The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2019–2038
https://doi.org/10.1214/22-AOAS1705

DATA-ADAPTIVE DISCRIMINATIVE FEATURE LOCALIZATION WITH STATISTICALLY GUARANTEED INTERPRETATION

BY BEN DAI[1,a], XIAOTONG SHEN[2,b], LIN YEE CHEN[3,d], CHUNLIN LI[2,c] AND WEI PAN[4,e]

[1]*Department of Statistics, The Chinese University of Hong Kong,* [a]*bendai@cuhk.edu.hk*
[2]*School of Statistics, University of Minnesota,* [b]*xshen@umn.edu,* [c]*li000007@umn.edu*
[3]*Lillehei Heart Institute and Cardiovascular Division, University of Minnesota,* [d]*chenx484@umn.edu*
[4]*Division of Biostatistics, University of Minnesota,* [e]*panxx014@umn.edu*

In explainable artificial intelligence, discriminative feature localization is critical to reveal a black-box model's decision-making process from raw data to prediction. In this article we use two real datasets, the MNIST handwritten digits and MIT-BIH electrocardiogram (ECG) signals, to motivate key characteristics of discriminative features, namely, *adaptiveness*, *predictive importance* and *effectiveness*. Then we develop a localization framework, based on adversarial attacks, to effectively localize discriminative features. In contrast to existing heuristic methods, we also provide a statistically guaranteed interpretability of the localized features by measuring a generalized partial R^2. We apply the proposed method to the MNIST dataset and the MIT-BIH dataset with a convolutional autoencoder. In the first, the compact image regions localized by the proposed method are visually appealing. Similarly, in the second, the identified ECG features are biologically plausible and consistent with cardiac electrophysiological principles while locating subtle anomalies in a QRS complex that may not be discernible by the naked eye. Overall, the proposed method compares favorably with state-of-the-art competitors. Accompanying this paper is a Python library **dnn-locate** that implements the proposed approach.

1. Introduction. The empirical success of machine learning in real applications has profound impacts on many scientific and engineering areas, including image analysis (LeCun et al. (1989), He et al. (2016)), recommender systems (Wang, Wang and Yeung (2015)), natural language processing (Hochreiter and Schmidhuber (1997)), drug discovery (Vamathevan et al. (2019)) and protein structure prediction (Jumper et al. (2021), Evans et al. (2021)). However, the nature of a black-box model makes it challenging to interpret its decision-making process. The lack of interpretability hinders transparency, trust and understanding of scientific discovery. To meet challenges, explainable AI (XAI) is emerging, which includes localizing discriminative features attributing to a model's predictive performance, shaping or confirming human intuitions and knowledge, for instance, visual explanation on image recognition.

1.1. *Motivation*: *DL discriminative localization in the MIT-BIH ECG dataset*. Our investigation responds to the need for locating features that are most critical to a learning outcome. The present study is motivated by the MIT-BIH ECG dataset and the MNIST dataset. Specifically, the MNIST dataset serves as a benchmark for studying XAI methods (Lundberg and Lee (2017), Ribeiro, Singh and Guestrin (2016)), in part, because the results of localization could be easily evaluated by human intuition. As demonstrated in Figures 3 and 7, localized image pixels explain how a deep convolutional network differentiates digits "7" and "9"

Received March 2022; revised August 2022.
Key words and phrases. Explainable artificial intelligence, discriminative features, localization, generalized partial R^2, interpretability, regularization, deep learning.

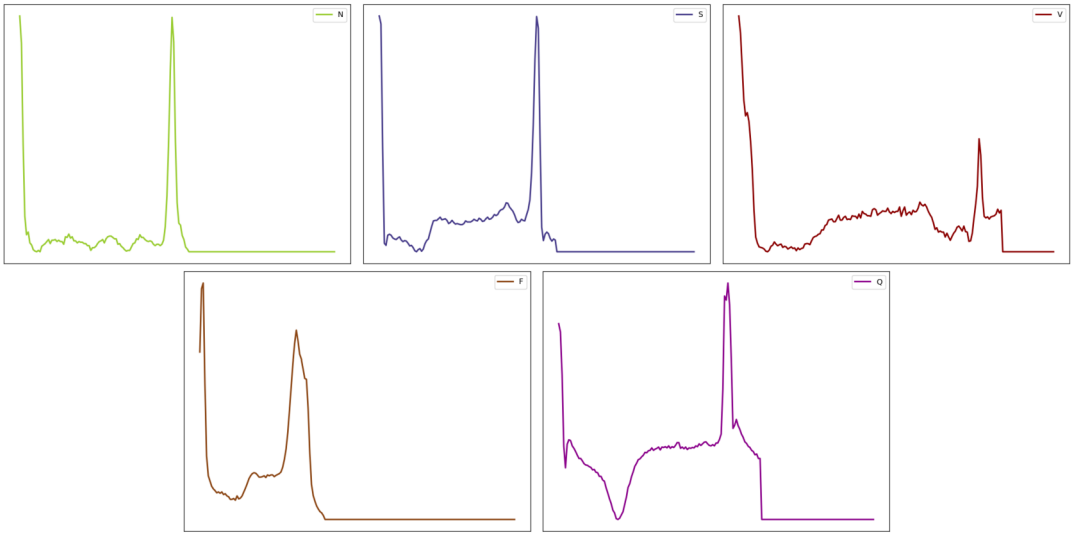

FIG. 1. *Five classes of ECG beat*: {*"N:" normal, left/right bundle branch block, atrial escape, nodal escape*}, {*"S:" atrial premature, aberrant atrial premature, nodal premature, supra-ventricular premature*}, {*"V:" premature ventricular contraction, ventricular escape*}, {*"F": fusion of ventricular and normal*}, {*"Q:" paced, fusion of paced and normal, unclassifiable*}.

on the MNIST data. A more substantial medical application is based on the MIT-BIH ECG dataset; this dataset is a commonly used ECG benchmark dataset which consists of ECG recordings from 47 different subjects recorded at the sampling rate of 360Hz by the BIH Arrhythmia Laboratory. Each beat is annotated into five different classes under the Association for the Advancement of Medical Instrumentation (AAMI) EC57 standard (Stergiou et al. (2018)): "N, S, V, F" and "Q." One random sample per class is demonstrated in Figure 1.

Broadly speaking, the existing ECG diagnosis methods in the literature can be categorized into two: conventional machine learning (ML) and deep learning (DL) methods. Conventional ML methods first extract manually-crafted features based on ECG background knowledge and some signal morphological technique, including the QRS complex, T wave, R-R interval, S-T interval (Wasimuddin et al. (2020)); see Figure 2. Next, conventional classification

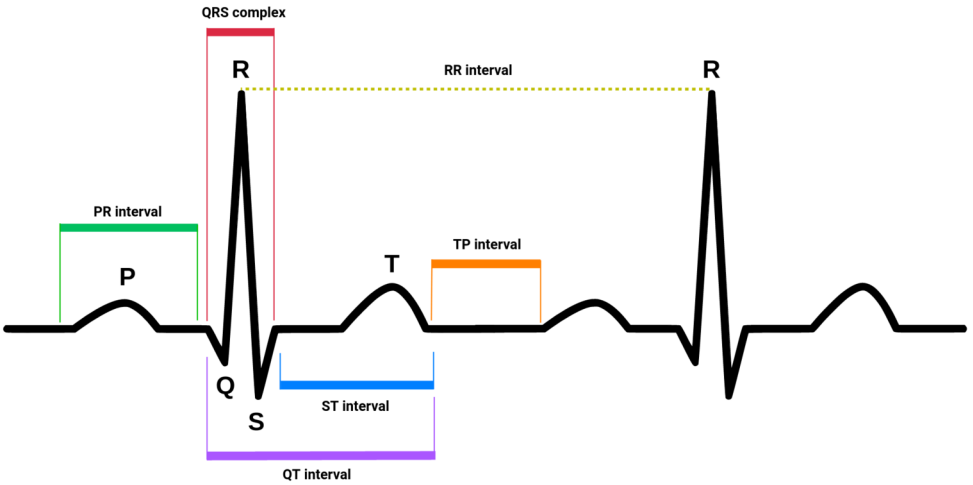

FIG. 2. *A typical ECG signal with its most common waveforms, where important points and intervals are marked.*

methods, such as support vector machines (SVMs; Cortes and Vapnik (1995)), random forest (Breiman (2001)) and gradient boosting (Friedman (2001)) can be used to implement ECG diagnosis under a supervised learning framework based on extracted features (Jambukia, Dabhi and Prajapati (2015)). However, conventional methods strongly depend on the quality of the manually-defined features which are limited by existing domain knowledge. Specifically, the manually-crafted morphological features may not be able to capture all predictive information in the original ECG signals (Bharti et al. (2021), Thygesen et al. (2007)). Moreover, it is also challenging to perfectly extract morphological features from ECG signals due to electrical noise caused by tray magnetic fields and accessories that vibrate (Elgendi (2013)). Therefore, certain biases may be introduced during feature engineering, thus hampering the accuracy of ECG diagnosis.

Recently, deep learning has garnered considerable success in ECG diagnosis. DL differs from conventional ML methods in directly fitting a neural network based on raw ECG signals without feature engineering to extract manual-crafted features. DL models have recently delivered superior performance in the classification of ECG diagnosis. For instance, existing convolutional neural networks (Attia et al. (2019), Ko et al. (2020), Rajpurkar et al. (2017)) achieved over 93% heartbeat classification accuracy. In contrast to conventional ML methods, DL models can effectively and adaptively extract the underlying information from raw data. Alternatively, the DL models may localize some novel discriminative features that even ECG experts may not be aware of nor can discern. However, despite their merits, DL models are often referred to as a black box, referring to the seeming mystery of their decision-making processes. The lack of interpretable features relevant to the prediction stands out as a significant barrier to the clinical use of their routine. Therefore, our primary goal is to develop a localization framework to unmask unknown discriminative features of black-box models to help bridge the bench-to-bedside gap and explore the domain knowledge of interpreting ECGs.

Discriminative feature localization for DL models is important but challenging. The major difficulties include: (i) Discriminative features are data-dependent on an input instance. For example, in the MNIST or ECG dataset the location of discriminative features may differ with inputs; see Figures 8 and 10. On this premise classical variable selection methods, based on tabular data, are unsuitable without modification; instead, it requires *data-adaptive* feature selection. (ii) A reliable statistical measure supported by theory is required to quantify *predictive importance* of any discriminative feature. Most existing methods are heuristic and fail to interpret the localized features. (iii) As indicated in Figure 3, the localized features should effectively explain the discrimination of different outcomes. Hence, *effectiveness* and *predictive importance* should be simultaneously considered for selecting sensible discriminative features.

1.2. *Prior work and our contributions*. Three major approaches have emerged for discriminative feature localization, including two-stage methods, feature-importance-based methods and backtracking methods. Specifically, two-stage methods use a simple explainable model, such as a local linear model, to approximate a complex black-box model and then to extract discriminative features. In particular, a method called local interpretable model-agnostic explanations (LIME) (Ribeiro, Singh and Guestrin (2016)) approximates a classification model by a local sparse linear model, based on a kernel smoother as in Davis, Lii and Politis (2011), then highlights those features with positive linear coefficients. Deep-Taylor (Montavon et al. (2017)) expands and decomposes a neural network output in terms of its input variables and generates a heatmap by back-propagating explanations from output to input. Feature-importance-based methods rank each feature's contribution by its importance based on an approximating model in a two-stage method. For example, SHAP (SHapley Additive exPlanations) (Lundberg and Lee (2017)) develops a kernel method integrating LIME

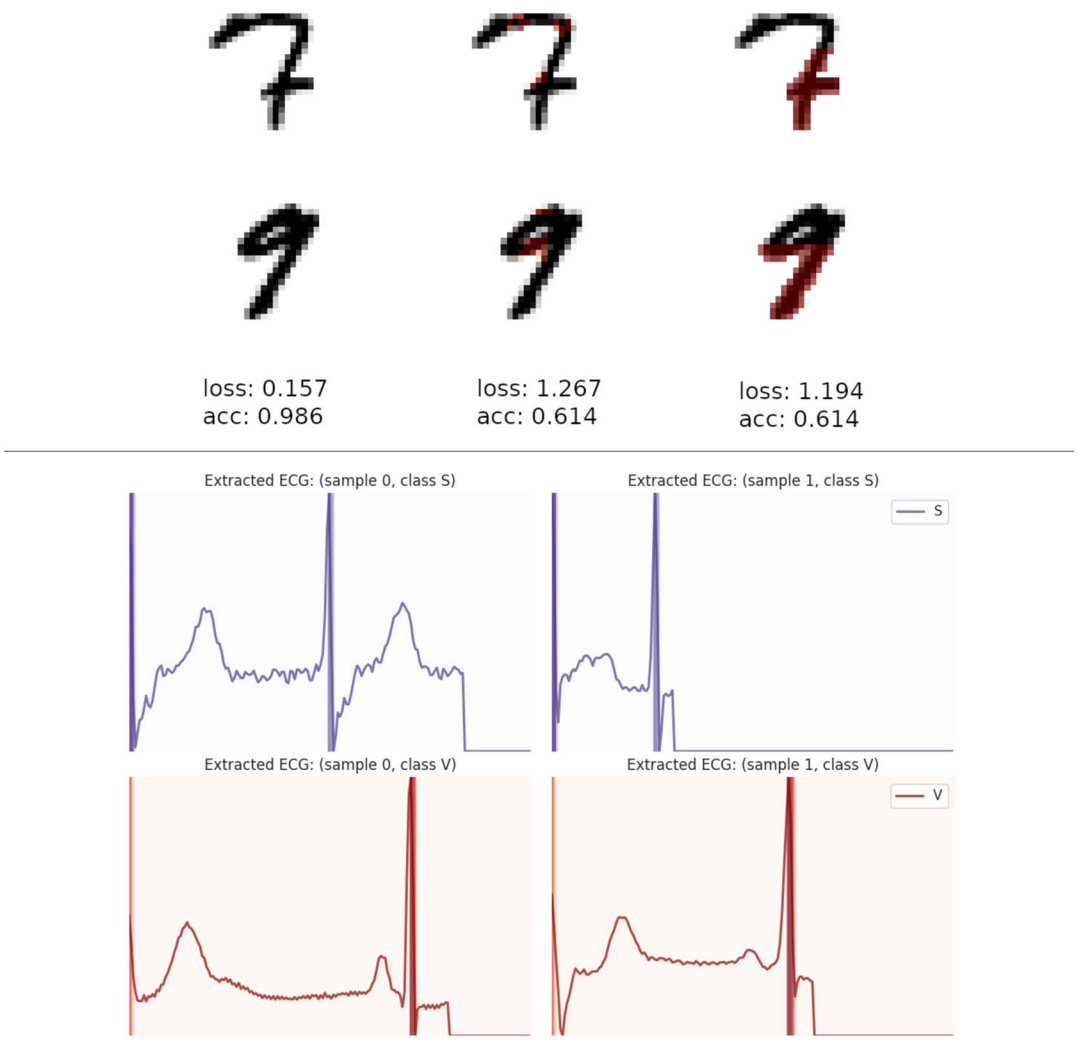

FIG. 3. *Examples to illustrate the concepts of* adaptiveness, predictive importance and effectiveness *of discriminative features on the MNIST and MIT-BIH data*. Upper panel. *The left represents raw images of digits* 7 *and* 9, *the middle represents images with localized pixels* (*marked in red*) *by the proposed method and the right represents images with localized pixels from row* 15 *to* 28. *Here "loss" and "acc" denote the cross-entropy and classification accuracy of a conventional neural network for each of the original and two disrupted dataset* (*by removing the localized regions of all images*). Lower panel. *The top and bottom each show two extracted ECGs and their localized regions from class "S"* (*blue*) *or "V"* (*red*) *in the MIT-BIH data. Specifically, the red/blue solid lines are the extracted ECG signals, and the highlighted vertical blue/red bars are localized regions by the proposed method. More discussion can be found in Sections* 5 *and* 6.

with the SHAP-value as the kernel weights and feature importance to quantify the contribution of features in an approximating local linear model. The backtracking methods map the activation layers of a neural network back to the input feature space, identifying which input patterns contribute more to prediction. In particular, Zhou et al. (2016) uses the global average pooling (GAP) together with class activation mapping (CAM) at the last layer of a convolutional neural network (CNN). Then it backtracks discriminative regions at the previous convolutional layers to the predicted scores. Gradient-CAM (Selvaraju et al. (2017)) extends GAP to a general CNN model by computing the gradient of a decision score concerning the feature activation maps of a convolutional layer. DeconvNet (Zeiler and Fergus (2014)) and Layerwise Relevance Propagation (LRP) (Bach et al. (2015)) perform backtracking with a deconvolution and conservative relevance redistribution, respectively. Finally, Patternnet

(Kindermans et al. (2017)) identify discriminative features by localizing the signal and noise directions for each neuron of a neural network.

Despite their merits, issues remain. First, a two-stage approach does not directly interpret an original model since discriminative features are localized by a simple approximation. For example, discriminative features generated by a linear approximation model (Lundberg and Lee (2017), Ribeiro, Singh and Guestrin (2016)) may be neither discriminative nor interpretable in the original model. Second, most existing methods are heuristic. As argued in Tjoa and Guan (2019), an intermediate backtracking process for GAP, Gradient-CAM and LRP are not amenable to scrutiny. Moreover, DeconvNet and LRP fail to produce a theoretically correct explanation, even for a linear model (Kindermans et al. (2017)). Finally, the above methods usually provide a dense representation of discriminative features, as suggested in Figure 9, yielding less effective interpretation.

There are three key contributions of our work in this paper:

- We propose a generalized partial R^2 in Definition 2.1 to quantify the degree of *predictive importance* of discriminative features so that they can be interpreted similarly, as in classical statistical analysis.
- The proposed localization framework (5) is able to simultaneously consider both *predictive importance* and *effectiveness*. Specifically, as illustrated in Figures 7 and 10, it provides a *flexible* framework to localize discriminative features corresponding to a certain amount of accuracy, as measured by an R^2.
- Through numerical experiments in Section 5 (the MNIST dataset), the localized discriminative features not only confirm the visual intuition but also are more efficient than the other existing methods. The numerical experiments in Section 6 suggest that localized ECG features are biologically plausible and consistent with cardiac electrophysiological principles, while locating subtle anomalies in sinus rhythm that may not be discernible by the naked eyes.

2. Generalized partial R^2 for discriminative localization.

In this section we introduce generalized partial R^2 to quantify the degree of *predictive importance* of discriminative features.

2.1. *Motivation*.

In a learning paradigm, a prediction function d is trained to predict an outcome Y for a given instance $\boldsymbol{X}$, where $\boldsymbol{X} = (X_1, \ldots, X_p)^\intercal$ is a p-dimensional continuous feature vector. Without loss of generality, each feature component X_j is rescaled to $[0, 1]$. For example, in the MNIST dataset, $\boldsymbol{X}$ is a gray-scale image, and Y is its associated digit label (LeCun and Cortes (2010)). To assess the performance, a loss function $L(\cdot, \cdot)$ is used, such as the cross-entropy loss $L(d(\boldsymbol{X}), Y) = -\mathbf{1}_Y^\intercal \log(\operatorname{softmax}(d(\boldsymbol{X})))$, where $\mathbf{1}_Y$ is the one-hot encoding of Y and $\operatorname{softmax}(\boldsymbol{z}) = (\operatorname{softmax}(\boldsymbol{z})_1, \ldots, \operatorname{softmax}(\boldsymbol{z})_p)^\intercal$ with $\operatorname{softmax}(\boldsymbol{z})_i = \exp(z_i)/\sum_j \exp(z_j)$.

Our goal is to identify discriminative features that effectively disrupt or deteriorate the prediction performance of a given learner d. To proceed, we highlight three distinctive characteristics of discriminative features motivated from real applications, namely, *adaptiveness*, *predictive importance* and *effectiveness*. As an illustrative example based on the MNIST dataset, consider two localized feature sets in the left panel of Figure 3. The feature set removed in the middle or right panel decreases the predictive accuracy of d by the same amount from 0.986 to 0.614 which suggests that the discriminate features should contribute largely to the predictive performance of d. Moreover, with the same amount of deterioration of performance, the highlighted features in the middle panel appear more compact, which we call more *effective* in the sequel, and thus more preferred as discriminative features. Furthermore, key characteristics are also captured by the MIH-BIH data. In particular, the amplitudes and locations

of the QRS complexes (Kusumoto (2020)) as well as of P and T waves, varying across ECG signals even of the same class, dictate that the discriminative features should be *adaptive* to the input ECG signals, as shown in the right panel of Figure 3. Note that the QRS complex corresponds to the spread of a stimulus through the ventricles and is usually the most visually important part of an ECG tracing (Kusumoto (2020)). Moreover, ion channel aberrations and structural abnormalities in the ventricles can affect electrical conduction in the ventricles (Rudy (2004)), manifesting with *subtle* anomalies in the QRS complex in sinus rhythm that may not be discernible by the naked eyes, yielding sparse or *effective* discriminative features.

In summary, three distinctive characteristics of discriminative features are desired:

- *Adaptiveness*. Discriminative feature extraction has to be *adaptive* to an input instance and a specific learner d. For example, in the MNIST/MIT-BIH dataset the location of discriminative features may differ with input images/signals.
- *Predictive importance*. The prediction accuracy of a learner d would significantly deteriorate without discriminative features. Alternatively, discriminative features can explain a large proportion of its predictive performance.
- *Effectiveness*. Discriminative features should effectively describe the discrimination of the outcome. Therefore, under the same *predictive importance*, the number/amount of localized discriminative features should be as small as possible. For example, compact localized pixels in the MNIST dataset or compact and accurate location of QRS complexes of ECG signals in the MIT-BIH ECG dataset.

To address *adaptiveness*, we introduce a *localizer* $\boldsymbol{\delta}(\boldsymbol{x}) = (\delta_1(\boldsymbol{x}), \ldots, \delta_p(\boldsymbol{x}))^\intercal : \mathbb{R}^p \to \mathbb{R}^p$ to produce a disruption adaptively based on an instance $\boldsymbol{x}$ to yield disrupted features $\boldsymbol{x}_{\boldsymbol{\delta}} = \boldsymbol{x} - \boldsymbol{\delta}(\boldsymbol{x})$. Without loss of generality, assume that each $|\delta_j(\boldsymbol{x})| \leq 1$ because x_j is rescaled to be in $[0, 1]$. In practice, the restriction $|\delta_j(\boldsymbol{x})| \leq 1$ is usually met by construction, for example, in an autoencoder in image classification; see Section 3.2 for illustration.

2.2. *Generalized partial R^2.* To measure the degree of *predictive importance* of a localizer $\boldsymbol{\delta}(\cdot)$, we introduce a generalized partial R^2 which mimics the partial R^2 in regression (Nagelkerke (1991)) and McFadden's R^2 (McFadden et al. (1973)) in classification. Specifically, the main idea of the partial R^2 is one minus the ratio of the full-model risk to the partial-model risk. On this ground we generalize the partial R^2 to black-box models in Definition 1.

DEFINITION 2.1 (Generalized partial R^2). Given a predictive model d, we define the generalized partial R^2 based on a localizer $\boldsymbol{\delta}(\cdot)$ as

$$R^2(d, \boldsymbol{\delta}) = 1 - \frac{\mathbb{E}(L(d(\boldsymbol{X}), Y))}{\mathbb{E}(L(d(\boldsymbol{X}_{\boldsymbol{\delta}}), Y))}. \tag{1}$$

If $R^2(d, \boldsymbol{\delta}) \geq r^2$, we say that the localized features by $\boldsymbol{\delta}(\cdot)$ is r^2-discriminative.

The generalized partial R^2 is one minus the proportion of the risk on full features $\boldsymbol{X}$ over that of the disrupted features $\boldsymbol{X}_\delta = \boldsymbol{X} - \boldsymbol{\delta}(\boldsymbol{X})$. It is a natural and clear criterion to extend the classical R^2 and to measure the *predictive importance* of the features disrupted by a localizer. Specifically, a higher R^2 yields stronger *predictive importance* of the localized discriminative features. When $\boldsymbol{\delta}(\boldsymbol{x})$ does not affect the performance of d, that is, $\mathbb{E}(L(d(\boldsymbol{X}_{\boldsymbol{\delta}}), Y)) = \mathbb{E}(L(d(\boldsymbol{X}), Y))$, or $R^2(d, \boldsymbol{\delta}) = 0$, the localized features contain no information for prediction. On the other hand, $r^2_{\max} = \max_{\boldsymbol{\delta}} R^2(d, \boldsymbol{\delta})$ the largest R^2 among all possible localizers, gives an upper bound of R^2. For instance, a localizer with each $\delta_j(\boldsymbol{x}) = \boldsymbol{x}_j$ disrupts extremely by removing all features which forces a learner d to predict without features. In general, $0 < R^2 < r^2_{\max}$ indicates the percentage of performance explained by $\boldsymbol{\delta}$.

3. Methods. Our main idea of identifying effective discriminative features is to seek a localizer $\boldsymbol{\delta}(\boldsymbol{x})$ yielding the most effective disruption of the features to reduce the prediction accuracy of a learner d.

3.1. *A discriminative localization framework.* In Figure 3 the r^2-discriminative localizer in the right panel is ineffective, although it also affects the same amount of prediction accuracy. Therefore, discriminative features should have an *effective* (or compact) representation, in addition to their contribution to a learner's prediction accuracy.

To achieve this goal, we introduce an activity L_1-regularizer $J(\boldsymbol{\delta})$ to quantify the *effectiveness* of a localizer,

$$J(\boldsymbol{\delta}) = \sup_{\boldsymbol{x}} \sum_{j=1}^{p} \big|\delta_j(\boldsymbol{x})\big|. \tag{2}$$

The benefits of this regularizer are twofold. First, it coincides with greedy feature selection results, as indicated in Appendix A (Dai et al. (2023)). Second, the supremum in (2) makes the localized features more balanced over an entire sample, as suggested in Section 5. Moreover, we specify $\|\boldsymbol{\delta}(\boldsymbol{x})\|_\infty \le 1$, for any $\boldsymbol{x}$, to control the magnitude of the disruption. This requirement can be trivially satisfied, for instance, using the proposed truncated rectified linear unit (TReLU) or Tanh as an activation function in the output layer of any deep neural network; see (9) in Section 3.2.

Next, we define an effective r^2-discriminative localizer $\boldsymbol{\delta}^0$ as the one minimizing $J(\boldsymbol{\delta})$ among all r^2-discriminative localizers. Then $\boldsymbol{\delta}^0$ can be regarded as an optimal localizer for identifying discriminative features to *interpret* a learner's predictability through *effective* disruption.

DEFINITION 3.1 (Effective $\boldsymbol{r}^2$-discriminative). For $0 \le r^2 \le r^2_{\max}$, an effective r^2-discriminative localizer to d is defined as

$$\boldsymbol{\delta}^0 \in \operatorname*{argmin}_{\boldsymbol{\delta}\in\mathcal{H}_b: R^2(d,\boldsymbol{\delta})\ge r^2} J(\boldsymbol{\delta}), \tag{3}$$

where $\mathcal{H}_b$ is a candidate collection of localizers such that $\sup_{\boldsymbol{x}} \|\boldsymbol{\delta}(\boldsymbol{x})\|_\infty \le 1$, and we say that the localized features by $\boldsymbol{\delta}^0(\cdot)$ is effective r^2-discriminative.

As noted in Definition 3.1, $\boldsymbol{\delta}^0$ is a most effective localizer that minimizes the regularization $J(\cdot)$ among all r^2-discriminative localizers. Without loss of generality, we assume that $\boldsymbol{\delta}^0$ always exists[1] but may not be unique in the sequel. Note that, in the presence of multiple global minimizers in (3), each of them could be useful, since our goal is to estimate such an effective r^2-discriminative localizer.

To identify an effective discriminative localizer for a learner d, we maximize $R^2(d,\boldsymbol{\delta})$ or the prediction risk $\mathbb{E}(L(d(\boldsymbol{X}-\boldsymbol{\delta}(\boldsymbol{X})),Y))$ with respect to $\boldsymbol{\delta}$ under the restriction of $J(\boldsymbol{\delta})$. This leads to our proposed framework,

$$\max_{\boldsymbol{\delta}\in\mathcal{H}_b} \mathbb{E}\big(L\big(d\big(\boldsymbol{X}-\boldsymbol{\delta}(\boldsymbol{X})\big),Y\big)\big), \quad \text{subject to } J(\boldsymbol{\delta}) \le \tau, \tag{4}$$

where $\tau > 0$ is a tuning parameter to balance the objective of deteriorating the prediction performance and magnitude of a localizer $\boldsymbol{\delta}(\cdot)$. To make the constraint sensible, we let $\tau \le p$ since $\sup_{\boldsymbol{\delta}\in\mathcal{H}_b} J(\boldsymbol{\delta}) = p$. As shown in Lemma 3.2, a most effective r^2-discriminative localizer $\boldsymbol{\delta}^0$ can be identified by (4).

[1] Otherwise, the definition can be adapted to an ε-global minimizer, where the difference between its minimum value and the global minimum is no less than or equal to ε.

LEMMA 3.2. *Let $\boldsymbol{\delta}_\tau^0$ be a global maximizer of (4), and*

$$\tau^0 = \min\{\tau \in (0, p] : R^2(d, \boldsymbol{\delta}_\tau^0) \geq r^2\};$$

then $\boldsymbol{\delta}_{\tau^0}^0$ is an effective r^2-discriminative localizer with $J(\boldsymbol{\delta}_{\tau^0}^0) = \tau^0$.

Lemma 3.2 says that (4) recovers an effective r^2-discriminative localizer, defined in Definition 3.1, in a similar fashion as Fisher consistency in classification (Bartlett, Jordan and McAuliffe (2006), Lin (2004)).

Given a training sample $(\boldsymbol{x}_i, y_i)_{i=1}^n$, we propose an empirical risk function to estimate $\boldsymbol{\delta}_\tau^0$ and τ^0,

$$\max_{\boldsymbol{\delta} \in \mathcal{H}_b} L_n(d, \boldsymbol{\delta}) = \frac{1}{n} \sum_{i=1}^n L\big(d\big(\boldsymbol{x}_i - \boldsymbol{\delta}(\boldsymbol{x}_i)\big), y_i\big), \quad \text{subj to, } J(\boldsymbol{\delta}) \leq \tau. \tag{5}$$

Denote $\widehat{\boldsymbol{\delta}}_\tau$ as a maximizer of (5) for a given τ. In view of Lemma 3.2, our final estimate of $(\boldsymbol{\delta}_{\tau^0}^0, \tau^0)$ is

$$\widehat{\boldsymbol{\delta}}_{\widehat{\tau}} \text{ is a maximizer of (5)}, \quad \text{where } \widehat{\tau} = \min\big\{\tau \in (0, p] : R^2(d, \widehat{\boldsymbol{\delta}}_\tau) \geq r^2\big\}. \tag{6}$$

In practice, $\tau \in (0, p]$ is replaced by $\tau \in \boldsymbol{\tau}$, where $\boldsymbol{\tau}$ is the candidate set of the tuning parameter τ, as some grid points for positive real numbers, and the estimated R^2 is evaluated, based on an independent test sample $\mathcal{D}_{\text{test}} = (\boldsymbol{x}_i, y_i)_{i=n+1}^{n+m}$,

$$\widehat{R}^2(d, \widehat{\boldsymbol{\delta}}_\tau; \mathcal{D}_{\text{test}}) = 1 - \frac{\sum_{(\boldsymbol{x}, y) \in \mathcal{D}_{\text{test}}} L(d(\boldsymbol{x}), y)}{\sum_{(\boldsymbol{x}, y) \in \mathcal{D}_{\text{test}}} L(d(\boldsymbol{x} - \widehat{\boldsymbol{\delta}}_\tau(\boldsymbol{x})), y)}. \tag{7}$$

Taken together, we iteratively solve (5) for $\tau \in \boldsymbol{\tau}$ from the smallest to the largest via a grid search (Bergstra and Bengio (2012)), and it terminates once $\widehat{R}^2(d, \widehat{\boldsymbol{\delta}}_\tau; \mathcal{D}_{\text{test}})$ exceeds a prespecified target r^2.

3.2. *A convolutional autoencoder discriminative localizer.* The proposed framework (5) admits a general localizer, such as a deep neural network. In practice, a network architecture incorporating data structure would be preferred (Bengio (2012)). For example, for the image-to-image localization in the MNIST dataset or the sequence-to-sequence localization in the ECG dataset, convolutional autoencoder architectures are natural options to impose a "local smoothing" structure of the localized features. Therefore, this section illustrates the localizer $\boldsymbol{\delta}$ as a convolutional autoencoder. It is noted that the network architecture of a discriminative model sets a standard for designing a localizer's architecture.

Consider a localizer of the form $\boldsymbol{\delta}(\boldsymbol{x}) = \boldsymbol{x} \odot \boldsymbol{\pi}(\boldsymbol{x})$; $\boldsymbol{x}$ is an image, where $\odot$ is the elementwise product and $0 \leq \boldsymbol{\pi}(\boldsymbol{x}) \leq 1$ represents the percentage of image features that a localizer removes from the original feature $\boldsymbol{x}$.

Subsequently, we implement our proposed localizer by taking an image $\boldsymbol{x}$ as input and giving output as disruption proportion $\boldsymbol{\pi}(\boldsymbol{x})$. Specifically, we build a convolutional autoencoder discriminative localizer, based on a convolutional autoencoder network (CAE; Masci et al. (2011), Rumelhart, Hinton and Williams (1985)), which is composed of three components: Encoder-CNN (E-CNN), hidden neural network (HNN) and Decoder-CNN (D-CNN), as illustrated in Figure 4. Besides, on the CAE backend model we introduce a TReLU-softmax or Tanh+softmax activation function to control the activity L_1-regularizer of the localizer. On this ground we consider a localizer class,

$$\mathcal{H}_b^\tau = \big\{\boldsymbol{\delta}_\tau(\boldsymbol{x}) = \boldsymbol{x} \odot \boldsymbol{\pi}_{\boldsymbol{\theta}}^\tau(\boldsymbol{x}) : \boldsymbol{\pi}_{\boldsymbol{\theta}}^\tau(\boldsymbol{x}) = A_\tau\big(\text{CAE}_{\boldsymbol{\theta}}(\boldsymbol{x})\big); \boldsymbol{\theta} \in \boldsymbol{\Theta}\big\}, \tag{8}$$

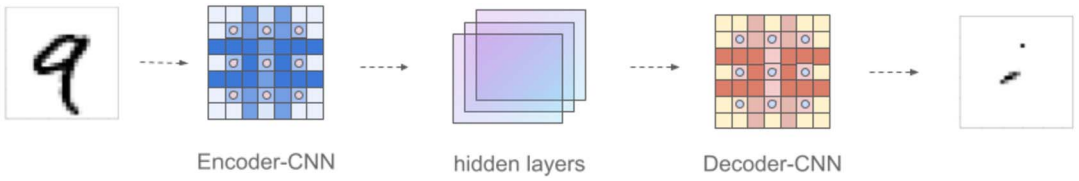

FIG. 4. *Our localization network structure, based on a convolutional autoencoder, which is composed of three components*: *Encoder-CNN* (*E-CNN*), *hidden neural network* (*HNN*) *and Decoder-CNN* (*D-CNN*).

where $\mathrm{CAE}_{\boldsymbol{\theta}}(\boldsymbol{x})$ is a convolutional autoencoder with $\boldsymbol{\theta} \in \mathbb{R}^q$ denoting its parameters, $\boldsymbol{\Theta}$ is a parameter space of $\boldsymbol{\theta}$ and $A_\tau(\cdot)$ is a structured activation function, such as

$$A_\tau(z) = \mathrm{TReLu}\big(\tau \cdot \mathrm{softmax}(z)\big), \quad \text{or} \quad A_\tau(z) = \mathrm{Tanh}\big(\tau \cdot \mathrm{softmax}(z)\big), \tag{9}$$

where $\mathrm{TReLu}(\boldsymbol{u}) = \min(\boldsymbol{u}_+, \mathbf{1})$ is the truncated ReLU function.

Note that for any $\boldsymbol{\delta} \in \mathcal{H}_b^\tau$, based on the definition of A_τ, the following conditions are automatically satisfied: (i) $\sup_{\boldsymbol{x}} \|\boldsymbol{\delta}(\boldsymbol{x})\|_\infty \le 1$; (ii) $J(\boldsymbol{\delta}) = \sup_{\boldsymbol{x}} \|\boldsymbol{\delta}(\boldsymbol{x})\|_1 \le \tau$. Therefore, the constraints in (5) can be removed, given $\mathcal{H}_b^\tau$, and the optimization of (5) becomes

$$\max_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} L\big(d\big(\boldsymbol{x}_i - \boldsymbol{x}_i \odot \boldsymbol{\pi}_{\boldsymbol{\theta}}^\tau(\boldsymbol{x}_i)\big), y_i\big) \tag{10}$$

which can be solved by gradient descent (GD) or stochastic gradient descent (SGD; Raginsky, Rakhlin and Telgarsky (2017)). The GD solution of (10) attains a local maximizer of (10) under some mild assumptions (Lee et al. (2016)). Note that the convergence result can be extended to SGD, as in (Ge et al. (2015)), and a global maximizer may be obtained by GD or SGD with additional assumptions (Raginsky, Rakhlin and Telgarsky (2017)). Once $\widehat{\boldsymbol{\theta}}$ is obtained, the estimated localizer is specified as

$$\widehat{\boldsymbol{\delta}}_\tau(\boldsymbol{x}) = \boldsymbol{x} \odot A_\tau\big(\mathrm{CAE}_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{x})\big). \tag{11}$$

3.3. *Interpretation uncertainty.* Robustness is a general challenge to existing interpretation approaches. For example, Ghorbani, Abid and Zou (2019) indicate that systematic perturbations can lead to dramatically different interpretations without changing the label. To distinguish the interpretability and robustness for the proposed framework, we propose an *unexplainable* R^2 as a confidence interval for the generalized partial R^2 to distinguish the prediction deterioration caused by discriminative features from model instability. In particular, given a learner d and a localizer $\widehat{\boldsymbol{\delta}}_\tau$, we construct a confidence interval for $R^2(d, \widehat{\boldsymbol{\delta}}_\tau)$ via bootstrap on a test sample.

First, we generate a bootstrap sample $\mathcal{D}_{\mathrm{test}}^{(b)}$ by drawing B independent observations from the test data $\mathcal{D}_{\mathrm{test}}$ with replacement. Then the unexplainable R^2 for $R^2(d, \widehat{\boldsymbol{\delta}}_\tau)$ is obtained using the sampling distribution of the bootstrapped estimates $\{\widehat{R}^2(d, \widehat{\boldsymbol{\delta}}_\tau; \mathcal{D}_{\mathrm{test}}^{(B)})\}_{b=1}^B$. For example, for the MNIST dataset we obtain a 95% confidence interval of $R^2(d, \widehat{\boldsymbol{\delta}}_\tau)$ by computing the $\lfloor 0.025B \rfloor$th and $\lfloor 0.975B \rfloor$th ordered estimated R^2 on the bootstrap samples, as indicated in Figure 5. More detail can be found in Section 5.

4. Theoretical guarantee. This section indicates that the proposed framework yields discriminative features attaining a target R^2 with optimal effectiveness asymptotically.

To proceed, let $\boldsymbol{\delta}_\tau^0$ be a global maximizer of (4) over a function class $\mathcal{H}_b = \{\boldsymbol{\delta} \in \mathcal{H} : \sup_{\boldsymbol{x}} \|\boldsymbol{\delta}(\boldsymbol{x})\|_\infty \le 1\}$. Without loss of generality, assume that $0 \le L(d(\boldsymbol{x}_{\boldsymbol{\delta}}), Y) \le U$ for a sufficiently large constant $U \ge 1$, for any $\boldsymbol{\delta} \in \mathcal{H}_b$ and $\boldsymbol{x} \in \mathbb{R}^p$ (Wu and Liu (2007)). To make the constraint sensible, we let $\tau \le p$ since $\sup_{\boldsymbol{\delta} \in \mathcal{H}_b} J(\boldsymbol{\delta}) = p$.

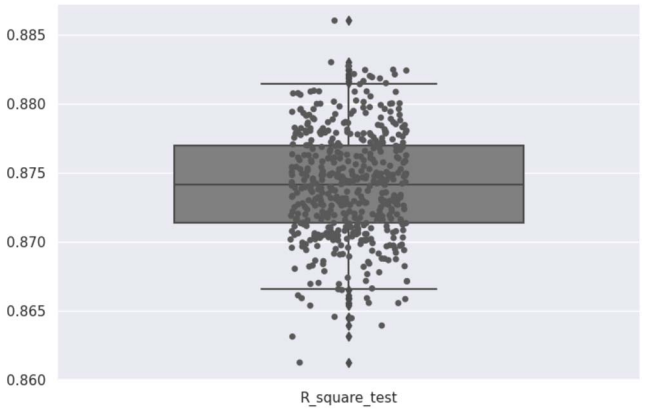

FIG. 5. *Boxplot of the estimated R^2 of the proposed method based on* 500 *bootstrap samples for the MNIST benchmark example. This example illustrates the concept of an unexplainable R^2.*

Denote the Rademacher complexity for the function class $\mathcal{H}_b$, as $\kappa_n = \mathbb{E}\mathcal{R}_n(\mathcal{H}_b) = \sup_{\boldsymbol{\delta}\in\mathcal{H}_b} n^{-1}\sum_{i=1}^n |\eta_i(L(d(\boldsymbol{X}_i - \boldsymbol{\delta}(\boldsymbol{X}_i)), Y))|$ and $\{\eta_i\}_{i=1}^n$ are i.i.d. Rademacher random variables with η_i taking the values $+1$ and -1 with probability 1/2 each. To make the constraint sensible, we let $\tau \le p$ since $\sup_{\boldsymbol{\delta}\in\mathcal{H}_b} J(\boldsymbol{\delta}) = p$. Theorem 4.1 gives a convergence rate for the discrepancy between $\boldsymbol{\delta}_\tau^0$ and $\widehat{\boldsymbol{\delta}}_\tau$ in terms of R^2 uniformly over $0 < \tau \le p$.

THEOREM 4.1 (Asymptotics of $\boldsymbol{R}^2$). *Let $\widehat{\boldsymbol{\delta}}_\tau$ be a global maximizer of* (5)*, for $\varepsilon_n \ge 8\kappa_n$ and any predictive model d, we have*

$$\mathbb{P}\Big(\sup_{0<\tau\le p} R^2(d,\boldsymbol{\delta}_\tau^0) - R^2(d,\widehat{\boldsymbol{\delta}}_\tau) \ge \varepsilon_n\Big) \le K\exp\Big(-\frac{n\varepsilon_n^2}{KU^2}\Big),$$

where $K > 0$ is a constant. Hence,

$$\sup_{0<\tau\le p}\big(R^2(d,\boldsymbol{\delta}_\tau^0) - R^2(d,\widehat{\boldsymbol{\delta}}_\tau)\big) = O_p\big(\max(\kappa_n, n^{-1/2})\big).$$

Note that the asymptotics of the Rademacher complexity κ_n for a candidate class $\mathcal{H}$ has been extensively investigated in the literature (Bartlett and Mendelson (2002), Bartlett, Bousquet and Mendelson (2005)). Therefore, the uniform convergence rate can be obtained for a generic candidate class by Theorem 4.1. Moreover, the asymptotics for a fixed τ is also provided in Appendix C, where the rate can be further improved.

Next, we show that $\widehat{\boldsymbol{\delta}}_{\widehat{\tau}}$ is an asymptotically effective r^2-discriminative localizer. Note that $\widehat{\boldsymbol{\delta}}_{\widehat{\tau}}$ already is an r^2-discriminative localizer, since $R^2(d,\widehat{\boldsymbol{\delta}}_{\widehat{\tau}}) \ge r^2$ by the definition of $\widehat{\tau}$ in (6). Therefore, it suffices to show *effectiveness*, that is, $|J(\widehat{\boldsymbol{\delta}}_{\widehat{\tau}}) - J(\boldsymbol{\delta}_{\tau^0}^0)| = |\widehat{\tau} - \tau^0| \xrightarrow{p} 0$. To proceed, we require a smoothness condition of $R^2(d,\boldsymbol{\delta}_\tau^0)$ over τ in Assumption A.

ASSUMPTION A (Smooth). Assume that $R^2(d,\boldsymbol{\delta}_\tau^0)$ is a continuous function in τ. Moreover, there exists a constant $\mu_0 > 0$ such that $|\tau_1 - \tau_2| \le \mu$ if $|R^2(d,\boldsymbol{\delta}_{\tau_1}^0) - R^2(d,\boldsymbol{\delta}_{\tau_2}^0)| \le c_0\mu^\alpha$ for any $\mu \le \mu_0$.

THEOREM 4.2 (Oracle property). *Let $\boldsymbol{\delta}^0$ be an effective r^2-discriminative localizer in Definition* 3.1 *and $\widehat{\boldsymbol{\delta}}_{\widehat{\tau}}$ be a global maximizer of* (6)*. Under Assumption* A*, for $\omega_n \ge 2(8\kappa_n/c_0)^{1/\alpha}$, we have*

$$\mathbb{P}(|\widehat{\tau} - \tau^0| \ge \omega_n) = \mathbb{P}(|J(\widehat{\boldsymbol{\delta}}_{\widehat{\tau}}) - J(\boldsymbol{\delta}^0)| \ge \omega_n) \le K'\exp\Big(-\frac{n\omega_n^{2\alpha}}{K'U^2}\Big),$$

where $K' > 0$ is a universal constant. Therefore,

$$\left|J(\widehat{\boldsymbol{\delta}}_{\widehat{\tau}}) - J(\boldsymbol{\delta}^0_{\tau^0})\right| = \left|\widehat{\tau} - \tau^0\right| \xrightarrow{p} 0,$$

and $\widehat{\boldsymbol{\delta}}_{\widehat{\tau}}$ is an asymptotically effective r^2-discriminative localizer.

Therefore, the proposed framework yields an effective r^2-discriminative localizer, as defined in (3.1), rendering theoretical reliable discriminative features for a target R^2. Moreover, the theorems are illustrated for the proposed convolutional autoencoder neural network (10) in Corollary B.1, where the convergence rates are computed depending on the sample size and the network architecture.

5. MNIST benchmark. This section examines the numerical performance and visualizes discriminative features generated from the proposed localizer for the MNIST handwritten digit dataset (LeCun and Cortes (2010)) (http://yann.lecun.com/exdb/mnist/). All empirical results are produced in our Python library **dnn-locate** (https://github.com/statmlben/dnn-locate).

For the MNIST data, we extract 14,251 images (28×28 field) from the dataset with labels "7" and "9." Our goal is to localize discriminative features for distinguishing digits "7" and "9" with a specific generalized partial R^2.

First, we train a decision function d as a CNN, where we regularize each parameter of the CNN by the L_1-norm with weight 0.001. Here the CNN model is optimized by the Adam algorithm with an initial learning rate of 0.001, early stopping based on the validation accuracy with patience as 10, and 20% of the training data as a validation set.

Then a convolutional autoencoder (CAE), as in (8) and Figure 4, is constructed as the localizer. For training we optimize the model by stochastic gradient descent with an initial learning rate of $10/\tau$ and reduce the learning rate by a factor of 0.382 (Bengio (2012)), when the validation loss has stopped improving. Moreover, early stopping is conducted based on validation accuracy with patience as 15 (Raskutti, Wainwright and Yu (2014)).

For the proposed method, we implement (10) based on $\tau = 4, 6, 8, 10, 12, 14, 18, 20$ and the relation between τ and its corresponding estimated R^2s are demonstrated in Figure 6.

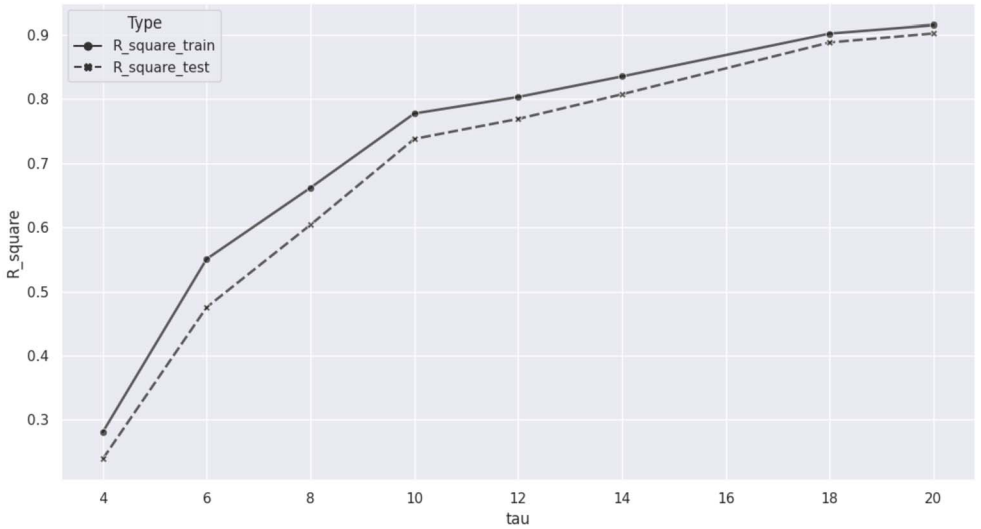

FIG. 6. *Training and testing estimated R^2s for the proposed framework in handwritten digit dataset with $\tau = 6, 8, 10, 12, 14, 16, 18, 20$ which indicates that the R^2 increases as the magnitude for an estimated localizer becoming large.*

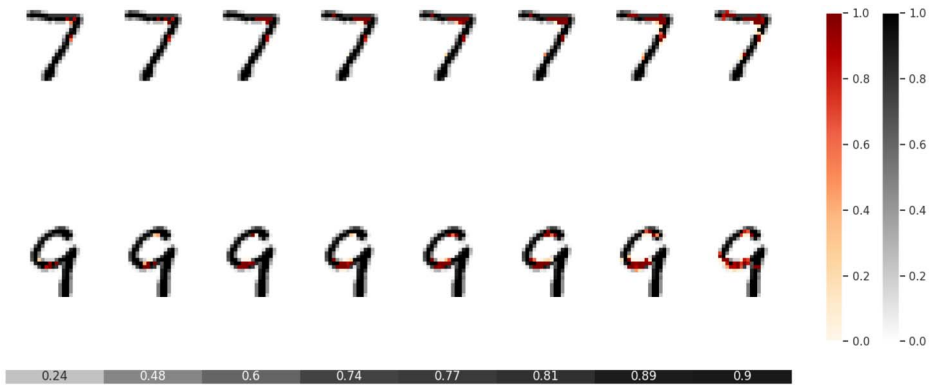

FIG. 7. *Illustrative instances of localized discriminative features* (*red*) *by the proposed method for "7" and "9" digits* (*black*) *as well as their corresponding estimated* R^2*s* (*the heatmap in x-axis*). *The gray color bar indicates gray scale of original images*, *and the red color bar indicates the proportion of removing features*, *that is*, $\boldsymbol{\pi}(\boldsymbol{x})$ *in* (10).

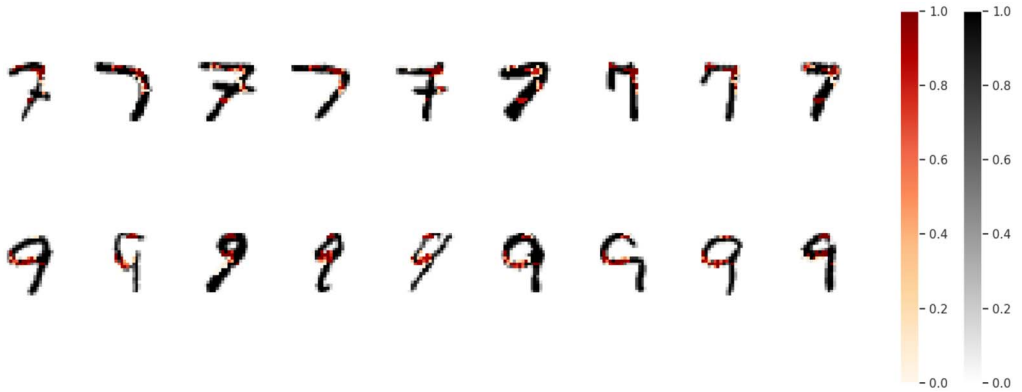

FIG. 8. *Data-adaptive localized discriminative features* (*red*) *for the proposed method with* $\tau = 20$ *based on different "7" and "9" digits* (*black*). *The gray color bar indicates gray scale of original images*, *and the red color bar indicates the proportion*.

Note that the estimated R^2 increases as the activity L_1-norm of the localizer becoming large. Furthermore, the discriminative features, identified by the proposed method for two illustrative instances of "7" and "9," are visualized in Figure 7. Specifically, as the estimated R^2 becomes larger, the disrupted instance labeled as "9" becomes more and more like "7."

As illustrated by the boxplot (Figure 5), a 95% confidence interval [0.867, 0.882] for the $R^2(d, \widehat{\boldsymbol{\delta}}_\tau)$ indicates some uncertainty with the fitted localizer ($\tau = 17$), where the R^2 is categorized as unexplainable if it falls inside the confidence interval.

Next, we compare the proposed method with five state-of-the-art methods by both human visual and numerical evaluations, including deep Taylor explainer (Montavon et al. (2017)), gradient-based explainer (Selvaraju et al. (2017)), lrp.z (Bach et al. (2015)), DeconvNet (Zeiler and Fergus (2014)) and pattern.net (Kindermans et al. (2017)). All competitors are implemented by the Python library `innvestigate` (https://github.com/albermax/innvestigate), and the `batch size` is set as 64 for pattern.net. In particular, a heatmap of discriminative features produced by each method is validated by a visual inspection and by a numerical comparison based on the estimated R^2, given the same amount/magnitude of feature disruption.

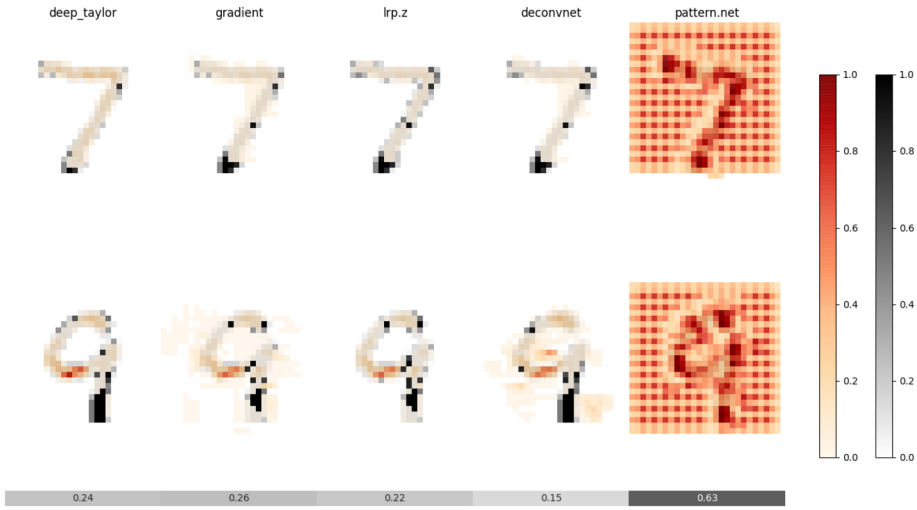

FIG. 9. *Illustrative instances of localized discriminative features* (*red*), *based on five competitors*, *for "7" and "9" digits* (*black*), *and their corresponding generalized partial* R^2*s* (*the heatmap in x-axis*). *The gray color bar indicates gray scale of original images*, *and the red color bar indicates importance of pixels produced by a localizer.*

5.1. *Visual comparison*. As displayed in Figure 7, the proposed method produces more compact discriminative features. By comparison, the other competitors yield dense image features spreading over the entire digits. Moreover, the proposed method gives roughly equal attention to two images in discriminating digits "7" from "9" which conforms with human intuition. However, as depicted in Figure 9, all competitors generate imbalanced discriminative features that are more in one of the two images of "7" and "9," as shown in Figure 9. As a result, the proposed method is more conducive for label-specific analysis.

5.2. *Numerical comparison*. To make a fair comparison, we conduct a pairwise comparison between the proposed localizer and each competitor under the same magnitude of $J(\cdot)$. Specifically, we compute the value of $J(\cdot)$ and the estimated R^2 of detected regions by a competitor. To fairness, we chose our tuning parameter τ to be the same as the $J(\cdot)$ of the competitor. Then compare the R^2s for the proposed method and the corresponding competitor.

As indicated in Table 1, under the same magnitude $J(\cdot)$ the proposed localizer outperforms all competitors in terms of R^2, where the amounts of improvement are 58.47%, 147.1%, 146.5%, 308.0% and 44.14%.

In summary, the proposed method has significant benefits. First, as illustrated in Figure 7, it provides a flexible framework to localize desirable discriminative features to explain a certain

TABLE 1

A pairwise comparison for the proposed framework and five existing methods based on 10*-fold cross validation. Here* $J(\cdot)$ *is the activity* L_1*-regularizer, as defined in* (2), *and the estimated* R^2, *as in* (7)

	Activity L_1-norm $J(\cdot)$	$\widehat{R}^2$ (competitor in the first column)	$\widehat{R}^2$ (our method)
deep-Taylor	11.698(0.228)	0.236(0.016)	0.374(0.084)
gradient	26.028(0.319)	0.289(0.012)	0.714(0.033)
lrp.z	14.689(0.219)	0.256(0.014)	0.631(0.077)
DeconvNet	27.832(0.955)	0.175(0.015)	0.714(0.023)
pattern.net	374.709(2.762)	0.648(0.006)	0.934(0.001)

amount of predictive performance as measured by an R^2. Second, the visual and numerical results in Figures 7 and 9 and Table 1 suggest that the proposed method can produce compact and effective discriminative features which are consistent with human visual judgment.

6. ECG data analysis. Finally, we present the results of applying our method to the MIT-BIH arrhythmia electrocardiogram (ECG) dataset for heartbeat classification (Moody and Mark (1990)). The MIT-BIH dataset consists of ECG recordings from 47 different subjects recorded at the sampling rate of 360 Hz by the BIH Arrhythmia Laboratory. Each beat is annotated into five different classes by following the Association for the Advancement of Medical Instrumentation (AAMI) EC57 standard: labeled as "N, S, V, F" and "Q." The preprocessed dataset is publicly available at https://www.kaggle.com/shayanfazeli/heartbeat. The MIT-BIH ECG dataset has been extensively studied, including using deep convolutional neural networks (Kachuee, Fazeli and Sarrafzadeh (2018), Martis et al. (2013), Acharya et al. (2017)). In spite of the impressive predictive performance obtained by the devised networks (with more than 93% classification accuracy), it is unknown why and how the networks achieved their good performance. To advance our understanding and possibly offering new insights, our goal is to localize discriminative signal fragments, based on the deep CNN developed in Kachuee, Fazeli and Sarrafzadeh (2018), which is one of the state-of-the-art ECG classification methods.

For implementation we build a localizer by using a convolutional autoencoder structure in Figure 4 with two convolutional layers as an encoder and two transposed convolution layers as a decoder. For training, we use the SGDW optimizer with "`learning_rate=.1`, `weight_decay=1e-4`,'`momentum=.9`." Besides, a reducing learning rate scheme is used with "`factor=.382`" and "`patience=3`," and early stopping is adopted with "`patience=20`." Moreover, we tune the hyperparameter τ to achieve various R^2s: 10%, 50%, 60%, 70%, 75%. Training one network takes less than half an hour on a GeForce GTX 2060Ti GPU. All the Python codes are publicly available in https://github.com/statmlben/dnn-locate.

To demonstrate our localization results, we concentrate on the localized ECG signals under the label "S" (including atrial premature, aberrant atrial premature, nodal premature, and supra-ventricular premature) and the label "V" (including premature ventricular contraction, and ventricular escape).

As shown in the lower panel of Figure 10, the localized regions (highlighted by the red bars) of ECG complexes in sinus rhythm are most informative in distinguishing presence of ventricular ectopic beats from supraventricular ectopic beats in a particular individual. The localized regions lie in the QRS complex which correlates with ventricular depolarization or electrical propagation in the ventricles (Mirvis and Goldberger (2001)). Ion channel aberrations and structural abnormalities in the ventricles can affect electrical conduction in the ventricles (Rudy (2004)), manifesting with subtle anomalies in the QRS complex in sinus rhythm that may not be discernible by the naked eye but is detectable by the convolutional autoencoder. Of note, as the R^2 increases from 10% to 88%, the highlighted color bar is progressively broader, covering a higher proportion of the QRS complex. The foregoing observations are sensible: the regions of interest resided in the QRS complex are biologically plausible and consistent with cardiac electrophysiological principles.

As shown in the upper panel of Figure 10, similarly, the regions of interest (highlighted by the blue bars) of ECG complexes in sinus rhythm are most informative in distinguishing the presence of supraventricular ectopic beats from ventricular ectopic beats in a particular individual. As in the left panel, the regions of interest lies in the QRS complex, which is intuitive and biologically plausible, as explained above.

As shown in the last three figures in the upper panel of Figure 10 for supraventricular complexes, as the R^2 increases from 80% to 84% and finally 88%, the blue bar progressively

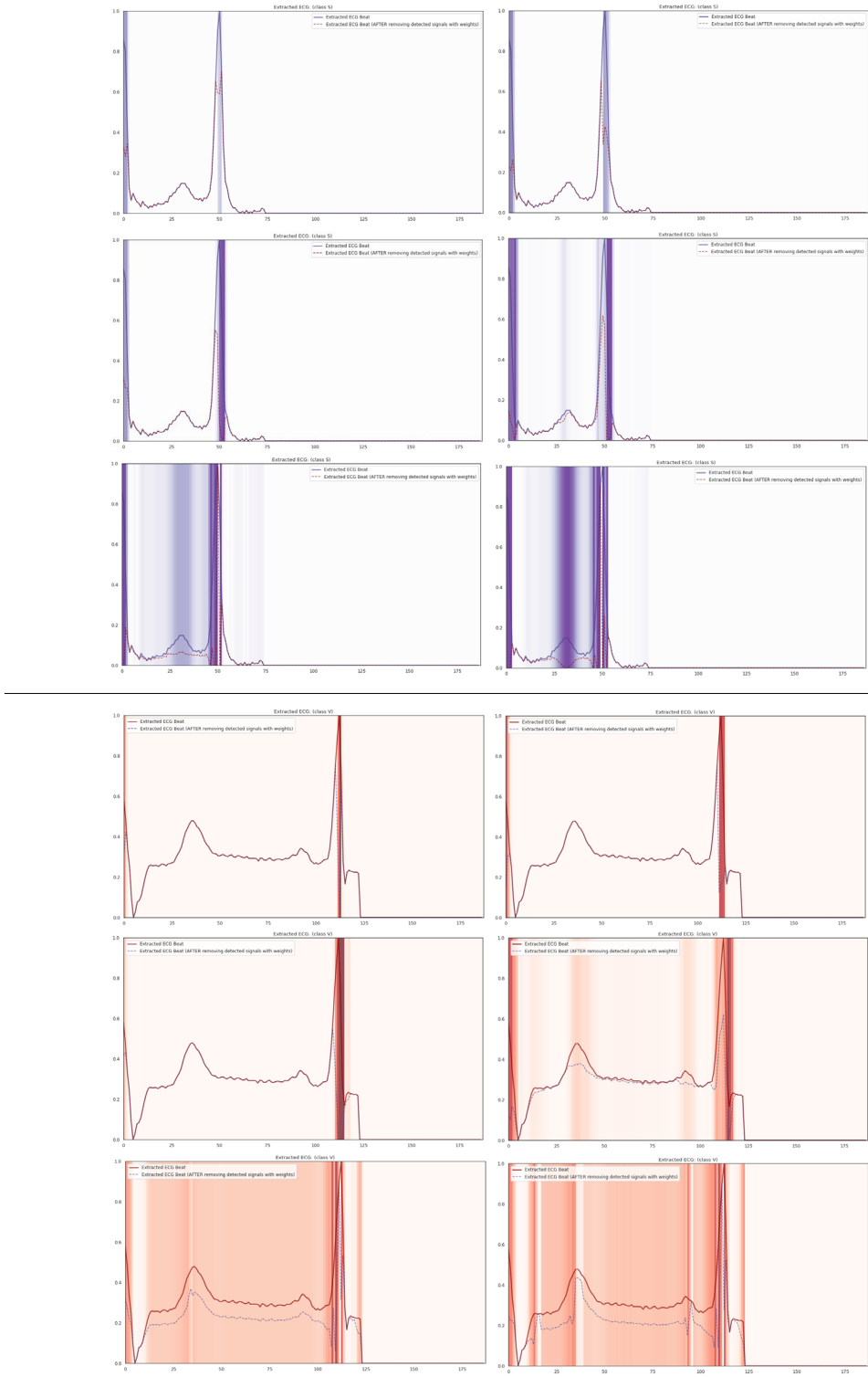

FIG. 10. *The proposed method reveals subtle discriminative features used by a deep convolutional neural network for ECG classification on the MIT-BIH dataset. The left/right panel highlights the localized features for a particular individual under label 'V'/'S'. Note that the* R^2 *increases from* 10% *to* 88% *for both upper and lower panels. The medical literature provided in Section* 6 *gives supporting evidence for biological plausibility of the localized features*; *see more discussion in Section* 6.

TABLE 2
R^2s for the proposed framework with different network architectures. Here "CAE" indicates a convolutional autoencoder, "MLP" indicates a multilayer perceptron and the estimated R^2 is computed, as in (7), based on 10-fold cross-validation

CAE64	CAE128	CAE256	MLP512	MLP1024	MLP2048
0.816(0.028)	0.872(0.011)	0.872(0.015)	0.141(0.252)	0.156(0.255)	0.133(0.242)

highlights the P wave of ECG complexes in sinus rhythm. This observation is consistent with our understanding of the mechanistic underpinnings of atrial depolarization which correlates with the P wave. Ion channel alterations and structural changes in the atria can affect electrical conduction in the atria (Rudy (2004)), manifesting with subtle anomalies in the P wave in sinus rhythm that may not be discernible by the naked eye but are detectable by the convolutional autoencoder.

Collectively, the examples above underscore the fact that the discriminative regions of interest identified by our proposed method are biologically plausible and consistent with cardiac electrophysiological principles while locating subtle anomalies in the P wave and QRS complex that may not be discernible by the naked eye. By inspecting our results with an ECG clinician (Dr. Chen in the authorship), the localized discriminative features of the ECG are consistent with medical interpretation in ECG diagnosis.

6.1. *Robustness against localization network architecture.* This section examines the robustness of the proposed framework against network architectures. We use the same implementation configuration with $\tau = 0.05$ and examine CAE network architectures with different numbers of neurons, denoted as `CAE64`, `CAE128`, `CAE256` and `CAE512`, where `CAE64` is constructed as: Conv1D(64)+Conv1D(32)+Conv1DTranspose(32)+Conv1DTranspose(64), and other CAE networks are defined likewise. Moreover, we also implement a localizer with a multilayer perceptron (MLP) structure: `MLP256`, `MLP512`, `MLP1024` and `MLP2048`. For example, `MLP256` is constructed as: Dense(256)+Dense(128)+Dense(64)+Dense(187), and other MLP networks are defined likewise. As indicated in Table 2, R^2s of the localized discriminative features provided by convolutional autoencoders are significantly higher and more stable than those produced by MLPs. In particular, for CAE-based networks larger networks generally improve the performance. The localization results by the CAE networks are illustrated in Figure 11: the localized discriminative features are fairly consistent with different CAE-based network architectures.

7. Discussion. XAI methods have gained prominence in many scientific domains, for example, medical diagnostics which require both interpretability and predictive accuracy. To identify discriminative features, we quantify the quality of interpretability by a generalized partial R^2 while measuring the interpretation effectiveness by an activity L_1-norm. On this ground we construct a localizer by disrupting the original features and seek a localizer yielding the most deteriorated performance of a learner while having the smallest activity norm for minimal feature disruption. Theoretically, we show that the proposed localization method identifies discriminative features asymptotically. Moreover, we apply the proposed framework to the MNIST and MIT-BIH ECG datasets to interpret a learning outcome of a convolutional autoencoder neural network. Numerical results suggest that the proposed localizer compares favorably with state-of-the-art competitors in the literature while identifying discriminative regions that are not only visually/biologically plausible but also concise. Furthermore, it is of interest to know if any localized features are genuinely important for which hypothesis testing targeting a data-adaptive localizer, as in Dai, Shen and Pan (2022), would be needed as a possible extension of our framework.

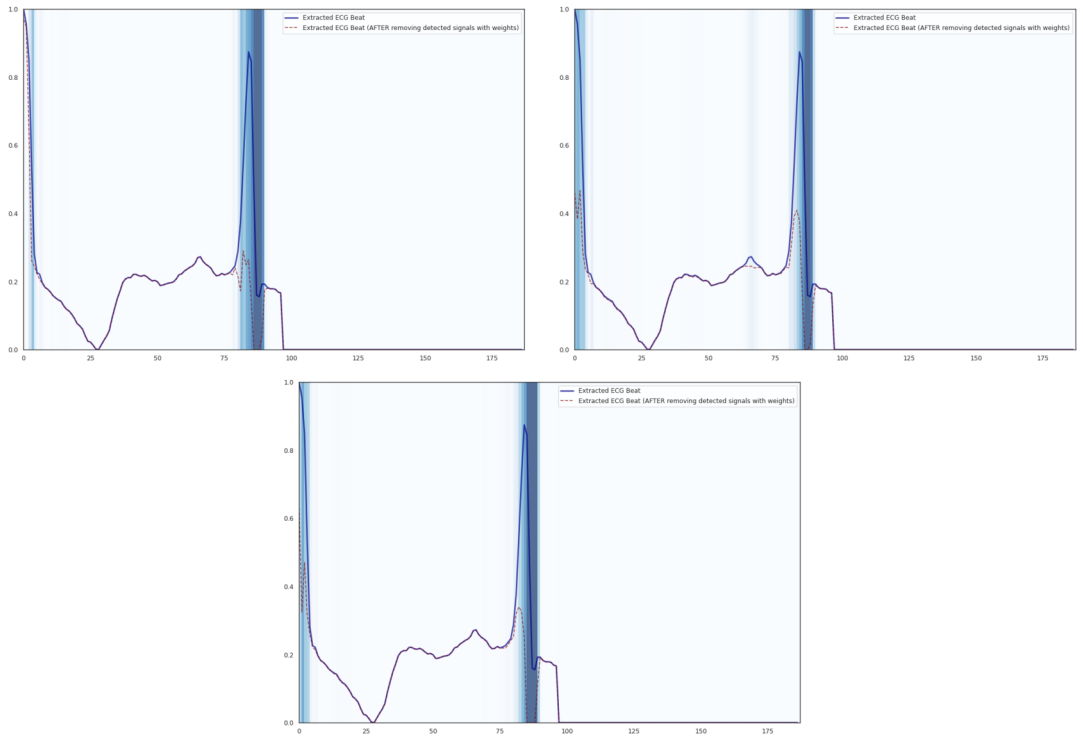

FIG. 11. *The localized discriminative features of one ECG signal in the MIT-BIH dataset based on the proposed framework with different CAE network architectures*: `CAE64` - `CAE256`.

Acknowledgments. The corresponding authors for this work are Ben Dai and Wei Pan. The authors would like to thank the referees, the Associate Editor and the Editor for the constructive feedback which greatly improved this work.

Funding. We would like to acknowledge support for this project from RGC-ECS 24302422, the CUHK direct grant, NSF DMS-1712564, DMS-1721216, DMS-1952539 and NIH grants R01GM126002, R01AG069895, R01AG065636, R01AG074858, R01AG074858, U01AG073079 and RF1 AG067924.

SUPPLEMENTARY MATERIAL

Supplement to "Data-adaptive discriminative feature localization with statistically guaranteed interpretation" (DOI: 10.1214/22-AOAS1705SUPPA; .pdf). The supplementary materials consist of: Appendix A indicates that the proposed framework incorporates greedy feature selection for a linear regression model and a piecewise linear regression model; Appendix B provides details of assumptions and asymptotic results for the proposed framework; Appendix C refines the asymptotic results of the proposed framework based on a fixed τ; Appendix D provides the technical proofs.

Python package dnn-locate (DOI: 10.1214/22-AOAS1705SUPPB; .zip). The Python package **dnn-locate** is available in PyPi (https://pypi.org/project/dnn-locate/). For the most recent version of the package, see https://github.com/statmlben/dnn-locate.

REFERENCES

ACHARYA, U. R., OH, S. L., HAGIWARA, Y., TAN, J. H., ADAM, M., GERTYCH, A. and SAN TAN, R. (2017). A deep convolutional neural network model to classify heartbeats. *Comput. Biol. Med.* **89** 389–396.

ATTIA, Z. I., NOSEWORTHY, P. A., LOPEZ-JIMENEZ, F., ASIRVATHAM, S. J., DESHMUKH, A. J., GERSH, B. J., CARTER, R. E., YAO, X., RABINSTEIN, A. A. et al. (2019). An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: A retrospective analysis of outcome prediction. *Lancet* **394** 861–867.

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** e0130140. https://doi.org/10.1371/journal.pone.0130140

BARTLETT, P. L., BOUSQUET, O. and MENDELSON, S. (2005). Local Rademacher complexities. *Ann. Statist.* **33** 1497–1537. MR2166554 https://doi.org/10.1214/009053605000000282

BARTLETT, P. L., JORDAN, M. I. and MCAULIFFE, J. D. (2006). Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.* **101** 138–156. MR2268032 https://doi.org/10.1198/016214505000000907

BARTLETT, P. L. and MENDELSON, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.* **3** 463–482. MR1984026 https://doi.org/10.1162/153244303321897690

BENGIO, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In *Neural Networks*: *Tricks of the Trade* 437–478. Springer, Berlin.

BERGSTRA, J. and BENGIO, Y. (2012). Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **13** 281–305. MR2913701

BHARTI, R., KHAMPARIA, A., SHABAZ, M., DHIMAN, G., PANDE, S. and SINGH, P. (2021). Prediction of heart disease using a combination of machine learning and deep learning. *Comput. Intell. Neurosci.* **2021** 8387680. https://doi.org/10.1155/2021/8387680

BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.

CORTES, C. and VAPNIK, V. (1995). Support-vector networks. *Mach. Learn.* **20** 273–297.

DAI, B., SHEN, X., CHEN, L. Y, LI, C. and PAN, W. (2023). Supplement to "Data-adaptive discriminative feature localization with statistically guaranteed interpretation." https://doi.org/10.1214/22-AOAS1705SUPPA, https://doi.org/10.1214/22-AOAS1705SUPPB

DAI, B., SHEN, X. and PAN, W. (2022). Significance tests of feature relevance for a black-box learner. *IEEE Trans. Neural Netw. Learn. Syst.*

DAVIS, R. A., LII, K.-S. and POLITIS, D. N. (2011). Remarks on some nonparametric estimates of a density function. In *Selected Works of Murray Rosenblatt* 95–100. Springer, Berlin.

ELGENDI, M. (2013). Fast QRS detection with an optimized knowledge-based method: Evaluation on 11 standard ECG databases. *PLoS ONE* **8** e73557. https://doi.org/10.1371/journal.pone.0073557

EVANS, R., O'NEILL, M., PRITZEL, A., ANTROPOVA, N., SENIOR, A. W., GREEN, T., ŽÍDEK, A., BATES, R., BLACKWELL, S. et al. (2021). Protein complex prediction with AlphaFold-Multimer. *Biorxiv*.

FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **29** 1189–1232. MR1873328 https://doi.org/10.1214/aos/1013203451

GE, R., HUANG, F., JIN, C. and YUAN, Y. (2015). Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory* 797–842.

GHORBANI, A., ABID, A. and ZOU, J. (2019). Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence* **33** 3681–3688.

HE, K., ZHANG, X., REN, S. and SUN, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 770–778.

HOCHREITER, S. and SCHMIDHUBER, J. (1997). Long short-term memory. *Neural Comput.* **9** 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

JAMBUKIA, S. H., DABHI, V. K. and PRAJAPATI, H. B. (2015). Classification of ECG signals using machine learning techniques: A survey. In 2015 *International Conference on Advances in Computer Engineering and Applications* 714–721. IEEE, Ghaziabad.

JUMPER, J., EVANS, R., PRITZEL, A., GREEN, T., FIGURNOV, M., RONNEBERGER, O., TUNYASUVUNAKOOL, K., BATES, R., ŽÍDEK, A. et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* **596** 583–589.

KACHUEE, M., FAZELI, S. and SARRAFZADEH, M. (2018). Ecg heartbeat classification: A deep transferable representation. In 2018 *IEEE International Conference on Healthcare Informatics* (*ICHI*) 443–444. IEEE, New York.

KINDERMANS, P.-J., SCHÜTT, K. T., ALBER, M., MÜLLER, K.-R., ERHAN, D., KIM, B. and DÄHNE, S. (2017). Learning how to explain neural networks: Patternnet and patternattribution. ArXiv preprint. Available at arXiv:1705.05598.

KO, W.-Y., SIONTIS, K. C., ATTIA, Z. I., CARTER, R. E., KAPA, S., OMMEN, S. R., DEMUTH, S. J., ACKERMAN, M. J., GERSH, B. J. et al. (2020). Detection of hypertrophic cardiomyopathy using a convolutional neural network-enabled electrocardiogram. *J. Am. Coll. Cardiol.* **75** 722–733.

KUSUMOTO, F. (2020). *ECG Interpretation*: *From Pathophysiology to Clinical Application*. Springer Nature, Berlin.

LeCun, Y. and Cortes, C. (2010). MNIST handwritten digit database.
LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Comput*. **1** 541–551.
Lee, J. D., Simchowitz, M., Jordan, M. I. and Recht, B. (2016). Gradient descent only converges to minimizers. In *Conference on Learning Theory* 1246–1257.
Lin, Y. (2004). A note on margin-based loss functions in classification. *Statist. Probab. Lett*. **68** 73–82. MR2064687 https://doi.org/10.1016/j.spl.2004.03.002
Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* 4765–4774.
Martis, R. J., Acharya, U. R., Lim, C. M., Mandana, K., Ray, A. K. and Chakraborty, C. (2013). Application of higher order cumulant features for cardiac health diagnosis using ECG signals. *Int. J. Neural Syst*. **23** 1350014.
Masci, J., Meier, U., Cireşan, D. and Schmidhuber, J. (2011). Stacked convolutional auto-encoders for hierarchical feature extraction. In *International Conference on Artificial Neural Networks* 52–59. Springer, Berlin.
McFadden, D. et al. (1973). Conditional logit analysis of qualitative choice behavior.
Mirvis, D. M. and Goldberger, A. L. (2001). Electrocardiography. In *Heart Disease* 82–128. W. B. Saunders, Philadelphia.
Montavon, G., Lapuschkin, S., Binder, A., Samek, W. and Müller, K.-R. (2017). Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognit*. **65** 211–222.
Moody, G. B. and Mark, R. G. (1990). The MIT-BIH arrhythmia database on CD-ROM and software for use with it. In [1990] *Proceedings Computers in Cardiology* 185–188. IEEE, Chicago.
Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika* **78** 691–692. MR1130937 https://doi.org/10.1093/biomet/78.3.691
Raginsky, M., Rakhlin, A. and Telgarsky, M. (2017). Non-convex learning via stochastic gradient Langevin dynamics: A nonasymptotic analysis. ArXiv preprint. Available at arXiv:1702.03849.
Rajpurkar, P., Hannun, A. Y., Haghpanahi, M., Bourn, C. and Ng, A. Y. (2017). Cardiologist-level arrhythmia detection with convolutional neural networks. ArXiv preprint. Available at arXiv:1707.01836.
Raskutti, G., Wainwright, M. J. and Yu, B. (2014). Early stopping and non-parametric regression: An optimal data-dependent stopping rule. *J. Mach. Learn. Res*. **15** 335–366. MR3190843
Ribeiro, M. T., Singh, S. and Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135–1144.
Rudy, Y. (2004). Ionic mechanisms of cardiac electrical activity: A theoretical approach. In *Cardiac Electrophysiology: From Cell to Bedside* 255–266. Elsevier, Philadelphia.
Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1985). Learning internal representations by error propagation Technical Report California Univ. San Diego La Jolla Inst. for Cognitive Science.
Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* 618–626.
Stergiou, G. S., Alpert, B., Mieke, S., Asmar, R., Atkins, N., Eckert, S., Frick, G., Friedman, B., Grassl, T. et al. (2018). A universal standard for the validation of blood pressure measuring devices: Association for the advancement of medical instrumentation/European society of hypertension/international organization for standardization (AAMI/ESH/ISO) collaboration statement. *J. Hypertens*. **71** 368–374.
Thygesen, K., Alpert, J. S., White, H. D. and Joint ESC/ACCF/AHA/WHF Task Force for the Redefinition of Myocardial Infarction (2007). Universal definition of myocardial infarction. *J. Am. Coll. Cardiol*. **50** 2173–2195.
Tjoa, E. and Guan, C. (2019). A survey on explainable artificial intelligence (XAI): Towards medical XAI. ArXiv preprint. Available at arXiv:1907.07374.
Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P. et al. (2019). Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **18** 463–477.
Wang, H., Wang, N. and Yeung, D.-Y. (2015). Collaborative deep learning for recommender systems. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1235–1244.
Wasimuddin, M., Elleithy, K., Abuzneid, A.-S., Faezipour, M. and Abuzaghleh, O. (2020). Stages-based ECG signal analysis from traditional signal processing to machine learning approaches: A survey. *IEEE Access* **8** 177782–177803.
Wu, Y. and Liu, Y. (2007). Robust truncated hinge loss support vector machines. *J. Amer. Statist. Assoc*. **102** 974–983. MR2411659 https://doi.org/10.1198/016214507000000617

ZEILER, M. D. and FERGUS, R. (2014). Visualizing and understanding convolutional networks. In *European Conference on Computer Vision* 818–833. Springer, Berlin.

ZHOU, B., KHOSLA, A., LAPEDRIZA, A., OLIVA, A. and TORRALBA, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2921–2929.