Explaining the Behavior of Black-Box Prediction Algorithms with Causal Learning

Numair Sani, Daniel Malinsky, and Ilya Shpitser

***Abstract*—We propose to explain the behavior of black-box prediction methods (e.g., deep neural networks trained on image pixel data) using causal graphical models. Specifically, we explore learning the structure of a causal graph where the nodes represent prediction outcomes along with a set of macro-level "interpretable" features, while allowing for arbitrary unmeasured confounding among these variables. The resulting graph may indicate which of the interpretable features, if any, are possible causes of the prediction outcome and which may be merely associated with prediction outcomes due to confounding. The approach is motivated by a counterfactual theory of causal explanation wherein good explanations point to factors which are "difference-makers" in an interventionist sense. The resulting analysis may be useful in algorithm auditing and evaluation, by identifying features which make a causal difference to the algorithm's output.**

I. Introduction

In recent years, black-box artificial intelligence (AI) or machine learning (ML) prediction methods have exhibited impressive performance in a wide range of prediction tasks. In particular, methods based on deep neural networks (DNNs) have been successfully used to analyze high-dimensional data in settings such as healthcare and social data analytics [1], [2]. An important obstacle to widespread adoption of such methods, particularly in socially-impactful settings, is their black-box nature: it is not obvious, in many cases, how to explain the predictions produced by such algorithms when they succeed (or fail), given that they find imperceptible patterns among high-dimensional sets of features. Moreover, the relevant "explanatory" or "interpretable" units may not coincide nicely with the set of raw features used by the prediction method (e.g., image pixels). Here we present an approch to post-hoc explanation of algorithm behavior which builds on ideas from causality and graphical models. We propose that to explain *post hoc* the output of a black-box method is to understand which variables, from among a set of interpretable features, make a causal difference to the output. That is, we ask which potential targets of manipulation may have non-zero intervention effects on the prediction outcome.

There have been numerous approaches to explainability of machine learning algorithms [3], [4], [5], [6], [7], [8], [9], [10], [11]. Many have focused on (what may be broadly called) "feature importance" measures. Feature importance measures standardly approach explanation from a purely associational standpoint: features ranked as highly "important" are typically inputs that are highly correlated with the predicted outcome (or prediction error) in the sense of having a large regression coefficient or perturbation gradient, perhaps in the context of a local and simple approximating model class (e.g., linear regression, decision trees, or rule lists). However, the purely associational "importance" standpoint has at least two shortcomings. First, the inputs to a DNN (e.g., individual pixels) are often at the wrong level of description to capture a useful or actionable explanation. For example, an individual pixel may contribute very little to the output of a prediction method but contribute a lot in aggregate – higher-level features comprised of many pixels or patterns across individual inputs (e.g., differences between or variances among collections of lower-level attributes) may be the appropriate ingredients of a more useful explanation. Second, features (at whatever level of description) may be highly associated with outcomes without causing them. Two variables may be highly associated because they are both determined by a common cause that is not among the set of potential candidate features. That is, if the black-box algorithm is in fact tracking some omitted variable that is highly correlated with some input feature, the input feature may be labelled "important" in a way that does not support generalization or guide action. We propose to use causal discovery methods (a.k.a. causal structure learning) to determine which interpretable features, from among a pre-selected set of candidates, may plausably be causal deteriminants of the outcome behavior, and distinguish these causal features from variables that are associated with the behavior due to confounding.

We begin by providing some background on causal explanation and the formalism of causal inference, including causal discovery. We then describe our proposal for explaining the behaviors of black-box prediction algorithms and present a simulation study that illustrates our ideas. We also apply a version of our proposal to two datasets: annotated image data for bird classification and annotated chest X-ray images for pneumonia detection. Finally, we discuss some applications, limitations, and future directions of this work.

II. Causal Explanation

A. Explaining Algorithm Behaviors

There is a long history of debate in science and philosophy over what properly constitues an explanation of some phenomenon. (In our case, the relevant phenomenon will be the output of a prediction algorithm.) A connection between explanation and "investigating causes" has been influential, in Western philosophy, at least since Aristotle [12]. More

N.S. and I.S. are with the Department of Computer Science, Johns Hopkins University, Baltimore, MD USA, email: {snumair1,ilyas.cs.}@jhu.edu. D.M. is with the Department of Biostatistics, Columbia University, New York, NY USA, email: dsm2128@cumc.columbia.edu. N.S. and D.M. contributed equally.

recently, scholarship on *causal explanation* [13], [14], [15], [16] has highlighted various benefits to pursuing understanding of complex systems via causal or counterfactual knowledge, which may be of particular utility to the machine learning community. We focus here primarily on some relevant ideas discussed by Woodward [15] to motivate our perspective in this paper, though similar issues are raised elsewhere in the literature.

In influential 20th-century philosophical accounts, explanation was construed via applications of deductive logical reasoning (i.e., showing how observations could be derived from physical laws and background conditions) or simple probabilistic reasoning [17]. One shortcoming – discussed by several philosophers in the late 20th-century – of all such proposals is that explanation is intuitively asymmetric: the height of a flagpole explains the length of its shadow (given the sun's position in the sky) but not vice versa; the length of a classic mechanical pendulum explains the device's period of motion, but not vice versa. Logical and associational relationships do not exhibit such asymmetries. Moreover, some statements of fact or strong associations seem explanatorily irrelevant to a given phenomenon, as when the fact that somebody neglected to supply water to a rock "explains" why it is not living. (An analogous fact may have been more relevant for a plant, which in fact needs water to live.) Woodward argues that "explanatory relevance" is best understood via counterfactual contrasts and that the asymmetry of explanation reflects the role of causality.

On Woodward's counterfactual theory of causal explanation, explanations answer *what-would-have-been-different* questions. Specifically, the relevant counterfactuals describe the outcomes of interventions or manipulations. X helps explain Y if, under suitable background conditions, some intervention on X produces a change in the distribution of Y. (Here we presume the object of explanation to be the random variable Y, not a specific value or event $Y = y$. That is, we choose to focus on *type-level* explanation rather *token-level* explanations of particular events.) This perspective has strong connections to the literature on causal models in artificial intelligence and statistics [18], [19], [20]. A causal model for outcome Y precisely stipulates how Y would change under various interventions. So, to explain black-box algorithms we endeavour to build causal models for their behaviors. We propose that such causal explanations can be useful for algorithm evaluation and informing decision-making. In contrast, purely associational measures will be symmetric, include potentially irrelevant information, and fail to support (interventionist) counterfactual reasoning.[1]

[1]Some approaches to explainability focus on a different counterfactual notion: roughly, they aim to identify values in the input space for which a prediction decision changes ("minimum-edits"), assuming all variables are independent of each other [21], [22], [23]. In most settings of interest, the relevant features are not independent of each other, as will become clear in our examples below. Some promising recent work has combined causal knowledge with counterfactual explanations of this sort [24], focusing on counterfactual input values that are consistent with background causal relationships. While interesting, such work is orthogonal to our proposal here, which focuses on type-level rather than token-level explanation, operates on a different set of features than the ones used to generate the prediction, and does not presume that causal relationships among variables are known a priori.

Despite a paucity of causal approaches to explainability in the ML literature (with some exceptions, discussed later), survey research suggests that causal explanations are of particular interest to industry practitioners; [25] quote one chief scientist as saying "Figuring out causal factors is the holy grail of explainability," and report similar sentiments expressed by many organizations.

B. Causal Modeling

Next we provide some background to make our proposal more precise. Throughout, we use uppercase letters (e.g., X, Y) to denote random variables or vectors and lowercase (x, y) to denote fixed values.

We use causal graphs to represent causal relations among random variables [18], [19]. In a causal directed acyclic graph (DAG) $\mathcal{G} = (V, E)$, a directed edge between variables $X \to Y$ $(X, Y \in V)$ denotes that X is a direct cause of Y, relative to the variables on the graph. Direct causation may be explicated via a system of nonparametric structural equations (NPSEMs) a.k.a. a structural causal model (SCM). The distribution of Y given an intervention that sets X to x is denoted $p(y \mid \mathrm{do}(x))$ by Pearl [19]. Causal effects are often defined as interventional constrasts, e.g., the average causal effect (ACE): $\mathbb{E}[Y \mid \mathrm{do}(x)] - \mathbb{E}[Y \mid \mathrm{do}(x')]$ for values x, x'. Equivalently, one may express causal effects within the formalism of potential outcomes or counterfactual random variables, c.f. [26].

Given some collection of variables V and observational data on V, one may endeavor to learn the causal structure, i.e., to select a causal graph supported by the data. We focus on learning causal relations from purely observational (non-experimental) data here, though in some ML settings there exists the capacity to "simulate" interventions directly, which may be even more informative. There exists a significant literature on selecting causal graphs from a mix of observational and interventional data, e.g. [27], [28], and though we do not make use of such methods here, the approach we propose could be applied in those mixed settings as well.

There are a variety of algorithms for causal structure learning, but what most approaches share is that they exploit patterns of statistical constraints implied by distinct causal models to distinguish among candidate graphs. One paradigm is constraint-based learning, which will be our focus. In constraint-based learning, the aim is to select a causal graph or set of causal graphs consistent with observed data by directly testing a sequence of conditional independence hypotheses – distinct models will imply distinct patterns of conditional independence, and so by rejecting (or failing to reject) a collection of independence hypotheses, one may narrow down the set of models consistent with the data. (These methods typically rely on some version of the faithfulness assumption [18], which stipulates that all observed independence constraints correspond to missing edges in the graph.) For example, a classic constraint-based method is the PC algorithm [18], which aims to learn an equivalence class of DAGs by starting from a fully-connected model and removing edges when conditional independence constraints are discovered via statistical tests. Since multiple DAGs may imply the same set of conditional

independence constraints, PC estimates a CPDAG (completed partial DAG), a mixed graph with directed and undirected edges that represents a Markov equivalence class of DAGs. (Two graphs are called Markov equivalent if they imply the same conditional independence constraints.) Variations on the PC algorithm and related approaches to selecting CPDAGs have been thoroughly studied in the literature [29], [30].

In settings with unmeasured (latent) confounding variables, it is typical to study graphs with bidirected edges to represent dependence due to confounding. For example, a partial ancestral graph (PAG) [31] is a graphical representation which includes directed edges ($X \rightarrow Y$ means X is a causal ancestor of Y), bidirected edges ($X \leftrightarrow Y$ means X and Y are both caused by some unmeasured common factor(s), e.g., $X \leftarrow U \rightarrow Y$), and partially directed edges ($X \circ\!\!\rightarrow Y$ or $X \circ\!\!-\!\!\circ Y$) where the circle marks indicate ambiguity about whether the endpoints are arrows or tails. Generally, PAGs may also include additional edge types to represent selection bias, but this is irrelevant for our purposes here. PAGs inherit their causal interpretation by encoding the commonalities among a set of underlying causal DAGs with latent variables. A bit more formally, a PAG represents an equivalence class of maximal ancestral graphs (MAGs), which encode the independence relations among observed variables when some variables are unobserved [32], [31]. The FCI algorithm [18], [33] is a well-known constraint-based method, which uses sequential tests of conditional independence to select a PAG from data. Similarly to the PC algorithm, FCI begins by iteratively deleting edges from a fully-connected graph by a sequence of conditional independence tests. However, the space of possible models (mixed graphs with multiple types of edges) is much greater for FCI as compared with PC, so the rules linking patterns of association to possible causal structures are more complex. Though only independence relations among observed variables are testable, it can be shown formally that certain patterns of conditional independence among observed variables are incompatable with certain latent structures, so some observed patterns will rule out unmeasured confounding and other observed patterns will on the contrary suggest the existence of unmeasured confounders. Details of the FCI algorithm may be found in the literature [33], [34]. Variations on the FCI algorithm and alternative PAG-learning procedures have also been studied [34], [35], [36], [37].

III. Explaining Black-Box Predictions

Consider a supervised learning setting with a high-dimensional set of "low-level" features (e.g., pixels in an image) $X = (X_1, ..., X_q)$ taking values in an input space $\mathcal{X}^q$ and outcome Y taking values in $\mathcal{Y}$. A prediction or classification algorithm (e.g., DNN) learns a map $f : \mathcal{X}^q \mapsto \mathcal{Y}$. Predicted values from this function are denoted $\widehat{Y}$. To explain the predictions $\widehat{Y}$, we focus on a smaller set of "high-level" features $Z = (Z_1, ..., Z_p)$ $(p \ll q)$ which are "interpretable" in the sense that they correspond to human-understandable and potentially manipulable elements of the application domain of substantive interest, e.g., "effusion in the lung," "presence of tumor in the upper lobe," and so on in lung imaging or

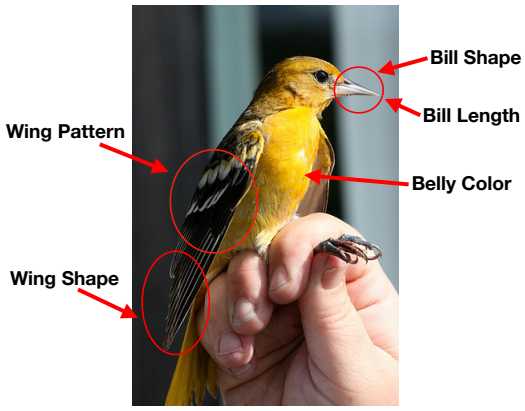


Fig. 1: An image of a Baltimore Oriole annotated with interpretable features.

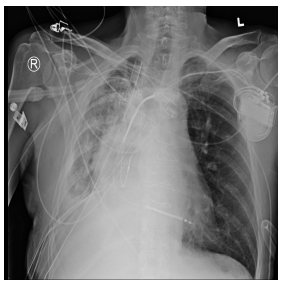


Fig. 2: A chest X-ray from a pneumonia patient. The image was annotated by a radiologist to indicate `Effusion`, `Pneumothorax`, and `Pneumonia`.

"has red colored wing feathers," "has a curved beak," and so on in bird classification. For now, we assume that the interpretable feature set is predetermined and available in the form of an annotated dataset, i.e., that each image is labeled with interpretable features. Later, we discuss the possibility of automatically extracting a set of possibly interpretable features from data which lacks annotations. We allow that elements of Z are statistically and causally interrelated, but assume that Z contains no variables which are deterministically related (e.g., where $Z_i = z_i$ always implies $Z_j = z_j$ for some $i \neq j$), though we also return to this issue later. Examples from two image datasets that we discuss later, on bird classification and lung imaging, are displayed in Figs. 1 and 2: X consists of the raw pixels and Z includes the interpretable annotations identified in the figures.

Our proposal is to learn a causal PAG $\widehat{\mathcal{G}}$ among the variable set $V = (Z, \widehat{Y})$, with the minimal background knowledge imposed that the predicted outcome $\widehat{Y}$ is a causal non-ancestor of Z (there is no directed path from $\widehat{Y}$ into any element of Z). That is, we allow that the content of an image may cause the output of a prediction algorithm trained on that image, but not vice versa. Additional background knowledge may also be imposed if it is known, for example, that none of the elements of Z may cause each other (they are only dependent due to latent common causes) or there are groups of variables which precede others in a causal ordering. If in $\widehat{\mathcal{G}}$, there is

PAG Edge	Meaning
$Z_i \to \widehat{Y}$	Z_i is a cause of $\widehat{Y}$
$Z_i \leftrightarrow \widehat{Y}$	Z_i and $\widehat{Y}$ share an unmeasured common cause $Z_i \leftarrow U \to \widehat{Y}$
$Z_i \circ\!\!\to \widehat{Y}$	Either Z_i is a cause of $\widehat{Y}$ or there is unmeasured confounding, or both

TABLE I: Types of edges relating some interpretable feature Z_i and the predicted outcome $\widehat{Y}$. No edge between Z_i and $\widehat{Y}$ means they are conditionally independent given some subset of the measured variables.

a directed edge $Z_i \to \widehat{Y}$, then Z_i is a cause (definite causal ancestor) of the prediction, if instead there is a bidirected edge $Z_i \leftrightarrow \widehat{Y}$ then Z_i is *not* a cause of the prediction but they are dependent due to common latent factors, and if $Z_i \circ\!\!\to \widehat{Y}$ then Z_i is a possible cause (possible ancestor) of the prediction but unmeasured confounding cannot be ruled out. These edge types are summarized in Table I. The reason it is important to search for a PAG and not a DAG (or equivalence class of DAGs) is that Z will in general *not* include all possibly relevant variables. Even if only interpretable features strongly associated with prediction outcomes are pre-selected, observed associations may be attributed in part or in whole to latent common causes.

We emphasize that the graphical representation $\widehat{\mathcal{G}}$ is an intentionally crude approximation to the real prediction process. By assumption, $\widehat{Y}$ is truly generated from $X_1, ..., X_q$, not $Z_1, ..., Z_p$. However, we view $\widehat{\mathcal{G}}$ as a useful approximation insofar as it captures some of the salient "inner workings" of the opaque and complicated prediction algorithm. Prediction algorithms such as DNNs will have some lower-dimensional internal representations which they learn and exploit to classify an input image. We would hope, for example, that an accurate bird classification procedure would internally track important aspects of bird physiology (e.g., wing shape, color patterns, head or bill shapes and sizes). If in fact the DNN's internal representation is entirely disjoint from (functions of) our chosen $Z_1, ..., Z_p$, we expect to find that these features are not causes; if some subset of $Z_1, ..., Z_p$ nearly correspond to important features in the internal representation we expect to find that those will be picked out as (possible) causes of the DNN's output. The theoretical properties of PAG-learning algorithms such as FCI [33] imply that, in the large-sample limit, the causal determinants of $\widehat{Y}$ will by identified by $Z_i \to \widehat{Y}$ or $Z_i \circ\!\!\to \widehat{Y}$ edges in $\widehat{\mathcal{G}}$, whereas features connected by $Z_i \leftrightarrow \widehat{Y}$ are associated due to confounding and features with no paths into $\widehat{Y}$ are causally irrelevant.

Two causality-inspired approaches to explanation are worth mentioning here. [38] propose a causal attribution method for neural networks. They estimate the ACE of each input neuron on each output neuron under a model assumption motivated by the network structure: they assume if input neurons are dependent, it is only because of latent common causes. The "contribution" of each input neuron is then quantified by the magnitude of this estimated causal effect. This is similar in spirit to our approach, but we do not take the input neurons as the relevant units of analysis (instead focusing on substantively interpretable "macro" features which may be complex aggregates), nor do we assume the causal structure is fixed a priori. Our proposal is also not limited to prediction models based on neural networks or any particular model architecture. [39] introduce an approach (CXPlain) which is model-agnostic but similar to [38] in taking the "low-level" input features X as the relevant units of analysis (in contrast with the "high-level" features Z). CXPlain is based on Granger causality and their measure effectively quantifies the change in prediction error when an input feature is left out. Unlike our proposal, this measure does not have an interventionist interpretation except under the restrictive assumption that all relevant variables have been measured, i.e., no latent confounding.

Next we illustrate a basic version of our procedure with a simulation study.

IV. A Simulation Study

Consider the following experiment, inspired by and modified from a study reported in [40]. A research subject is presented with a sequence of 2D black and white images, and responds ($Y = 1$) or does not respond ($Y = 0$) with some target behavior for each image presented. The raw features X are then the $d \times d$ image pixels. The images may contain several shapes (alone or in combination) – a horizontal bar (H), vertical bar (V), circle (C), or rectangle (R) – in addition to random pixel noise; see Fig. 3. The target behavior Y is caused only by the presence of verticle bars and circles, in the sense that manipulating the image pixel arrangement to contain a verticle bar or a circle (or both) makes $Y = 1$ much more likely, whereas manipulating the presence of the other shapes does not change the distribution of Y at all. In our simulations, this is accomplished by sampling the target behavior $Y = 1$ with probability depending monotonically only on V and C. However, the various shapes are not independent. Circles and horizontal bars also cause rectangles to be more likely. R would thus be associated with the outcome Y, though conditionally independent given C. H would be marginally independent of Y but dependent given R. The details of the simulation are given below, as well as summarized by the DAG in Fig. 4(a).

$$
\begin{aligned}
U_1 &\sim \text{Uniform}(0,1) \quad U_2 \sim \text{Uniform}(0,1) \\
H &\sim \text{Bernoulli}(U_1) \quad C \sim \text{Bernoulli}(U_2) \\
V &\sim \text{Bernoulli}(1 - U_1) \\
R &\sim \text{Bernoulli}(\text{expit}(0.75H + 0.5C)) \\
Y &\sim \text{Bernoulli}(\text{expit}(-0.5 + 2.5V + 1.75C))
\end{aligned}
$$

Using the above process, 5000 images are generated. Next, we train a deep convolutional neural network (CNN) on the raw image pixels X according to the standard supervised paradigm. We use Resnet18 [41] and initialize the weights with values estimated by pre-training the network on ImageNet [42]. This is implemented using the PyTorch framework.[2] The model performs quite well – 81% accuracy on a held

[2]The software is freely available at: `https://pytorch.org`

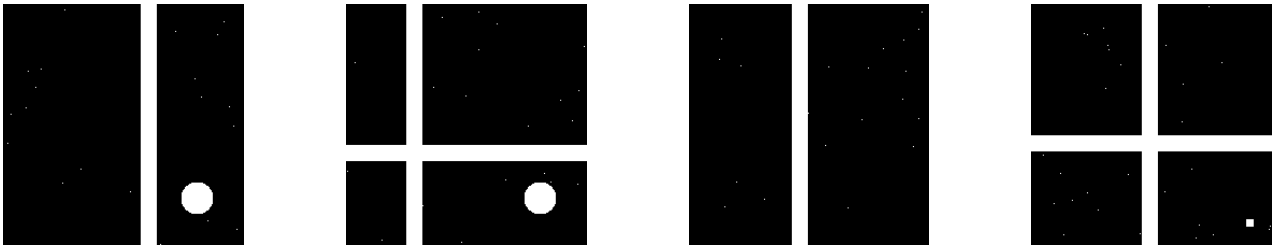


Fig. 3: Simulated image examples with horizontal bars, vertical bars, circles, and rectangles.

Fig. 4: (a) A causal diagram representing the true data generating process. (b) The PAG learned using FCI with output $\widehat{Y}$ from a convolutional neural network.

out test set of 2000 images – so we may reasonably hope that these predictions track the underlying mechanism by which the data was generated. Our interpretable features Z are indicators of the presence of the various shapes in the image. Since the true underlying behavior is causally determined by V and C, we expect V and C to be "important" for the predictions $\widehat{Y}$, but due to the mutual dependence the other shapes are also highly correlated with $\widehat{Y}$. Moreover, we mimic a setting where C is (wrongfully) omitted from the set of candidate interpretable features; in practice, the feature set proposed by domain experts or available in the annotated data will typically exclude various relevant determinants of the underlying outcome. In that case, C is an unmeasured confounder. Applying the PAG-estimation algorithm FCI to variable set $(H, V, R, \widehat{Y})$ we learn the structure in Fig. 4(b), which indicates correctly that V is a (possible) cause, but that R is not: R is associated with the prediction outcomes only due to an unmeasured common cause. This simple example illustrates how the estimated PAG using incomplete interpretable features can be useful: we disentangle mere statistical associations from (potentially) causal relationships, in this case indicating correctly that interventions on V may make a difference to the prediction algorithm's behavior but interventions on R and H would not, and moreover that there is a potentially important cause of the output (C) that has been excluded from our set of candidate features, as evident from the bidirected edge $R \leftrightarrow \widehat{Y}$ learned by FCI.

V. Experiments: Bird Classification and Pneumonia Detection from X-rays

We conduct two real data experiments to demonstrate the utility of our approach. First, we study a neural network for bird classification, trained on the Caltech-UCSD 200-2011 image dataset [43]. It consists of 200 categories of birds, with 11,788 images in total. Each image comes annotated with 312 binary attributes describing interpretable bird characteristics like eye color, size, wing shape, etc. We build a black-box prediction model using raw pixel features to predict the class of the bird and then use FCI to explain the output of the model. Second, we follow essentially the same procedure to explain the behavior of a pneumonia detection neural network, trained on a subset of the ChestX-ray8 dataset [44]. The dataset consists of 108,948 frontal X-ray images of 32,217 patients, along with corresponding free-text X-ray reports generated by certified radiologists. Both data sources are publicly available online.

A. Data Preprocessing & Model Training

Bird Image Data: Since many of the species classifications have few associated images, we first broadly group species into 9 coarser groups. For example we group the Baird Sparrow and Black Throated Sparrow into one Sparrow class. Details on the grouping scheme can be found in the supplementary material. This leads to 9 possible outcome labels and 5514 images (instances that do not fit into one of these categories are excluded from analysis). The number of images across each class is not evenly balanced. So, we subsample overrepresented classes in order to get roughly equal number of images per class. This yields a dataset of 3538 images, which is then partitioned into a training, validation, and testing datasets of 2849, 520, and 529 images respectively. We train the ResNet18 architecture (pre-trained on ImageNet) on our dataset and achieve an accuracy of 86.57% on the testing set. For computational convenience and statistical efficiency, we consolidate the 312 available binary attributes into ordinal attributes. (With too many attributes, some combinations of values may rarely or never co-occur, which leads to data fragmentation.) For

example, four binary attributes describing "back pattern" (solid, spotted, striped, or multi-colored), are consolidated into one attribute `Back Pattern` taking on five possible values: the 4 designations above and an additional category if the attribute is missing. Even once this consolidation is performed, there still exist many attributes with a large number of possible values, so we group together similar values. For example we group dark colors such as gray, black, and brown into one category and warm colors such as red, yellow, and orange into another. Other attributes are consolidated similarly, as described in the supplementary material. After the above preprocessing, for each image we have a predicted label from the CNN and 26 ordinal attributes.

Chest X-ray Data: From the full ChestX-ray8 dataset, we create a smaller dataset of chest X-rays labeled with a binary diagnosis outcome: pneumonia or not. The original dataset contained 120 images annotated by a board-certified radiologist to contain pneumonia. To obtain "control" images, we randomly select 120 images with "no findings" reported in the radiology report. (One of the images is excluded from analysis due to missing information.) To generate interpretable features, we choose the following 7 text-mined radiologist finding labels: `Cardiomegaly`, `Atelectasis`, `Effusion`, `Infiltration`, `Mass`, `Nodule`, and `Pneumothorax`. If the radiology report contained one of these findings, the corresponding label was given a value 1 and 0 otherwise. This produces an analysis dataset of 239 images: 120 pneumonia and 119 "control" with each image containing 7 binary interpretable features. Using the same architecture as for the previous experiment and reserving 50 images for testing, ResNet18 achieves an accuracy of 74.55%.

B. Structure Learning

Structure learning methods such as FCI produce a single estimated PAG as output. In applications, it is common to repeat the graph estimation on multiple bootstraps or subsamples of the original data, in order to control false discovery rates or to mitigate the cumulative effect of statistical errors in the independence tests [45]. We create 50 bootstrapped replicates of the bird image dataset and 100 replicates of the X-ray dataset. We run FCI on each replicate with independence test rejection threshold (a tuning parameter) set to $\alpha = 10^{-5}$ and $\alpha = .05$ for the birds and X-ray datasets, respectively, with the knowledge constraint imposed that outcome $\widehat{Y}$ cannot cause any of the interpretable features.[3] Here FCI is used with the χ^2 independence test, and we limit the maximum conditioning set size to 4 for computational tractability in the birds dataset. (No limit is necessary in the smaller X-ray dataset.)

C. Results

We compute the relative frequency over bootstrap replicates of $Z_i \to \widehat{Y}$ and $Z_i \circ\!\!\to \widehat{Y}$ edges from all attributes. This represents the frequency with which an attribute is determined to be a cause or possible cause of the predicted outcome label and constitutes a rough measure of "edge stability" or confidence in that attribute's causal status. The computed edge stabilities are presented in the Fig. 5. In the bird classification experiment, we find that the most stable (possible) causes include `Size`, `Eye Color`, and `Bill Shape`, which are intuitively salient features for distinguishing among bird categories. We have lower confidence that the other features are possible causes of $\widehat{Y}$. In the X-ray experiment, we find that edges from `Atelectasis`, `Infiltration`, and `Effusion` are stable possible causes of the `Pneumonia` label. This is reassuring, since these are clinically relevant conditions for patients with pneumonia. An estimated PAG from one of the bootstrapped subsamples is displayed in Fig. 6. The FCI procedure has detected a variety of edge types, including bidirected ($\leftrightarrow$) edges to indicate unmeasured confounding among some variables, but edges from `Atelectasis`, `Infiltration`, and `Effusion` are consistently either directed ($\to$) or partially directed ($\circ\!\!\to$) into the predicted outcome.

[3]We use the command-line interface to the TETRAD freeware: `https://github.com/cmu-phil/tetrad`

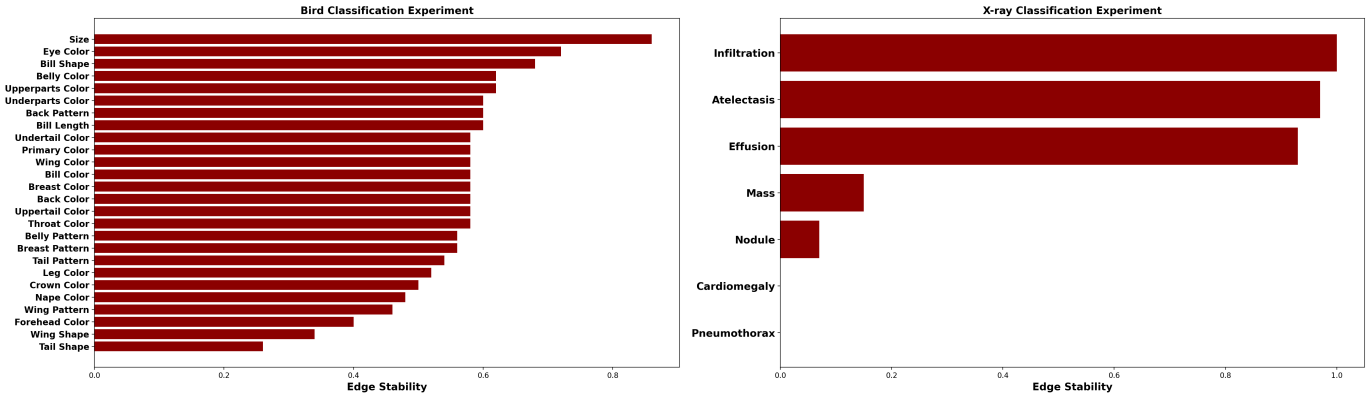
VI. Discussion

We have presented an analytical approach (based on existing tools) to support explaining the behavior of black-box prediction algorithms. Below we discuss some potential uses and limitations.

A. Algorithm Auditing and Evaluation

One important goal related to building explainable AI systems is the auditing and evaluation of algorithms post-hoc. If a prediction algorithm appears to perform well, it is important to understand why it performs well before deploying the algorithm. Users will want to know that the algorithm is "paying attention to" the right aspects of the input and not tracking spurious artifacts [25]. This is important both from the perspective of generalization to new domains as well as from the perspective of fairness. To illustrate the former, consider instances of "dataset bias" or "data leakage" wherein an irrelevant artifact of the data collection process proves very important to the performance of the prediction method. This may impede generalization to other datasets where the artifact is absent or somehow the data collection process is different. For example, [46] study the role of violet image highlighting on dermoscopic images in a skin cancer detection task. They find that this image highlighting significantly affects the likelihood that a skin lesion is classified as cancerous by a commerical CNN tool. (The researchers are able to diagnose the problem because they have access to the same images pre- and post-highlighting: effectively, they are able to simulate an intervention on the highlighting.)

To illustrate the fairness perspective, consider a recent episode of alleged racial bias in Google's Vision Cloud API, a tool which automatically labels images into various categories [47]. Users found an image of a dark-skinned hand holding a non-contact digital thermometer that was labelled "gun" by the API, while a similar image with a light-skinned individual was labelled "electronic device." More tellingly, when the image with dark skin was crudely altered to contain light beige-colored skin (an intervention on "skin color"), the same object was



(a) (b)

Fig. 5: Results for two experiments. Potential causal determinants of (a) bird classifier output from bird images and (b) pneumonia classifier output from X-ray images.

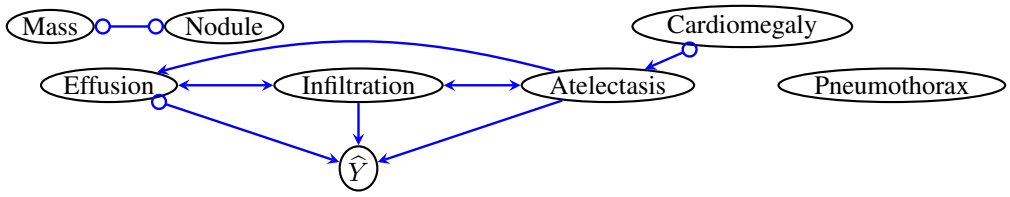


Fig. 6: A PAG estimated from the X-ray data using FCI. $\widehat{Y}$ denotes the output of the pneumonia classification neural network ($\widehat{Y} = 1$ for pneumonia cases, $\widehat{Y} = 0$ for controls). Directed and partially directed edges into $\widehat{Y}$ from `Atelectasis`, `Infiltration`, and `Effusion` indicate that these are likely causes of the prediction algorithm's output, whereas the other variables are mostly not causally relevant. Across subsamples of the data, edges from these three key variables are sometimes directed and sometimes partially directed.

labelled "monocular." This simple experiment was suggestive of the idea that skin color was inappropriately a cause of the object label and tracking biased or stereotyped associations. Google apologized and revised their algorithm, though denied any "evidence of systemic bias related to skin tone."

Auditing algorithms to determine whether inappropriate features have a causal impact on the output can be an important part of the bias-checking pipeline. Moreover, a benefit of our proposed approach is that the black-box model may be audited without access to the model itself, only the predicted values. This may be desirable in some settings where the model itself is proprietary.

B. Informativeness and Background Knowledge

It is important to emphasize that a PAG-based causal discovery analysis is informative to a degree that depends on the data (the patterns of association) and the strength of imposed background knowledge. Here we only imposed the minimal knowledge that $\widehat{Y}$ is not a cause of any of the image features and we allowed for arbitrary causal relationships and latent structure otherwise. Being entirely agnostic about the possibility of unmeasured confounding may lead, depending on the observed patterns of dependence and independence in the data, to only weakly informative results if the patterns of association cannot rule out possible confounding anywhere. If the data fails to rule out confounders and fails to identify definite causes of $\widehat{Y}$, this does not indicate that the analysis has failed but just that only negative conclusions are supported – e.g., the chosen set of interpretable features and background assumptions are not sufficiently rich to identify the causes of the output. It is standard in the causal discovery literature to acknowledge that the strength of supported causal conclusions depends on the strength of input causal background assumptions [18]. In some cases, domain knowledge may support restricting the set of possible causal structures, e.g., when it is believed the some relationships must be unconfounded or some correlations among Z may only be due to latent variables (because some elements of Z cannot cause each other).

C. Selecting or Constructing Interpretable Features

In our experiments we use hand-crafted interpretable features that are available in the form of annotations with the data. Annotations are not always available in applications. In such settings, one approach would be to manually annotate the raw data with expert evaluations, e.g., when clinical experts annotate medical images with labels of important features (e.g. "tumor in the upper lobe," "effusion in the lung," etc).

Alternatively, one may endeavor to extract interpretable features automatically from the raw data. Unsupervised learning

techniques may be used in some contexts to construct features, though in general there is no guarantee that these will be substantively (e.g., clinically) meaningful or correspond to manipulable elements of the domain. We explored applying some recent proposals for automatic feature extraction methods to our problem, but unfortunately none of the available techniques proved appropriate, for reasons we discuss.

A series of related papers [40], [48], [49] has introduced techniques for extracting *causal features* from "low-level" data, i.e., features that have a causal effect on the target outcome. In [40], the authors introduce an important distinction underlying these methods: observational classes versus causal classes of images, each being defined as an equivalence class over conditional or interventional distributions respectively. Specifically, let Y be a binary random variable and denote two images by $X = x, X = x'$. Then x, x' are in the same *observational* class if $p(Y \mid X = x) = p(Y \mid X = x')$. x, x' are in the same *causal* class if $p(Y \mid \text{do}(X = x)) = p(Y \mid \text{do}(X = x'))$. Under certain assumptions, the causal class is always a coarsening of the observational class and so-called "manipulator functions" can be learned that minimally alter an image to change the image's causal class. The idea is that relevant features in the image are first "learned" and then manipulated to map the image between causal classes. However, there are two important roadblocks to applying this methodology in our setting.

First, since our target behavior is the prediction $\widehat{Y}$ (which is some deterministic function of the image pixels), there is no observed or unobserved confounding between $\widehat{Y}$ and X. This implies our observational and causal classes are identical. Hence any manipulator function we learn would simply amount to making the minimum number of pixel edits to the image in order to alter its observational class, similar to the aforementioned "minimum-edit" counterfactual methods [23].

Second, even if we learn this manipulator function, the output is not readily useful. The function takes an image as input and produces as output another ("close") edited image with the desired observational class. It does not correspond to any label which we may apply to other images. (For example, taking a photo of an Oriole as input, the output would be an modified bird image with some different pixels, but these differences will not generally map on to anything we can readily identify with bird physiology and use to construct an annotated dataset.) This does not actually give us access to the "interpretable features" that were manipulated to achieve the desired observational class.

Alternative approaches to automatic feature slection also proved problematic for our purposes. In [50], the authors train an object detection network that serves as a feature extractor for future images, and then they use these extracted features to infer directions of causal effects. However, this approach relies on the availability of training data with a priori associated object categories: the Pascal Visual Object Classes, which are labels for types of objects (e.g., aeroplane, bicycle, bird, boat, etc.) [51]. This approach is thus not truly data-driven and relies on auxiliary information that is not generally available. In another line of work, various unsupervised learning methods (based on autoencoders or variants of independent component analysis) are applied to extract features from medical images to improve classification performance [52], [53], [54]. These approaches are more data-driven but do not produce interpretable features – they produce a compact lower-dimensional representation of the pixels in an abstract vector space, which does not necessarily map onto something a user may understand or recognize.

In general, we observe a tradeoff between the interpretability of the feature set versus the extent to which the features are learned in a data-driven manner. Specifically, "high-level" features associated with datasets (whether they are extracted by human judgement or automatically) may not always be interpretable in the sense intended here: they may not correspond to manipulable elements of the domain. Moreover, some features may be deterministically related (which would pose a problem for most structure learning algorithms) and so some feature pre-selection may be necessary. Thus, human judgement may be indispensable at the feature-selection stage of the process and indeed this may be desirable if the goal is to audit or evaluate potential biases in algorithms as discussed in Section 6.1.

VII. Conclusion

Causal structure learning algorithms – specifically PAG learning algorithms such as FCI and its variants – may be valuable tools for explaining black-box prediction methods. We have demonstrated the utility of using FCI in both simulated and real data experiments, where we are able to distinguish between possible causes of the prediction outcome and features that are associated due to unmeasured confounding. We hope the analysis presented here stimulates further cross-pollination between research communities focusing on causal discovery and explainable AI.

References

[1] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, no. 1, pp. 24–29, 2019.

[2] R. G. Guimaraes, R. L. Rosa, D. De Gaetano, D. Z. Rodriguez, and G. Bressan, "Age groups classification in social network using deep learning," *IEEE Access*, vol. 5, pp. 10 805–10 816, 2017.

[3] P. R. Rijnbeek and J. A. Kors, "Finding a short and accurate decision rule in disjunctive normal form by exhaustive search," *Machine Learning*, vol. 80, no. 1, pp. 33–62, 2010.

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, ""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*, 2016, pp. 1135–1144.

[5] H. Yang, C. Rudin, and M. Seltzer, "Scalable Bayesian rule lists," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 3921–3930.

[6] J. Zeng, B. Ustun, and C. Rudin, "Interpretable classification models for recidivism prediction," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 180, no. 3, pp. 689–722, 2017.

[7] H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec, "Interpretable & explorable approximations of black box models," *arXiv preprint arXiv:1707.01154*, 2017.

[8] P. Adler, C. Falk, S. A. Friedler, T. Nix, G. Rybeck, C. Scheidegger, B. Smith, and S. Venkatasubramanian, "Auditing black-box models for indirect influence," *Knowledge and Information Systems*, vol. 54, no. 1, pp. 95–122, 2018.

[9] S. Dash, O. Gunluk, and D. Wei, "Boolean decision rules via column generation," in *Advances in Neural Information Processing Systems*, 2018, pp. 4655–4665.

[10] C. Molnar, *Interpretable Machine Learning*, 2019, https://christophm.github.io/interpretable-ml-book/.

[11] T. Wang, "Gaining free or low-cost transparency with interpretable partial substitute," in *Proceedings of the 36th International Conference on Machine Learning*, 2019, pp. 6505–6514.

[12] Aristotle, *Posterior Analytics*.

[13] N. Cartwright, "Causal laws and effective strategies," *Noûs*, pp. 419–437, 1979.

[14] W. C. Salmon, *Scientific explanation and the causal structure of the world*. Princeton University Press, 1984.

[15] J. Woodward, *Making things happen: A theory of causal explanation*. Oxford University Press, 2005.

[16] T. Lombrozo and N. Vasilyeva, "Causal explanation," in *Oxford Handbook of Causal Reasoning*. Oxford University Press, 2017, pp. 415–432.

[17] C. G. Hempel, *Aspects of scientific explanation*. Free Press, 1965.

[18] P. L. Spirtes, C. N. Glymour, and R. Scheines, *Causation, prediction, and search*. MIT press, 2000.

[19] J. Pearl, *Causality*. Cambridge University Press, 2009.

[20] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms*. MIT Press, 2017.

[21] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harvard Journal of Law & Technology*, vol. 31, p. 841, 2017.

[22] S. Sharma, J. Henderson, and J. Ghosh, "Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models," *arXiv preprint arXiv:1905.07857*, 2019.

[23] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, "Counterfactual visual explanations," *arXiv preprint arXiv:1904.07451*, 2019.

[24] D. Mahajan, C. Tan, and A. Sharma, "Preserving causal constraints in counterfactual explanations for machine learning classifiers," *arXiv preprint arXiv:1912.03277*, 2019.

[25] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. Moura, and P. Eckersley, "Explainable machine learning in deployment," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 648–657.

[26] T. S. Richardson and J. M. Robins, "Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality," *Center for the Statistics and the Social Sciences, University of Washington Working Paper 128*, pp. 1–146, 2013.

[27] S. Triantafillou and I. Tsamardinos, "Constraint-based causal discovery from multiple interventions over overlapping variable sets," *Journal of Machine Learning Research*, vol. 16, pp. 2147–2205, 2015.

[28] Y. Wang, L. Solus, K. Yang, and C. Uhler, "Permutation-based causal inference algorithms with interventions," in *Advances in Neural Information Processing Systems*, 2017, pp. 5822–5831.

[29] D. Colombo and M. H. Maathuis, "Order-independent constraint-based causal structure learning," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3741–3782, 2014.

[30] R. Cui, P. Groot, and T. Heskes, "Copula PC algorithm for causal discovery from mixed data," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2016, pp. 377–392.

[31] J. Zhang, "Causal reasoning with ancestral graphs," *Journal of Machine Learning Research*, vol. 9, pp. 1437–1474, 2008.

[32] T. S. Richardson and P. Spirtes, "Ancestral graph Markov models," *Annals of Statistics*, vol. 30, no. 4, pp. 962–1030, 2002.

[33] J. Zhang, "On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias," *Artificial Intelligence*, vol. 172, no. 16-17, pp. 1873–1896, 2008.

[34] D. Colombo, M. H. Maathuis, M. Kalisch, and T. S. Richardson, "Learning high-dimensional directed acyclic graphs with latent and selection variables," *Annals of Statistics*, pp. 294–321, 2012.

[35] T. Claassen and T. Heskes, "A Bayesian approach to constraint based causal inference," in *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, 2012, pp. 207–216.

[36] J. M. Ogarrio, P. Spirtes, and J. Ramsey, "A hybrid causal search algorithm for latent variable models," in *Proceedings of the Eighth International Conference on Probabilistic Graphical Models*, 2016, pp. 368–379.

[37] K. Tsirlis, V. Lagani, S. Triantafillou, and I. Tsamardinos, "On scoring maximal ancestral graphs with the max–min hill climbing algorithm," *International Journal of Approximate Reasoning*, vol. 102, pp. 74–85, 2018.

[38] A. Chattopadhyay, P. Manupriya, A. Sarkar, and V. N. Balasubramanian, "Neural network attributions: A causal perspective," in *Proceedings of the 36th International Conference on Machine Learning*, 2019, pp. 981–990.

[39] P. Schwab and W. Karlen, "CXPlain: Causal explanations for model interpretation under uncertainty," in *Advances in Neural Information Processing Systems*, 2019, pp. 10 220–10 230.

[40] K. Chalupka, P. Perona, and F. Eberhardt, "Visual causal feature learning," in *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, 2015, pp. 181–190.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 630–645.

[42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, 2009, pp. 248–255.

[43] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011.

[44] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2097–2106.

[45] D. J. Stekhoven, I. Moraes, G. Sveinbjörnsson, L. Hennig, M. H. Maathuis, and P. Bühlmann, "Causal stability ranking," *Bioinformatics*, vol. 28, no. 21, pp. 2819–2823, 2012.

[46] J. K. Winkler, C. Fink, F. Toberer, A. Enk, T. Deinlein, R. Hofmann-Wellenhof, L. Thomas, A. Lallas, A. Blum, W. Stolz, and H. A. Haenssle, "Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition," *JAMA Dermatology*, vol. 155, no. 10, pp. 1135–1141, 2019.

[47] N. Kayser-Bril, "Google apologizes after its Vision AI produced racist results," *AlgorithmWatch*, 2020, https://algorithmwatch.org/en/story/google-vision-racism/.

[48] K. Chalupka, T. Bischoff, P. Perona, and F. Eberhardt, "Unsupervised discovery of el nino using causal feature learning on microlevel climate data," in *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence*, 2016, pp. 72–81.

[49] K. Chalupka, F. Eberhardt, and P. Perona, "Causal feature learning: an overview," *Behaviormetrika*, vol. 44, no. 1, pp. 137–164, 2017.

[50] D. Lopez-Paz, R. Nishihara, S. Chintala, B. Scholkopf, and L. Bottou, "Discovering causal signals in images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6979–6987.

[51] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015.

[52] J. Arevalo, A. Cruz-Roa, V. Arias, E. Romero, and F. A. González, "An unsupervised feature learning framework for basal cell carcinoma image analysis," *Artificial Intelligence in Medicine*, vol. 64, no. 2, pp. 131–145, 2015.

[53] C. T. Sari and C. Gunduz-Demir, "Unsupervised feature extraction via deep learning for histopathological classification of colon tissue images," *IEEE Transactions on Medical Imaging*, vol. 38, no. 5, pp. 1139–1149, 2018.

[54] J. Jiang, J. Ma, C. Chen, Z. Wang, Z. Cai, and L. Wang, "SuperPCA: A superpixelwise PCA approach for unsupervised feature extraction of hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 8, pp. 4581–4593, 2018.