Quantifying a Global Measure of Confidence Using the Simplex Method

Gurpreet Johl
DPhil Applicant
gurpreetjohl@gmail.com

Abstract

Understanding when a model is appropriate for a given data set is a key consideration when deploying ML models in practice, but this is not straightforward. Accuracy on a test data set is often used for this purpose, although it 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—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 extends the Simplex approach to provide a single global value that answers the question "How much did the model extrapolate in order to arrive at these outputs for this data set?" which can serve as a measure of confidence in the model for the given setting. If a model achieves good accuracy but with a high degree of extrapolation, we may want to exercise caution when using the results in practice. This gives practitioners another perspective on the applicability of a model for their data set. This paper proposes possible formulations of such a distance measure and illustrates their use in a medical application.

1 Introduction

A key consideration when applying machine learning models in real-world settings is model evaluation—how do we choose the best model and how confident can we be in its outputs? But this is an open question, with practitioners relying on heuristics and judgement based on a number of accuracy and confidence measures[1][2]. In healthcare settings, this is a particularly important consideration because data is often limited and the impact of decisions based on these models is significant. Having a better understanding of the applicability of a model for a given task, including its strengths and limitations, can help to reduce negative outcomes caused by 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.

In this paper, we aim 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. A real-world example of this question is investigated in Section 5.2—given a 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 its circle of competence? A

measure of extrapolation could highlight where extra caution should be exercised when following the model's outputs, even when the accuracy is high.

This has implications in model selection, transfer learning and data-centric AI among others. The benefit of such a confidence measure is that the same Simplex methodology can be used as an end-to-end solution for identifying and addressing issues with data and models—the global confidence measure can be used to identify suboptimal model-data combinations, which can then be understood at a more detailed level in terms of both features and examples in order to rectify any issues.

Related works The concept of latent space distances as a measure of interpretability is used in the *Deep k-Nearest Neighbours* (Deep KNN) approach [4] where it is applied in the context 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" [4, p.7]. The distance measures presented hereafter provide a measure of extrapolation which may therefore fit the definition of credibility more closely than confidence, although these are closely coupled. Distinguishing between confidence or credibility is not the focus of this work and confidence is used more widely in related literature, so these terms will be used interchangeably hereafter. The importance of an extrapolation measure was noted in the context of the Deep KNN approach: "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."[4, p.5]

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*, where "plug-in readiness" and "global quantification" are explicit goals of the research [5, p.2]. A measure based on the Simplex residuals has the former, by virtue of being a post-hoc explainability technique. In this paper, we focus 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 7. The relative importance assigned to performance and confidence measures when selecting between models is subjective. Rechkemmer and Yin observed that respondents claimed to base decisions on confidence but in fact follow the model with the highest accuracy, suggesting that model selection choices are often based on accuracy metrics and justified after the fact using confidence metrics [6].

The concept of a single global measure based on model extrapolation 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, which is discussed further in Section 7.

Contribution In this work, we explore different distance measures based on the Simplex approach, to act as a confidence measure for a given model applied to a particular data set, thereby aligning Simplex with other interpretability methods such as Deep KNN. By using Simplex as the foundation, this measure retains the benefits of Simplex as an interpretability method, namely, allowing a user-defined corpus. By aggregating it as a global measure, 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.

2 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, i.e. using its latent representation $\boldsymbol{h_{test_i}}$ produced by the model's last hidden layer, (2) an approximation of this latent representation $\tilde{\boldsymbol{h}}_{\boldsymbol{test_i}}$ can be defined in terms of a user-selected 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 following form 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 studies their empirical behaviour on real data.

3 Methodology

This section describes the experimental set up used to evaluate confidence measures. The lack of ground truth for such a measure means that quantitative evaluation can be challenging [4, p.8]. In the following experiments, a binary classification convolutional neural network (CNN) model is trained on a defined set of in-distribution data (IDD). We then evaluate the success of a confidence measure by observing its relation to the proportion of out-of-distribution data (OOD), i.e. unseen classes. This is applied in two cases: Section 3.1 uses MNIST image data to build an intuition of the latent space distance measures on a relatively simple, well-studied data set, and Section 3.2 applies the same approach to a real-world medical 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.

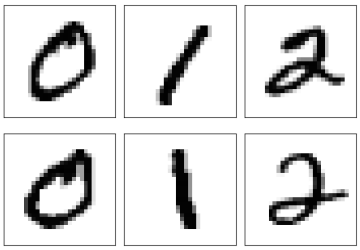


Figure 1: Example inputs from MNIST data set. The models are trained on zeros and ones (in-distribution), then shown varying quantities of twos (out-of-distribution) at test time.

3.1 MNIST binary classification

A subset of the MNIST handwritten digit classification data set [11] is used to train a binary classifier. A CNN is trained on a training set containing only the digits 0 and 1 (12000 instances total) with the goal of identifying ones. The digits 0 and 1 form the IDD data and the digit 2 is the OOD data. The CNN is used to classify unseen images from test sets containing varying proportions of OOD data, serving as a ground truth measure for the level of extrapolation. A range of accuracy and distance measures are evaluated on each test set. For a fair comparison, the positive and negative classes are kept balanced; an OOD proportion of 0 denotes a test set containing equal numbers of zeros and ones, and as the OOD proportion increases zeros are replaced by 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

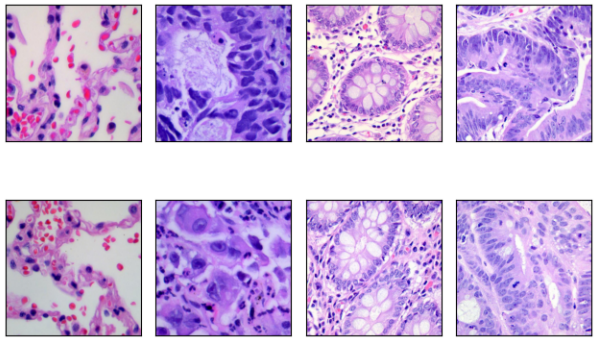


Figure 2: Example inputs from lung and colon histopathological image data set. Left to right: benign lung, cancerous lung, benign colon, cancerous colon. The models are trained on lung images (in-distribution), then shown varying quantities of colon images (out-of-distribution) at test time.

latent space and its Simplex approximations in Section 4.1. 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.

3.2 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 8000 images of 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 OOD images are introduced. 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 an OOD proportion of 1 corresponds to test data taken entirely from colon images, where classes are still balanced.

4 Distance measures

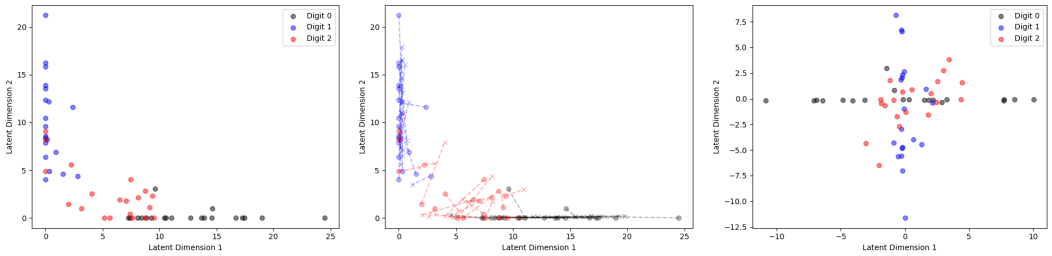
It is instructive to study the latent space of a simple model to gain an intuition of what this reveals of the model's behaviour. The CNN trained in Section 3.1 contains two neurons in its final hidden layer, so the positions of points can be visualised in 2D latent space. This is used to guide the design of distance measures in this section.

4.1 Interpreting the latent space

The positions of zeros and ones (IDD) and twos (OOD) in the 2-dimensional latent space of the CNN model are shown in Figure 3a. The model learns to separate the in-distribution samples; 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 a non-linear kernel to separate the classes in latent space, and the final softmax layer is a typical logistic regression [13].

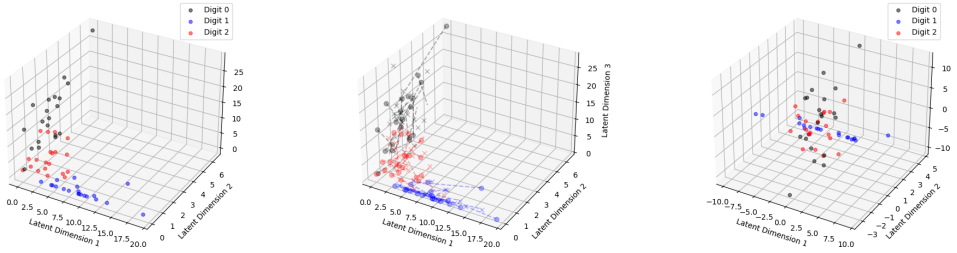
In contrast, the OOD data, i.e. the twos, lie in the middle of the latent space suggesting that the model is less able to distinguish these from the ones. Of note here is that they do generally lie closer to the cluster of zeros (the negative classification) than the cluster of ones (the positive classification). This is the correct classification for our model trained to identify ones, which implies that we may still get reasonable accuracy in practice despite having never seen similar data before, therefore relying on accuracy metrics alone could instil a misplaced sense of confidence in the model's appropriateness. The *confidence* in this model should be lower in this scenario even if the predictions are correct.

The effect of applying Simplex to this data is shown in Figure 3b. 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 nearest points in the corpus hull. In contrast, the out-of-distribution digits do not move in a consistent direction,



(a) (b) (c)

Figure 3: 2D latent space. Left to right: (a) sample input digits in the model's latent space; (b) residual shift of true points (circles) to their Simplex approximation (crosses); (c) residual vectors in latent space.



(a) (b) (c)

Figure 4: 3D latent space. Left to right: (a) sample input digits in the model's latent space; (b) residual shift of true points (circles) to their Simplex approximation (crosses); (c) residual vectors in latent space.

reflecting that the model is unable to reconstruct them consistently using the corpus, but the basis vectors provided by the corpus span the required latent space. This can be observed in Figure 3c which shows the residual vectors in latent space.

The same approach is repeated for the 3-dimensional case to verify that similar relations hold using a new CNN created with 3 neurons in its last hidden layer. Following the same approach, in Figure 4a the IDD points again form linearly separable clusters in 3D latent space, while the OOD data lies in between. In Figures 4b and 4c, it is again observed that the Simplex approximations bring points closer towards the centre of clusters. [1]

4.2 Distance measure formulations

A set of distance measures is proposed in this section, guided by the experimental insights above. These are then applied to the two data sets in Section 5 to establish their empirical 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 measure how 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 initially expect this to be larger for OOD data points, but as observed in Figures 3b and 3c, the IDD points move along the axis, whereas OOD points move more randomly and begin closer to

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

the centre of the latent space, so their residuals are actually smaller. This direction-specific movement is of particular interest, and forms the basis of distance measures introduced later in this section.

The next straightforward measure is to take the norm of the latent space. As observed in Figure 3a and 4a, the IDD points are 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 caused by OOD data.

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

A variation of this approach is introduced to address the fact 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 shrink 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, 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 for the true class 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 variance 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 KNN approach [4]. This allows for more interpretable scales—a value of 1 for h_{rel_norm} or $h_{rel\_norm\_classwise}$ 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 $h_{rel\_norm\_directionwise}$ the value should increase with OOD data as these are typically further out-of-plane.

$$h_{rel} = \frac{h_{test}}{\hat{h}_{validation}} \tag{6}$$

5 Results

The results of the proposed experiments are provided here, firstly for handwritten digit data in Section 5.1 and then to a real data set of histopathological images in Section 5.2.

5.1 MNIST classification experiment results

As a benchmark, common performance metrics accuracy and AUC are shown in Figure 5a, as well as the the mean of model probabilities which is commonly cited as a measure of confidence. It is observed that accuracy and AUC both decrease as the OOD proportion increases, which is expected, though their absolute values are still relatively high; a practitioner may well still choose such a model on the basis of these values, not knowing that their test data is so far removed from the model's training data. The model probabilities are high across all levels of OOD, corroborating the observations of Papernot and McDaniel in the context of Deep KNN that model probabilities are poor proxies of confidence, despite being widely used for this purpose [4]. The dispersion of these probabilities does show some relation to the OOD proportion, shown in Figure 5b.

The values of h_{norm} and $\hat{h}_{norm}$ are shown in Figure 5c, a residual-based measure r_{norm} in Figure 5d, and relative distance measures h_{rel_norm} and $h_{rel\_norm\_classwise}$ in Figure 5e. These all decrease as OOD points are added, in line with the understanding that OOD points lie in the middle of the space whereas IDD points are separated to the extremes. The natural scaling of h_{rel_norm} and $h_{rel\_norm\_classwise}$ is observed, as well as their similarity owing to the balanced classes of the experimental design. The behaviour in cases of imbalanced classes is a topic of further research. The direction-specific relative measure $h_{rel\_norm\_directionwise}$, which isolates the out-of-plane axis, is

shown in Figure 5f. This increases with OOD because IDD points typically lie in the plane whereas OOD points are more scattered.

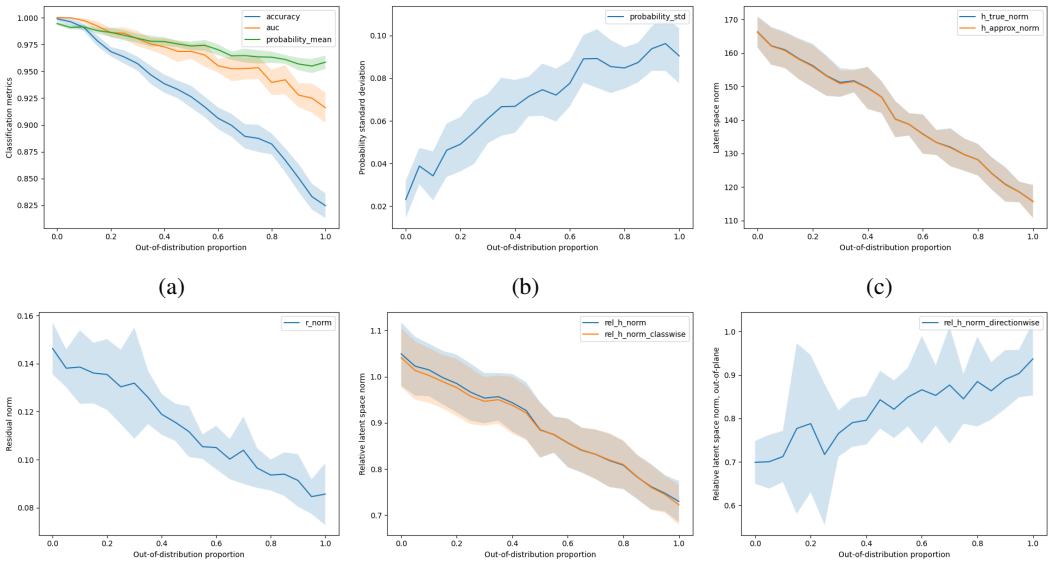


Figure 5: MNIST experiment results showing behaviour as OOD proportion increases: (a) Classification metrics; (b) Standard deviation of model probabilities; (c) Norm of true and approximate latent spaces; (d) Average norm of residual vectors; (e) Ratio of test and validation latent space norms; (f) Ratio of test and validation latent space norms in the out-of-plane axis.

5.2 Cancer classification experiment results

The results for a similar experiment on the tissue image data set shows largely consistent results, see Figure 6. This is encouraging, as it suggests that the conclusions drawn about Simplex-based confidence measures generalise to real-world data sets, and may therefore be useful to aid practitioners in understanding and applying models.

The only notable difference is the behaviour of the direction-specific latent space norm in Figure 6f. 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. This approach worked for the MNIST model with a 2D latent space. For higher dimensions, this may not be as appropriate if the IDD points lie in a plane that is not aligned with the latent space axes. 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 one or more axes to isolate. This is an area of further research.

6 Conclusions

The Simplex approach to the latent space was used to form several distance measures that can act as a measure of confidence in the model for the given application. An intuitive understanding of the latent space of a model was explored by examining simple models in 2D and 3D for in-distribution and out-of-distribution data. This latent space perspective helps to understand the behaviour of a model, which in turn guided the formulation of distance metrics which provide a global quantification of extrapolation.

These distance measures were applied to two binary classification tasks in an experimental setup that allowed us to vary the ground truth level of extrapolation. 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 insight can be used to supplement accuracy measurements to

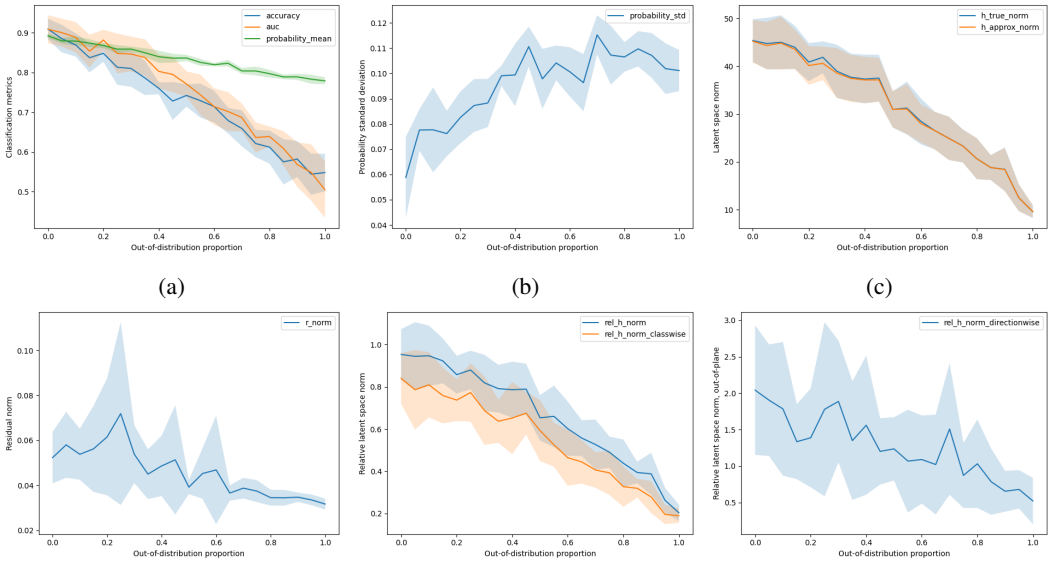


Figure 6: Histopathological cancer experiment results showing behaviour as OOD proportion increases: (a) Classification metrics; (b) Standard deviation of model probabilities; (c) Norm of true and approximate latent spaces; (d) Average norm of residual vectors; (e) Ratio of test and validation latent space norms; (f) Ratio of test and validation latent space norms in the out-of-plane axis.

guide practitioners when selecting a model and understanding when that model is operating beyond its intended use case.

Armed with this knowledge, the practitioner can 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 biases caused by underrepresented examples in the training data, both of which can exacerbate healthcare inequity. Thus, this approach could be a useful tool in the burgeoning field of data-centric AI [14].

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 formulations, tasks and domains, as well as linking with the related areas of model design, transfer learning and data-centric AI.

Variations and comparisons with other distance measures Alternative formulations of distance measures based on latent spaces and Simplex approximations could be explored, such as the covariance-based approach noted in Section 5.2. 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 sets other than that used to train the model could be investigated. This would allow the approach to be used in cases where the training data is not available, for example, where privacy concerns limit dissemination of the training data. 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 be benchmarked against other approaches, including Deep KNN, as well as more traditional measures of distance used in transfer learning [7], such as KL-divergence and Maximum Mean Discrepancy, both in the input space and latent space.

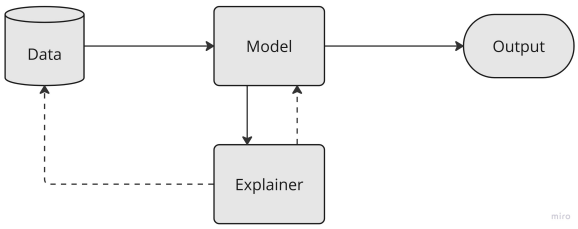


Figure 7: Using explainability methods to continuously improve models and data. The solid black arrows show the flow of data in current paradigms, where a model generates outputs from data with a post-hoc explainer. The dashed lines show how global explainability measures can facilitate a feedback loop where models and data can be improved based on the explainer.

Additional tasks and domains The experiments here only considered the task of 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 public 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 pertinent in healthcare settings.

Transfer learning Transfer learning is an increasingly important topic with the proliferation of pre-trained foundational models. The application of such models could provide powerful predictions on small data sets, provided the user can evaluate the confidence of the model. This could allow for more powerful individualised models, where a foundational model can be trained on the health records of a particular patient or group of patients. 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 Zhuang et al. is *interpretability* of transfer [7]. The approach explored in this paper is rooted in explainability methods and therefore naturally addresses this concern.

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.

Data-centric AI The approach outlined here can be used to identify data sets which are poorly described by existing models, i.e. where substantial extrapolation is required. These can be explored in-depth by explaining the poorly approximated test data in terms of features or examples, as in the original Simplex paper [3]. The feature-based approach can identify cases where features are not informative enough and may benefit from collecting additional fields. The example-based approach can identify cases where particular test instances are underrepresented in the test data, such as when specific ethnic groups are not adequately represented in the healthcare studies used for training data and would therefore be at risk of misdiagnosis by the model. This could also identify areas of the latent space where synthetic data could be utilised. Identifying these biases and having the tools to address them can play an important role in improving input data and addressing healthcare inequity.

Continuous improvement of model and data Taking the above two points of model design and data improvement 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 each in turn be addressed and revised. By linking explainability with model design and data collection, it can be framed as a key component in performance and ultimately healthcare outcomes. Not only does better explainability aid human understanding, it can be employed as a tactic in improving the models themselves and addressing biases in data, leading to improved outcomes.

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] Marie Guyomard, Susana Barbosa, and Lionel Fillatre. Kernel logistic regression approximation of an understandable ReLU neural network. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 12268–12291. PMLR, 23–29 Jul 2023.

[14] 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].