Learning How to MIMIC: Using Model Explanations to Guide Deep Learning Training

Matthew Watson, Bashar Awwad Shiekh Hasan, Noura Al Moubayed
Durham University
Durham, UK
{matthew.s.watson,bashar.awwad-shiekh-hasan,noura.al-moubayed}@durham.ac.uk

Abstract

Healthcare is seen as one of the most influential applications of Deep Learning (DL). Increasingly, DL models have been shown to achieve high-levels of performance on medical diagnosis tasks, in some cases achieving levels of performance on-par with medical experts. Yet, very few are deployed into real-life scenarios. One of the main reasons for this is the lack of trust in those models by medical professionals driven by the black-box nature of the deployed models. Numerous explainability techniques have been developed to alleviate this issue by providing a view on how the model reached a given decision. Recent studies have shown that those explanations can expose the models' reliance on areas of the feature space that has no justifiable medical interpretation, widening the gap with the medical experts. In this paper we evaluate the deviation of saliency maps produced by DL classification models from radiologist's eye-gaze while they study the MIMIC-CXR-EGD images, and we propose a novel model architecture that utilises model explanations during training only (i.e. not during inference) to improve the overall plausibility of the model explanations. We substantially improve the similarity between the model's explanations and radiologists' eye-gaze data, reducing Kullback-Leibler Divergence by 90% and increasing Normalised Scanpath Saliency by 216%. We argue that this significant improvement is an important step towards building more robust and interpretable DL solutions in healthcare.

1. Introduction

Applications of Deep Learning (DL) to healthcare have been growing rapidly in a wide range of medical scenarios; ranging from critical care [24] and diabetes risk prediction [1] to the diagnosis of chest x-rays (CXRs) [28]. This is partly driven by the rising accuracy of such models, with some beginning to achieve performance on-par with (or even exceeding) that of medical professionals [22]. However, despite these developments we are yet to see a similar growth in the number of DL models being deployed into real-world medical scenarios [2]. This is down to numerous limiting factors; most notably, before such techniques can become established in the medical field, they must be ethical in their decision-making, trustworthy, transparent and explainable [5, 12].

It is in these areas that many DL models can perform poorly. In particular, many models fail to accurately capture the causal relationships between input features and the output classification and rely instead on task irrelevant features. For example, a wide-ranging study on the use of Machine Learning (ML) and DL techniques for COVID-19 prediction from chest x-rays (CXRs) [17] has shown that many models are making spurious correlations, leading to the models being unable to accurately generalise. Furthermore, recent studies on the robustness of DL models have shown that changes to training hyperparameters can greatly affect the learned features [26] - this damages the trust between clinicians and DL techniques as it highlights just how sensitive to small changes the models are, even when those changes are independent of the medical questions the model is trying to answer.

Thus, the gold-standard for any ML model is to be able to achieve high-levels of performance whilst learning the concrete causal relationships present in the data. Unfortunately, the presence of learned causal features is extremely difficult to verify due to a lack of useful data supporting the task. Following practices in pedagogy, expert's Eye Gaze Data (EGD) can be used as a proxy for causal relationships [23, 19]. The release and initial analysis of the MIMIC-CXR-EGD dataset [15] showed that even current state-of-the-art CXR classification models fail to learn the same set of features as used by radiologists in their diagnoses.

In this paper, we present a novel deep learning architecture that learns a more consistent feature set than previous techniques. Using the MIMIC-CXR-EGD dataset, which to the best of our knowledge is the only large-scale image

dataset with accompanying expert eye-gaze data, we compare the similarity between explanations computed from DL models and the EGD from radiologists. We report that there is a significant increase in overlap (increasing from -0.4634 to 0.5410 when measured by Normalised Scanpath Saliency and improving from 9.1233 to 0.8398 when measured by Kullback-Leibler Divergence) between explanations from our proposed technique and the EGD than there is from any other model architecture tested; including current state-of-the-art methods specifically designed to combat this issue. We also show that our proposed architecture produces more consistent explanations than previous models, increasing explanation consistency [26] from 0.1785 to 0.5333 with no cost to model performance nor the need for specialist's EGD at inference time.

2. Related Work

In order to explain the decisions made by DL models, numerous explainability techniques have been developed with the aim of "opening up" the black-box architectures. In this paper we focus on two post-hoc techniques [13] that are designed to explain deep learning models; our aim is to compare the explanations from a variety of established architectures (as well as our novel models) and so the techniques used must be model-agnostic and easy to apply. SHAP [16] is a permutation-based approach which has theoretical groundings in game theory. Grad-CAM [18] is a gradient-based approach which uses the gradient of any target concept flowing into the final convolutional layer of a network to produce a saliency map. We focus on these two techniques in this paper as not only are they the current de-facto standards, but they can also both be applied to a wide-range of model architectures allowing for the easy comparison of explanations from varying model types.

Previous work has used these explainability techniques to investigate the robustness and adaptability of DL models [26, 8], finding that even small changes to the training procedure can result in significant changes to the learned features. These results, coupled with many network's susceptibility to issues such as adversarial attacks [10] and shortcut learning [9], suggest that many modern DL architectures are not necessarily learning causal relationships in the data to achieve high performance and might be relying on spurious correlations. It can be extremely difficult to verify that the learned features are indeed causal - there are only a limited number of mostly toy datasets that include descriptions of their causal relationships [3].

In the absence of such data, recent work has used EGD of experts making decisions on a visual task as a proxy for concrete causal relationships [15]. Such data can be used to determine whether models are learning features that domain experts would use in their assessment of the data - this use case has groundings from real-world applications, with similar techniques being used pedagogically in fields such as radiology [25]. The MIMIC-CXR-EGD dataset [15] is a subset of MIMIC-CXR [14], containing 1,083 CXR images from three classes (Pneumonia, Congestive Heart Failure and Normal). Accompanying the images are aligned EGD from a trained radiologist. Both raw eye gaze information and calculated fixation points are available for this EGD - we refer readers interested in the EGD collection process to [15]. Alongside the release of the dataset the authors also show that explanations from traditional classification models do not significantly overlap with the radiologist's EGD. They propose a multi-task UNet model which uses EGD at train-time to learn to both classify the CXR image and reproduce the ground-truth EGD in order to improve the similarity between model explanations and EGD. However, the results are not very convincing and the study lacked a verifiable method of comparing their model explanations and the EGD. Additionally, this technique requires the use of expert EGD during training which is costly and difficult to collect, especially in the medical domain. We compare our method against both the baseline models and the improved UNet architecture using static EGD heatmaps proposed in [15] resulting in significantly higher degree of similarity between model explanations and EGD across all tested metrics.

3. Method

Our proposed architecture consists of an ensemble architecture M made up of S sub-models (of any architecture) and a discriminator, D. We begin by describing the architecture of our model, and then detail its training procedure.

We define an explanation ensemble model as $M : \mathcal{X} \rightarrow \mathcal{Y}$, where $\mathcal{X}$ is the set of inputs, and $\mathcal{Y}$ the outputs. M consists of S sub-models $m_0, ..., m_S$, where $S \in \mathbb{N}$, each of which has the same architecture suited to the task. In our proposed network, each m_i is trained with a different hyperparameter setup - i.e. with different random seeds, or training data order. Architecture hyperparameters, such as layer size and learning rate, are kept constant. The final output of the explanation ensemble is the average output of all sub-models:

$$M(x) = \frac{\sum_{i \in [0,S]} m_i(x)}{S} \tag{1}$$

The network also adds a discriminator, $D : \mathcal{E} \rightarrow \mathcal{S}$, where $\mathcal{E}$ is the set of model explanations (calculated via any feature importance attribution method) and $\mathcal{S} = [0, S]$. We denote the explanations of sub-model m_i on the inputs x as $E_i(x)$. The discriminator is trained on the explanations produced by each of the S sub-models, with the aim of learning to identify which of the sub-models a given explanation originated from. As the task of the discriminator has been shown to be easily learned [26] the architecture of D should

be chosen carefully, ensuring it is not too complex that M is drastically overfitting.

The S sub-models and discriminator D are all trained together, optimising the loss function in Equation 2, where CELoss$(\cdot,\cdot)$ is cross-entropy loss and $\beta \in [0, \infty)$ is a hyperparameter weighting D's contribution during a training epoch. The subtraction of the discriminator loss in Eq. 2 ensures that the sub-model m_i "fools" the discriminator by learning to produce explanations that are similar to that of the other sub-models in the ensemble.

$$\text{loss} = \sum_i \texttt{CELoss}(m_i(x), y) - \beta \cdot \texttt{CELoss}(D(E_i(x)), i) \tag{2}$$

Every α epochs (where α is another tunable hyperparameter), the discriminator D is updated with respect to the loss function CELoss$(D(E_i(x)), i)$, without back-propagating through the sub-models, allowing D to learn how to effectively classify the explanations. This only needs to be done every α epochs due to the ease of the task [26]. This equates to the S sub-models and D being updated in a two-player minimax game - the goal of D is to learn to separate the sub-models' explanations, whereas the sub-models are aiming to fool the discriminator, all whilst also optimising $m_0, ..., m_i$ on the downstream task. The result is a set of S sub-models that produce similar explanations. The assumption here is that this learnt explanation is closer to representing the causal relationships and less reliant on the spurious correlations.

Training of this model can be unstable - this is a direct consequence of the discriminator and ensemble sub-models having opposing goals. For example, if each sub-model gives each feature of the input equal weight then the loss of the discriminator will be maximised, reducing Eq. 2. However, this would also result in the sub-model predicting the same class for every input. Training stability is linked to a "good" choice of α. This can be optimised like any hyperparameter (e.g. through a grid-search or random search), although we have empirically found through experimentation that $\alpha = 2$ provides stable training.

To summarise, the intuition behind our architecture is to train a discriminator D which encourages each of the S sub-models in an ensemble to learn a similar set of features. As each of the sub-models is trained with a different hyperparameter setup, they will each learn a slightly different set of features. As training progresses, D will learn to use the noisy features of each sub-model to (correctly) classify which sub-model explanations originate from - and in turn, the sub-models will learn to use different features for its classification, in order to fool D. The final result is an ensemble model that has learned to "ignore" a wide range of spurious features, with each of the sub-models only using features which all m_i agree are important. As multiple models must agree that any given feature is important for it to be used, it is more likely that these are causally related with the target, and thus is more likely to be included in an expert's eye-gaze data.

4. Experimental Setup

All experiments[1] are carried out on the MIMIC-CXR-EGD dataset [15]. The models are trained on the same 3-label classification task: given a CXR image, predict its diagnosis (Pneumonia, Congestive Heart Failure or Normal). We train three architectures to compare our explanation ensemble to: **1) baseline:** a standard UNet architecture trained with a learning rate (LR) of 0.003 with Adam optimiser, batch size 32, and pre-trained EfficientNet-b0 [21] as the encoder and bottleneck layers; **2) improved UNet:** the modified UNet architecture [15] using static heatmaps during training to both classify and reproduce the EGD given a CXR using identical hyperparameters; and **3) standard ensemble:** an ensemble architecture consisting of 10 UNet architectures identical to **2)**, trained with LR=0.003 using the Adam optimiser and batch size 4 [15]. A reduced batch was used due to memory constraints. Each experiment allows us to compare our results against a different standard of model: **1)** is a standard classification model and used as a baseline, **2)** is the SOTA for similarity between model explanations and EGD, and **3)** confirms that our results are not just a result of utilising an ensemble architecture (and instead are inherent to our proposed architecture and training procedure). UNet was chosen to allow for direct comparison with the current state of the art model on the MIMIC-CXR-EGD dataset in [15]. We also experimented with Vision Transformers [7], however due to the small size of MIMIC-CXR-EGD they are unable to gain levels of performance matching those of our baseline and so we do not include their results in this paper. Across all experiments the same 80/20 train-test split is used for the MIMIC-CXR-EGD dataset.

We train our proposed explanation ensembles using standard UNet with a classification head as our sub-models. Batch sizes of 4 and a learning rate of 0.00001 using the Adam optimiser are used. We use a CNN for our discriminator, with two convolution layers. Max pooling (with kernel size and stride of 2) and ReLU activations are used after each convolution layer. We set $\beta = 0.2$ to ensure the two parts of the main loss function are of the same order of magnitude. We use 10 sub-models per Explanation Ensemble (see the Supplementary Material for results on different numbers of sub-models). We report the accuracy (across all three labels) for all models as a performance metric.

In order to allow for direct comparison with [15], we compute the explanations for all models using Grad-CAM

[1]Code to reproduce these experiments can be found at: https://github.com/mattswatson/learning-to-mimic

[18] on the final convolution layer. We sampled examples from the test set for inspection. We compare the similarity of these explanations to EGD heatmaps generated from the eye-gaze fixations, which gives us scalar values of importance for each pixel based on the radiologist's eye gaze [15]. To measure similarity to the EGD heatmaps we follow standard practice of comparing saliency maps [4]; we report both the Kullback–Leibler Divergence (KLD) as a distribution-based metric, and the Normalised Saliency Scanpath (NSS) as a location-based metric. KLD is an information-theoretic measure of the difference between one probability distribution and another; importantly, note that it is a *divergence* metric, meaning smaller values indicate better similarity. NSS is designed to be used to compare saliency maps with a ground-truth, and is the normalised saliency at fixed locations. We note that metrics such as Intersection over Union (IoU) are not suited to comparing EGD and saliency heatmaps [4] as one must consider how much importance is placed on each pixel (by both the model and the expert), rather than treating explanations/EGD as binary heatmaps.

It is known that NSS is sensitive to false positives, however that is desirable here - we hypothesise that the (non-explanation ensemble) models are learning many noisy features which are not necessarily causally linked to output - we want to penalise the models if this is indeed the case. Negative NSS values highlight negative correlation, with chance at 0 and positive values indicating positive correlation.

Explanation consistency [26] measures the change in model explanations under different hyper parameter settings perpendicular to the task. Higher consistency is linked to explanations more robust to spurious correlations [26]. We would expect our explanation ensemble model to achieve higher explanation consistency than other models tested. For each architecture, 10 models are trained with different random seeds. The Grad-CAM explanations are generated on the test set for these 10 models, with these explanations also being used to calculate the explanation consistency C for each architecture. Following the methods of [26], we use a binary logistic regression classifier to measure the separability of two sets of explanations.

Furthermore, we confirm our results on Grad-CAM by repeating these experiments with SHAP. This confirms that our results are not limited to one explanation technique; if both explainability methods agree on the outcome, then we can conclude with increased certainty that the model is indeed learning "better" (i.e. similar, causal) features.

5. Results and Discussion

Table 1 reports the best model performance as well as summary statistics for both the KLD and NSS metrics used to compare the similarity between the model's Grad-CAM explanations and the EGD. Table 1 in the Supplementary Material reports the results for each training hyperparameter setup used. The performance of both the Baseline and Improved UNet models are equal to the results reported in [15], confirming that these models are behaving as expected. Furthermore, both ensembling techniques perform better than these two models; this is to be expected given that they are ensemble architectures [6]. Importantly, our Explanation Ensemble architecture is shown to improve upon the performance of the baseline models by 3.39% indicating that the models are not sacrificing model performance for improved explanations. Given that the explanations from Explanation Ensembles are shown to better align with radiologist EGD, this also suggests that features used by radiologists are better for disease classification than those learned by the baseline model.

Both Table 1 and Figure 1 report the Kullback-Leibler Divergence and Normalised Scanpath Saliency between the Grad-CAM explanations from each model architecture and the radiologist's EGD heatmaps (for details on EGD heatmap generation, see [15]). From Figure 1 we can see that our Explanation Ensemble model produces explanations that are more similar to the EGD than all other architectures tested, when measured by both a distribution-based measure (KLD) and a location-based metric (NSS). To confirm that these conclusions are statistically correct, we perform a Paired t-test at the $\alpha = 0.05$ significance level between the similarity metrics from the baseline and Explanation Ensemble models. Our null and alternative hypotheses are the same for both KLD and NSS: $H_0 : \mu_d = 0, H_1 : \mu_d \neq 0$, where μ_d is the mean of the differences between the KLD/NSS values for the two architectures. The distributions of the differences were confirmed to be normal before carrying out the t-test. Table 2 reports both the test statistics and p-values for each of our hypothesis tests. Given that all p-values are significantly less than α, we can conclude that our explanation ensemble architecture produces explanations that are statistically more similar to radiologist EGD than both baseline and current state-of-the-art techniques. Significantly, all models except explanation ensembles achieve negative NSS scores, showing anti-correspondence against the EGD [4] and making our explanation ensemble architecture the only method tested to use features that are positively correlated with those used by experts. This is further highlighted by the large reduction in KLD from our methods when compared with the baseline models tested; this underlines how significantly different the features used by current state-of-the-art models and medical experts are (and follows results suggesting that many networks suffer from shortcut learning [9] and spurious correlations [27]), and shows that our proposed method is a significant improvement. While we have focused on Explanation Ensembles of size 10 in this paper, the effect

Table 1. Table reporting the performance of the best-performing model for each architecture, alongside the similarity between the model Grad-CAM explanations and the EGD. Note that KLD is a divergence metric, and so smaller is better. Grad-CAM explanation consistency was calculated across all 10 training hyperparameter setups for each architecture.

		KLD		NSS		
Model	Accuracy	Mean ($\pm$ std. dev)	Median ($\pm$ IQR)	Mean ($\pm$ std. dev)	Median ($\pm$ IQR)	Consistency
Baseline	75.55%	14.4041 ± 7.6886	13.4535 ± 10.5240	-0.8579 ± 1.2345	-1.0391 ± 1.4737	0.1785
Improved UNet	76.51%	9.9371 ± 6.4179	9.1221 ± 8.4260	-0.3244 ± 1.5237	-0.4634 ± 1.9781	0.1596
Normal Ensemble	**79.86%**	3.8839 ± 3.2510	2.7740 ± 4.0799	-0.1646 ± 1.5721	-0.1307 ± 2.0840	0.3042
Explanation Ensemble (Ours)	78.94%	$\mathbf{0.8196 \pm 0.1273}$	$\mathbf{0.8398 \pm 0.1658}$	$\mathbf{0.6757 \pm 1.1178}$	$\mathbf{0.5410 \pm 1.5653}$	**0.5333**

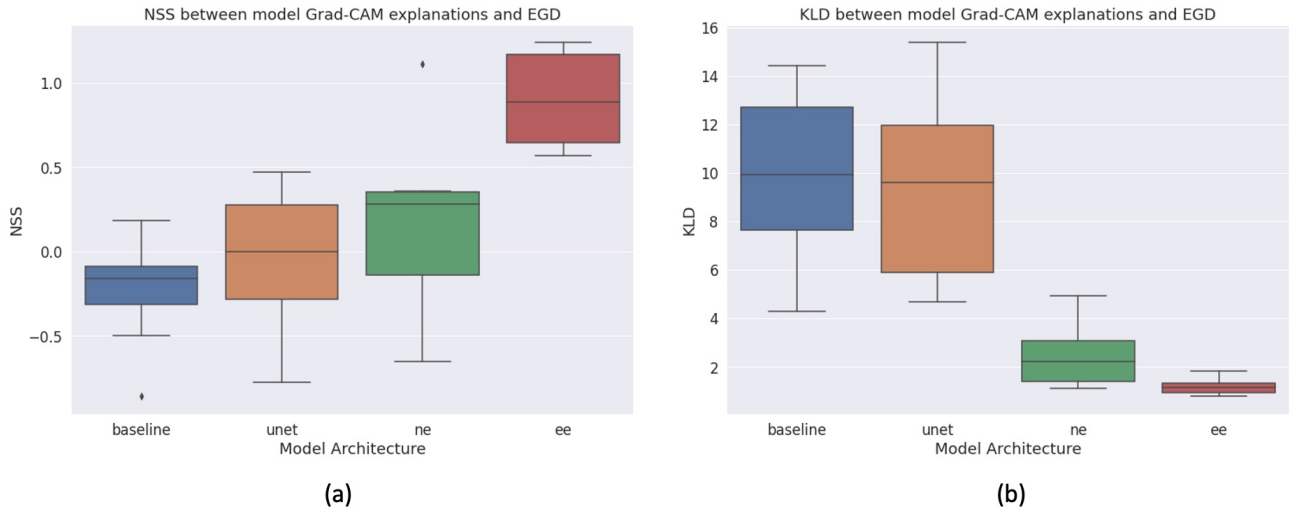

Figure 1. Boxplots of mean (a) NSS and (b) KLD between model Grad-CAM explanations and radiologist EGD, across each of the 10 training random seeds tested. Note that KLD is a divergence metric meaning smaller values are better.

of changing the number of sub-models is explored in Figure 1 of the Supplementary Material. These experiments show that as the number of sub-models increase so does the agreement between model explanations and the EGD - however, it is important to note the trade-off between training cost and increased performance as the Explanation Ensemble size increases.

In addition to improved similarity with expert EGD, explanation consistency (Table 1) is also significantly improved in our explanation ensemble model. This can also be seen by the significantly smaller range of NSS and KLD of the explanations from the explanation ensembles (as reported in Figure 1) when compared with other architectures tested. This inherently increases trust in the model, as it shows that our architecture is more robust than the others tested. It also further highlights how our network learns "better" (i.e. similar to those in EGD) features than the baseline models - our model is learning fewer noisy/spurious features and instead placing more importance on the features that have a higher probability of being causally related to the task.

We also investigate the similarity between SHAP values and the EGD data; this is shown in Figure 2. Similarly to the Grad-CAM results, we see that our proposed Explanation Ensemble architecture improves the similarity upon all other model architectures tested. Similar patterns can be seen between all 4 architectures tested across the KLD and NSS values on the Grad-CAM and SHAP results, with the boxplots highlighting that the level of improvement of our explanation ensemble architecture is at the same scale regardless of the explainablility technique used. As both the results of Grad-CAM and SHAP agree, we can conclude that our proposed model is learning to use features similarly to a radiologist. These results can also be seen from a visual comparison of explanations: Figure 3 shows example CXRs and their corresponding EGD and explanations from all models tested, showing that our explanation ensemble places much more importance on regions similar to the expert radiologist (i.e. around the lungs and heart) than both the baseline and current state of the art models. Notice how columns 2 (baseline Grad-CAM) and 3 (Improved UNet Grad-CAM) in Figure 3 show how much of the feature attribution is placed in spuriously correlated features (such as the top-left corner and the image borders). On the other hand, our explanation ensemble architecture learns a significantly different set of features (using features around

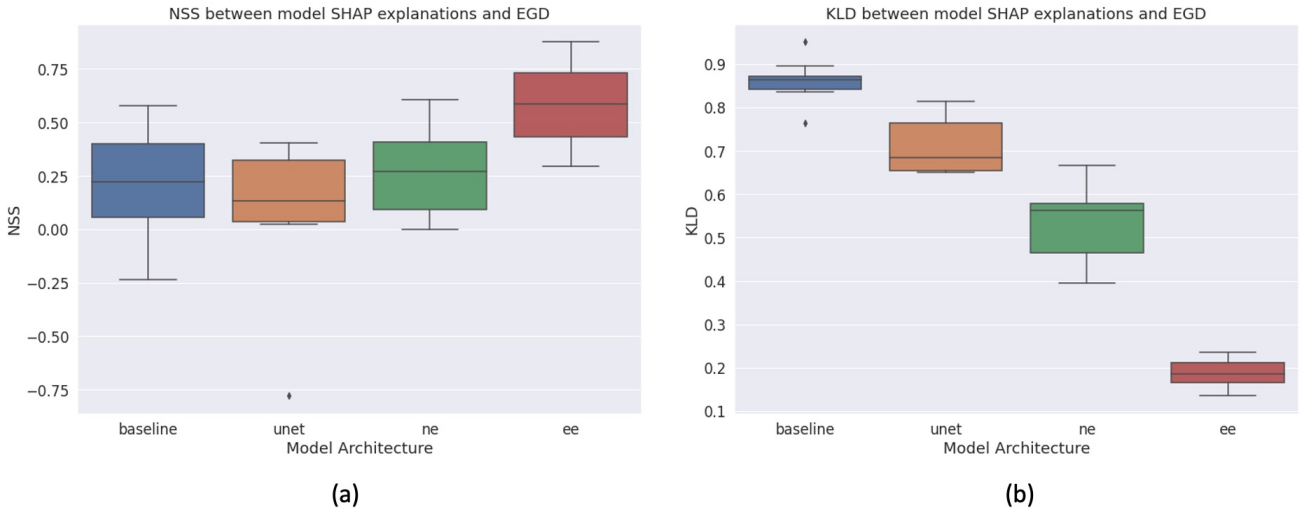

Figure 2. Boxplots showing the mean (a) NSS and (b) KLD between model SHAP explanations and radiologist EGD, across each of the 10 training random seeds tested. Note that KLD is a divergence metric meaning smaller values are better.

the lungs and heart, with these areas much more closely matching the areas shown in the EGD heatmap in the first column), further showing that our training technique has a notable affect on the representations learned by the model. This is desirable, as it highlights how our model is learning to use features similar to those used by experts, making it less likely that our model is over-reliant on spurious features.

Figure 4 shows how the learned features of our explanation ensemble model change as training progresses. Note that this figure shows only the most important pixels of each model - when showing the importance of all pixels, the heatmaps become difficult to analyse by eye. In particular, Figure 4 highlights how our training process (i.e. the discriminator and our loss function in Equation 2) encourages the sub-models of our ensemble to learn similar features as training progresses, despite the sub-models starting with vastly different sets of explanations. This verifies that our intuitive understanding of our explanation ensemble architecture, and most importantly our understanding of *why* it produces explanations closer to expert's EGD, is correct.

Table 2. Test statistics t and p-values for the Paired t-test performed between the Explanation Ensembles and Baseline (top) and the Explanation Ensembles and Improved UNet (bottom) models.

	Test Statistic	***p*-value**
KLD	18.005	6.8698×10^{-34}
NSS	-9.9137	5.7567^{-17}
	Test Statistic	***p*-value**
KLD	14.4617	7.5950×10^{-27}
NSS	-5.8058	3.5764×10^{-8}

6. Conclusion

Through the use of two explainability techniques and both distribution- and location-based metrics, we have shown that our Explanation Ensemble technique improves upon baseline models in both terms of performance and explanation similarity to EGD on the MIMIC-CXR-EGD dataset. Furthermore, we have shown that the Explanation Ensemble architecture also improves upon the current state-of-the-art models which share learned features with radiologist's EGD. In addition to improving agreement between model explanations and expert EGD, our proposed model architecture also improves classification performance and explanation consistency when compared with current state of the art techniques. Qualitative analysis of our results shows that our proposed architecture is a highly significant improvement upon current models, and whilst we do not claim that our results are yet perfect they are a huge improvement in what is a very difficult task. Furthermore, unlike the previous state of the art [15] technique, our proposed architecture does not require EGD heatmaps during training - due to the cost of collecting EGD (especially in fields such as medicine, where expert knowledge is required), we believe this is a significant advantage over previously proposed methods.

In future work, it would be interesting to perform an in depth causal analysis of the learned features of our model and compare this with a causal analysis of the learned features of baseline models. The improved performance, increased explanation consistency and improved better agreement with expert EGD suggests that our architecture may be learning more causal features than the baseline models,

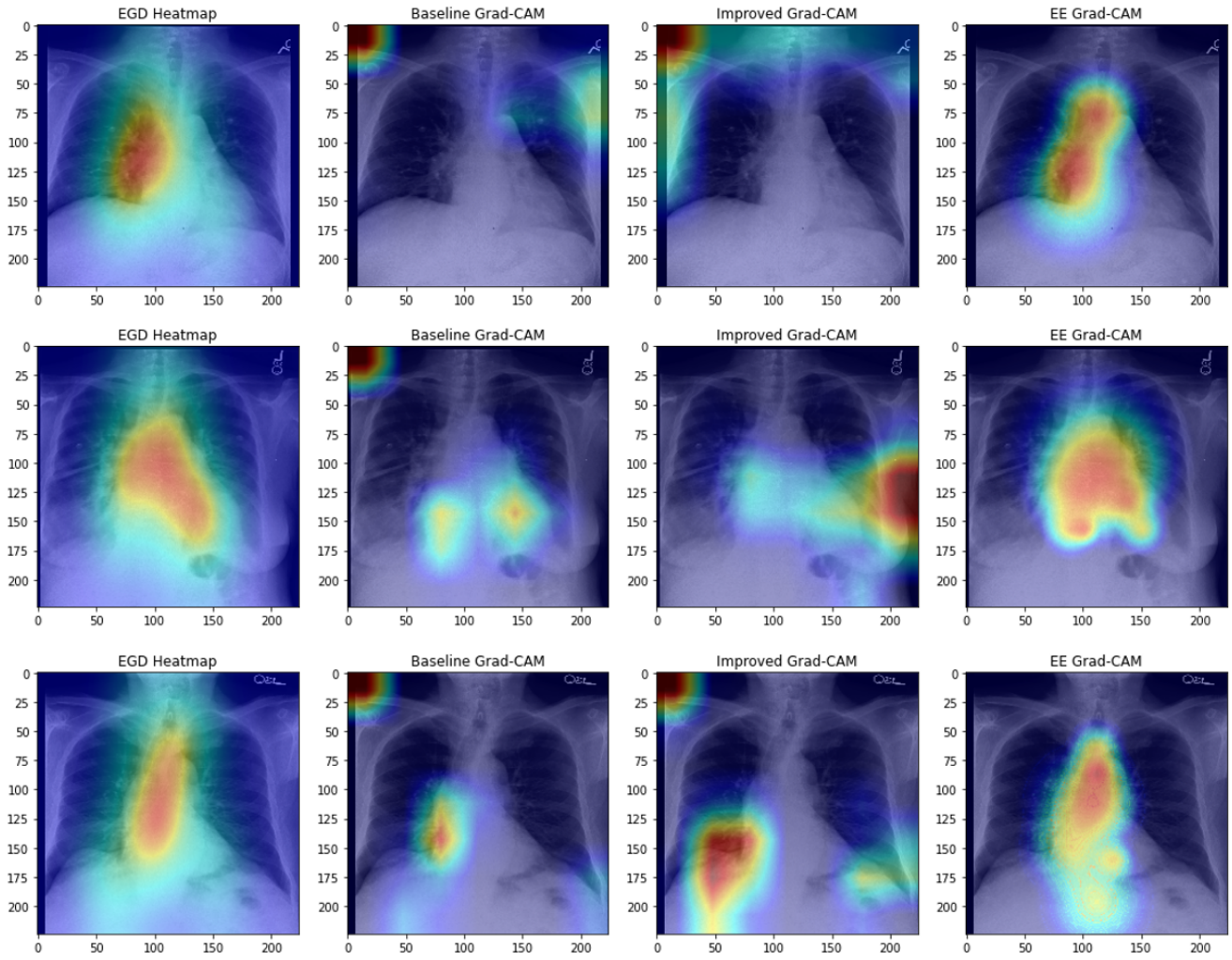

Figure 3. 3 samples from the MIMIC-CXR-EGD dataset, overlaid with the radiologist's EGD and Grad-CAM explanations from the baseline, improved UNet and Explanation Ensemble models.

with the baseline models possibly relying more on spurious features. We hypothesise this as one would only expect causal features to be those that are learned consistently across multiple variations of a well-performing model. Furthermore, the increased agreement with expert radiologists (whom you would expect to use causal features in their diagnoses) further supports this conclusion. However, to fully verify this hypothesis, an extensive causal analysis of the trained models, and their learned features, must be undertaken (using techniques such as those used in [20] and [11]) and so we leave this for future work.

Due to its increased similarity with a medical professional's decision making process, we believe that more trust will be placed in our model by clinicians than current state-of-the-art techniques. We hope that these results encourage the use of our architecture in other areas of medical practice, and other sensitive fields, as well as the release of further datasets similar to MIMIC-CXR-EGD which can facilitate this type of research.

Acknowledgements

This work is supported by grant 25R17P01847 from the European Regional Development Fund and Cievert Ltd.

References

[1] Zakhriya Alhassan, Matthew Watson, David Budgen, Riyad Alshammari, Ali Alessa, and Noura Al Moubayed. Improving current glycated hemoglobin prediction in adults: Use of machine learning algorithms with electronic health records. *JMIR Med Inform*, 9(5):e25237, May 2021.

[2] Stan Benjamens, Pranavsingh Dhunnoo, and Bertalan Meskó. The state of artificial intelligence-based fda-

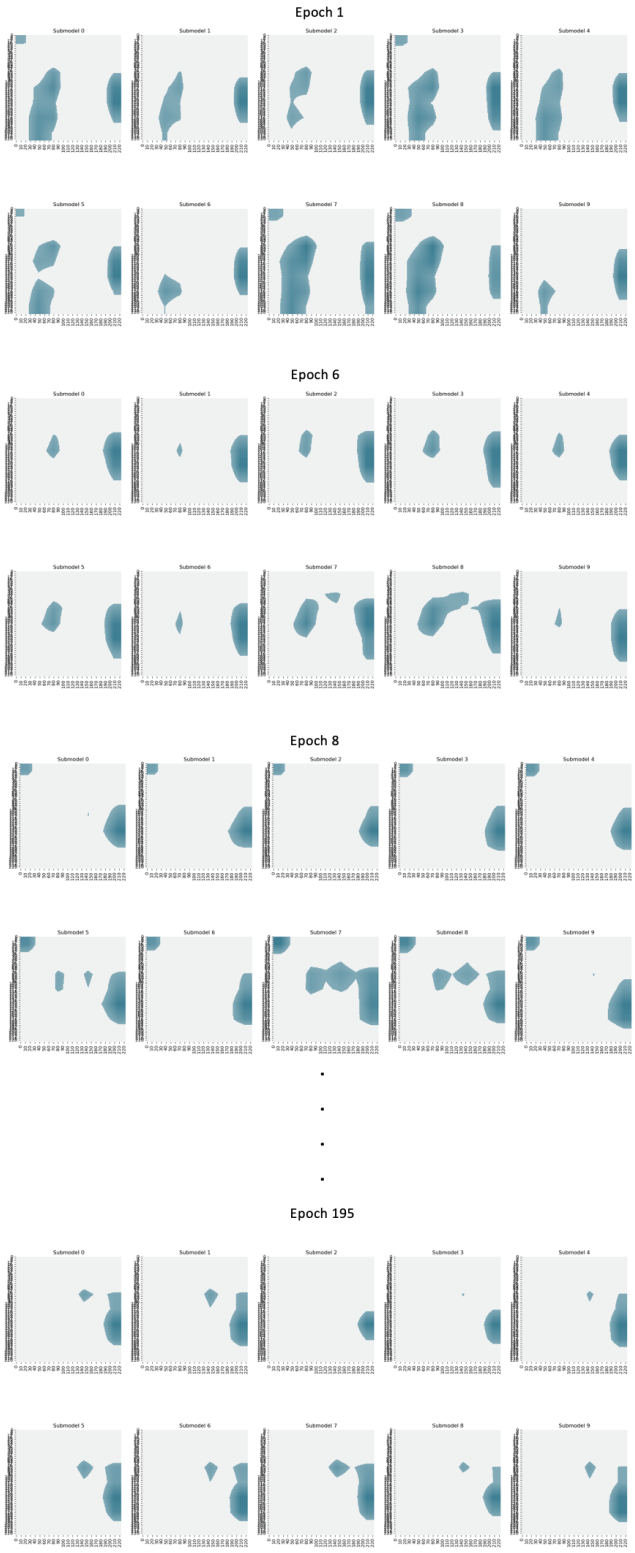

Figure 4. Average GradCAM values (across the validation split) of each sub-model of our Explanation Ensemble model, as training progresses. To aid with visualisation, only the most important 50% of pixels are shown. Sub-models start training with vastly different learned features, and as training progresses our training procedure encourages the sub-models to learn similar features. A fully animated version of this figure, and code to reproduce it on other models, will be released upon publication.

approved medical devices and algorithms: an online database. *npj Digital Medicine*, 3(1):118, Sep 2020.

[3] Bradley Butcher, Vincent S. Huang, Christopher Robinson, Jeremy Reffin, Sema K. Sgaier, Grace Charles, and Novi Quadrianto. Causal datasheet for datasets: An evaluation guide for real-world data analysis and data collection design using bayesian networks. *Frontiers in Artificial Intelligence*, 4, 2021.

[4] Zoya Bylinskii, Tilke Judd, Aude Oliva, et al. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2018.

[5] D. S. Char, M. D. Abràmoff, and C. Feudtner. Identifying Ethical Considerations for Machine Learning Healthcare Applications. *Am J Bioeth*, 20(11):7–17, 11 2020.

[6] Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. A survey on ensemble learning. *Frontiers of Computer Science*, 14(2):241–258, 2020.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[8] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[9] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *CoRR*, abs/2004.07780, 2020.

[10] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[11] Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*, 2019.

[12] Joshua James Hatherley. Limits of trust in medical ai. *Journal of Medical Ethics*, 46(7):478–481, 2020.

[13] Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis, and Douglas B. Kell. What do we need to build explainable AI systems for the medical domain? *CoRR*, abs/1712.09923, 2017.

[14] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, Dec 2019.

[15] Alexandros Karargyris, Satyananda Kashyap, Ismini Lourentzou, et al. Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for ai development. *Scientific Data*, 8(1):92, Mar 2021.

[16] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774, 2017.

[17] Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, Jonathan R. Weir-McCall, Zhongzhao Teng, Effrossyni Gkrania-Klotsas, Alessandro Ruggiero, Anna Korhonen, Emily Jefferson, Emmanuel Ako, Georg Langs, Ghassem Gozaliasl, Guang Yang, Helmut Prosch, Jacobus Preller, Jan Stanczuk, Jing Tang, Johannes Hofmanninger, Judith Babar, Lorena Escudero Sánchez, Muhunthan Thillai, Paula Martin Gonzalez, Philip Teare, Xiaoxiang Zhu, Mishal Patel, Conor Cafolla, Hojjat Azadbakht, Joseph Jacob, Josh Lowe, Kang Zhang, Kyle Bradley, Marcel Wassin, Markus Holzer, Kangyu Ji, Maria Delgado Ortet, Tao Ai, Nicholas Walton, Pietro Lio, Samuel Stranks, Tolou Shadbahr, Weizhe Lin, Yunfei Zha, Zhangming Niu, James H. F. Rudd, Evis Sala, Carola-Bibiane Schönlieb, and AIX-COVNET. Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. *Nature Machine Intelligence*, 3(3):199–217, Mar 2021.

[18] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[19] Shyamli Sindhwani, Gregory Minissale, Gerald Weber, Christof Lutteroth, Anthony Lambert, Neal Curtis, and Elizabeth Broadbent. A multidisciplinary study of eye tracking technology for visual intelligence. *Education Sciences*, 10(8), 2020.

[20] Sumedha Singla, Stephen Wallace, Sofia Triantafillou, and Kayhan Batmanghelich. Using causal analysis for conceptual deep learning explanation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 519–528. Springer, 2021.

[21] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[22] Eric J. Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, Jan 2019.

[23] A van der Gijp, C J Ravesloot, H Jarodzka, M F van der Schaaf, I C van der Schaaf, J P J van Schaik, and Th J Ten Cate. How visual search relates to visual diagnostic performance: a narrative systematic review of eye-tracking research in radiology. *Adv. Health Sci. Educ. Theory Pract.*, 22(3):765–787, Aug. 2017.

[24] Alfredo Vellido, Vicent Ribas, Carles Morales, Adolfo Ruiz Sanmartín, and Juan Carlos Ruiz Rodríguez. Machine learning in critical care: state-of-the-art and a sepsis case

study. *BioMedical Engineering OnLine*, 17(1):135, Nov 2018.

[25] Stephen Waite, Arkadij Grigorian, Robert G. Alexander, Stephen L. Macknik, Marisa Carrasco, David J. Heeger, and Susana Martinez-Conde. Analysis of perceptual expertise in radiology – current knowledge and a new perspective. *Frontiers in Human Neuroscience*, 13, 2019.

[26] Matthew Watson, Bashar Awwad Shiekh Hasan, and Noura Al Moubayed. Agree to disagree: When deep learning models with identical architectures produce distinct explanations. *CoRR*, abs/2105.06791, 2021.

[27] Yao-Yuan Yang and Kamalika Chaudhuri. Understanding rare spurious correlations in neural networks, 2022.

[28] Erdi Çallı, Ecem Sogancioglu, Bram van Ginneken, Kicky G. van Leeuwen, and Keelin Murphy. Deep learning for chest x-ray analysis: A survey. *Medical Image Analysis*, 72:102125, 2021.