Explaining Models by Propagating Shapley Values

Hugh Chen
Paul G. Allen School of CSE
University of Washington
Seattle, WA, USA

Scott Lundberg
Microsoft Research
Redmond, WA, USA

Su-In Lee
Paul G. Allen School of CSE
University of Washington
Seattle, WA, USA

Abstract

In healthcare, making the best possible predictions with complex models (e.g., neural networks, ensembles/stacks of different models) can impact patient welfare. In order to make these complex models explainable, we present DeepSHAP for mixed model types, a framework for layer wise propagation of Shapley values that builds upon DeepLIFT (an existing approach for explaining neural networks). We show that in addition to being able to explain neural networks, this new framework naturally enables attributions for stacks of mixed models (e.g., neural network feature extractor into a tree model) as well as attributions of the loss. Finally, we theoretically justify a method for obtaining attributions with respect to a background distribution (under a Shapley value framework).

Introduction

Neural networks and ensembles of models are currently used across many domains. For these complex models, explanations accounting for how features relate to predictions is often desirable and at times mandatory (Goodman and Flaxman 2017). In medicine, explainable AI (XAI) is important for scientific discovery, transparency, and much more (Holzinger et al. 2017). One popular class of XAI methods is per-sample feature attributions (i.e., values for each feature for a given prediction).

In this paper, we focus on SHAP values (Lundberg and Lee 2017) – Shapley values (Shapley 1953) with a conditional expectation of the model prediction as the set function. Shapley values are the only additive feature attribution method that satisfies the desirable properties of local accuracy, missingness, and consistency. In order to approximate SHAP values for neural networks, we fix a problem in the original formulation of DeepSHAP (Lundberg and Lee 2017) where previously it used $E[x]$ as the reference and theoretically justify a new method to create explanations relative to background distributions. Furthermore, we extend it to explain stacks of mixed model types as well as loss functions rather than margin outputs.

Popular model agnostic explanation methods that also aim to obtain SHAP values are KernelSHAP (Lundberg and Lee 2017) and IME (Štrumbelj and Kononenko 2014). The downside of most model agnostic methods are that they are sampling based and consequently high variance or slow.

Alternatively, local feature attributions targeted to deep networks has been addressed in numerous works: Occlusion (Zeiler and Fergus 2014), Saliency Maps (Simonyan, Vedaldi, and Zisserman 2013), Layer-Wise Relevance Propagation (Bach et al. 2015), DeepLIFT, Integrated Gradients (IG) (Sundararajan, Taly, and Yan 2017), and Generalized Integrated Gradients (GIG) (Merrill et al. 2019).

Of these methods, the ones that have connections to the Shapley Values are IG and GIG. IG integrates gradients along a path between a baseline and the sample being explained. This explanation approaches the Aumann-Shapley value. GIG is a generalization of IG to explain losses and mixed model types – a feature DeepSHAP also aims to provide. IG and GIG have two downsides: 1.) integrating along a path can be expensive or imprecise and 2.) the Aumann-Shapley values fundamentally differ to the SHAP values we aim to approximate. Finally, DASP (Ancona, Öztireli, and Gross 2019) is an approach that approximates SHAP values for deep networks. This approach works by replacing point activations at all layers by probability distributions and requires many more model evaluations than DeepSHAP. Because DASP aims to obtain the same SHAP values as in DeepSHAP it is possible to use DASP as a part of the DeepSHAP framework.

Approach

Propagating SHAP values

DeepSHAP builds upon DeepLIFT; in this section we aim to better understand how DeepLIFT's rules connect to SHAP values. This has been briefly touched upon in (Shrikumar, Greenside, and Kundaje 2017) and (Lundberg and Lee 2017), but here we explicitly define the relationship.

DeepSHAP is a method that explains a sample (foreground sample), by setting features to be "missing". Missing features are set to corresponding values in a baseline sample (background sample). Note that DeepSHAP generally uses a background distribution, however focusing on a single background sample is sufficient because we can rewrite the SHAP values as an average over attributions with respect to a single background sample at a time (see next section for

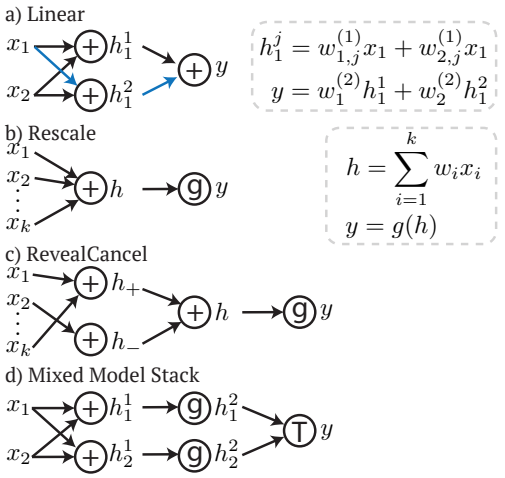

Figure 1: *Visualization of models for understanding DeepLIFT's connection to SHAP values*. In the figure g is a non-linear function and T is a non-differentiable tree model.

more details). In this section, we define a foreground sample to have features xf_{x_i} and neuron values f_h (obtained by a forward pass) and a background sample to have b_{x_i} or b_h. Finally we define $\phi(\cdot)$ to be attribution values.

If our model is **fully linear** as in Figure 1a, we can get the exact SHAP values for an input x_i by summing the attributions along all possible paths between that input x_i and the model's output y. Therefore, we can focus on a particular path (in blue). Furthermore, the path's contribution to $\phi(x_i)$ is exactly the product of the weights along the path and the difference in x_1: $w_2^{(2)} w_{1,2}^{(1)} (f_{x_1} - b_{x_1})$, because we can rewrite the layers of linear equations in 1a as a single linear equation. Note that we can derive the attribution for x_1 in terms of the attribution of intermediary nodes (as in the chain rule):

$$\phi(h_1^2) = w_2^{(2)} (f_{h_1^2} - b_{h_1^2})$$
$$\phi(x_1) = \frac{\phi(h_1^2)}{f_{h_1^2} - b_{h_1^2}} w_{1,2}^{(1)} (f_{x_1} - b_{x_1}) \quad (1)$$

Next, we move on to reinterpreting the two variants of DeepLIFT: the Rescale rule and the RevealCancel rule. First, a gradient based interpretation of the **Rescale rule** has been discussed in (Ancona et al. 2018). Here, we explicitly tie this interpretation to the SHAP values we hope to obtain.

For clarity, we consider the example in Figure 1b. First, the attribution value for $\phi(h)$ is $g(f_h) - g(b_h)$ because SHAP values maintain local accuracy (sum of attributions equals $f_y - b_y$) and g is a function with a single input. Then, under the Rescale rule, $\phi(x_i) = \frac{\phi(h)}{f_h - b_h} w_i (f_{x_i} - b_{x_i})$ (note the resemblance to Equation (1)). Under this formulation it is easy to see that the Rescale rule first computes the exact SHAP value for h and then propagates it back linearly. In other words, the the non-linear and linear functions are treated as separate functions. Passing back nonlinear attributions linearly is clearly an approximation, but confers two benefits: 1.) fast computation on order of a backward pass and 2.) a guarantee of local accuracy.

Next, we describe how the **RevealCancel rule** (originally formulated to bring DeepLIFT closer to SHAP values) connects to SHAP values in the context of Figure 1c. RevealCancel partitions x_i into positive and negative components based on if $w_i(f_{x_i} - b_{x_i}) < t$ (where t=0), in essence forming nodes h_+ and h_-. This rule computes the exact SHAP attributions for h_+ and h_- and then propagates the resultant SHAP values linearly. Specifically:

$$\phi(g_+) = \frac{1}{2}((g(f_{h_+} + f_{h_-}) - g(b_{h_+} + f_{h_-}) + (g(f_{h_+} + b_{h_-}) - g(b_{h_+} + b_{h_-}))$$
$$\phi(g_+) = \frac{1}{2}((g(f_{h_+} + f_{h_-}) - g(f_{h_+} + b_{h_-}) + (g(b_{h_+} + f_{h_-}) - g(b_{h_+} + b_{h_-}))$$
$$\phi(x_i) = \begin{cases} \frac{\phi_{h_+}}{f_{h_+} - b_{h_+}} w_i (f_{x_i} - b_{x_i}), & \text{if } w_i(f_{x_i} - b_{x_i}) > t \\ \frac{\phi_{h_-}}{f_{h_-} - b_{h_-}} w_i (f_{x_i} - b_{x_i}), & \text{otherwise} \end{cases}$$

Under this formulation, we can see that in contrast to the Rescale rule that explains a linearity and nonlinearity by exactly explaining the nonlinearity and backpropagating, the RevealCancel rule exactly explains the nonlinearity and a partition of the inputs to the linearity as a single function prior to backpropagating. The RevealCancel rule incurs a higher computational cost in order to get a an estimate of $\phi(x_i)$ that is ideally closer to the SHAP values.

This reframing naturally motivates explanations for **stacks of mixed model types**. In particular, for Figure 1d, we can take advantage of fast, exact methods for obtaining SHAP values for tree models to obtain $\phi(h_j^2)$ using Independent Tree SHAP (Lundberg et al. 2018). Then, we can propagate these attributions to get $\phi(x_i)$ using either the Rescale or RevealCancel rule. This argument extends to explaining losses rather than output margins as well.

Although we consider specific examples here, the linear propagation described above will generalize to arbitrary networks if SHAP values can be computed or approximated for individual components.

SHAP values with a background distribution

Note that many methods (Integrated Gradients, Occlusion) recommend the utilization of a single background/reference sample. In fact, DeepSHAP as previously described in (Lundberg and Lee 2017) created attributions with respect to a single reference equal to the expected value of the inputs. However, in order to obtain SHAP values for a given background distribution, we prove that the correct approach is as follows: obtain SHAP values for each baseline in your background distribution and average over the resultant attributions. Although similar methodologies have been used heuristically (Shrikumar, Greenside, and Kundaje 2017; Erion et al. 2019), we provide a theoretical justification in Theorem 1 in the context of SHAP values.

Theorem 1. *The average over single reference SHAP values approaches the true SHAP values for a given distribution.*

Proof. Define D to be the data distribution, N to be the set of all features, and f to be the model being explained. Ad-

ditionally, define $\mathcal{X}(x, x', S)$ to return a sample where the features in S are taken from x and the remaining features from x'. Define C to be all combinations of the set $N \setminus \{i\}$ and P to be all permutations of $N \setminus \{i\}$. Starting with the definition of SHAP values for a single feature: $\phi_i(x)$

$$= \sum_{S \in C} W(|S|, |N|)(\mathbb{E}_D[f(X)|x_{S\cup\{i\}}] - \mathbb{E}_D[f(X)|x_S])$$

$$= \frac{1}{|P|} \sum_{S \subseteq P} \mathbb{E}_{\mathcal{D}}[f(x)|\text{do}(x_{S\cup\{i\}})] - \mathbb{E}_{\mathcal{D}}[\text{do}(f(x)|x_S)]$$

$$= \frac{1}{|P|} \sum_{S \subseteq P} \frac{1}{|D|} \sum_{x' \in D} f(\mathcal{X}(x, x', S \cup \{i\})) - f(\mathcal{X}(x, x', S))$$

$$= \frac{1}{|D|} \sum_{x' \in D} \underbrace{\frac{1}{|P|} \sum_{S \subseteq P} f(\mathcal{X}(x, x', S \cup \{i\})) - f(\mathcal{X}(x, x', S))}_{\text{single reference SHAP value}}$$

where the second step depends on an interventional conditional expectation (Janzing, Minorics, and Blöbaum 2019) which is very close to Random Baseline Shapley in (Sundararajan and Najmi 2019)). □

Experiments

Background distributions avoid bias

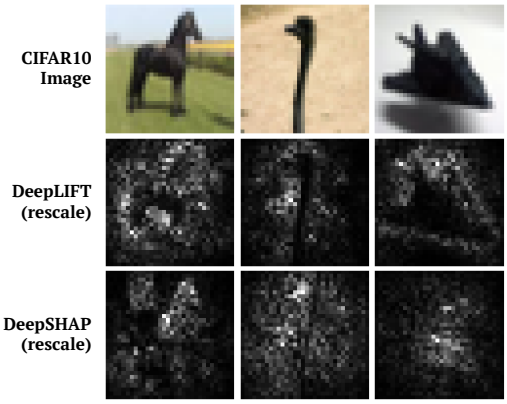

Figure 2: *Using a single baseline leads to bias in explanations.*

In this section, we utilize the popular CIFAR10 dataset (Krizhevsky, Hinton, and others 2009) to demonstrate that single references lead to bias in explanations. We train a CNN that achieves 75.56% test accuracy and evaluate it using either a zero baseline as in DeepLIFT or with a random set of 1000 baselines as in DeepSHAP.

In Figure 2, we can see that for these images drawn from the CIFAR10 training set, DeepLIFT has a clear bias that results in low attributions for darker regions of the image. For DeepSHAP, having multiple references drawn from a background distribution solves this problem and we see attributions in sensical dark regions in the image.

Explaining mortality prediction

In this section, we validate DeepSHAP's explanations for an MLP with 82.56% test accuracy predicting 15 year mortality. The dataset has 79 features for 14,407 individuals released by (Lundberg et al. 2018) based on NHANES I Epidemiologic Followup Study (Cox et al. 1997).

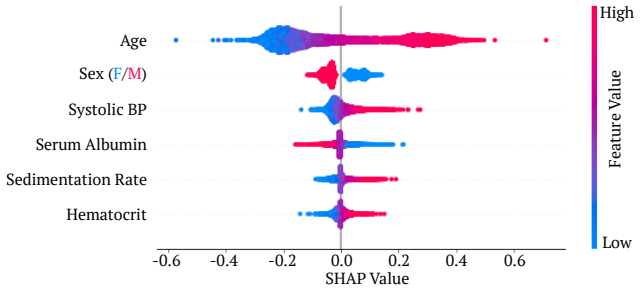

Figure 3: *Summary plot of DeepSHAP attribution values.* Each point is the local feature attribution value, colored by feature value. For brevity, we only show the top 6 features.

In Figure 3, we plot a summary of DeepSHAP (with 1000 random background samples) attributions for all NHANES training samples (n=8023) and notice a few trends. First, Age is predictably the most important and old age contributes to a positive mortality prediction (positive SHAP values). Second, the Sex feature validates a well-known difference in mortality (Gjonça et al. 1999). Finally, the trends linking high systolic BP, low serum albumin, high sedimentation rate, and high hematocrit to mortality have been independently discovered (Port et al. 2000; Goldwasser and Feldman 1997; Paul et al. 2012; Go et al. 2016).

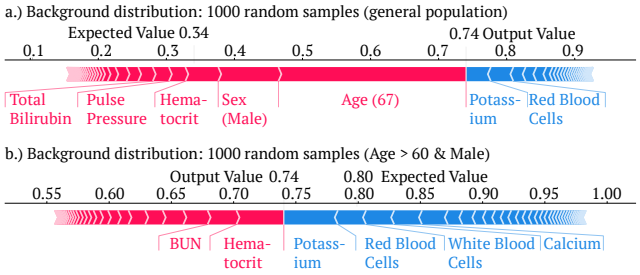

Figure 4: *Explaining an individual's mortality prediction for different backgrounds distributions.*

Next, we show the benefits of being able to specify a background distribution. In Figure 4a, we see that explaining an individual's mortality prediction with respect to a general population emphasizes that the individual's age and gender are driving a high mortality prediction. However, in practice doctors are unlikely to compare a 67-year old male to a general population that includes much younger individuals. In Figure 4b, being able to specify a background distribution allows us to compare our individual against a more relevant distribution of males over 60. In this case, gender and age are naturally no longer important, and the individual actually may not have cause for concern.

Interpreting a stack of mixed model types

Stacks, and more generally ensembles, of models are increasingly popular for performant predictions (Bao,

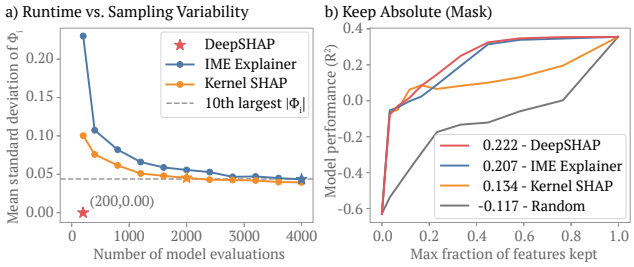

Figure 5: *Ablation test for explaining an LSTM feature extractor fed into an XGB model.* All methods used background of 20 samples obtained via kmeans. [a.] Convergence of methods for a single explanation. [b.] Model performance versus # features kept for DeepSHAP (rescale), IME Explainer (4000 samples), KernelSHAP (2000 samples) and a baseline (Random) (AUC in the legend).

Bergman, and Thompson 2009; Güneş, Wolfinger, and Tan 2017; Zhai and Chen 2018). In this section, our aim is to evaluate the efficacy of DeepSHAP for a neural network feature extractor fed into a tree model. For this experiment, we use the Rescale rule for simplicity and Independent TreeSHAP to explain the tree model (Lundberg et al. 2018). The dataset is a simulated one called Corrgroups60. Features $X \in \mathbb{R}^{1000\times 60}$ have tight correlation between groups of features (x_i is feature i), where $\rho_{x_i,x_i}=1$, $\rho_{x_i,x_{i+1}}=\rho_{x_i,x_{i+2}}=\rho_{x_{i+1},x_{i+2}}=.99$ if $(i \bmod 3)=0$, and $\rho_{x_i,x_j}=0$ otherwise. The label $y \in \mathbb{R}^n$ is generated linearly as $y=X\beta+\epsilon$ where $\epsilon \sim \mathcal{N}_n(\mu=0, \sigma^2=10^{-4})$ and $\beta_i=1$ if $(i \bmod 3)=0$ and $\beta_i=0$ otherwise.

We evaluate DeepSHAP with an ablation metric called *keep absolute (mask)* (Lundberg et al. 2018). The metric works in the following manner: 1) Obtain the feature attributions for all test samples 2) Mask all features (by mean imputation) 3) Introduce one feature at a time (unmask) from largest absolute attribution value to smallest for each sample and measure R^2. The R^2 should initially increase rapidly, because we introduce the "most important" features first.

We compare against two sampling-based methods (a natural alternative for explaining mixed model stacks) that provide SHAP values in expectation: KernelSHAP and IME explainer. In Figure 5b, DeepSHAP (rescale) has no variability and requires a fixed number of model evaluations. IME Explainer and KernelSHAP, benefit from having more samples (and therefore more model evaluations). For the final comparison, we check the variability of the tenth largest attribution (absolute value) of the sampling based methods to determine "convergence" across different numbers of samples. Then, we use the number of samples at the point of "convergence" for the next figure.

In Figure 5c, we can see that DeepSHAP has a slightly higher performance than model agnostic methods. Promisingly, all methods demonstrate initial steepness in their performance; this indicates that the most important features had higher attribution values. We hypothesize that KernelSHAP and IME Explainer's lower performance is due in part to noise in their estimates. This highlights an important point: model agnostic methods often have sampling variability that makes determining convergence difficult. For a fixed background distribution, DeepSHAP does not suffer from this variability and generally requires fewer model evaluations.

Improving the RevealCancel rule

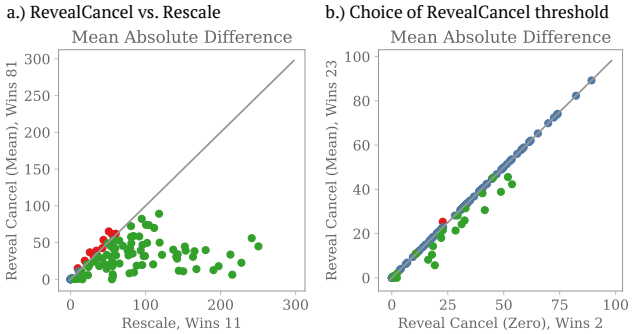

Figure 6: *Comparison of new RevealCancel^{Mean} rule for estimating SHAP values on a toy example.* The axes correspond to mean absolute difference from the SHAP values (computed exactly). Green means RevealCancel$^{\text{Mean}}$ wins and red means it loses.

DeepLIFT's RevealCancel rule's connection to the SHAP values is touched upon in (Shrikumar, Greenside, and Kundaje 2017). Our SHAP value framework explicitly defines this connection. In this section, we propose a simple improvement to the RevealCancel rule. In DeepLIFT's RevealCancel rule the threshold t is set to 0 (for splitting h_- and h_+). Our proposed rule RevealCancel$^{\text{Mean}}$ sets the threshold to the mean value of $w_i(f_{x_i}-b_{x_i})$ across i. Intuitively, splitting by the mean better separates x_i nodes, resulting in a better approximation than splitting by zero.

We experimentally validate RevealCancel$^{\text{Mean}}$ in Figure 6, explaining a simple function: $\text{ReLU}(x_1+x_2+x_3+x_4+100)$. We fix the background to zero: $b_{x_i}=0$ and draw 100 foreground samples from a discrete uniform distribution: $f_{x_i} \sim U\{-1000, 1000\}$.

In Figure 6a, we show that RevealCancel$^{\text{Mean}}$ offers a large improvement for approximating SHAP values over the Rescale rule and a modest one over the original RevealCancel rule (at no additional asymptotic computational cost).

Conclusion

In this paper, we improve the original DeepSHAP formulation (Lundberg and Lee 2017) in several ways: we 1.) provide a new theoretically justified way to provide attributions with a background distribution 2.) extend DeepSHAP to explain stacks of mixed model types 3.) present improvements of the RevealCancel rule.

Future work includes more quantitative validation on different data sets and comparison to more interpretability methods. In addition, we primarily used Rescale rule for many of these evaluations, but more empirical evaluations of RevealCancel are also important.

References

[Ancona et al. 2018] Ancona, M.; Ceolini, E.; Oztireli, C.; and Gross, M. 2018. Towards better understanding of gradient-based attribution methods for deep neural networks. In *6th International Conference on Learning Representations (ICLR 2018)*.

[Ancona, Öztireli, and Gross 2019] Ancona, M.; Öztireli, C.; and Gross, M. 2019. Explaining deep neural networks with a polynomial time algorithm for shapley values approximation. *arXiv preprint arXiv:1903.10992*.

[Bach et al. 2015] Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 10(7):e0130140.

[Bao, Bergman, and Thompson 2009] Bao, X.; Bergman, L.; and Thompson, R. 2009. Stacking recommendation engines with additional meta-features. In *Proceedings of the third ACM conference on Recommender systems*, 109–116. ACM.

[Cox et al. 1997] Cox, C. S.; Feldman, J. J.; Golden, C. D.; Lane, M. A.; Madans, J. H.; Mussolino, M. E.; and Rothwell, S. T. 1997. Plan and operation of the nhanes i epidemiologic followup study, 1992. *Vital and Health Statistics*.

[Erion et al. 2019] Erion, G.; Janizek, J. D.; Sturmfels, P.; Lundberg, S.; and Lee, S.-I. 2019. Learning explainable models using attribution priors. *arXiv preprint arXiv:1906.10670*.

[Gjonça et al. 1999] Gjonça, A.; Tomassini, C.; Vaupel, J. W.; et al. 1999. *Male-female differences in mortality in the developed world*. Citeseer.

[Go et al. 2016] Go, D. J.; Lee, E. Y.; Lee, E. B.; Song, Y. W.; Konig, M. F.; and Park, J. K. 2016. Elevated erythrocyte sedimentation rate is predictive of interstitial lung disease and mortality in dermatomyositis: a korean retrospective cohort study. *Journal of Korean medical science* 31(3):389–396.

[Goldwasser and Feldman 1997] Goldwasser, P., and Feldman, J. 1997. Association of serum albumin and mortality risk. *Journal of clinical epidemiology* 50(6):693–703.

[Goodman and Flaxman 2017] Goodman, B., and Flaxman, S. 2017. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine* 38(3):50–57.

[Güneş, Wolfinger, and Tan 2017] Güneş, F.; Wolfinger, R.; and Tan, P.-Y. 2017. Stacked ensemble models for improved prediction accuracy. In *SAS Conference Proceedings*.

[Holzinger et al. 2017] Holzinger, A.; Biemann, C.; Pattichis, C. S.; and Kell, D. B. 2017. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*.

[Janzing, Minorics, and Blöbaum 2019] Janzing, D.; Minorics, L.; and Blöbaum, P. 2019. Feature relevance quantification in explainable ai: A causality problem. *arXiv preprint arXiv:1910.13413*.

[Krizhevsky, Hinton, and others 2009] Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. Technical report, Citeseer.

[Lundberg and Lee 2017] Lundberg, S. M., and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 4765–4774.

[Lundberg et al. 2018] Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; and Lee, S. 2018. Explainable ai for trees: From local explanations to global understanding. *CoRR* abs/1905.04610.

[Merrill et al. 2019] Merrill, J.; Ward, G.; Kamkar, S.; Budzik, J.; and Merrill, D. 2019. Generalized integrated gradients: A practical method for explaining diverse ensembles. *CoRR* abs/1909.01869.

[Paul et al. 2012] Paul, L.; Jeemon, P.; Hewitt, J.; McCallum, L.; Higgins, P.; Walters, M.; McClure, J.; Dawson, J.; Meredith, P.; Jones, G. C.; et al. 2012. Hematocrit predicts long-term mortality in a nonlinear and sex-specific manner in hypertensive adults. *Hypertension* 60(3):631–638.

[Port et al. 2000] Port, S.; Demer, L.; Jennrich, R.; Walter, D.; and Garfinkel, A. 2000. Systolic blood pressure and mortality. *The Lancet* 355(9199):175–180.

[Shapley 1953] Shapley, L. S. 1953. A value for n-person games. *Contributions to the Theory of Games* 2(28):307–317.

[Shrikumar, Greenside, and Kundaje 2017] Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 3145–3153. JMLR. org.

[Simonyan, Vedaldi, and Zisserman 2013] Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

[Štrumbelj and Kononenko 2014] Štrumbelj, E., and Kononenko, I. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems* 41(3):647–665.

[Sundararajan and Najmi 2019] Sundararajan, M., and Najmi, A. 2019. The many shapley values for model explanation. *arXiv preprint arXiv:1908.08474*.

[Sundararajan, Taly, and Yan 2017] Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*.

[Zeiler and Fergus 2014] Zeiler, M. D., and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, 818–833. Springer.

[Zhai and Chen 2018] Zhai, B., and Chen, J. 2018. Development of a stacked ensemble model for forecasting and analyzing daily average pm 2.5 concentrations in beijing, china. *Science of the Total Environment* 635:644–658.