance space is only about 40%. This is enough that photos can indeed be helpful where training data is very scarce, but the benefit is relatively quickly subsumed by a relatively modest quantity of domain-specific data.

We also note from these results that the assumption of negligible PX^- is reasonable in the case of DenseNet features (indeed the lower bound on PX^- is marginally negative), but it is clearly not valid for RGB5. It is perhaps not surprising that there is a higher scope for contradictory cues in such a relatively crude feature space.

5 Discussion

X-ray image data is much less readily available than photographic data for training deep learning models, and it is more difficult and expensive to produce. It therefore makes sense to try to exploit the high availability of photographic data. In this study we have performed experiments using a newly acquired dataset, COMPASS-XP, consisting of matched pairs of photographic and X-ray images of the same objects in the same poses, in order to investigate the transferability of image data between the two modalities. Our specific focus is on automated threat detection in the context of aviation security.

These experiments exploited photographic data in two very distinct ways. The first, and more ob-

viously successful, was a representation transfer approach in which a pre-trained deep convolutional neural network was used for feature extraction. The transfer source in this case was ∼1.2 million ImageNet photographs; the domain-specific photographs from the matched pairs dataset were used only for performance comparison, not for learning transfer. The model's pre-trained filter parameters were used to map the images into a lower-dimensional, higher-salience appearance space within which the domain-specific X-ray threat detection learning was able to perform better than in the raw image space. As has previously been found in numerous other domains [23], this approach provided a clear benefit, in this case increasing AUC percentage results from the low 80s to the low 90s. While still some way short of practical deployability in a busy airport, this is certainly a significant improvement.

The second approach—built on top of the first—was one of functional transfer, in which we attempted to make use of up to ∼1500 domain-specific photos from the matched pairs dataset to assist the learning of danger cues applicable to X-ray images. The use of matched pairs meant that functional differences would be driven primarily by differences between the two modalities rather than by differences in subject matter or context. This transfer approach produced some benefit in cases where the supply of X-ray training data was highly constrained, but was outpaced by even a relatively small addition of real X-ray data—of the order of 50–100 images in our tests.

We interpret this as evidence that the shared pool of cues across the two modalities that are useful for a task such as classifying benign vs danger is relatively small. We developed a simple model of this cueing and estimated that the overlap is no more than about 40%. This disjunction places limits on what can usefully be learned without access to a reasonable quantity of data from the target X-ray domain. An overlap of 40% is in some ways remarkable for such different imaging modalities, and is enough to produce significantly better results than would be possible in the absence of such overlap, but the missing 60% is an insurmountable obstacle. Having X-ray data for training gives access to the full range of cues, including the overlap. **It is clear that learning all the available cues imperfectly is better than learning only 40% with higher precision.**

In an ideal world, the scarcity of domain-specific training data would be addressed and there would be no need for any form of transfer learning. This is unlikely to be the case any time soon, so strategies for dealing with this problem can be expected to remain relevant for some time. On the basis of our results, we make the following recommendations:

- Unless training data in the target modality is very abundant (e.g. >10,000 images), it is better to use a pre-trained photographic CNN for image features. The risk of lost sensitivity to modality-

specific features is almost certainly outweighed by the benefit of access to rich generic ones. We found DenseNet-201 to be the most performant model for our X-ray images, but CNN differences were marginal compared to not using one at all.

- If at all possible, acquire domain-specific training data in the target modality. The more the better, but even 10 or 20 target domain X-rays can substantially improve performance on a binary classification task.

- Domain-specific photographic images can usefully supplement very scarce X-ray images (e.g. $<$100), but the benefits are marginal otherwise.

Actual performance gains will vary significantly with the problem domain, so it is difficult to draw more general conclusions. However, it seems likely that the majority of the benefit to be had from transfer from photographic to other imaging modalities will already have been captured in the feature extraction of a pre-trained CNN, and the available improvements beyond that may be quite limited.

It is worth remarking that human performance on our X-ray data is better than that of our trained models. At some level, this performance must be making use of cues present in the X-rays, and the interpretation of those cues must in some sense be transferred from outside the target domain—the subjects were not trained on domain-specific data and were not experienced X-ray readers. But this lack of domain training was also true of subjects looking for threats in our *photographic* data. Clearly, these viewers were bringing much larger experiential and cognitive frameworks to bear in order to make contextual judgements regarding danger. Such frameworks are far beyond the scope of current DL models.

Acknowledgments

This work was funded by the UK Government Defence and Security Accelerator as part of the Future Aviation Security Solutions programme (https://www.gov.uk/government/groups/future-aviation-security-solutions-programme).

References

[1] S. Akçay, M. E. Kundegorski, M. Devereux, and T. P. Breckon, Transfer learning using convolutional neural networks for object classification within x-ray baggage security imagery, *IEEE*

International Conference on Image Processing (2016), 1057–1061.

[2] S. Akcay, M. E. Kundegorski, C. G. Willcocks, and T. P. Breckon, Using Deep Convolutional Neural Network Architectures for Object Classification and Detection Within X-Ray Baggage Security Imagery, *IEEE Transactions on Information Forensics and Security* **13** (2018), 2203–2215.

[3] A. T. Biggs and S. R. Mitroff, Improving the Efficacy of Security Screening Tasks: A Review of Visual Search Challenges and Ways to Mitigate Their Adverse Effects, *Applied Cognitive Psychology* **29** (2014), 142–148.

[4] A. P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern Recognition* **30** (1997), 1145–1159.

[5] M. Caldwell, M. Ransley, T. W. Rogers, and L. D. Griffin, Transferring X-ray based automated threat detection between scanners with different energies and resolution, in: *Counterterrorism, Crime Fighting, Forensics, and Surveillance*, H. Bouma, F. Carlysle-Davies, R. J. Stokes, and Y. Yitzhaky, eds., SPIE, 2017, pp. 104410F-1–10.

[6] F. Chollet, Xception: Deep Learning with Depthwise Separable Convolutions, *arXiv.org* (2016), 1–8.

[7] M. Christopher, A. Belghith, C. Bowd, J. A. Proudfoot, M. H. Goldbaum, R. N. Weinreb, C. A. Girkin, J. M. Liebmann, and L. M. Zangwill, Performance of Deep Learning Architectures and Transfer Learning for Detecting Glaucomatous Optic Neuropathy in Fundus Photographs., *Scientific Reports* **8** (2018), 16685-1–13.

[8] G.-B. Cui, D. Zhao, and W. Wang, Glioma Grading on Conventional MR Images: A Deep Learning Study With Transfer Learning, *Frontiers in Neuroscience* **12** (2018), 1–10.

[9] V. Cutler and S. Paddock, Use of threat image projection (TIP) to enhance security performance, *43rd Annual 2009 International Carnahan Conference on Security Technology* (2009), 46–51.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2009, pp. 248–255.

[11] C. Galea and R. A. Farrugia, Matching Software-Generated Sketches to Face Photographs With a Very Deep CNN, Morphed Faces, and Transfer Learning, *IEEE Transactions on Information Forensics and Security* **13** (2018), 1421–1431.

[12] K. He, X. Zhang, S. Ren, and J. Sun, Deep Residual Learning for Image Recognition, *arXiv.org* (2015), 1–12.

[13] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger, Densely Connected Convolutional Networks, *arXiv.org* (2018), 1–9.

[14] N. Jaccard, T. W. Rogers, E. J. Morton, and L. D. Griffin, Tackling the x-ray cargo inspection challenge using machine learning, in: *SPIE Defense + Security*, A. Ashok, M. A. Neifeld, and M. E. Gehm, eds., SPIE, 2016, pp. 98470N-1–13.

[15] N. Jaccard, T. W. Rogers, E. J. Morton, and L. D. Griffin, Detection of concealed cars in complex cargo X-ray imagery using deep learning, *Journal of X-ray Science and Technology* **25** (2017), 323–339.

[16] F. A. A. Kingdom and N. Prins, *Psychophysics: A Practical Introduction*, Academic Press, San Diego, second edition, 2016.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, Imagenet classification with deep convolutional neural networks, *Advances in Neural Information Processing Systems* **25** (2012), 1097–1105.

[18] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, T. Duerig, and V. Ferrari, The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale, *arXiv.org* (2018), 1–20.

[19] M. Lagunas and E. Garces, Transfer Learning for Illustration Classification, *arXiv.org* (2018), 1–9.

[20] D. Mery, V. Riffo, U. Zscherpel, G. Mondragón, I. Lillo, I. Zuccar, H. Lobel, and M. Carrasco, GDXray: The database of X-ray images for nondestructive testing, *Journal of Nondestructive Evaluation* **34** (2015), 1–12.

[21] D. Mery, E. Svec, M. Arias, V. Riffo, J. M. Saavedra, and S. Banerjee, Modern Computer Vision Techniques for X-Ray Testing in Baggage Inspection, *IEEE Transactions on Systems, Man, and Cybernetics: Systems* **47** (2017), 682–692.

[22] R. Raina, A. Y. Ng, and D. Koller, Constructing informative priors using transfer learning, in: *ICML '06: Proceedings of the 23rd International Conference on Machine Learning*, ACM Press, New York, USA, 2006, pp. 713–720.

[23] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, CNN Features off-the-shelf: an Astounding Baseline for Recognition, in: *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, 2016, pp. 512–519.

[24] T. W. Rogers, N. Jaccard, and L. D. Griffin, A deep learning framework for the automated inspection of complex dual-energy x-ray cargo imagery, *SPIE Defense + Commercial Sensing* (2017), 101870L-1–12.

[25] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, *International Journal of Computer Vision* **115** (2015), 211–252.

[26] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, MobileNetV2: Inverted Residuals and Linear Bottlenecks, *arXiv.org* (2018), 1–14.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, Going Deeper with Convolutions, *arXiv.org* (2014), 1–9.

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, Rethinking the Inception Architecture for Computer Vision, *arXiv.org* (2015), 1–10.

[29] S. Thrun, Is Learning The n-th Thing Any Easier Than Learning The First?, *Advances in Neural Information Processing Systems* (1996), 640–646.

[30] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, How transferable are features in deep neural networks?, *arXiv.org* (2014), 1–9.

[31] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, Learning Transferable Architectures for Scalable Image Recognition, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, 2018, pp. 8697–8710.

	Classes			Instances			Matched Pairs		
	Threat	Benign	Total	Threat	Benign	Total	Threat	Benign	Total
ImageNet	11	176	187	26	224	250	93	845	938
Custom	24	158	182	43	208	251	165	798	963
Total	35	334	369	69	432	501	258	1643	1901

Table 1: Summary of the COMPASS-XP dataset.

Model	Reference	Input Size	Parameters	Features
DenseNet-201	[13]	224×224	18.3m	1920
Inception V3	[27, 28]	299×299	21.8m	2048
MobileNet V2 (alpha=1.4)	[26]	224×224	4.4m	1792
NASNet (Large)	[31]	331×331	84.9m	4032
Xception	[6]	299×299	20.9m	2048

Table 2: Summary of the deep networks for which we have tested pre-trained appearance features.

Target	Training	Median AUC (%)			Cue Contributions
		Human	DenseNet	RGB5	
Colour	—	95.9			$X^0 + PX^+ + PX^- + HX$
	Colour		90.6	82.0	$X^0 + PX^+ + PX^-$
	Colour + Photo		90.9	78.8	$X^0 + PX^+$
	Colour + Inv Photo		87.5	76.7	$X^0 + PX^-$
	Photo		76.3	70.5	$PX^+ - PX^-$
Photo	—	97.6			$P^0 + XP^+ + XP^- + HP$
	Photo		92.0	70.5	$P^0 + XP^+ + XP^-$
	Photo + Colour		91.2	69.1	$P^0 + XP^+$
	Photo + Inv Colour		88.5	70.0	$P^0 + XP^-$
	Colour		75.9	54.4	$XP^+ - XP^-$

Table 3: Summary of feature space test results.

	Targeting X-rays				Targeting Photos			
	$\leq \frac{PX^+}{X^0} \leq$		$\frac{PX^-}{X^0} \geq$	$\frac{HX}{X} \geq$	$\leq \frac{XP^+}{P^0} \leq$		$\frac{XP^-}{P^0} \geq$	$\frac{HP}{P} \geq$
DenseNet	0.08	0.70	-0.01	0.13	0.09	0.67	0.02	0.13
RGB5	0.23	0.78	0.14	0.43	0.03	0.22	0.07	1.32

Table 4: Estimated bounds on relative sizes of danger cue sets.

FIGURE CAPTIONS

Figure 1: Elements of the COMPASS-XP dataset. Each X-ray/photo matched pair is present in six different *image types*, of which five are X-ray variants. Each object class (in this case *carving knife*), may be represented by multiple *instances*, and each instance occurs in multiple *poses*.

Figure 2: Classification performance of human observers compared with logistic regression using engineered features. (75% training split, N=100 for engineered features; N=5 human subjects for each modality, distribution estimated by bootstrap resampling to N=100.)

Figure 3: Classification performance using pre-trained appearance features. Broken down by (a) image type; (b) classification algorithm; (c) appearance feature model. (75% training split, N=100.)

Figure 4: Effect of supplementing training data with domain-specific images from a different modality. Results in the left plot are for classifiers targeting colour X-ray data, those on the right are for classifiers targeting photos. Classifiers are trained using a *fraction* of the available training data in their correct target modality: a fraction of 0% means no training is performed with the correct modality, while a fraction of 100% means the whole training set is used. The training is optionally supplemented with the unused portion of the training data, but using the 'wrong' imaging modality (light bars). AUC performance in both cases is calculated for the target modality only, and the test sets are disjoint from training data in both modalities. Note the non-uniform sampling on the horizontal axis: more tests were performed at low levels of target data, where the effect of the transfer data is most pronounced. Results are aggregated over feature models. (LR, 80% overall training split, N=100.)

Figure 5: (a) Conceptual representation of the different sets of danger cues learnable in some feature space $\mathbf{F}$. $\mathbf{P^0}$ are photo cues that are neutral with respect to X-ray images, and $\mathbf{X^0}$ are likewise X-ray cues that are neutral for photos. $\mathbf{PX}^+$ are cues present in both image modalities that are *consistent* (i.e., indicate the same danger class in both modalities), while $\mathbf{PX}^-$ are present in both but *contradictory*. (b) Schematic of the relationship between measurable performance (AUC, normalised such that $[0.5, 1] \mapsto [0, 1]$) and the underlying information available from feature cues. The true relationship is

unknown, but we can derive some bounds on the relative contribution sizes given the constraint that the relationship is monotonically increasing and decelerating (as shown).

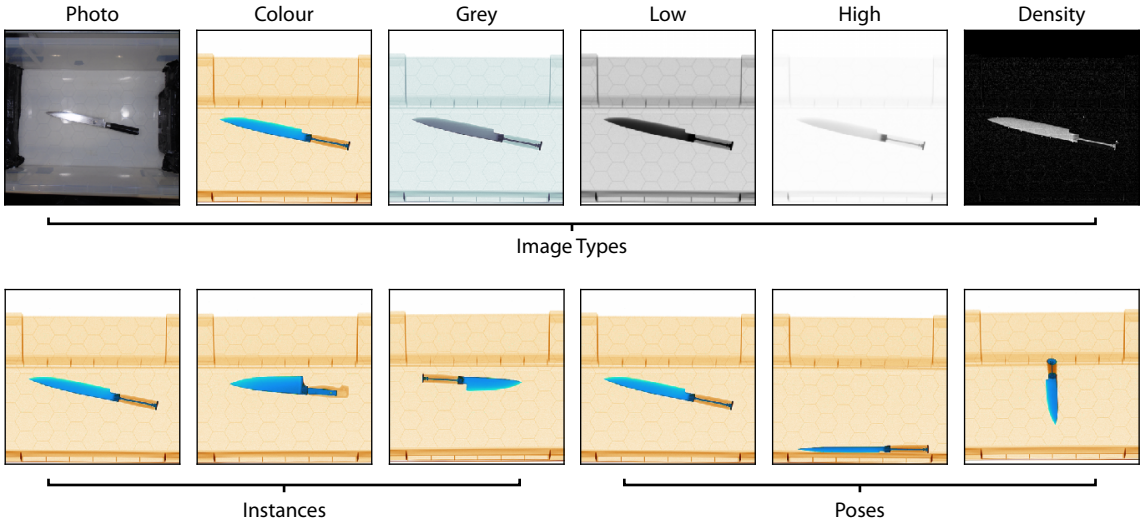
Photo
Colour
Grey
Low
High
Density
Image Types
Instances
Poses

Figure 1

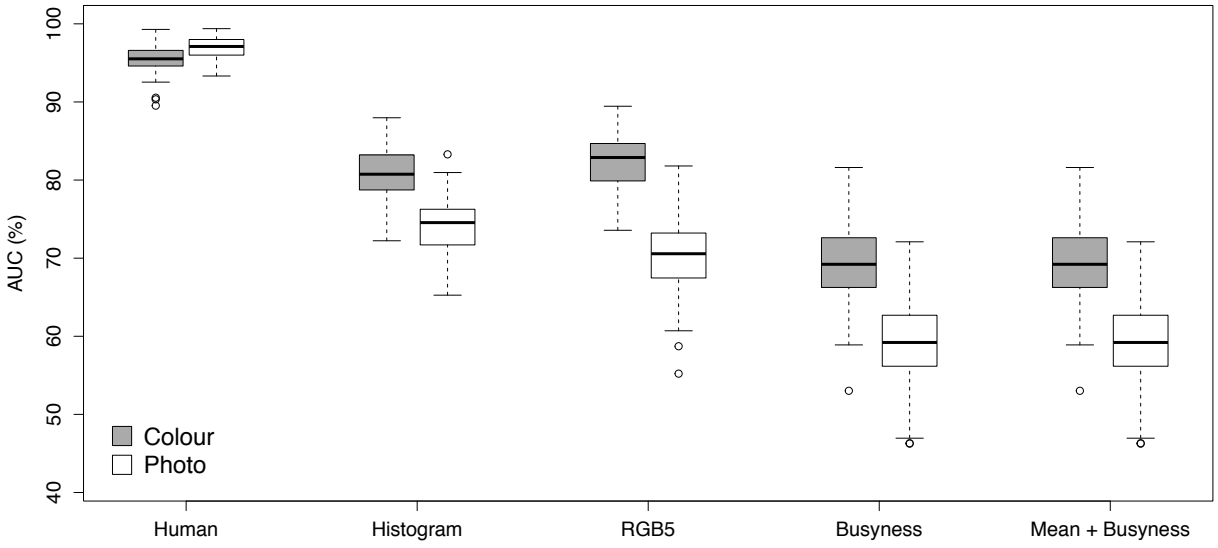
AUC (%)
40
50
60
70
80
90
100
Colour
Photo
Human
Histogram
RGB5
Busyness
Mean + Busyness

Figure 2

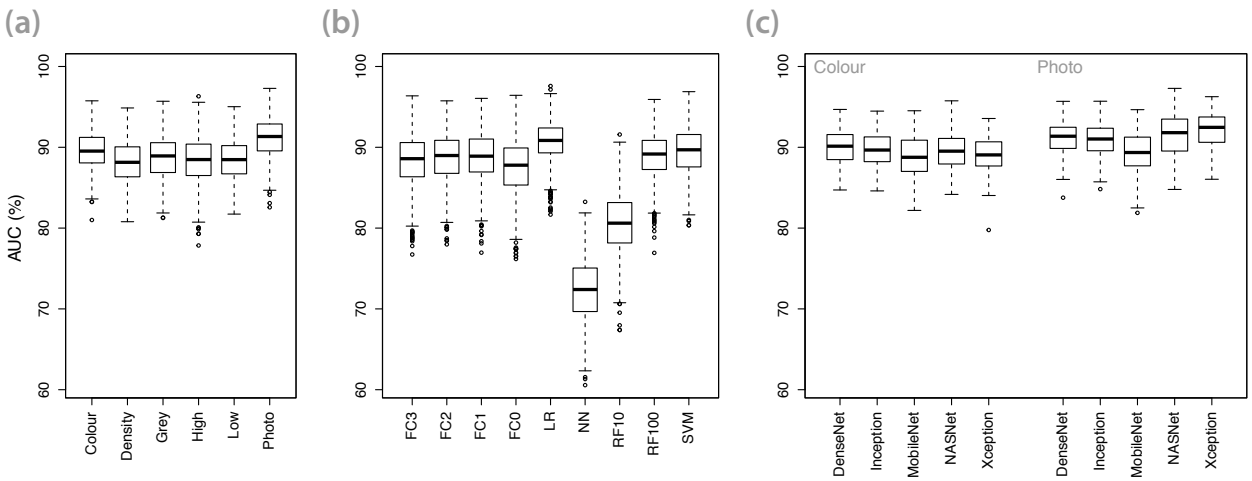
(a)
(b)
(c)
AUC (%)
100
90
80
70
60
Colour
Density
Grey
High
Low
Photo
FC3
FC2
FC1
FC0
LR
NN
RF10
RF100
SVM
Colour
Photo
DenseNet
Inception
MobileNet
NASNet
Xception

Figure 3

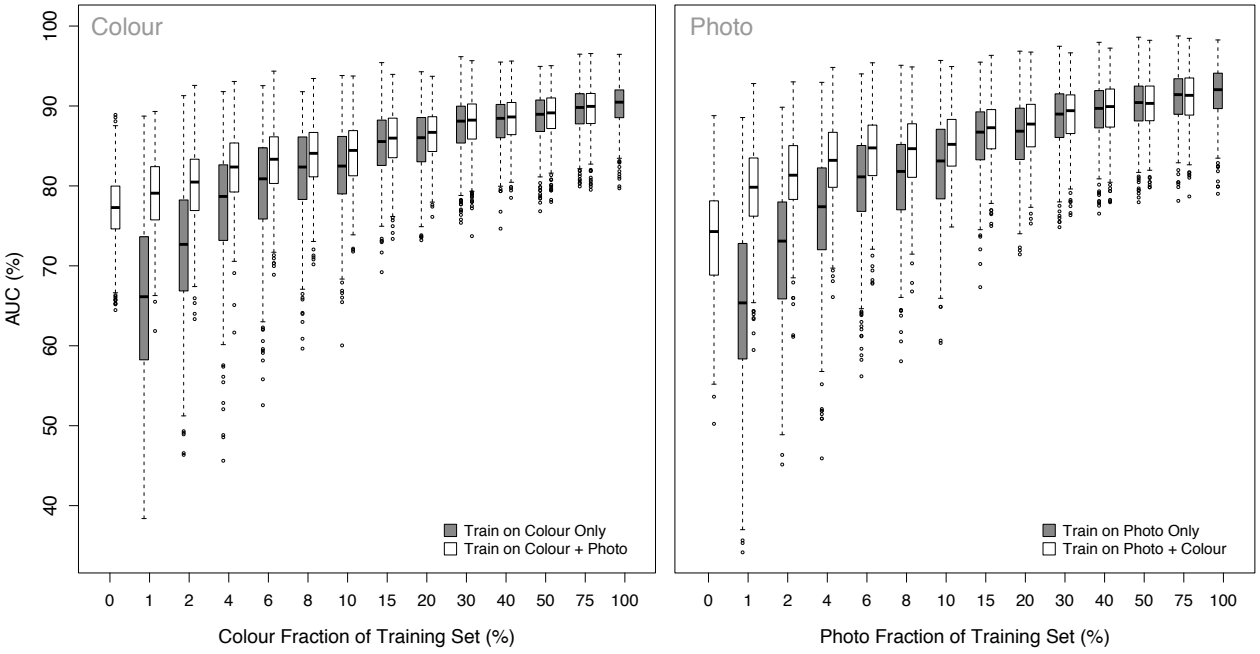
Colour
Photo
AUC (%)
40
50
60
70
80
90
100
Train on Colour Only
Train on Colour + Photo
Train on Photo Only
Train on Photo + Colour
0
1
2
4
6
8
10
15
20
30
40
50
75
100
Colour Fraction of Training Set (%)
Photo Fraction of Training Set (%)

Figure 4

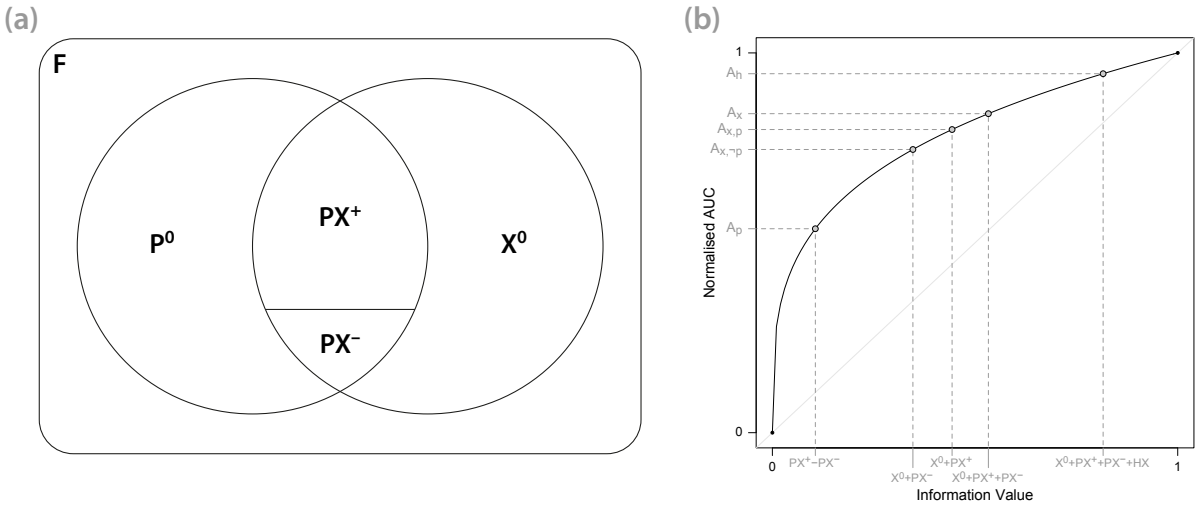
(a)
F
P0
PX+
PX−
X0
(b)
Normalised AUC
Information Value
1
0
Ah
Ax
Ax,p
Ax,¬p
Ap
0
PX+−PX−
X0+PX−
X0+PX+
X0+PX+PX−
X0+PX+PX−+HX
1

Figure 5