Extending Simplex Explainability From Individual Examples to Quantify a Global Measure of Credibility

Gurpreet Johl
DPhil Applicant
gurpreetjohl@gmail.com

Abstract

Model selection is a key aspect of deploying ML models in practice, though knowing whether the model is appropriate for the data set at hand is not straightforward. Accuracy on a test data set is often used for this purpose, but this does not take into account the confidence in those predictions; a model may be extrapolating far beyond the bounds of its training set and still achieve reasonable accuracy. This is especially true when data sets are limited in size, so by testing multiple models one may find a model with high test accuracy by luck. Simplex provides a method to explain a given test example in terms of a corpus of examples. This paper seeks to extend the Simplex approach to the overall dataset, to serve as a distance measure of how far a particular model needs to extrapolate in order to make predictions for the given data set. In doing so, this can establish a global quantification of extrapolation that can serve as measure of model confidence. This gives practitioners another perspective of the applicability of a model for their data set.

1 Introduction

A key consideration when applying machine learning models in real-world settings is model selection—how do we choose the best model and how confident can we be in its outputs? This is an open question, with practitioners relying on heuristics and judgement based on a number of accuracy and confidence measures[1][2]. This is particularly relevant in settings such as healthcare because data is often limited and the impact of decisions based on these models is significant. Having a better insight into the applicability of a model for a given task provides a better understanding of its strengths and limitations, so can help to reduce spurious predictions and biases against groups underrepresented in the training data, thus improving health inequality.

The Simplex approach [3] can be used to explain individual predictions of a particular model in terms of either examples or features. A particular application of this explored in the original paper is to identify examples from a different distribution using the residuals of the latent approximation, in effect measuring the degree of extrapolation required by the model in order to classify a particular instance.

This research aims to extend this method by aggregating this concept over the test data set, to provide a global quantification of the degree of extrapolation. For a given data set, to what extent does a particular model need to extrapolate beyond the data it was trained on? Intuitively, we may want to be more sceptical of models with a high degree of extrapolation, so this effectively provides a measure of confidence. This motivation is investigated with a real-world application in Section 5.4—given an cancer detection image classification model which, unbeknownst to us, has only been trained on images of lungs, can we identify when the model is operating outside of the circle of competence? This can guide practitioners in cases where extra caution should be exercised when following the model's outputs, even when the accuracy is high.

This has implications in model selection, detecting biases in data and transfer learning among others. The benefit of such a confidence measure is that the same Simplex methodology can be used to evaluate confidence at a high level for the overall data set, and then any discrepancies can be investigated further in terms of examples or features.

2 Related works

The concept of latent space distances as a measure of interpretability is used in the Deep k-Nearest approach [4] and this is used to address the related concepts of confidence and credibility. Confidence is defined as "how likely the prediction is to be correct according to the model's training set", while credibility quantifies "how relevant the training set is to make this prediction". The distance measures presented hereafter provide a measure of extrapolation which may therefore fit the definition of credibility more closely, although the concepts of confidence and credibility are closely coupled. Distinguishing between confidence or credibility is not the focus of this work, so these will be used interchangeably hereafter. In the discussion of the Deep kNN approach, Papernot and McDaniel noted the importance of an extrapolation measure: "when one wishes to deploy ML in settings where safety or security are critical, it becomes necessary to invent mechanisms suitable to identify when the model is extrapolating too much from the representations it has built with its training data."

The idea of a single scalar value encapsulating the interpretability of a model is discussed as a desired attribute in the context of *Concept Activation Vectors* [5], where "plug-in readiness" and "global quantification" are noted as desirable attributes of an interpretability measure. This motivates the goal of a single value encapsulating confidence in this paper. A measure based on the Simplex residuals has the former, by virtue of being a post-hoc explainability technique. This work focuses on extending Simplex to provide a "global quantification" measure of confidence.

Various evaluation measures are discussed by Ding et al. [1] While these primarily focus on the model's performance on a given task rather than interpretability or confidence, we later discuss possible avenues of future research that more closely align accuracy and confidence measures, as illustrated in Figure 21.

The relative importance assigned to performance and confidence measures when selecting between models is subjective. Rechkemmer and Yin identify a discrepancy between the model selection behaviour stated by practitioners and that observed in practice; respondents claimed to base decisions on confidence but in fact follow the model with the highest accuracy, suggesting that decisions are based on accuracy and justified after the fact using confidence [6].

The concept of a single global measure based on the amount a model must extrapolate bears similarities to discussions of a transferability metric in the context of foundational models and transfer learning [7] [8] [9]. This is an open question in the field; the ideas presented here may help to bridge the gap between transferability and explainability. This line of thought is discussed further in Section 7.

Contribution In this work, different distance measures based on the Simplex approach are explored which can serve as a confidence measure for a given model applied to a particular data set. This aligns it with other interpretability methods such as Deep k-Nearest Neighbours, which can be used for global confidence measures that help users understand when a model is appropriate to use in practice. By using Simplex as the foundation, this retains the benefits of Simplex as an interpretability method, for example, allowing the user to specify the corpus. By aggregating it as a global measure, it means that the same methodology can be used to interrogate data either as a whole data set, in terms of its examples, or in terms of its features.

3 Problem formalism

The similarity between training data and test data can be suboptimal in practice. It can be difficult to identify when a model is operating outside of its intended limits, as the predicted classifications may still be correct. This is particularly important when sample sizes are small, as is often the case in healthcare settings, because accuracy and traditional tests of significance can be misleading when evaluating multiple models [10]. A measure of extrapolation can help provide a more robust measure of a model's appropriateness in a given setting.

Specifically, we consider the scenario where we have a model M trained on data $\boldsymbol{D_{train}}$ and we want to determine how appropriate it is for the same task on an unseen data set $\boldsymbol{D_{test}}$. Intuitively, global confidence k should depend on all three of these

$$k = f(M, \boldsymbol{D_{train}}, \boldsymbol{D_{test}}) \tag{1}$$

It is also noted that if k has consistent scaling across models and data sets, then it can also be used to distinguish between competing models $M_1, ..., M_m$.

The Simplex approach can be used to explain a given test instance in terms of a corpus of examples C. The extension proposed in this paper focuses on three key insights of the original method: (1) the behaviour of a model M on an instance $\boldsymbol{d_{test_i}} \in \boldsymbol{D_{test}}$ can understood better in latent space than input space, that is, by using its latent representation $\boldsymbol{h_{test_i}}$ produced by the model's last hidden layer, (2) an approximation of this latent representation $\tilde{h}_{test_i}$ can be defined in terms of a user-defined corpus of examples C, and (3) the residual $\boldsymbol{r_{test_i}} = ||\boldsymbol{h_{test_i}} - \tilde{\boldsymbol{h}}_{\boldsymbol{test_i}}||_2$ can identify cases where the input $\boldsymbol{d_{test_i}}$ differs from C.

Combining these insights, we can conclude that by selecting a corpus $C \subset \boldsymbol{D_{train}}$, a function of the form Equation 2 satisfies the requirements of Equation 1, and could therefore provide an insightful global confidence measure.

$$k = g(\boldsymbol{H_{train}}, \boldsymbol{H_{test}}) \tag{2}$$

The remainder of this paper explores possible formulations of such distance measures and their applications to real data.

4 Methodology

This section describes the experimental set up to evaluate confidence measures. The lack of ground truth for such a measure means that quantitative evaluation of such a measure can be challenging [4]. In the following experiments, a binary classification model is trained on a defined set of in-distribution data, then we evaluate the success of a confidence measure by observing its relation to the proportion of out-of-distribution (OOD) data, i.e. unseen classes. This is applied in two cases: Section 4.1 uses MNIST image data to build an intuition of the latent space measures on a relatively simple, well-studied data set, and Section 4.2 applies the same approach to a real-world healthcare use case. Examples of the inputs for these two experiments are given in Figures 1 and 2. All experiments are repeated 10 times and results have been replicated on different machines.

4.1 MNIST binary classification

A subset of the MNIST handwritten digit classification data set [11] is used to train a binary classifier. A training set containing only the digits 0 and 1 is used to train a convolutional neural network (CNN) with the goal of identifying ones; the digits 0 and 1 form the in-distribution data and the digit 2 is the OOD data introduced at test time. The CNN is used to classify images from test sets containing varying proportions of OOD data, serving as a ground truth measure for the level of extrapolation required. The accuracy on each test set is evaluated, along with a range of distance measures. For a fair comparison, the positive and negative classes are kept balanced; OOD=0 denotes a test set containing equal numbers of zeros and ones, as OOD increases zeros are replaced by twos, until OOD=1 where the test set contains equal numbers of ones and twos.

The CNN used for this task has two neurons in its last hidden layer. This allows the latent representation of inputs to be visualised in 2D space, which we use to gain an intuitive understanding of the latent space and its Simplex approximations. This is detailed in Section 5.1 This is repeated for the 3-dimensional case to verify that the observed behaviour does not drastically change as the dimensionality of the latent space increases.

4.2 Cancer binary classification

A similar approach is used to analyse the behaviour on a real-world health data set of lung and colon histopathological images labelled as cancerous or benign [12]. A CNN is trained on images of

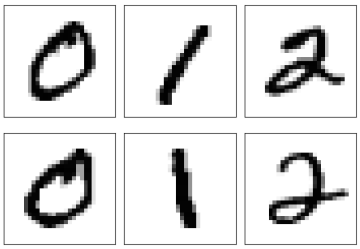

Figure 1: Example inputs from MNIST data set. The models are trained on zeros and ones (in-distribution), then shown varying quantities of twos (out-of-distribution) at test time.

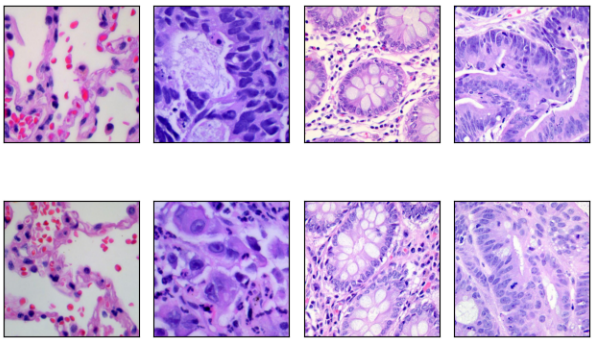

Figure 2: Example inputs from lung and colon tissue data set. Left to right: Benign lung, cancerous lung, benign colon, cancerous colon. The models are trained on lung images (in-distribution), then shown varying quantities of colon images (out-of-distribution) at test time.

lungs with the binary classification goal of identifying cancer; these are the in-distribution data. The out-of-distribution data in this case is the set of colon images. At test time, varying proportions of lung and colon images are introduced, with the model's accuracy and distance metrics evaluated for each instance. One difference to the previous experiment is the OOD data can contain positive and negative classes, i.e. colon images with cancer and without, so OOD=1 corresponds to test data taken entirely from the OOD colon images, where classes are still balanced.

5 Results

The results of the proposed experiments are provided here. Section 5.1 studies the latent space behaviour, which is used to inform distance measures described in Section 5.2. These are applied in two experiments, firstly on handwritten digit data in Section 5.3 then to a real data set of tissue images in Section 5.4.

5.1 Interpreting the latent space

It is instructive to study the latent space of a simple model to gain an intuition of what this reveals of the model's behaviour. The CNN contains two neurons in its final hidden layer, so the positions of points can be visualised in 2D latent space.

The positions of zeros and ones (in-distribution) and twos (out-of-distribution) are shown in Figure 3. We observe that the model learns to separate the in-distribution samples, the digits 0 and 1; zeros are aligned with the x-axis and ones with the y-axis. This is consistent with the interpretation of a neural network trained for binary classification as a kernel logistic regression model, where the hidden layers effectively learn an appropriate kernel to separate the classes in latent space, and the final softmax layer is a typical logistic regression.

In contrast, the out-of-distribution data, i.e. the twos, lie in the middle of the latent space, suggesting that the model is less able to distinguish these from the ones. Of note here is that they do generally lie closer to the cluster of zeros (the negative classification) than the cluster of ones (the positive

classification). This is the correct classification for our model trained to identify ones, which implies that we may still get believable results despite having never seen similar data before, therefore relying on accuracy metrics alone could be mislead one into a false sense of confidence in the model's appropriateness; the *confidence* in this model should be lower even if the predictions are correct.

The effect of applying Simplex to this data is shown in Figure 4. We observe that the in-distribution samples are shifted along their respective axes towards the centre of the cluster of that digit. This is consistent with the idea behind Simplex that "outlier" points are approximated by a combination of points in the corpus hull. In contrast, the out-of-distribution digits do not move in a consistent direction, reflecting that the model is unable to reconstruct them adequately using the corpus. This can be clearly observed in Figure 5 which shows the residual vectors in latent space.

The same approach is repeated for the 3-dimensional case to verify that similar relations hold. A new CNN is created with 3 neurons in the last hidden latent layer. Following the same approach, in Figure 6 the in-distribution data points again form linearly separable clusters in 3D latent space, while the out-of-distribution data lies in between. In Figures 7 and 8, it is again observed that the Simplex approximations bring points closer towards the centre of clusters. [1]

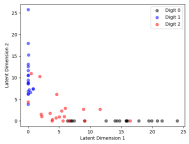

Figure 3: Distribution of a sample of digits in the model's latent space.

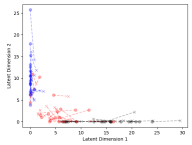

Figure 4: Residual shift of true values (circles) to their Simplex approximation (crosses).

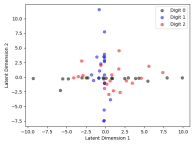

Figure 5: Residual vectors in the model's latent space.

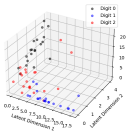

Figure 6: Distribution of a sample of digits in the model's latent space.

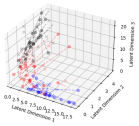

Figure 7: Residual shift of true values (circles) to their Simplex approximation (crosses).

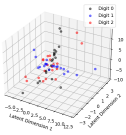

Figure 8: Residual vectors in the model's latent space.

5.2 Distance measures

A set of distance measures is proposed, guided by the experimental insights detailed in Section 5.1. These are then applied to the two data sets to establish their relation to the OOD proportion, i.e. the level of extrapolation

A natural starting point for a distance measure is to use the residual vectors of the Simplex approximation, which give a measure of how well closely a test example can be decomposed as a combination of corpus inputs. This is the motivation behind the r_norm distance measure, which calculates the vector norm of each residual example, and averages this over the test data set.

$$r_{norm} = \frac{\sum_i^n ||\boldsymbol{r_{test_i}}||_2}{n} \forall i \in \boldsymbol{D_{test}} \tag{3}$$

One may naively expect this to be larger for OOD data points, but as observed in Figures 4 and 5, the in-distribution points move along the axis, whereas OOD points move more randomly and begin

[1] Interactive versions of these 3D plots illustrate this more clearly. These are available in the Github repository for the latent space distribution , the residual distribution and an interactive time lapse animation of the residual shift.

closer to the centre of the latent space, so their residuals are actually smaller. This direction-dependent observation is key, and forms the basis of later distance measures.

The next straightforward measure is to take the norm of the latent space. As observed in Figure 3 and 6, the in-distribution points a separated by the classifier, whereas OOD points cluster towards the centre of the latent space. Therefore, this should provide some some measure of the shrinkage of the latent space

$$h_{norm} = ||\boldsymbol{H_{test}}||_2 \tag{4}$$

A variation of this approach notes that if one class has OOD points but the other does not (as in the MNIST experiment where we have OOD negative classifications but not positive classifications) then the latent space occupied by each predicted classification may grow at different rates. If the number of samples in each class was imbalanced, this observation could contain useful information. The next measure separates points by their predicted classification before calculating the norm of each, then returns the weighted average of class-wise norms.

$$h_{norm\_classwise} = \sum_c \frac{|\boldsymbol{D_{test_{\hat{y}=c}}}|}{|\boldsymbol{D_{test}}|} \cdot ||\boldsymbol{H_{test_{\hat{y}=c}}}||_2 \forall c \in [0, 1] \tag{5}$$

A direction-wise variant of this measure, $h_{norm\_directionwise}$, is calculated by noting that the in-distribution points of a given class generally occupy a particular hyperplane of the latent space, so the out-of-plane axis is generally noise. We expect this noise to be higher for OOD points than in-distribution points. The out-of-plane axis is determined by selecting the axis with the smallest standard deviation for the validation set. The calculation is then identical to Equation 5.

Relative variants of each of the above measures are obtained by taking the ratio of h_{norm} values between the test set and the latent approximation of the validation set under Simplex. Using a validation set in this way resembles the calibration set used for the deep k-nearest neighbours approach [4]. This allows for more interpretable scales—a value of 1 for h_{norm} or $h_{norm\_classwise}$ indicates that the test set came from a similar distribution as the in-distribution validation set, whereas a smaller value suggests the test contains more OOD data, which shrinks the latent space towards the centre. For $h_{norm\_directionwise}$ the value should increase with OOD data as these are typically further out-of-plane.

$$h_{rel\_norm} = \frac{h_{norm_{test}}}{\hat{h_{norm_{validation}}}} \tag{6}$$

5.3 MNIST classification experiment results

Common performance metrics accuracy and AUC are shown in Figure 9, as well as the the mean of the probabilities output by the model which are commonly cited as a measure of confidence. It is observed that accuracy and AUC both decrease as the out-of-distribution proportion increases as expected, though it should be noted that their absolute values are still relatively high; a practitioner may well still choose such a model on the basis of these values despite being so far removed from the model's training data. The probabilities are high across all levels of OOD, corroborating the observations of Papernot and McDaniel in the context of Deep kNN that model probabilties are poor proxies of confidence, despite being widely used for this purpose [4]. The dispersion of probabilities does show some relation to the OOD proportion, shown in Figure 10.

The h-space norms are shown in Figure 11. These decrease as OOD points are added, in line with the understanding that OOD points lie in the middle of the space whereas in-distribution points a separated to the extremes.

A residual-based distance measure is shown in Figure 12. This also decreases with the OOD proportion because, as observed in Figure 5, the in-distribution points are at the extremes of the space and are clustered toward the centre of their respective classes under the Simplex approximation. The OOD points are already closer to the centre of the space. It is also noted that the direction is informative; in-distribution points lie in a plane which they move within under Simplex, whereas OOD points do not move in a consistent direction. This provides the motivation for the direction-specific distance measures below.

Relative measures of the latent space are shown in Figures 13 and 14. These calculate the ratio of norms between the test data latent space and that of a validation set containing only in-distribution

points unseen at training time. The regular and classwise variants in 13 decrease with OOD for similar reasons as discussed, as the OOD points cluster in the middle of the latent space. These two variants are very similar, owing to the balanced classes of the experimental design. The behaviour in cases of imbalanced classes is a topic of further research. Of note about this formulation is that it has an interpretable scale: values close 1 denote the test and validation sets have similar distributions whereas smaller values denote test data that differs from the validation data and the model is therefore extrapolating beyond the limits of its training data. Another observation about these relative distance measures is the resemblance to the calibration set approach used in [4].

A third variation of this relative measure is shown in Figure 14, where the class-specific out-of-plane axis is isolated before calculating the ratio of norms. This increases with OOD because the in-distribution data points lie on the same plane so have small values in the out-of-plane axis, whereas OOD points are more scattered.

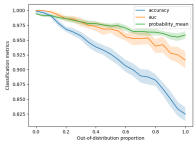

Figure 9: Classification metrics for MNIST data.

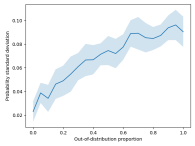

Figure 10: Standard deviation of model probabilities.

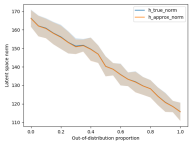

Figure 11: Norm of latent space.

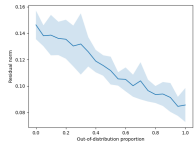

Figure 12: Norm of residuals.

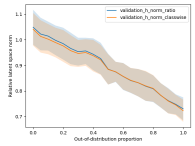

Figure 13: Ratio of test and validation latent space norms.

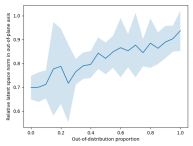

Figure 14: Ratio of test and validation latent space norms in the out-of-plane axis.

5.4 Tissue classification experiment results

The results for a similar experiment on the data set of tissue images shows largely consistent results with the previous experiment, see Figures 15 to 20. This is encouraging, as it shows the conclusions drawn generalise to real-world data sets, and may therefore be useful to aid practitioners in understanding and applying models.

The only notable difference is in the behaviour of the direction-specific latent space norm in Figure 20. This may be due to the different dimensionality of the latent space of the model used for this experiment. For each class in the data, this measure isolates the dimension with the smallest variance in the validation data as the noise axis for this class. This approach worked for the MNIST model with 2D latent space. For higher dimensions, this may not be as appropriate. An alternative methodology may be more robust in higher dimensions, where the covariance matrix of each class in the validation set is used to identify a variable number of axes to isolate. This is an area of further research.

6 Conclusions

We discuss the merits of exploring the latent space learned by a model to understand its behaviour for in-distribution and out-of-distribution data. The distribution of points in latent space, as well as the residuals under the Simplex approximation, are informative in distinguishing out-of-distribution data points. This was visualised for 2D and 3D cases.

This guides the design of distance metrics which provide a global quantification of extrapolation which can serve as a confidence measure. These distance measures were applied to two binary classification

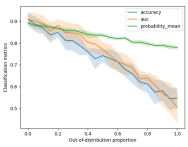

Figure 15: Classification metrics for tissue images.

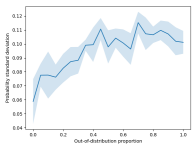

Figure 16: Standard deviation of model probabilities.

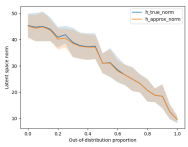

Figure 17: Norm of latent space.

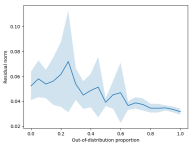

Figure 18: Norm of residuals.

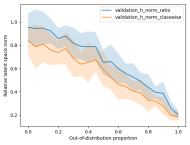

Figure 19: Ratio of test and validation latent space norms

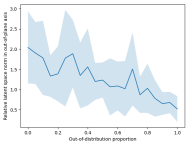

Figure 20: Ratio of test and validation latent space norms in the out-of-plane axis.

tasks in an experimental setup that allowed us to vary the ground truth level of extrapolation. The distance metrics showed consistent relations with the proportion of OOD data, and could therefore be useful in identifying when a model is operating beyond its intended use case. In both experiments, we show that the distance measures correspond with the level of extrapolation by varying the degree of out-of-distribution samples in the test set. This can be used to supplement accuracy measurements to guide practitioners when selecting a model and understanding its limitations.

By identifying cases where data sets require a high degree of extrapolation we give the practitioner the opportunity to either select a more appropriate model or address the cause of extrapolation, which can be done using the original Simplex method to understand individual instances either in terms of their features or corpus examples. This can identify aleatoric uncertainty caused by missing features or underrepresented examples in the training data that could cause biases exacerbating healthcare inequity, thus could be a useful approach in the burgeoning field of data-centric AI [13]. The consequence of the extension proposed in this work is that the Simplex approach can be applied top-down to the data set at a high-level, and then used to investigate further on the basis of features or examples, providing a holistic solution to understanding the interaction between data and models.

7 Future work

The approach of using latent space representations and approximations as a measure of confidence can be extended further across different distance formaulations, tasks and domains, as well as linking with the related areas of model design, transfer learning and data-centric AI.

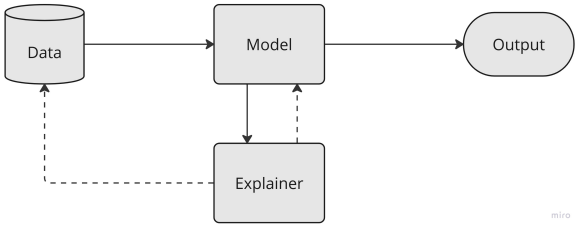

Figure 21: Using explainability methods to continuously improve models and data. The solid black arrows show the flow of data in current paradigms, where a model generates outputs from data and a post-hoc explainer can interpret the model's behaviour. The dashed lines show how global explainability measures can facilitate a feedback loop where models and data can be improved based on the explainer.

Variations and comparisons with other distance measures Alternative formulations of distance measures based on latent spaces and Simplex approximations could be explored, some noted in previous sections. Calibrating these measures experimentally on different models and data sets, as well as theoretical guarantees for such measures, are areas of further research. The effectiveness of using validation sets from data other than that used to train the model could be explored. This would allow the approach to be used in cases where the training data is not available, such as where privacy concerns limit the dissemination of the training set. The approach could also be applied over multiple hidden layers of the model and aggregated, as in the Deep KNN approach, rather than using only the final hidden layers. The measures here could aslo be benchmarked against other approaches, including Deep KNN, as well as more traditional measures of distance often used in transfer learning [7], such as KL-divergence and Maximum Mean Discrepancy, both in the input space and latent space.

Additional tasks and domains The experiments here only considered the task on binary classification. Further work could explore the latent space for other supervised tasks (multiclass classification, regression) as well as unsupervised tasks. The experiments in this paper used image data, owing to the availability of image data sets and well-established performance of CNNs on these tasks. None of the concepts discussed are specific to computer vision, so further work could explore other domains such as text data, tabular data and time series data which is particularly applicable in healthcare settings.

Transfer learning Transfer learning is an increasingly important topic with the proliferation of pre-trained foundation models. A method of measuring the transferability of a model across domains is an open question [9] [7]. The intuitive requirements for such a measure are that it be a model-specific measure between a source data set D_{source} and target data set D_{target}. This therefore bears a close resemblance to 1. The open question of identifying and explaining negative transfer could also be explored by researching behaviour in the latent space. A related research area noted by [7] is *interpretability* of transfer. The approach explored in this paper is rooted in explainability methods and therefore naturally addresses this concern. This has implications in healthcare, where foundational models could provide powerful predictions on small data sets, provided the user can evaluate the confidence of such a model. This could allow for for more powerful individualised models, where a foundational model can be trained on the health records of a particular patient or group of patients.

Model design With appropriate scaling of the distance measures proposed in this paper, we can compare confidence between different models to answer the question of which model best explains a given data set. Where no satisfactory models are found, this can influence the design of model architecture to better capture the unexplained information.

Bias detection The approach outlined here could be used to identify data sets which a poorly described by existing models, i.e. where substantial extrapolation is required. These can be explored in more detail by explaining the poorly approximated test data in terms of examples or features, as in the original paper [3]. The example-based approach can identify cases where particular test instances are underrepresented in the test data, for example when specific ethnic groups are not represented in the training data and therefore would be at risk of misclassification by the model. The feature-based approach can identify cases where features are not informative enough and may benefit from collecting additional fields; this addresses the issue of aleatoric uncertainty noted by Papernot and McDaniel [4]. Identifying these biases and having the tools to address them can play an important role in addressing healthcare inequity.

Continuous improvement of model and data Considering the above two points of model design and bias detection in tandem suggests an iterative approach to applying machine learning models whereby explanations of the model on a data set reveal shortcomings of both model and data, which can in turn be addressed and revised. This process can be repeated to iteratively improve models and data.

By linking explainability with model design and data collection, explainability can be framed as a key component in performance and ultimately healthcare outcomes. Not only does it provide additional

information and aid human-computer interaction, it can be employed as a tactic in improving the models themselves and addressing biases in data. This can be as a tool towards data-centric AI [13].

References

[1] Jie Ding, Vahid Tarokh, and Yuhong Yang. Model Selection Techniques: An Overview. *IEEE Signal Processing Magazine*, 35(6):16–34, November 2018. Conference Name: IEEE Signal Processing Magazine.

[2] Sebastian Raschka. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning, November 2020. arXiv:1811.12808 [cs, stat].

[3] Jonathan Crabbe, Zhaozhi Qian, Fergus Imrie, and Mihaela van der Schaar. Explaining Latent Representations with a Corpus of Examples. In *Advances in Neural Information Processing Systems*, volume 34, pages 12154–12166. Curran Associates, Inc., 2021.

[4] Nicolas Papernot and Patrick McDaniel. Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning, March 2018. arXiv:1803.04765 [cs, stat].

[5] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV), June 2018. arXiv:1711.11279 [stat].

[6] Amy Rechkemmer and Ming Yin. When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models. CHI '22, pages 1–14, New York, NY, USA, April 2022. Association for Computing Machinery.

[7] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A Comprehensive Survey on Transfer Learning. *Proceedings of the IEEE*, 109(1):43–76, January 2021.

[8] Manuel Weber, Maximilian Auch, Christoph Doblander, Peter Mandl, and Hans-Arno Jacobsen. Transfer Learning With Time Series Data: A Systematic Mapping Study, 2021.

[9] Sinno Jialin Pan and Qiang Yang. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, October 2010. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

[10] Campbell R. Harvey and Yan Liu. Multiple Testing in Economics, November 2013.

[11] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[12] L. Brannon Thomas Catherine P. Wilson Lauren A. DeLand Stephen M. Mastorides Andrew A. Borkowski, Marilyn M. Bui. Lc25000 lung and colon histopathological image dataset.

[13] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. Data-centric Artificial Intelligence: A Survey, June 2023. arXiv:2303.10158 [cs].