AID: Active Distillation Machine to Leverage Pre-Trained Black-Box Models in Private Data Settings

ABSTRACT

This paper presents an active distillation method for a local institution (e.g., hospital) to find the best queries within its given budget to distill an on-server black-box model's predictive knowledge into a local surrogate with transparent parameterization. This allows local institutions to understand better the predictive reasoning of the black-box model in its own local context or to further customize the distilled knowledge with its private dataset that cannot be centralized and fed into the server model. The proposed method thus addresses several challenges of deploying machine learning (ML) in many industrial settings (e.g., healthcare analytics) with strong proprietary constraints. These include: (1) the opaqueness of the server model's architecture which prevents local users from understanding its predictive reasoning in their local data contexts; (2) the increasing cost and risk of uploading local data on the cloud for analysis; and (3) the need to customize the server model with private onsite data. We evaluated the proposed method on both benchmark and real-world healthcare data where significant improvements over existing local distillation methods were observed. A theoretical analysis of the proposed method is also presented.

KEYWORDS

disease risk prediction, deep learning, model distillation

ACM Reference Format:

. 2021. AID: Active Distillation Machine to Leverage Pre-Trained Black-Box Models in Private Data Settings. In *Proceedings of WWW '21: The Web Conference (WWW '21).* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Leveraging and/or customizing pre-trained black-box models for transparent predictive analysis in local data context has become increasingly important for local institutions (e.g., hospitals) which often do not have sufficient data for training accurate predictive models. In application domains such as healthcare, private models trained using massive proprietary data would be released as a black-box service on the cloud due to the restrictions on sharing proprietary information (e.g., the disease model and patient data used in model training) [22]. For example, Google Health [1], Azure for Healthcare Cloud [2] and IQVIA Human Data Science Cloud [3] provide such machine learning services. Local institutions/users can subscribe to such service to analyze their local data. However, to use such services, local institutions still need to upload their private patient data to the service provider's cloud which increases both the risk of leaking sensitive information and the cost of protecting them from cyber attacks.

Furthermore, another concern of using such pre-trained service is that there is often no transparency regarding its prediction which prevents local institutions from understanding and inspecting whether its reasoning mechanism has accounted for potential biases (e.g., age, ethnicity and demographic regions) in their local data. Depending on such assessment, further customization of the pre-trained model might be necessary to gear it better towards the local context. Unfortunately, this is not possible since both the model and local data are proprietary and cannot be put together.

To sidestep these challenges, efforts have been made in both *model distillation* and *knowledge distillation*. Model distillation methods including [3–6, 13, 15, 16, 19–21, 23] aim to construct simple models with human-understandable features to explain the prediction of a sophisticated model for each data point locally. On the other hand, knowledge distillation [10] and/or mimic learning [1] methods that aims to distil the entire model [12, 14, 17, 24]. We will provide a more detailed discussion in Section. 2. Here we summarize the limitations they face in reliably leveraging private black-box models.

(1) **Inaccessible Black-Box Model:** For model distillation, some algorithms [6, 20, 21] require access to the architecture of the black-box model, which are not realistic in many real-world applications including healthcare. For knowledge distillation, they also need data access which is infeasible in our setting.
(2) **High Query Cost:** Existing models especially [2, 13, 15, 16] require scoring many local samples (e.g., patient) with the black-box model first in order to distill a local model for a local sample, which is computationally intensive and incurs extra cost to protect the privacy.
(3) **Lack of Multi-level Rationalization:** The above works can rationalize the black box's prediction at each data point well but do not aim to distill the black box in its entirety (e.g., to distill the associated risk factors for the disease across the entire patient population), despite that it is desirable especially for model customization. For example, when clinicians inspect that the model does not fit on a certain data demography (which might be under-represented in the provider's training data), they might want to add their extra domain information (e.g., patient-specific risks) into the existing model to improve its performance, which is currently prohibited.

[1] https://cloud.google.com/healthcare

[2] https://azure.microsoft.com/en-us/industries/healthcare/

[3] https://www.iqvia.com/solutions/human-data-science-cloud

WWW '21, April 19-23, 2021, Ljubljana, Slovenia

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/1122445.1122456

In this paper, we develop an ActIve Distillation Machine (AID) framework to accurately and efficiently distill a black-box model's predictive knowledge into a local surrogate with multi-level rationalization. This is enabled by the following technical contributions.

(1) **Active Sample Selection Strategy for Low Cost Learning**. Instead of sending all local data (e.g., sensitive patient records) to the black box service, AID has a sequential selection strategy that identifies the most informative data points as queries to the black-box model (Section 3.2.2), which reduces the query cost and the risk of leaking sensitive information.
(2) **Theoretically Guaranteed Distillation without Accessing Black-box Models**. AID distills a black-box model's predictive knowledge into a local surrogate with transparent parameterization(Section 3.2.1). We developed a formal theoretical analysis to derive guarantees for the distillation quality of AID, which is defined as the probability that the surrogate's prediction on a random data point agrees with that of the black-box model (Section 4).
(3) **Multi-level Knowledge Distillation**. The surrogate model of AID provides instance-wise relative importance of input features as local rationales; and a set of universal rules as global rationales that identifies relevant features for each target disease across the entire patient population (Section 3.2.1). This allows the proposed surrogate model to provide both local rationales which are patient-specific, and global rationales about a disease population which generalizes to unseen patients accurately.

We evaluate AID on several datasets including one real-world EHR dataset to demonstrate its effectiveness and efficiency (Section 5). The reported results show that AID achieves up to 13.80% average performance improvement over the best baseline and interpretability metrics, and up to 4.75× speed-up over the fastest baseline.

2 RELATED WORKS

Model distillation can be categorized into attention-based (white-box) methods and black-box methods. First, attention-based methods often exploit the attention weights [5, 6, 23] to distil a local model for each subject (e.g., patient). This, however, requires access to the server model's architecture, which is not possible in most healthcare setting where such information is both proprietary and vulnerable against cyber attacks [8]. Then, on the other hand, there are also black-box methods including [2, 13, 15, 16] which were proposed to distill DL models without accessing their architecture. However, these methods often map given inputs to human-understandable vectors in local space, where each vector comprises a set of interpretable features engineered by the domain experts (e.g., patches in images, or text phrases). Then, for an input data point, an additive linear model is defined on the human-interpretable space with the objective to match the prediction from the black-box model on the same data point. The resulting additive linear model can then be used as an explanation.

More recently, Chen et al. revisited model interpretation as an instance-wise feature selection framework named L2X, that learns a common function to generate local interpretation of the model's prediction at any data point. However, these methods still provide only local distillation per data point and do not aim to distil the black-box model in its entirety. Furthermore, they have also overlooked the concerns of potential large query cost.

Knowledge distillation [10] or mimic learning [1] aim to transfer the predictive power from a high-capacity but expensive DL model to a simpler model such as shallow neural networks for ease of deployment [12, 14, 17, 24]. However, assume that both models operate on the same domain and have access to the same data or at least similar datasets. This is however infeasible in our setting.

Active learning refers to a series of models that have active learner to ask queries in the form of unlabeled instances to be labeled by an oracle (e.g., a human annotator) [18]. However, standard active learning often have (hard) class labels as feedback, which cannot encode multi-level rationalization. In this work, we take a different approach to use (soft) class distribution as the feedback, which contains multi-level rationalization distilled from the black-box model, and is efficiently encoded into a local surrogate.

3 METHOD

3.1 Definitions and Notations

Definition 1 (Health Risk Prediction Task). Given a dataset $\mathbf{D} \triangleq \{(\mathbf{x}_t, \mathbf{c}_t)\}_{t=1}^{k}$ where $(\mathbf{x}_t, \mathbf{c}_t)$ denote the medical record of patient t, the diagnosing task is to compute a disease prediction vector $\mathbf{c} \triangleq [c^{(1)} \dots c^{(n)}]$ for an unseen patient record $\mathbf{x} \triangleq [x^{(1)} \dots x^{(m)}]$ where the binary feature $x^{(i)}$ indicates whether the patient is observed with medical code i [4]. This is achieved by learning the probabilities $\mathbb{P}(c^{(i)} = 1|\mathbf{x})$ and $\mathbb{P}(c^{(i)} = 0|\mathbf{x})$ that $\mathbf{x}$ has and does not have this disease. All patients are indexed by the same set of ICD-9 codes.

Definition 2 (Black-Box Model). A predictive model $\mathbb{P}(\mathbf{c}|\mathbf{x}) \triangleq \prod_{i=1}^{n} \mathbb{P}(c^{(i)}|\mathbf{x})$ takes patient data $\mathbf{x}$ as input and outputs a set of probability scores $\mathbf{o} \triangleq \{\mathbb{P}(c^{(i)}|\mathbf{x}) \mid c^{(i)} \in \{0, 1\}\}_{i=1}^{n}$. These scores indicate the likelihood that the patient has a particular disease $i \in \{1 \dots n\}$ but does not explain why.

Definition 3 (Global Rationale). For disease i, a global rationale is a vector $\mathbf{b}_i \triangleq [b_i^{(1)} \dots b_i^{(m)}]$ that indicates whether a medical code ι is relevant to disease i: relevant if $b_i^{(\iota)} = 1$. Otherwise, $b_i^{(\iota)} = 0$. To avoid trivial explanations, e.g., $b_i^{(\iota)} = 1 \; \forall \; (i, \iota)$, we restrict $\|\mathbf{b}_i\|_0 \leq \ell$ where ℓ is the maximum size of an explanation, as defined by a domain practitioner.

Definition 4 (Local Rationale). A local rationale for a patient record $(\mathbf{x}, \mathbf{c})$ is a vector function $\mathbf{e}(\mathbf{x}) \triangleq [e^{(1)}(\mathbf{x}) \dots e^{(m)}(\mathbf{x})]$ where $e^{(i)}(\mathbf{x}) \in \mathbb{R}$ indicates how strong the influence of feature i is on the model's prediction $\mathbf{c}$ of all target diseases for patient $\mathbf{x}$. Intuitively, it would assign large weights to medical codes that have strong influence on at least one target disease i of patient $\mathbf{x}$.

[4] The inputs are introduced as binary vectors only to be consistent with the binary data derived from MIMIC-III dataset. Our AID framework below only needs access to probabilistic outputs of the black box (regardless of the format of its input). Both AID and its theoretical analysis do not require the inputs to be binary. Our synthetic experiments in Section 4 and the appendices were performed on non-binary data.

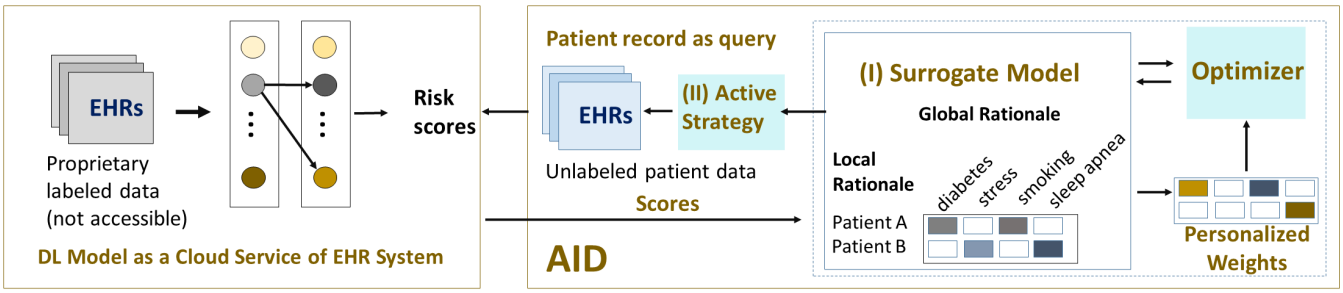

Figure 1: The AID frameworks helps local institution to leverage DL model trained on large proprietary health data without the need for accessing both the proprietary data and model, as well as generate multi-level distilled rationales. This is enabled by the following modules. Given a black-box model trained on private EHR data, AID (I) learns a surrogate model that incorporates global and local interpretable representations; and (II) takes an active strategy to sequentially query the black-box model for labeled data to distill its knowledge into these representations. The queried feedback is then used to update the representation parameters and the search focus of the active component.

A local rationale is therefore patient-specific and in the below specification, it is further parameterized via an explanation function $\mathbf{e}_{\mathbf{w}}(\mathbf{x}) \triangleq [e_{\mathbf{w}}^{(1)}(\mathbf{x}) \dots e_{\mathbf{w}}^{(m)}(\mathbf{x})]$ characterizes the importance weights $e_{\mathbf{w}}^{(i)}(\mathbf{x})$ of each feature $x^{(i)}$ and is parameterized by $\mathbf{w}$. For example, $e_{\mathbf{w}}^{(i)}(\mathbf{x}) = \exp(p^{(i)})/\sum_{t=1}^{m}\exp(p^{(t)})$ where the logit vector $\mathbf{p} = [p^{(1)} \dots p^{(m)}] \triangleq \mathbf{w}\mathbf{x}$ where $\mathbf{w} \in \mathbb{R}^{m\times m}$.

Finally, the dot product $\langle \mathbf{b}_i \circ \mathbf{e}_{\mathbf{w}}(\mathbf{x}) \circ \mathbf{x}, \boldsymbol{\alpha}\rangle$ combines the effect of all components to produce the probability of $c^{(i)} = 1$. In this formulation, the weight vector $\boldsymbol{\alpha}$ is learnable to combine the weighted features into the logit signal appropriately.

3.2 The AID Framework

The AID frameworks helps local institution to leverage a DL model trained on large proprietary health data without the need for accessing both the proprietary data and model, as well as generate multi-level rationales. This is enabled by the following modules. Given a model trained on private EHR data, AID (I) learns a surrogate model that incorporates global and local distilled representations; and (II) takes an active strategy to sequentially query the black-box model for labeled data to distill its knowledge into these representations. The queried feedback is used to update the representation parameters and the search focus of the active component.

3.2.1 **Module I. Surrogate Model Parameterization.** The surrogate mimics the behavior of the black box and provides global and local rationales, which is formally defined below:

$$\mathbb{P}_{\boldsymbol{\alpha}}\Big(\mathbf{c}|\mathbf{B}, \mathbf{e}_{\mathbf{w}}(\mathbf{x})\Big) \triangleq \prod_{i=1}^{n} \mathbb{P}_{\boldsymbol{\alpha}}\Big(c^{(i)}|\mathbf{b}_i, \mathbf{e}_{\mathbf{w}}(\mathbf{x})\Big) \tag{1}$$

where $\mathbf{B} \triangleq [\mathbf{b}_1 \dots \mathbf{b}_n]$ denotes the collection of global explanations (one explanation vector per disease for n target diseases) and $\boldsymbol{\alpha}$ denote the parameter of the surrogate as detailed next. In particular, the factor surrogates for individual diseases are parameterized as

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\alpha}}\Big(c^{(i)} = 1|\mathbf{b}_i, \mathbf{e}_{\mathbf{w}}(\mathbf{x})\Big) &\triangleq \Phi(r^{(i)}) \quad \text{and}\\ \mathbb{P}_{\boldsymbol{\alpha}}\Big(c^{(i)} = 0|\mathbf{b}_i, \mathbf{e}_{\mathbf{w}}(\mathbf{x})\Big) &\triangleq 1 - \Phi(r^{(i)}),\end{aligned} \tag{2}$$

with $r^{(i)} \triangleq h_{\boldsymbol{\alpha}}(\mathbf{b}_i \circ \mathbf{e}_{\mathbf{w}}(\mathbf{x}) \circ \mathbf{x})$ and $\Phi(r^{(i)}) \triangleq 1/(1+\exp(-r^{(i)}))$ denote the response function $r^{(i)}$ and the logistic function over $r^{(i)}$, respectively. We define $h_{\boldsymbol{\alpha}}(\mathbf{p}) \triangleq \langle \boldsymbol{\alpha}, \mathbf{p}\rangle$ as the dot product between two vectors and $\circ$ as the element-wise multiplication. Thus, $\boldsymbol{\alpha}$ parameterizes the response function.

Intuitively, the surrogate incorporates a neural network $\mathbf{e}_{\mathbf{w}}(\mathbf{x})$ (with softmax activator) parameterized by $\mathbf{w}$, which are learned to extract relevant instance-wise features for each data point, which provides in extra a local probabilistic explanation on the influence of $\mathbf{x}$'s individual features on its prediction. Its output is used to weigh the relevance of the features in $\mathbf{x}$. A point-wise multiplication with $\mathbf{b}_i$ helps remove irrelevant features to increase its interpretability. The results are fed into a logistic regression (LR) model parameterized by $\boldsymbol{\alpha}$ to predict whether the patient has a disease. This is chosen due to the self-interpretability of LR.

Distilling Black-Box Knowledge. Given the surrogate and black-box models, the distillation task can be framed as fitting the surrogate to match the black box via below[5]:

$$\begin{aligned}\underset{\mathbf{B},\mathbf{w},\boldsymbol{\alpha}}{\text{minimize}} \quad & \sum_{t=1}^{k}\sum_{i=1}^{n} \mathbf{D}_{\mathrm{KL}}\Big(\mathbb{P}_{\boldsymbol{\alpha}}(c_t^{(i)}|\mathbf{b}_i, \mathbf{e}_{\mathbf{w}}(\mathbf{x}_t))\|\mathbb{P}(c_t^{(i)}|\mathbf{x}_t)\Big)\\ \text{subject to} \quad & \|\mathbf{b}_i\|_0 \le \ell\,.\end{aligned} \tag{3}$$

where $\mathbf{b}_i \in \{0,1\}^m$, $\mathbf{x}_t$ represents the input features of patient $t = 1 \dots k$ and $c_t^{(i)}$ denotes a binary random variable that indicates if the patient is exposed to disease i.

In particular, Eq. (3) characterizes the divergence between the two distributions $\mathbb{P}_{\boldsymbol{\alpha}}(\mathbf{c}_t|\mathbf{B}, \mathbf{e}_{\mathbf{w}}(\mathbf{x}_t))$ and $\mathbb{P}(\mathbf{c}_t|\mathbf{x}_t)$ and is derived from their factorization in Section 3.1 and Eq. (1). This is with respect to a set of feedback $\{\mathbf{x}_t, \mathbf{o}_t\}_{t=1}^{k}$:

$$\forall(t,i):\ \mathbf{o}_t \triangleq \Big\{\mathbb{P}\Big(c_t^{(i)}|\mathbf{x}\Big) \mid c^{(i)} \in \{0,1\}\Big\}_{i=1}^{n}$$

[5]Optimizing this KL divergence in either forward or backward direction inspire the same practical approximation in Eq.(4) below.

where $\{\mathbf{x}_t\}_{t=1}^k$ are drawn i.i.d. Then, let Φ^{-1} denotes the inverse logistic function[6]. If we can find $\boldsymbol{\alpha}^*$, $\mathbf{w}^*$ and $\mathbf{b}_i^* \in \{0,1\}^m$ so that:

$$\forall(t,i):\ r_t^{(i)} \ = \ \Phi^{-1}\Big(\mathbb{P}(c_t^{(i)}=1|\mathbf{x}_t)\Big)$$

and hence, $\mathbb{P}_{\boldsymbol{\alpha}}\Big(c_t^{(i)}=1|\mathbf{b}_i,\mathbf{e}_{\mathbf{w}}(\mathbf{x}_t)\Big) \ = \ \mathbb{P}\Big(c_t^{(i)}=1|\mathbf{x}_t\Big)$.

Then, the above KL divergence in Eq. (3) becomes zero and this makes $(\boldsymbol{\alpha}^*, \mathbf{B}^*, \mathbf{w}^*)$ with $\mathbf{B}^* \triangleq \{\mathbf{b}_i^*\}_{i=1}^n$ its optimizer. In practice, however, the optimal tuple $(\boldsymbol{\alpha}^*, \mathbf{B}^*, \mathbf{w}^*)$ that makes the KL divergence become zero might not exist.

Practical Optimization. It is thus more practical to search for $(\boldsymbol{\alpha}^*, \mathbf{B}^*, \mathbf{w}^*)$ that instead minimizes the discrepancy between $q_t^{(i)} \triangleq \Phi^{-1}(\mathbb{P}(c_t^{(i)}=1|\mathbf{x}_t))$ and $r_t^{(i)} \triangleq h_{\boldsymbol{\alpha}}(\mathbf{e}_{\mathbf{w}}(\mathbf{x}_t) \circ \mathbf{b}_i \circ \mathbf{x}_t)$, which is a lower-bound (up to a constant) of the objective in Eq. (3) (via Pinkser inequality). That is, we want to minimize $\mathcal{G}(\mathbf{B},\mathbf{w},\boldsymbol{\alpha})$ with respect to $(\mathbf{B},\mathbf{w},\boldsymbol{\alpha})$ subject to $\|\mathbf{b}_i\|_0 \le \ell$ where

$$\mathcal{G}(\mathbf{B},\mathbf{w},\boldsymbol{\alpha}) \ \triangleq \ \sum_{t=1}^{k}\sum_{i=1}^{n}\Big(q_t^{(i)} - h_{\boldsymbol{\alpha}}(\mathbf{b}_i \circ \mathbf{e}_{\mathbf{w}}(\mathbf{x}_t) \circ \mathbf{x}_t)\Big)^2 \qquad (4)$$

Eq. (4) can then be optimized numerically via alternating minimization where we alternate between fixing $\mathbf{B}$ and optimizing $(\boldsymbol{\alpha},\mathbf{w})$ and vice versa. In particular, given $\mathbf{B}$, Eq. (4) is continuous in $(\boldsymbol{\alpha},\mathbf{w})$ and can be optimized via gradient descent.

On the other hand, $\mathbf{B} = [\mathbf{b}_1 \dots \mathbf{b}_n]$ needs to be sparse binary to be interpretable. Otherwise, optimizing Eq. (4) could result in a dense continuous vector which is not interpretable. To avoid this, we optimized an unconstrained, continuous proxy $\mathbf{z}_i \in [0,1]^m$ of $\mathbf{b}_i$, which can be minimized via gradient descent. The optimized unconstrained proxy $\mathbf{z}_i$ can also be transformed back into a binary vector $\mathbf{b}_i$ via top-ℓ binarization.

3.2.2 **Module II. Active Knowledge Distillation.** This section develops an active learning algorithm that finds the best queries to distill knowledge in the black-box model into an interpretable surrogate. In active setting, this is constrained by a budget since each query would incur (economical or computational) cost and possibly leak of sensitive data. To achieve this, our intuition is patient queries that yield the most information gain would often fall in the *disagreement region* between the current surrogate and black-box models. Formally, for a target disease i, we define the *disagreement region* $\mathbf{H}_k^{(i)}$ of the surrogate model as

$$\mathbf{H}_k^{(i)} \ \triangleq \ \Big\{\mathbf{x} \ \Big| \ \mathbf{D}_{\mathrm{KL}}\Big(\mathbb{P}_{\boldsymbol{\alpha}^*}^k\Big(c^{(i)}|\mathbf{x}\Big) \,\|\, \mathbb{P}\Big(c^{(i)}|\mathbf{x}\Big)\Big) > \zeta\Big\}\ , \qquad (5)$$

where we use $\mathbb{P}_{\boldsymbol{\alpha}^*}^k(c^{(i)}|\mathbf{x})$ as a short-hand notation for the surrogate model $\mathbb{P}_{\boldsymbol{\alpha}^*}(c^{(i)}|\mathbf{b}_i^*,\mathbf{e}_{\mathbf{w}^*}(\mathbf{x}))$ which is built using the first k queried data points $\{(\mathbf{x}_t,\mathbf{o}_t)\}_{t=1}^k$ where $\mathbf{o}_t$ is the (soft[7]) class distribution feedback (see Section 3.2.1). This region includes patient data that induces divergence above a threshold ζ between the surrogate and black-box models, which will likely result in their prediction disagreement. Note that ζ can be set algorithmically in Section 4.2.

An efficient learning strategy therefore samples only data inside the disagreement region, which requires an accurate identification of such a region. Deciding whether a datum belongs to the disagreement region however requires asking the black box's prediction at that datum, which raises a dilemma because we only want to query patient datum that belongs to the disagreement region. To circumvent this dilemma, we represent this region as a latent function $g_k^{(i)}: \mathbf{X} \to \{0,1\}$ that maps from patient data to a binary outcome such that $g_k^i(\mathbf{x}) = 1$ implies $\mathbf{x} \in \mathbf{H}_k^{(i)}$. Otherwise, $g_k^i(\mathbf{x}) = 0$. Learning $g_k^i(\mathbf{x})$ can then be made possible by leveraging the previously queried data $\{(\mathbf{x}_t,\mathbf{o}_t)\}_{t=1}^{k-1}$ as training examples since for each such datum $(\mathbf{x}_t,\mathbf{o}_t)$, if

$$\mathbf{D}_{\mathrm{KL}}\Big(\mathbb{P}_{\boldsymbol{\alpha}_*}^k\Big(c_t^{(i)}|\mathbf{x}_t\Big) \,\|\, \mathbb{P}\Big(c_t^{(i)}|\mathbf{x}_t\Big)\Big) \ > \ \zeta\,, \qquad (6)$$

we know that $\mathbf{x}_t \in \mathbf{H}_k^{(i)}$. Otherwise, $\mathbf{x}_t \notin \mathbf{H}_k^{(i)}$. Using these induced examples, one can use any of the existing off-the-shelf classifier to learn $g_k^{(i)}(\mathbf{x})$. Note that (a) entropy-based and/or expected model-change methods that sample patient based on criteria derived from the surrogate's predictive uncertainty are not applicable since patients who cause high divergence (see Eq. (6)) between the surrogate's and black box's cannot be identified accurately using surrogate's uncertainty alone; and (b) the disagreement at each stage k changes due to new observations and we need to learn a new classifier for each k.

4 THEORETICAL ANALYSIS

This section starts with a simple analysis of passive interpretation with i.i.d. feedback (Section 4.1). The developed results are then extended to active settings with non i.i.d. feedback (Section 4.2).

4.1 Passive Distillation Analysis

Minimizing the Kullback-Leibler (KL) divergence between black-box model $\mathbb{P}(\mathbf{c}|\mathbf{x})$ and its distilled surrogate $\mathbb{P}_{\boldsymbol{\alpha}}(\mathbf{c}|\mathbf{B},\mathbf{e}_{\mathbf{w}}(\mathbf{x}))$ can be achieved by solving a surrogate optimization objective with respect to a set of sampled data $\mathbf{U} = \{\mathbf{x}_t,\mathbf{o}_t\}_{t=1}^k$. Note that $\mathbf{U}$ is different from the proprietary dataset $\mathbf{D}$ (Section 3.1) with hard labels, which was used to train the black-box model.

This raises two questions: (**Q1**) how well does the resulting surrogate model mimic the black-box model of the training data; and (**Q2**) interestingly, how good is the surrogate model on unseen data? To address these questions, we first put forward the following definitions and assumptions:

Definition 1. Let $\mathcal{P}(\mathbf{x})$ denote the population distribution of unlabeled data $\mathbf{x}$. The interpretation quality $\mathbf{I}(\mathbf{B},\mathbf{w},\boldsymbol{\alpha})$ of $\mathbb{P}_{\boldsymbol{\alpha}}(\mathbf{c}|\mathbf{B},\mathbf{e}_{\mathbf{w}}(\mathbf{x}))$ is formally defined as:

$$\mathbf{I}(\mathbf{B},\mathbf{w},\boldsymbol{\alpha}) \ \triangleq \ \mathbb{E}_{\mathcal{P}}\Big[\mathbf{D}_{\mathrm{KL}}\left(\mathbb{P}_{\boldsymbol{\alpha}}(\mathbf{c}|\mathbf{B},\mathbf{e}_{\mathbf{w}}(\mathbf{x}))\|\mathbb{P}(\mathbf{c}|\mathbf{x})\right)\Big] \qquad (7)$$

[6] $\Phi^{-1}(a) = \log(a/(1-a))$ for $a \in (0,1)$.

[7] Soft feedback contains more information than the (hard) class label feedback in standard active learning.

where the expectation is over $\mathbf{x} \sim \mathcal{P}(\mathbf{x})$. This characterizes the average distributional divergence between the black box and surrogate. Using the factorization form of the black-box and surrogate models in Section 3.1 and Eq. (1) and the linearity of expectation, we have $\mathbf{I}(\mathbf{B}, \mathbf{w}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \mathbf{I}(\mathbf{b}_i, \mathbf{w}, \boldsymbol{\alpha})$ where

$$\mathbf{I}(\mathbf{b}_i, \mathbf{w}, \boldsymbol{\alpha}) \triangleq \mathbb{E}_{\mathcal{P}}\left[\mathbf{D}_{\mathrm{KL}}\left(\mathbb{P}_{\boldsymbol{\alpha}}(c^{(i)}|\mathbf{b}_i, \mathbf{e}_{\mathbf{w}}(\mathbf{x}) \| \mathbb{P}(c^{(i)}|\mathbf{x})\right)\right].$$

Then, let $\mathbf{I}_t(\mathbf{b}_i, \mathbf{w}, \boldsymbol{\alpha}) \triangleq \mathbf{D}_{\mathrm{KL}}(\mathbb{P}^{(i)}_{\boldsymbol{\alpha}}(\mathbf{x}_t) \| \mathbb{P}^{(i)}(\mathbf{x}_t))$ denote the distillation quality at $\mathbf{x}_t \sim \mathcal{P}(\mathbf{x})$ where $\mathbb{P}^{(i)}_{\boldsymbol{\alpha}}(\mathbf{x}_t)$ and $\mathbb{P}^{(i)}(\mathbf{x}_t)$ are short-hands for $\mathbb{P}_{\boldsymbol{\alpha}}(c_t^{(i)}|\mathbf{b}_i, \mathbf{e}_{\mathbf{w}}(\mathbf{x}_t))$ and $\mathbb{P}(c_t^{(i)}|\mathbf{x}_t)$. It follows that

$$\mathbf{I}(\mathbf{b}_i, \mathbf{w}, \boldsymbol{\alpha}) = \frac{1}{k}\mathbb{E}_{\mathcal{P}}\left[\sum_{t=1}^{k} \mathbf{I}_t(\mathbf{b}_i, \mathbf{w}, \boldsymbol{\alpha})\right] \quad (8)$$

To address (**Q1**) above, we bound $\mathbf{I}_t(\mathbf{b}_i, \mathbf{w}, \boldsymbol{\alpha})$ in terms of the following quantities that characterize the noise of the black box and the solution quality of the proposed algorithm (see Lemma 1 below).

Definition 2. Let the prediction uncertainty of the black-box model be denoted by

$$\nu \triangleq \min_{i=1}^{n} \min_{t=1}^{k} \min\left(\mathbb{P}\left(c_t^{(i)} = 0|\mathbf{x}_t\right), \mathbb{P}\left(c_t^{(i)} = 1|\mathbf{x}_t\right)\right).$$

That is, its prediction confidence is always less than $1 - \nu$ (see Appendix D for more discussion).

Definition 3. Let $\epsilon \triangleq \mathcal{G}(\mathbf{B}^*, \mathbf{w}^*, \boldsymbol{\alpha}^*)$ denotes the quality of fitting the surrogate to the black box in Section 3.2.1 where $(\mathbf{B}^*, \mathbf{w}^*, \boldsymbol{\alpha}^*)$ is the solution found by optimizing Eq. (4).

Lemma 1. Let $\phi \triangleq \epsilon_t^{(i)}$ denote the individual error yielded by $(\mathbf{B}^*, \mathbf{w}^*, \boldsymbol{\alpha}^*)$ at $\mathbf{x}_t$ and $c^{(i)}$:

$$\epsilon_t^{(i)} \triangleq \left(q_t^{(i)} - h_{\boldsymbol{\alpha}^*}(\mathbf{b}_i^* \circ \mathbf{e}_{\mathbf{w}^*}(\mathbf{x}_t) \circ \mathbf{x}_t)\right)^2 \quad (9)$$

where $q_t^{(i)} \triangleq \Phi^{-1}(\mathbb{P}(c_t^{(i)} = 1|\mathbf{x}_t))$. Then, Appendix B shows that $\mathbf{I}_t\left(\mathbf{b}_i^*, \mathbf{w}^*, \boldsymbol{\alpha}^*\right) \leq \phi^2/(8\nu(1 - \phi(1 - \nu))^2)$ with ν in Definition 2. Using the result of Lemma 1, we are now ready to address **Q2** via Theorems 1 and 2 below:

Theorem 1. Let $\boldsymbol{\psi}(\epsilon, \nu) \triangleq \epsilon^2/(8\nu(1-\epsilon(1-\nu))^2)$ with ϵ in Definition 3 and $\delta \in (0, 1)$. Then, it can be shown that (Appendix C) with probability at least $1 - \delta$,

$$\mathbf{I}\left(\mathbf{b}_i^*, \mathbf{w}^*, \boldsymbol{\alpha}^*\right) \leq \boldsymbol{\psi}(\epsilon, \nu)\left(1 + \left(\frac{1}{2k}\log\left(\frac{2}{\delta}\right)\right)^{\frac{1}{2}}\right). \quad (10)$$

Using Theorem 1, we can now bound the chance that the surrogate misinterprets the black box over unseen samples, which addresses **Q2**. Let $\mathbf{E}$ be the event that $\mathbb{P}_{\boldsymbol{\alpha}^*}(c^{(i)}|\mathbf{b}_i^*, \mathbf{e}_{\mathbf{w}^*}(\mathbf{x}))$ disagrees with $\mathbb{P}_i(c^{(i)}|\mathbf{x}))$ on the most probable label (i.e., the prediction) for a random input $\mathbf{x} \sim \mathcal{P}(\mathbf{x})$. A stronger version of Theorem 1 that bounds the chance that $\mathbf{E}$ happens is derived below.

Definition 4. Let $\gamma \triangleq \min_{\mathbf{x},i} |\mathbb{P}(0|\mathbf{x}) - \mathbb{P}(1|\mathbf{x})|$ denote the prediction robustness of the black-box model. That is, if we subtract and add $\gamma' \leq 0.5\gamma$ from $\mathbb{P}(0|\mathbf{x})$ to $\mathbb{P}(1|\mathbf{x})$, or vice versa, the label with the highest probability score will not change, and as such, the prediction will not change.

Theorem 2. Let $k = \frac{1}{2\gamma^4}\log\frac{2}{\delta}$ with $0 < \delta < 1$ where k denotes the size of the sampled data and γ is defined in Definition 4. Appendix E shows that with probability $1 - \delta$,

$$\mathcal{P}(\mathbf{E}) \leq \boldsymbol{\psi}(\epsilon, \nu)\left(\frac{1}{\gamma^2} + 1\right), \quad (11)$$

with ν and ϵ defined previously in Definitions 2 and 3.

Discussion. In our analysis (see Theorems 1 and 2), ν can be treated as a hardness constant that characterizes how difficult an distillation task is. If the optimization error ϵ is sufficiently small such that $\epsilon \leq \nu$, then $\boldsymbol{\psi}(\epsilon, \nu) \leq \nu\epsilon/(8\nu(1 - \epsilon(1 - \nu))^2) = \epsilon/8(1 - \epsilon + \epsilon\nu)^2 \leq \epsilon/(8(\epsilon^2 - \epsilon + 1)^2) \leq \epsilon/(8((\epsilon - 0.5)^2 + 0.75)^2) \leq \epsilon/(8 \times 0.75^2) = 2\epsilon/9 \leq \epsilon/3$. This implies $\mathcal{P}(\mathbf{E}) \leq \epsilon/3(1 + 1/\gamma^2)$, which means the disagreement rate between the surrogate and black box no longer depends on ν, thus *overcoming* it. On the other hand, $\epsilon > \nu$ might happen when the black-box is overfitted, which causes ν to be very small, as discussed in Appendix E.

Note that, for the rest of this paper, we assume that the optimization error is always smaller than ν, thus asserting the above simplified bound on $\mathcal{P}(\mathbf{E})$.

4.2 Active Distillation Analysis

The results of Section 4.1 do not apply directly to the active scenario here since the sequentially selected data points are no longer independent, which is a key assumption of Theorems 1 and 2. To begin, recall from Eq. (5) that $\mathbf{H}_q^{(i)}$ denote the region of inputs that induce high divergence between the surrogate (fitted using the first q queried data points) and black box. It can be shown that when choosing the divergence threshold $\zeta = \gamma^2$ with γ defined in Definition 4, inputs in this high-divergence region will also induce prediction disagreement between the surrogate and black-box model (see Appendix F). This is a key ingredient to establish the result in Theorem 3 below, which shows that by focusing on drawing only data from the disagreement region, one can achieve the same disagreement probability of the passive method in Section 4.1 with fewer samples.

To do this, we assume that at iteration $q + 1$, we can draw samples from $\mathbf{H}_q^{(i)}$, which are distributed by $\mathcal{Q}(\mathbf{x}) \triangleq \mathcal{P}(\mathbf{x})/\mathcal{P}(\mathbf{x} \in \mathbf{H}_q^{(i)})$ and $\mathbf{H}_{q+1}^{(i)} \subseteq \mathbf{H}_q^{(i)}$ for all q.

Our key objective here is to show that there exists q^* for which $\mathcal{P}(\mathbf{x} \in \mathbf{H}_{q^*}^{(i)}) \leq (\epsilon/3)(1/\gamma^2 + 1)$ and $q^* < k = \frac{1}{2\gamma^4}\log\frac{2}{\delta}$ where k is the sample complexity for the passive algorithm to achieve the above misclassification rate (Theorem 2). This is achieved via Theorem 3 below.

Theorem 3. For any target condition $c^{(i)}$, let $q^* = \frac{r}{2\gamma^2}\log\frac{2r}{\delta}$ with

$$r = \log\left(\frac{\epsilon}{3\gamma^2}\left(1+\gamma^2\right)\right) \Big/ \log\left(\frac{\epsilon}{3\gamma^2}\left(1+\gamma\right)\right)$$
$$= \log\left(\frac{\epsilon}{3}\left(1+\frac{1}{\gamma^2}\right)\right) \Big/ \log\left(\frac{\epsilon}{3\gamma^2}\left(1+\gamma\right)\right).$$

Then, Appendix F shows that with probability at least $1-\delta$, $\mathcal{P}(\mathbf{E}^{(i)}_{q^*}) \leq (\epsilon/3)(1+1/\gamma^2)$. This achieves the same mis-distillation rate of passive interpretation but requires fewer samples (i.e., $q^* < k = \frac{1}{2\gamma^4}\log\frac{2}{\delta}$ – see Theorem 2), thus demonstrating active distillation's theoretical advantage.

5 EXPERIMENTS

5.1 Experimental Setting

Synthetic Datasets. These are the synthetic benchmark datasets used in the recent work of [4] including: (1) the **XOR** dataset which is a collection of data tuples $(\mathbf{x}, \mathbf{c})$ where $\mathbf{c} = 1$ if $1/(1+\exp(\mathbf{x}_1\mathbf{x}_2)) > 0.5$ and $\mathbf{c} = 0$ otherwise; (2) **Nonlinear Additive** dataset which has $\mathbf{c} = 1$ if $1/1 + (-100\sin 2\mathbf{x}_1 + 2|\mathbf{x}_2| + \mathbf{x}_3 + \exp(-\mathbf{x}_4)) > 0.5$ and $\mathbf{c} = 0$ otherwise; (3) **Fusion Feature** dataset, which is also a dataset comprising of multiple $(\mathbf{x}, \mathbf{c})$ tuples for which $\mathbf{c} = [c_1\ c_2]$ is a 2-dimensional multi-hot encoding where c_1, c_2 are generated using $\mathbf{x}_1, \mathbf{x}_2$ from **XOR** and $\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6$ from **Nonlinear Additive**, respectively. In each synthetic dataset, 10,000 input $\mathbf{x}$ are randomly generated from a 10-dimensional standard Gaussian.

MIMIC-III Dataset. The MIMIC-III dataset [11] comprises of 46,433 in-hospital medical health records of different patients. Each health record is a sequence of medical events. Each event is encoded as a multi-hot, binary vector indexed by a set of $m = 1000$ most frequent medical codes (e.g., drugs, lab tests etc.) in healthcare. The input vector indicates which medical codes appear in the patient's record. The corresponding output is a 11-dimensional binary vector that indicates the mortality status (dead or alive) (DD) of the patient as well as whether he/she has any of the following 10 diseases, which include essential hypertension (EH), congestive heart failure (HF), atrial fibrillation (AF), coronary atherosclerosis of native coronary artery (CA), acute kidney failure (KF), diabetes (DI), hyperlipidemia (HA), acute respiratory failure (RF), urinary tract infection (UR), esophageal reflux (ER).

Baseline Methods We compare AID against the following baselines: L2X [4], SHAP [13], LIME [15] and ANCHOR [16].

Metrics. We follow [9] to use the metrics below.

- **Fidelity:** The distillation fidelity measure $\mathbb{F} = \kappa^{-1}\sum_{t=1}^{\kappa}\mathbb{I}(\mathbf{A}(\mathbf{x}_t) = \mathbf{B}(\mathbf{x}_t))$ of a method $\mathbf{A}$ is the percentage of test on which its explanation model agrees with the prediction of the black box $\mathbf{B}$.
- **Accuracy:** this is measured by the areas under the precision-recall (PR-AUC) and receiver operating characteristic curves (ROC-AUC) based on the method's true and false positive rate.
- **Efficiency:** this is measured by the method's incurred time to predict and explain all test samples.
- **Generalizability** $\mathbb{G}(\mathbf{e}) \triangleq |\mathbb{N}(\mathbf{x})|^{-1}\sum_{\mathbf{x}'}\mathbb{I}(\mathbf{e}(\mathbf{x}') = c'(\mathbf{x}'))$ of a local explanation $\mathbf{e}$ (generated to explain $\mathbf{x}$) is the percentage of data point $\mathbf{x}'$ (labeled with $c'(\mathbf{x}')$) in $\mathbf{x}$'s pre-defined neighborhood $\mathbb{N}(\mathbf{x})$ for which $\mathbf{e}$ can be applied on $\mathbf{x}'$ to generate feature weights that help the surrogate model predicts its label $c'(\mathbf{x}')$ correctly (i.e., $\mathbb{I}(\mathbf{e}(\mathbf{x}'), c'(\mathbf{x}')) = 1$).

Implementation Details. For synthetic experiment, we first build a target black-box model as a multi-layer perceptron (MLP) with 2 dense layers (with rectified linear unit (ReLU) activation) trained on the entire training data with 100 hidden units each. We also add dropout and L_2 regularization on these dense layers to avoid over-fitting. The interpretable models are then learned by randomly sampling 500 samples from the training set and sending those to the black-box for queried feedback.

For experiments with the MIMIC dataset, we build a black-box model using a similar but larger neural network architecture, which has 256 and 64 hidden units on its dense layers respectively. Again, we let each interpretation method samples a subset of 2500 samples from the training set and send those to the black-box for feedback. Our code is publicly available at https://github.com/weare-anonymous/AID.

Evaluation Strategy. For each dataset, we partition it into training (80%), validation (10%) and test (10%) sets. For each experiment, a state-of-the-art black-box model is constructed using the training data. Each distillation method then samples a small fraction of training data to interpret the black-box. To evaluate their interpretation qualities, we use the corresponding distillation model[8] to make predictions on an unseen test set.

5.2 Results

Exp 1. AID is more accurate and faster on benchmark and real world data.

We first compare AID with existing model distillation methods based on a budgeted data setting. Here, we set data budget of the tested methods to be 500 samples. The distillation task thus becomes much harder due to the limited amount of data. Results on all benchmark datasets and the real world MIMIC-III data are shown in Table 1. We also show the ROC curves for all methods on benchmark datasets in Fig. 2 and on MIMIC-III dataset on Fig. 5.

In particular, Table 1 shows that AID achieves significantly better performance than all baselines. For example, on Fusion Feature dataset, AID achieves 22.70%, 11.40% and 7.29% improvement in fidelity, accuracy (PR-AUC) and generalizability (hence, 13.80% improvement on average), respectively, over the best baseline. This is expected since AID has a global understanding of each prediction target, which allows it to decide if a local rationale can be applied correctly to each test instance.

In contrast, existing baselines use a common instance-specific rationale function to explain all test points (L2X) or simply match a test point with a nearest local rationale (LIME, SHAP and ANCHOR) and risk incurring inaccurate predictions. On real world EHR data (MIMIC-III dataset), the reported results also show that

[8] Surrogate or local model that explains the closest data point.

Table 1: Performance Comparison

	Fidelity	ROC-AUC	PR-AUC	Generalizability	Time (sec)
Performance (mean ± std) on **Fusion Feature**					
L2X	0.531 ± 0.078	0.523 ± 0.060	0.521 ± 0.060	0.521 ± 0.062	000.328
LIME	0.495 ± 0.009	0.500 ± 0.002	0.500 ± 0.002	0.478 ± 0.038	536.396
SHAP	0.636 ± 0.077	0.584 ± 0.053	0.582 ± 0.053	0.617 ± 0.020	658.090
ANCHOR	0.727 ± 0.016	0.731 ± 0.022	0.728 ± 0.014	0.550 ± 0.026	202.010
AID	**0.892 ± 0.030**	**0.809 ± 0.033**	**0.811 ± 0.033**	**0.662 ± 0.054**	**000.069**
Black Box	-	0.998	0.998	-	-
Performance (mean ± std) on **XOR**					
	Fidelity	ROC-AUC	PR-AUC	Generalizability	Time (sec)
L2X	0.521 ± 0.036	0.537 ± 0.039	0.517 ± 0.026	0.508 ± 0.020	000.134
LIME	0.509 ± 0.047	0.508 ± 0.051	0.501 ± 0.023	0.501 ± 0.007	179.364
SHAP	0.601 ± 0.082	0.599 ± 0.063	0.577 ± 0.063	0.525 ± 0.011	268.835
ANCHOR	0.517 ± 0.032	0.515 ± 0.033	0.517 ± 0.032	0.500 ± 0.023	190.087
AID	**0.658 ± 0.141**	**0.614 ± 0.095**	**0.616 ± 0.095**	**0.532 ± 0.032**	**000.050**
Black Box	-	0.956	0.960	-	-
Performance (mean ± std) on **Nonlinear Additive**					
	Fidelity	ROC-AUC	PR-AUC	Generalizability	Time (sec)
L2X	0.627 ± 0.153	0.596 ± 0.116	0.594 ± 0.119	0.587 ± 0.118	000.263
LIME	0.487 ± 0.024	0.495 ± 0.021	0.497 ± 0.018	0.477 ± 0.039	179.976
SHAP	0.707 ± 0.100	0.644 ± 0.090	0.639 ± 0.085	0.634 ± 0.018	397.069
ANCHOR	0.706 ± 0.022	0.717 ± 0.023	0.707 ± 0.023	0.570 ± 0.038	052.956
AID	**0.851 ± 0.109**	**0.732 ± 0.151**	**0.730 ± 0.150**	**0.702 ± 0.171**	**000.062**
Black Box	-	0.997	0.997	-	-
Performance (mean ± std) on **MIMIC-III**					
	Fidelity	ROC-AUC	PR-AUC	Generalizability	Time (sec)
L2X	0.926 ± 0.007	0.601 ± 0.006	0.260 ± 0.005	0.801 ± 0.006	0000.387
LIME	0.554 ± 0.003	0.515 ± 0.008	0.203 ± 0.003	0.789 ± 0.008	3600.000
SHAP	N.A	N.A	N.A	N.A	N.A
ANCHOR	N.A	N.A	N.A	N.A	N.A
AID	**0.929 ± 0.014**	**0.647 ± 0.016**	**0.297 ± 0.016**	**0.804 ± 0.001**	**0000.213**
Black Box	-	0.793	0.499	-	-

AID performs significantly better than baseline models in general, and mildly worse than the black box that was trained using all data.

AID also runs significantly faster than the baselines as shown in the reported efficiency in Table 1 and Fig. 3. Among all baselines, both SHAP and ANCHOR are time-consuming strategies that generate a new local model for each new data point. In contrast, AID constructs a surrogate model once and can rationalize test data without generating new models for them, and is more efficient (e.g., 4.75× faster than the fastest baseline on the Fusion Feature dataset).

Exp 2. The advantage of active distillation in budgeted data setting

To demonstrate the advantage of active distillation, we compare AID with (a) its passive variant AID- across multiple datasets; and (b) its oracle variant AID+ that is also passive but allowed to query the entire dataset. In particular, AID- has the same budget as AID but selects queries randomly instead of strategically like AID. On the other hand, AID+ is allowed to query the entire dataset (instead of up to 5000 data points like AID) and thus, achieves performance very close to that of the black box. That is, AID+ represents the upper-bound that AID seeks to approach via active distillation when there is a budget on the amount of data that can be queried. The results of these comparisons are plotted in Fig. 4 and Fig. 5, respectively.

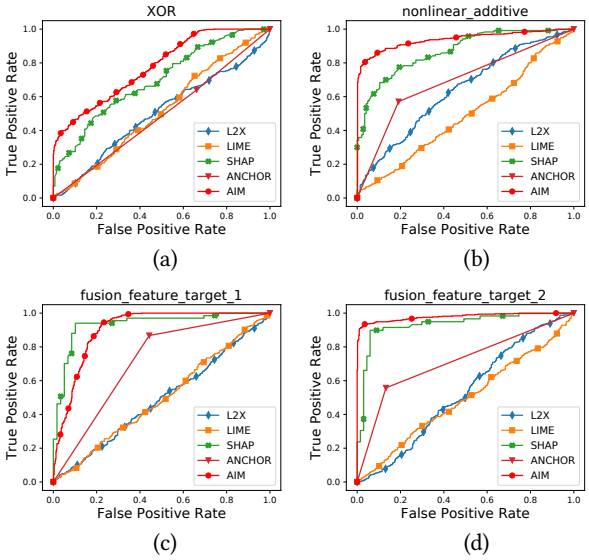

Figure 2: Receiver Operating Characteristic (ROC) performance curves of tested methods on (a) XOR, (b) Nonlinear Additive and (c,d) Fusion Feature datasets.

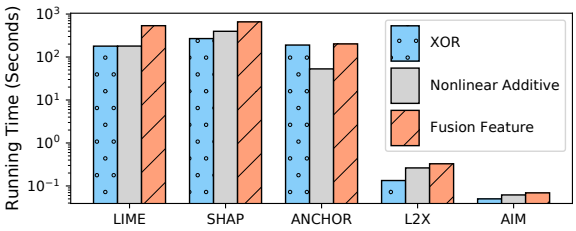

Figure 3: The processing time incurred by AID and baselines to explain 1000 samples in the test sets of XOR, Nonlinear Additive and Fusion Feature datasets.

First, from Fig. 4, we can see that AID performs better and much more stable than AID- (on a wide variety of disease prediction tasks) thanks to its ability to select data actively and strategically to maximize the information gain. In contrast, AID- selects data randomly and exhibits less stable performance. This corroborates our observations in the previous experiment and further demonstrates the effectiveness of active distillation in data-limited settings.

Second, to demonstrate how effective active distillation (AID) is in reducing the performance gap between its passive variant (AID-) and the oracle upper-bound (AID+), we plot their ROC-AUC performance curves against the amount of data AID and AID- are allowed to query (up to 5000 points) in Fig. 5 below. The results show that AID performs worse than AID+ as expected since its data budget is much less than that of AID+. It can, however, be observed in the same plot that its performance improves radically as we increase its data budget, which showcases the practical efficiency of AID in Section 3.2.2. In contrast, AID- has the same budget as AID. In

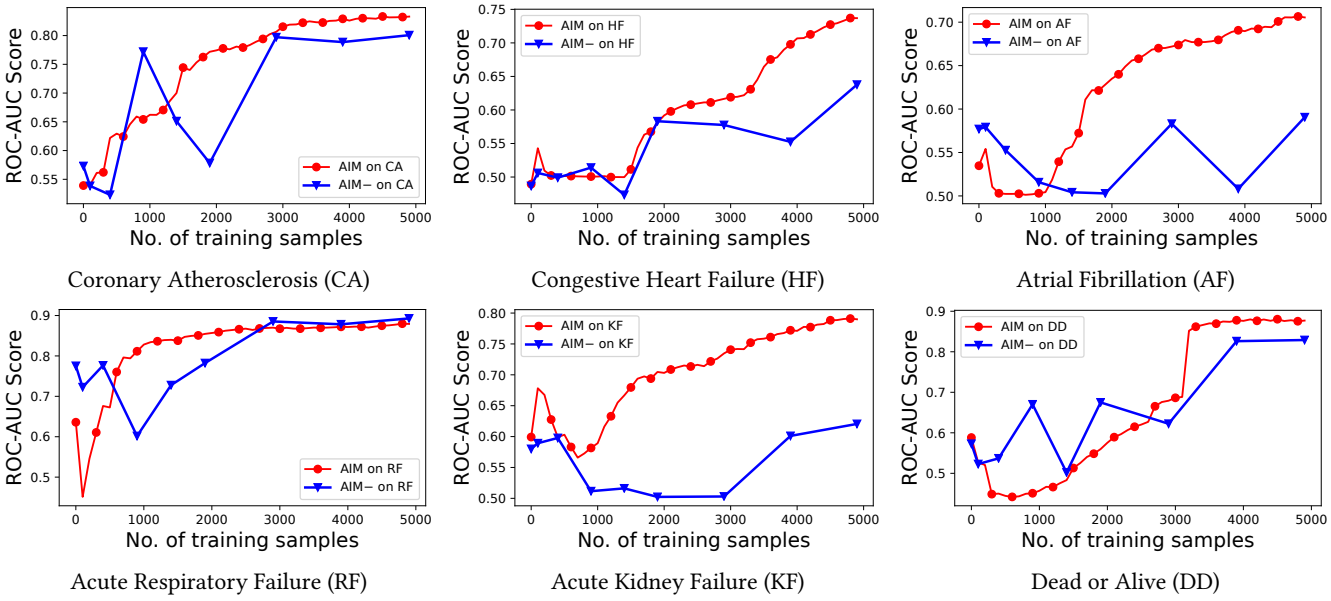

Figure 4: Performance comparison between AID− (passive) and AID (active) in scenarios where both methods are allowed the same data budget for interpretation.

this case, we observe that AID quickly outperforms AID− as the data budget increases. This is not surprising since AID− selects data randomly while AID selects data strategically, which widens their performance gap when the budget increases.

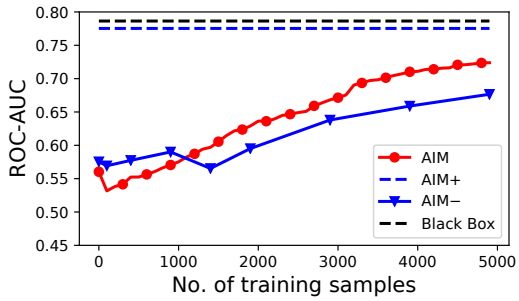

Figure 5: Performance (ROC-AUC) comparison on MIMIC-III dataset between AID and two passive variants: (a) AID+ is allowed to query the entire dataset; and (b) AID− which has the same budget as AID. The black box is trained on the entire dataset and serves as a performance upper-bound.

Exp 3. Case study of AID in patient subtyping.

To demonstrate potential uses of the AID's multi-level distillation on **MIMIC-III** dataset for patient subtyping, we extract its corresponding global rationale $\mathbf{b}_i$ for disease i (e.g., diabetes), which selects $\ell = 20$ codes from a pool of $m = 1000$ to explain the disease. Table 2 details the top 8 medical codes that best explain each target disease and mortality. As confirmed by clinicians whom we consulted with, these codes align with their domain knowledge.

AID's local explanations can then be leveraged to help categorize patients into sub-groups (i.e., subtyping) of the same disease to provide better personalized treatment. In particular, we extract AID's local explanation function $\mathbf{e}_\mathbf{w}(\mathbf{x})$ for the disease of interest (i.e., prediction target), and apply it on a representative set of 100 patients to generate their personalized weights for the above 20 medical codes. This results in a 20×100 weight matrix that characterizes the relevance of each medical code on each patient.

This is demonstrated via the heat-map plots in Fig. 6 below. In each plot, the dark-colored pixel represents the strong influence of a medical code on a patient's disease outcome. The personalized weight vector is projected onto a 2-dimensional space (using t-SNE) and the projected vectors are clustered into sub-groups (see scatter plots in Fig. 6) using Gaussian mixture model (GMM), which associate them with different disease subtypes (see Appendix G.3 for more results). The subtypes of new patients can therefore be identified by mapping their projected vectors to the nearest cluster.

This case study essentially demonstrates that AID's multi-level explanations can be used for patient subtyping in black-box setting. This is not possible in previous works which, due to the lack of global explanation, might mistake a feature relevant to disease A as also relevant to disease B (though it is not) if A and B co-occur in the same patient. This emphasizes on the necessity of having both global and local rationales in model distillation.

6 CONCLUSION

This paper introduces a black-box interpretation framework that optimizes a surrogate model to distill latent knowledge from a black box into its local and global interpretable representations. We

Target	Code 1	Code 2	Code 3	Code 4	Code 5	Code 6	Code 7	Code 8
EH	Apnea Time Interval	Inspiratory Time	Mean Airway Pressure	Hematocrit	Insulin	Creatine Kinase	Glucose	Metoprolol Tartrate
HF	Phosphorous	Abnormal Respiratory Pattern	Exercise Tolerance Test	Pantoprazole Sodium	Low Insp. Pressure	Warfarin	Anti-Embolism	Myositis Damage Index (MDI)
AF	Phosphorous	Total Bilirubin	Pneumococcal	Phenylephrine	Waveform-Vent	Respiratory Effort	Peak Insp. Pressure	Glucagon
CA	5% Dextrose	Red Blood Cells	Gastric Meds	pCO2	Ectopy Frequency	OxycoDONE	Amylase	Respiratory Rate
KF	High Blood Pressure	Ectopy Frequency	pCO2	SaO2	Sputum [Color]	High Resp. Rate	Arterial Blood Pressure systolic	Differential-Basos
DI	5% Dextrose	SVR	Leukocytes	Heparin	Creatine Kinase MB Isoenzyme	Neosynephrine-k	Insulin	Activity Tolerance
HA	5% Dextrose	Lorazepam	MCHC	Motor Response	Polychromasia	SpO2	Pantoprazole (Protonix)	Respiratory Rate
RF	Red Blood Cells	Lorazepam	Phosphorous	Motor Response	SVR	Low O2 Saturation	Heart rate Alarm - High	Mean Airway Pressure
UR	Urinary Tract Infections	Red Blood Cells	Lorazepam	Fentanyl Citrate	Flatus	Ventilator Type	Vancomycin	Respiratory Rate
ER	Nitroglycerin	Lorazepam	Propofol	WBC (4-11000)	Albuterol 0.083% Neb Soln	Prothrombin time	Lactate Dehydrogenase	CT 1 Drainage
DD	Allergy 1	CPK	SVR	WBC	INV Line#2SiteAppear	Ventilator Mode	Vti High	Urinal/Bedpan

Table 2: Examples of the top 8 medical codes that best explain each target disease and mortality.

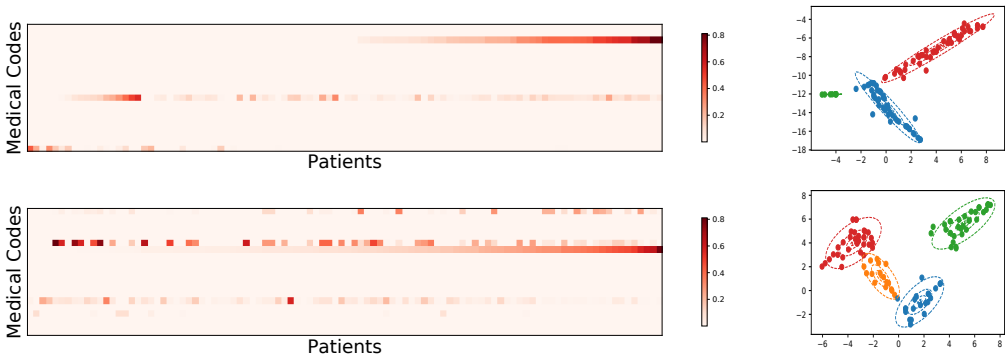

Figure 6: Heat-map (left) and scatter (right) plots of patients and their subtypes respectively for (upper plots) Diabetes (DI); and (lower plots) Coronary Atherosclerosis (CA).

develop an active interpretation algorithm (AID) to extract the most informative data from the black box for model interpretation using as few queries as permitted by a budget. AID is analyzed in both theory and practice, which shows promising results.

7 APPENDIX

7.1 Optimizing Eq. (4)

Optimizing $(\boldsymbol{\alpha}, \mathbf{w})$. Since $h_{\boldsymbol{\alpha}}(\mathbf{p}) \triangleq \langle \boldsymbol{\alpha}, \mathbf{p} \rangle$ is differentiable with respect to $\boldsymbol{\alpha}$, we can solve for $\boldsymbol{\alpha}$ straight-forwardly via gradient descent if $\mathbf{B} = \{\mathbf{b}_i\}_{i=1}^n$ is fixed. The gradient of $\boldsymbol{\alpha}$ is given below:

$$\begin{aligned} \nabla_{\boldsymbol{\alpha}} \mathcal{G} &= 2 \sum_{i=1}^{n} \sum_{t=1}^{k} \left(q_t^{(i)} - h_{\boldsymbol{\alpha}}(\mathbf{b}_i \circ \mathbf{e}_{\mathbf{w}}(\mathbf{x}_t) \circ \mathbf{x}_t) \right) \\ &\times \left(\mathbf{b}_i \circ \mathbf{e}_{\mathbf{w}}(\mathbf{x}_t) \circ \mathbf{x}_t \right). \end{aligned} \tag{12}$$

Likewise, we also have the gradient for $\mathbf{w}$:

$$\begin{aligned} \nabla_{\mathbf{w}} \mathcal{G} &= 2 \sum_{i=1}^{n} \sum_{t=1}^{k} \left(q_t^{(i)} - h_{\boldsymbol{\alpha}}(\mathbf{b}_i \circ \mathbf{e}_{\mathbf{w}}(\mathbf{x}_t) \circ \mathbf{x}_t) \right) \\ &\times \nabla_{\mathbf{w}} \Big(h_{\boldsymbol{\alpha}}(\mathbf{b}_i \circ \mathbf{e}_{\mathbf{w}}(\mathbf{x}_t) \circ \mathbf{x}_t) \Big) \end{aligned} \tag{13}$$

where $\nabla_{\mathbf{w}} \left(h_{\boldsymbol{\alpha}}(\mathbf{p}) \right) = \nabla_{\mathbf{p}} h_{\boldsymbol{\alpha}}(\mathbf{p}) \times \nabla_{\mathbf{w}} \mathbf{p} = \boldsymbol{\alpha} \times \nabla_{\mathbf{w}} \mathbf{p}$ with $\mathbf{p} = (\mathbf{b}_i \circ \mathbf{e}_{\mathbf{w}}(\mathbf{x}_t) \circ \mathbf{x}_t)$. In particular, we parameterize $\mathbf{e}_{\mathbf{w}}(\mathbf{x})$ with $\mathbf{w} \in \mathbb{R}^{m \times m}$ such that $\mathbf{e}_{\mathbf{w}}(\mathbf{x}) \triangleq [e_{\mathbf{w}}^{(1)}(\mathbf{x}) \dots e_{\mathbf{w}}^{(m)}(\mathbf{x})]$ and $e_{\mathbf{w}}^{(i)}(\mathbf{x})$ is defined in Section 3.2.1.

Although the above derivation seems to assume that $h_{\boldsymbol{\alpha}}$ and $\mathbf{e}_{\mathbf{w}}$ are specified in closed-form, this is not necessary. More broadly, we can characterize $h_{\boldsymbol{\alpha}}(\mathbf{b}_i \circ \mathbf{e}_{\mathbf{w}}(\mathbf{x}_t) \circ \mathbf{x}_t)$ using a neural network parameterized with $(\mathbf{w}, \boldsymbol{\alpha})$ for high flexibility. The resulting function is no longer analytic but its derivative can still be computed via back-propagation, which is sufficient to solve (12).

Optimizing B. To optimize the discrete variables $\mathbf{B}$, we use $\mathbf{z}_i \in [0,1]^m$ as a continuous proxy for $\mathbf{b}_i \in \{0,1\}^m$: $\mathbf{b}_i$ is replaced by $\mathbf{z}_i$ in (4) and $\mathbf{z}_i$ is optimized via gradient descent. The optimized unconstrained proxy $\mathbf{z}_i$ can then be transformed into a valid binary vector $\mathbf{b}_i$ via top-ℓ binarization: the corresponding entries in $\mathbf{b}_i$ to the ℓ largest components in $\mathbf{z}_i$ are set to 1 while the rest is set to 0. Here $\mathbf{z}_i$ can be viewed as a vector of scores that ranks the impact of features on predictive outcome. In practice, $\mathbf{z}_i$ can also be parameterized as neural networks for efficient implementation [7].

7.2 Proof to Lemma 1

Lemma 1. Let $\phi \triangleq \epsilon_t^{(i)}$ denotes the individual error yielded by $(\mathbf{B}^*, \mathbf{w}^*, \boldsymbol{\alpha}^*)$ at data point $\mathbf{x}_t$ and target condition $c^{(i)}$. Then, we have $\mathbf{I}_t \left(\mathbf{b}_i^*, \mathbf{w}^*, \boldsymbol{\alpha}^* \right) \leq \phi^2 / (8\nu(1 - \phi(1-\nu))^2)$ where ν is the black-box model's fitting noise (Definition 2).

Proof. Since ϕ is the individual error made by a surrogate parameterized with $(\mathbf{B}^*, \mathbf{w}^*, \boldsymbol{\alpha}^*)$ at data point $\mathbf{x}_t$ and target condition $c^{(i)}$, we have

$$r_t^{(i)} \leq \Phi^{-1} \left(\mathbb{P} \left(c_t^{(i)} | \mathbf{x}_t \right) \right) + \epsilon_t^{(i)}. \tag{14}$$

Applying Φ on both sides of the inequality thus yields

$$\Phi \left(r_t^{(i)} \right) \leq \Phi \left(\Phi^{-1} \left(\mathbb{P} \left(c_t^{(i)} | \mathbf{x}_t \right) \right) + \epsilon_t^{(i)} \right). \tag{15}$$

To avoid cluttering the notations, let us use q and p as short-hand notations for $\Phi(r_t^{(i)})$ and $\mathbb{P}(c_t^{(i)}|\mathbf{x}_t)$ in the remaining of the proof. The above can thus be rewritten concisely as $q \leq \Phi(\Phi^{-1}(p)+\phi)$.

Then, using the analytic forms of $\Phi(\mathbf{a}) = 1/(1+\exp(-\mathbf{a}))$ and $\Phi^{-1}(\mathbf{a}) = \log(\mathbf{a}/(1-\mathbf{a}))$, we can equivalently rewrite the above inequality $q \leq \Phi(\Phi^{-1}(p)+\phi)$ as $q \leq p/(p+\exp(-\phi)(1-p)) \leq p/(p+(1-\phi)(1-p))$ where the last inequality is due to the fact that $1-\phi \leq \exp(-\phi)$.

This implies $q-p \leq \phi p(1-p)/(1-\phi(1-p)) \leq \phi/(4(1-\phi(1-p))) \leq \phi/(4(1-\phi(1-\nu)))$ where the second last and last inequalities follow from the facts that (a) $p(1-p) \leq (p+1-p)^2/4 = 1/4$ and (b) $1-p \leq 1-\nu$ (which follows from the definition of ν), respectively.

Taking square on both sides of $q-p \leq \phi/(4(1-\phi(1-\nu)))$ yields $(q-p)^2 \leq \phi^2/(16(1-\phi(1-\nu))^2)$. Multiplying both sides with $2/\nu$ yields $2(q-p)^2/\nu \leq \phi^2/(8\nu(1-\phi(1-\nu))^2)$. Finally, note that by applying Pinsker inequality (the upper-bound version),

$$\begin{aligned}\mathbf{I}_t(\mathbf{b}_i^*, \mathbf{w}^*, \boldsymbol{\alpha}^*) &= \mathbf{D}_{\mathrm{KL}}\left(\mathbb{P}_{\boldsymbol{\alpha}^*}\left(c_t^{(i)}|\mathbf{b}_i^*, \mathbf{e}_{\mathbf{w}^*}(\mathbf{x}_t)\right) \| \mathbb{P}\left(c_t^{(i)}|\mathbf{x}_t\right)\right) \\ &\leq 2(q-p)^2/\nu \leq \phi^2/(8\nu(1-\phi(1-\nu))^2)\,.\end{aligned}$$

This completes our proof of Lemma 1.

7.3 Proof of Theorem 1

Let $\widehat{\mathbf{I}}(\mathbf{b}_i^*, \mathbf{w}^*, \boldsymbol{\alpha}^*) \triangleq (1/k)\sum_{t=1}^k \mathbf{I}_t(\mathbf{b}_i^*, \mathbf{w}^*, \boldsymbol{\alpha}^*)$. Applying Lemma 1 independently for each pair of (i,t) yields

$$\mathbf{I}_t(\mathbf{b}_i^*, \mathbf{w}^*, \boldsymbol{\alpha}^*) \leq \epsilon_t^{(i)^2}/\left(8\nu(1-\epsilon_t^{(i)}(1-\nu))^2\right) \leq \epsilon^2/\left(8\nu(1-\epsilon(1-\nu))^2\right)$$

where ϵ is defined in Definition 3. This implies $0 \leq \widehat{\mathbf{I}}(\mathbf{b}_i^*, \mathbf{w}^*, \boldsymbol{\alpha}^*) \leq \epsilon^2/(8\nu(1-\epsilon(1-\nu))^2)$. Then, to bound the difference between $\mathbf{I}(\mathbf{b}_i^*, \mathbf{w}^*, \boldsymbol{\alpha}^*)$ (i.e., true mean) and $\widehat{\mathbf{I}}(\mathbf{b}_i^*, \mathbf{w}^*, \boldsymbol{\alpha}^*)$ (i.e., empirical mean), we exploit the following concentration inequality:

Lemma 2 **[Hoeffding Inequality].** Let $\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_k$ be independent samples drawn from an arbitrary distribution $\mathcal{P}$ such that $0 \leq \mathbf{I}_i \leq \ell$. Let $\widehat{\mathbf{I}} = k^{-1}\sum_{t=1}^k \mathbf{I}_t$ and $\mathbf{I} = \mathbb{E}[\mathbf{I}_t]$,

$$\mathcal{P}\left(|\mathbf{I}-\widehat{\mathbf{I}}| \leq \theta\right) \geq 1 - 2\exp\left(-\frac{2k\theta^2}{\ell^2}\right) \tag{16}$$

Proof. Omitted

.

Using Lemma 2 above we are now ready to establish the following key result:

Theorem 1. Let $\boldsymbol{\psi}(\epsilon,\nu) \triangleq \epsilon^2/(8\nu(1-\epsilon(1-\nu))^2)$ and $\delta \in (0,1)$. Then, with probability at least $1-\delta$, we have:

$$\mathbf{I}\left(\mathbf{b}_i^*, \mathbf{w}^*, \boldsymbol{\alpha}^*\right) \leq \boldsymbol{\psi}(\epsilon,\nu)\left(1+\sqrt{\frac{1}{2k}\log\left(\frac{2}{\delta}\right)}\right) \tag{17}$$

Proof. Let $\mathbf{I}_t = \mathbf{I}_t(\mathbf{b}_i^*, \mathbf{w}^*, \boldsymbol{\alpha}^*)$, and $\widehat{\mathbf{I}} = \widehat{\mathbf{I}}(\mathbf{b}_i^*, \mathbf{w}^*, \boldsymbol{\alpha}^*)$, $\mathbf{I} = \mathbf{I}(\mathbf{b}_i^*, \mathbf{w}^*, \boldsymbol{\alpha}^*)$ and $\ell = \boldsymbol{\psi}(\epsilon,\nu) \triangleq \epsilon^2/(8\nu(1-\epsilon(1-\nu))^2)$, Lemma 2 guarantees that with probability at least $1-2\exp(-2k\theta^2/\boldsymbol{\psi}^2(\epsilon,\nu))$,

$$|\mathbf{I}(\mathbf{b}_i^*, \mathbf{w}^*, \boldsymbol{\alpha}^*) - \widehat{\mathbf{I}}(\mathbf{b}_i^*, \mathbf{w}^*, \boldsymbol{\alpha}^*)| \leq \theta \tag{18}$$

which implies $\mathbf{I}(\mathbf{b}_i^*, \mathbf{w}^*, \boldsymbol{\alpha}^*) \leq \widehat{\mathbf{I}}(\mathbf{b}_i^*, \mathbf{w}^*, \boldsymbol{\alpha}^*) + \theta \leq \boldsymbol{\psi}(\epsilon,\nu) + \theta$. Now, setting $\delta = 2\exp(-2k\theta^2/\boldsymbol{\psi}^2(\epsilon,\nu))$ and solving for θ yields

$$\theta = \boldsymbol{\psi}(\epsilon,\nu)\sqrt{\frac{1}{2k}\log\left(\frac{2}{\delta}\right)}\,. \tag{19}$$

Plugging this into the previous inequality yields (17).

7.4 Proof of Theorem 2

To derive the key result in Theorem 2, we first establish the following intermediate result:

Lemma 3. Let γ denote the prediction robustness of the black-box as in Definition 4. If we have

$$\mathbf{D}_{\mathrm{KL}}\left(\mathbb{P}_{\boldsymbol{\alpha}^*}\left(c^{(i)}|\mathbf{b}_i^*, \mathbf{e}_{\mathbf{w}^*}(\mathbf{x})\right) \| \mathbb{P}(c^{(i)}|\mathbf{x})\right) \leq \gamma^2\,, \tag{20}$$

then $\mathbb{P}_{\boldsymbol{\alpha}^*}\left(c^{(i)}|\mathbf{b}_i^*, \mathbf{e}_{\mathbf{w}^*}(\mathbf{x})\right)$ and $\mathbb{P}\left(c^{(i)}|\mathbf{x}\right)$ yield the same prediction.

Proof. Let us denote $\mathbb{Q}(0) \triangleq \mathbb{P}_{\boldsymbol{\alpha}^*}(c^{(i)}=0|\mathbf{b}_i^*, \mathbf{e}_{\mathbf{w}^*}(\mathbf{x}))$ and $\mathbb{Q}(1) \triangleq \mathbb{P}_{\boldsymbol{\alpha}^*}(c^{(i)}=1|\mathbf{b}_i^*, \mathbf{e}_{\mathbf{w}^*}(\mathbf{x}))$. Likewise, let $\mathbb{P}(0) \triangleq \mathbb{P}(c^{(i)}=0|\mathbf{x})$ and $\mathbb{P}(1) \triangleq \mathbb{P}(c^{(i)}=1|\mathbf{x})$. By Pinkser inequality (the lower-bound version), we have

$$\begin{aligned}2\left((\mathbb{P}(0)-\mathbb{Q}(0))^2 + (\mathbb{P}(1)-\mathbb{Q}(1))^2\right) &\leq \mathbf{D}_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) \\ &\leq \gamma^2.\end{aligned} \tag{21}$$

Furthermore, note that $(\mathbb{P}(0)-\mathbb{Q}(0))^2 = (\mathbb{P}(1)-\mathbb{Q}(1))^2$ since $1 = \mathbb{P}(0)+\mathbb{P}(1) = \mathbb{Q}(0)+\mathbb{Q}(1)$. Thus, the above inequality can be rewritten more concisely as

$$\left(\mathbb{P}(0)-\mathbb{Q}(0)\right)^2 = \left(\mathbb{P}(1)-\mathbb{Q}(1)\right)^2 \leq \frac{\gamma^2}{4}\,. \tag{22}$$

This implies $\mathbb{P}(0)-\gamma/2 \leq \mathbb{Q}(0) \leq \mathbb{P}(0)+\gamma/2$ and $\mathbb{P}(1)-\gamma/2 \leq \mathbb{Q}(1) \leq \mathbb{P}(1)+\gamma/2$. Now, let c^* be the prediction made by $\mathbb{P}$, e.g., $\mathbb{P}(c^*) > \mathbb{P}(1-c^*)+\gamma$ (by Definition 4). Thus, we have $\mathbb{Q}(c^*) \geq \mathbb{P}(c^*)-\gamma/2 > \mathbb{P}(1-c^*)+\gamma-\gamma/2 \geq \mathbb{Q}(1-c^*)-\gamma/2+\gamma-\gamma/2 = \mathbb{Q}(1-c^*)$. Hence, c^* is also the prediction made by $\mathbb{Q}$ since $\mathbb{Q}(c^*) > \mathbb{Q}(1-c^*)$.

Using the result of Lemma 3 above and let the event that the model $\mathbb{P}_{\boldsymbol{\alpha}^*}(c^{(i)}|\mathbf{b}_i^*, \mathbf{e}_{\mathbf{w}^*}(\mathbf{x}))$ disagrees with $\mathbb{P}(c^{(i)}|\mathbf{x}))$ on a random input $\mathbf{x} \sim \mathcal{P}(\mathbf{x})$ be denoted by $\mathbf{E}$, we can then derive a stronger version of Theorem 1 that explicitly bounds the chance that $\mathbf{E}$ happens (i.e., the chance surrogate mis-distil the black-box) via Theorem 2.

Theorem 2. Let $k = (1/(2\gamma^4))\log(2/\delta)$ with $0 < \delta < 1$ where k denotes the size of the sampled data and γ is defined in Definition 4. Then, with probability at least $1-\delta$,

$$\mathcal{P}(\mathbf{E}) \leq \boldsymbol{\psi}(\epsilon,\nu)\left(\frac{1}{\gamma^2}+1\right)\,, \tag{23}$$

with ν and ϵ defined previously in Definitions 2 and 3.

Proof. Applying the result of Theorem 1, it follows that with probability at least $1-\delta$,

$$\begin{aligned}\mathbf{I}(\mathbf{b}_i^*, \mathbf{w}^*, \boldsymbol{\alpha}^*) &\leq \boldsymbol{\psi}(\epsilon,\nu)\left(1+\left(\frac{1}{2k}\right)^{\frac{1}{2}}\log\left(\frac{2}{\delta}\right)^{\frac{1}{2}}\right) \\ &= \boldsymbol{\psi}(\epsilon,\nu)\left(\gamma^2+1\right)\end{aligned} \tag{24}$$

where the second step follows by plugging in $k = (1/(2\gamma^4))\log(2/\delta)$. Then, by Markov inequality,

$$\begin{aligned}&\mathcal{P}\left(\mathbf{D}_{\mathrm{KL}}\left(\mathbb{P}_{\boldsymbol{\alpha}^*}\left(c^{(i)}|\mathbf{b}_i^*, \mathbf{e}_{\mathbf{w}^*}(\mathbf{x})\right)\|\mathbb{P}\left(c^{(i)}|\mathbf{x}\right)\right) > \gamma^2\right)\\ \leq\ & \mathbf{I}(\mathbf{b}_i^*, \mathbf{w}^*, \boldsymbol{\alpha}^*)\gamma^{-2} \leq \boldsymbol{\psi}(\epsilon, \nu)\left(1+\frac{1}{\gamma^2}\right)\end{aligned} \tag{25}$$

Note that the Markov inequality $\mathcal{P}(\mathbf{r} > a) \leq \mathbb{E}(\mathbf{r})/a$ applies to the above because $\mathbf{I}(\mathbf{b}_i^*, \mathbf{w}^*, \boldsymbol{\alpha}^*)$ is defined to be the expectation over $\mathbf{x}$ of the above $\mathbf{D}_{\mathrm{KL}}$ term (see Section 4 above).

On the other hand, let $\bar{\mathbf{E}}$ denote the event that the surrogate interprets the black-box model correctly. By Lemma 3,

$$\mathbf{D}_{\mathrm{KL}}\left(\mathbb{P}_{\boldsymbol{\alpha}^*}\left(c^{(i)}|\mathbf{b}_i^*, \mathbf{e}_{\mathbf{w}^*}(\mathbf{x})\right)\|\mathbb{P}\left(c^{(i)}|\mathbf{x}\right)\right) \leq \gamma^2 \tag{26}$$

implies $\bar{\mathbf{E}}$. As a result, we have $1 - \mathcal{P}(\mathbf{E}) = \mathcal{P}(\bar{\mathbf{E}}) \geq$

$$\mathcal{P}\left(\mathbf{D}_{\mathrm{KL}}\left(\mathbb{P}_{\boldsymbol{\alpha}^*}\left(c^{(i)}|\mathbf{b}_i^*, \mathbf{e}_{\mathbf{w}^*}(\mathbf{x})\right)\|\mathbb{P}\left(c^{(i)}|\mathbf{x}\right)\right) \leq \gamma^2\right) =$$

$$1 - \mathcal{P}\left(\mathbf{D}_{\mathrm{KL}}\left(\mathbb{P}_{\boldsymbol{\alpha}^*}\left(c^{(i)}|\mathbf{b}_i^*, \mathbf{e}_{\mathbf{w}^*}(\mathbf{x})\right)\|\mathbb{P}\left(c^{(i)}|\mathbf{x}\right)\right) > \gamma^2\right) \tag{27}$$

This essentially implies $\mathcal{P}(\mathbf{E})$ is bounded above by

$$\begin{aligned}&\mathcal{P}\left(\mathbf{D}_{\mathrm{KL}}\left(\mathbb{P}_{\boldsymbol{\alpha}^*}\left(c^{(i)}|\mathbf{b}_i^*, \mathbf{e}_{\mathbf{w}^*}(\mathbf{x})\right)\|\mathbb{P}\left(c^{(i)}|\mathbf{x}\right)\right) > \gamma^2\right)\\ &\leq \boldsymbol{\psi}(\epsilon, \nu)\left(1+\frac{1}{\gamma^2}\right) \leq \frac{\epsilon}{3}\left(1+\frac{1}{\gamma^2}\right)\end{aligned} \tag{28}$$

which yields Eq. (23) and completes our proof.

7.5 Proof of Theorem 3

To derive the key result in Theorem 3, we establish first the following intermediate result:

Lemma 4. For any $\lambda \in \mathbb{N}^*$, let $\mathbf{H}_{q+\lambda}^{(i)}$ and $\mathbf{H}_q^{(i)}$ denote, respectively, the disagreement regions between the surrogate and black-box models at iterations $q+\lambda$ and q. Then, with probability at least $1-\delta$, we have

$$\mathcal{P}\left(\mathbf{H}_{q+\lambda}^{(i)}\right) \leq \mathcal{P}\left(\mathbf{H}_q^{(i)}\right)\frac{\epsilon}{3\gamma^2}\left(1+\sqrt{\frac{1}{2\lambda}\log\left(\frac{2}{\delta}\right)}\right) \tag{29}$$

Proof. First, let's recall that

$$\mathbf{H}_\ell^{(i)} \triangleq \left\{\mathbf{x}\,\middle|\,\mathbf{D}_{\mathrm{KL}}\left(\mathbb{P}_{\boldsymbol{\alpha}^*}^{\ell}\left(c^{(i)}|\mathbf{x}\right)\|\mathbb{P}\left(c^{(i)}|\mathbf{x}\right)\right) > \gamma^2\right\} \tag{30}$$

denote the disagreement region between the surrogate and black-box after ℓ iterations of active interpretation. Again, we use $\mathbb{P}_{\boldsymbol{\alpha}^*}^{\ell}(c^{(i)}|\mathbf{x})$ as a short-hand notation for $\mathbb{P}_{\boldsymbol{\alpha}^*}^{\ell}(c^{(i)}|\mathbf{b}_i^*, \mathbf{e}_{\mathbf{w}^*}(\mathbf{x}))$. The superscript ℓ indicates that the surrogate has been fitted using the first ℓ queried data points.

In the scope of this lemma, we are interested in the cases where $\ell = q$ and $\ell = q+\lambda$. For $\ell = q+\lambda$, the probability of $\mathbf{x} \in \mathbf{H}_{q+\lambda}^{(i)}$ can be expressed as

$$\begin{aligned}\mathcal{P}\left(\mathbf{H}_{q+\lambda}^{(i)}\right) &= \left(\int_{\mathbf{x}\in\mathbf{H}_{q+\lambda}^{(i)}}\frac{\mathcal{P}(\mathbf{x})}{\mathcal{P}\left(\mathbf{H}_q^{(i)}\right)}\mathrm{d}\mathbf{x}\right)\mathcal{P}\left(\mathbf{H}_q^{(i)}\right)\\ &\leq \left(\int_{\mathbf{x}\in\mathbf{H}_q^{(i)}}\frac{\mathcal{P}(\mathbf{x})}{\mathcal{P}\left(\mathbf{H}_q^{(i)}\right)}\mathrm{d}\mathbf{x}\right)\mathcal{P}\left(\mathbf{H}_q^{(i)}\right)\\ &= \mathcal{P}\left(\mathbf{x}\in\mathbf{H}_{q+\lambda}^{(i)}|\mathbf{x}\in\mathbf{H}_q^{(i)}\right)\mathcal{P}\left(\mathbf{H}_q^{(i)}\right)\end{aligned} \tag{31}$$

where the first inequality follows from our assumption earlier that $\mathbf{H}_{q+\lambda}^{(i)} \subseteq \mathbf{H}_q^{(i)}$. Note that the first factor in the RHS of the last equality above is essentially the probability that

$$\mathbf{D}_{\mathrm{KL}}\left(\mathbb{P}_{\boldsymbol{\alpha}^*}\left(c^{(i)}|\mathbf{b}_i^*, \mathbf{e}_{\mathbf{w}^*}(\mathbf{x})\right)\|\mathbb{P}\left(c^{(i)}|\mathbf{x}\right)\right) > \gamma^2 \tag{32}$$

if the surrogate is fitted using λ input samples drawn independently from $\mathbf{H}_q^{(i)}$ via an augmented data distribution $\mathcal{Q}(x) = \mathcal{P}(x)/\mathcal{P}(\mathbf{H}_q^{(i)})$. As such, we can define a new interpretation quality $\mathbf{I}'(\mathbf{b}_i^*, \mathbf{w}^*, \boldsymbol{\alpha}^*)$, which is the same as $\mathbf{I}(\mathbf{b}_i^*, \mathbf{w}^*, \boldsymbol{\alpha}^*)$, except that the data distribution is $\mathcal{Q}(\mathbf{x})$ instead of $\mathcal{P}(\mathbf{x})$. We can reuse Theorem 1 to bound $\mathbf{I}'(\mathbf{b}_i^*, \mathbf{w}^*, \boldsymbol{\alpha}^*)$ with probability at least $1-\delta$:

$$\begin{aligned}\mathbf{I}'\left(\mathbf{b}_i^*, \mathbf{w}^*, \boldsymbol{\alpha}^*\right) &\leq \boldsymbol{\psi}(\epsilon, \nu)\left(1+\sqrt{\frac{1}{2\lambda}\log\left(\frac{2}{\delta}\right)}\right)\\ &\leq \frac{\epsilon}{3}\left(1+\sqrt{\frac{1}{2\lambda}\log\left(\frac{2}{\delta}\right)}\right)\end{aligned} \tag{33}$$

where the second inequality follows from our assumption that $\epsilon \leq \nu$, which in turns implies $\psi(\epsilon, \nu) \leq \epsilon/3$. Then, since $\mathbf{I}'(\mathbf{b}_i^*, \mathbf{w}^*, \boldsymbol{\alpha}^*)$ is effectively the expectation of the KL term in (32), it can be exploited to bound the probability that the KL term exceeds a threshold of γ^2 via Markov inequality as detailed below:

$$\begin{aligned}&\mathcal{P}\left(\mathbf{D}_{\mathrm{KL}}\left(\mathbb{P}_{\boldsymbol{\alpha}^*}\left(c^{(i)}|\mathbf{b}_i^*, \mathbf{e}_{\mathbf{w}^*}(\mathbf{x})\right)\|\mathbb{P}\left(c^{(i)}|\mathbf{x}\right)\right) > \gamma^2\right)\\ \leq\ & \mathbf{I}'(\mathbf{b}_i^*, \mathbf{w}^*, \boldsymbol{\alpha}^*)\gamma^{-2} \leq \frac{\epsilon}{3\gamma^2}\left(1+\sqrt{\frac{1}{2\lambda}\log\left(\frac{2}{\delta}\right)}\right)\end{aligned} \tag{34}$$

where the last inequality follows from Eq. (33) above. Lastly, by definition of the data distribution $\mathcal{Q}(\mathbf{x})$, the LHS of the above equation is effectively the probability that $\mathbf{x} \in \mathbf{H}_{q+\lambda}^{(i)}$ given that $\mathbf{x} \in \mathbf{H}_q^{(i)}$. Thus, the above inequality can be rewritten as:

$$\mathcal{P}\left(\mathbf{x}\in\mathbf{H}_{q+\lambda}^{(i)}|\mathbf{x}\in\mathbf{H}_q^{(i)}\right) \leq \frac{\epsilon}{3\gamma^2}\left(1+\sqrt{\frac{1}{2\lambda}\log\left(\frac{2}{\delta}\right)}\right)$$

Plugging this into Eq. (31) above completes our proof. Using this result, we are now ready to prove Theorem 3.

Theorem 3. For any target $c^{(i)}$, let $q^* = (r/(2\gamma^2))\log(2r/\delta)$ with

$$r = \frac{\log\left(\frac{\epsilon}{3\gamma^2}\left(1+\gamma^2\right)\right)}{\log\left(\frac{\epsilon}{3\gamma^2}\left(1+\gamma\right)\right)} = \frac{\log\left(\frac{\epsilon}{3}\left(1+\frac{1}{\gamma^2}\right)\right)}{\log\left(\frac{\epsilon}{3\gamma^2}\left(1+\gamma\right)\right)}, \tag{35}$$

and let $\mathbf{E}^{(i)}_{q^*}$ denote the event that the surrogate fitted on q^* queried data points disagrees with the black-box on a random data sample $\mathbf{x}$. Then, with probability at least $1-\delta$, $\mathcal{P}(\mathbf{E}^{(i)}_{q^*}) \leq (\epsilon/3)(1+1/\gamma^2)$, which achieves the same mis-distillation rate of its passive version while requiring fewer samples, i.e. $q^* < k = (1/(2\gamma^4))\log(2/\delta)$.

Proof. Choose $q = 0$ and apply Lemma 4 independently for $\ell + u\lambda$ with $u \in \{0, \ldots, r-1\}$ and δ/r, we have

$$\mathcal{P}\left(\mathbf{H}^{(i)}_{(u+1)\lambda}\right) \quad \leq \quad \mathcal{P}\left(\mathbf{H}^{(i)}_{u\lambda}\right) \frac{\epsilon}{3\gamma^2}\left(1+\sqrt{\frac{1}{2\lambda}\log\left(\frac{2r}{\delta}\right)}\right)$$

held independently for each u with probability at least $1-\delta/r$. Thus, by the union bound, the probability the above holds simultaneously for $u \in \{0, \ldots, r-1\}$ is at least $1 - r \times (\delta/r) = 1-\delta$. When that happens, we can chain those inequalities together to yield

$$\begin{aligned}\mathcal{P}\left(\mathbf{H}^{(i)}_{r\lambda}\right) \quad &\leq \quad \mathcal{P}\left(\mathbf{H}^{(i)}_{0}\right)\left(\frac{\epsilon}{3\gamma^2}\right)^r\left(1+\sqrt{\frac{1}{2\lambda}\log\left(\frac{2r}{\delta}\right)}\right)^r \\ &\leq \quad \left(\frac{\epsilon}{3\gamma^2}\right)^r\left(1+\sqrt{\frac{1}{2\lambda}\log\left(\frac{2r}{\delta}\right)}\right)^r \end{aligned} \tag{36}$$

where the last inequality simply follows from the fact that $\mathcal{P}(\mathbf{H}^{(i)}_0) \leq 1$. To guarantee that the active interpretation algorithm achieves the same mis-interpretation rate as its passive version, we set the RHS of the above inequality to $\epsilon/3(1+1/\gamma^2)$ and solve for r and λ. In particular, to reduce the order of sample complexity, we can set $\gamma = \sqrt{\frac{1}{2\lambda}\log\frac{2r}{\delta}}$ and solve for λ, which yields $\lambda = (1/(2\gamma^2))\log(2r/\delta)$. Plugging this into

$$\left(\frac{\epsilon}{3\gamma^2}\right)^r\left(1+\sqrt{\frac{1}{2\lambda}\log\left(\frac{2r}{\delta}\right)}\right)^r \quad = \quad \frac{\epsilon}{3}\left(1+\frac{1}{\gamma^2}\right) \tag{37}$$

and solving for r yields

$$r \quad = \quad \frac{\log\left(\frac{\epsilon}{3\gamma^2}\left(1+\gamma^2\right)\right)}{\log\left(\frac{\epsilon}{3\gamma^2}\left(1+\gamma\right)\right)} = \frac{\log\left(\frac{\epsilon}{3}\left(1+\frac{1}{\gamma^2}\right)\right)}{\log\left(\frac{\epsilon}{3\gamma^2}\left(1+\gamma\right)\right)}. \tag{38}$$

Thus, choosing $\lambda = (1/(2\gamma^2))\log(2r/\delta)$ and r as above yields

$$\mathcal{P}\left(\mathbf{E}^{(i)}_{r\lambda}\right) \quad \leq \quad \mathcal{P}\left(\mathbf{H}^{(i)}_{r\lambda}\right) \quad \leq \quad \frac{\epsilon}{3}\left(1+\frac{1}{\gamma^2}\right), \tag{39}$$

where the first inequality follows because by Lemma 3, we have $\neg\mathbf{H}^{(i)}_{r\lambda}$ implies $\neg\mathbf{E}^{(i)}_{r\lambda}$ which means $\mathcal{P}(\neg\mathbf{H}^{(i)}_{r\lambda}) \leq \mathcal{P}(\neg\mathbf{E}^{(i)}_{r\lambda})$ and consequently, $\mathcal{P}(\mathbf{H}^{(i)}_{r\lambda}) \geq \mathcal{P}(\mathbf{E}^{(i)}_{r\lambda})$. Finally, choose $q^* = r\lambda$ yields $\mathcal{P}(\mathbf{E}^{(i)}_{q^*}) \leq (\epsilon/3)(1+1/\gamma^2)$.

REFERENCES

[1] Jimmy Ba and Rich Caruana. 2014. Do deep nets really need to be deep?. In *NIPS*. 2654–2662.

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

[3] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. MAzller. 2010. How to explain individual classification decisions. *JMLR* 11 (2010), 1803–1831.

[4] Jianbo Chen, Le Song, Martin J Wainwright, and Michael I Jordan. 2018. Learning to Explain: An Information-Theoretic Perspective on Model Interpretation. *arXiv preprint arXiv:1802.07814* (2018).

[5] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. 2017. GRAM: Graph-based Attention Model for Healthcare Representation Learning. In *KDD*.

[6] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *NIPS*.

[7] Matthieu Courbariaux and Yoshua Bengio. 2016. BinaryNet: Training Deep Neural Networks with Weights and Activations Constrained to +1 or -1. *CoRR* abs/1602.02830 (2016). arXiv:1602.02830 http://arxiv.org/abs/1602.02830

[8] Samuel G Finlayson, Isaac S Kohane, and Andrew L Beam. 2018. Adversarial Attacks Against Medical Deep Learning Systems. *arXiv:1804.05296* (2018).

[9] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* (2018).

[10] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[11] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* (2016).

[12] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. 2015. Unifying distillation and privileged information. *ICLR* (2015).

[13] S. M. Lundberg and S.-I. Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *NIPS*. 4768–4777.

[14] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. 2017. Data distillation: Towards omni-supervised learning. *arXiv preprint arXiv:1712.04440* (2017).

[15] M. T. Ribeiro, S. Singh, and C. Guestrin. 2016. Why should I trust you?: Explaining the predictions of any classifier. In *SIGKDD*. 1135–1144.

[16] M. T. Ribeiro, S. Singh, and C. Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanation. In *AAAI*.

[17] Bharat Bhusan Sau and Vineeth N Balasubramanian. 2016. Deep model compression: Distilling knowledge from noisy teachers. *arXiv:1610.09650* (2016).

[18] Burr Settles. 2012. *Active Learning*. Morgan & Claypool Publishers.

[19] A. Shrikumar, P. Greenside, and A. Kundaje. 2017. Learning important features through propagating activation differences. In *ICML*. 3145–3153.

[20] K. Simonyan, A. Vedaldi, and A. Zisserman. 2013. Deep inside convolutional networks: Visualizing image classification models and saliency maps. http://arxiv.org/abs/1312.6034

[21] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. 2014. Striving for simplicity: The all convolutional net. http://arxiv.org/abs/1412.6806

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

[23] Yanbo Xu, Siddharth Biswal, Shriprasad R Deshpande, Kevin O Maher, and Jimeng Sun. 2018. RAIM: Recurrent Attentive and Intensive Model of Multimodal Patient Monitoring Data. In *KDD*.

[24] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. 2017. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *CVPR*, Vol. 2.