Received October 5, 2019, accepted October 19, 2019, date of current version November 1, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2949286

Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View

OCTAVIO LOYOLA-GONZÁLEZ
Tecnologico de Monterrey, Puebla 72453, México
e-mail: octavioloyola@tec.mx

ABSTRACT Nowadays, in the international scientific community of machine learning, there exists an enormous discussion about the use of black-box models or explainable models; especially in practical problems. On the one hand, a part of the community defends that black-box models are more accurate than explainable models in some contexts, like image preprocessing. On the other hand, there exist another part of the community alleging that explainable models are better than black-box models because they can obtain comparable results and also they can explain these results in a language close to a human expert by using patterns. In this paper, advantages and weaknesses for each approach are shown; taking into account a state-of-the-art review for both approaches, their practical applications, trends, and future challenges. This paper shows that both approaches are suitable for solving practical problems, but experts in machine learning need to understand the input data, the problem to solve, and the best way for showing the output data before applying a machine learning model. Also, we propose some ideas for fusing both, explainable and black-box, approaches to provide better solutions to experts in real-world domains. Additionally, we show one way to measure the effectiveness of the applied machine learning model by using expert opinions jointly with statistical methods. Throughout this paper, we show the impact of using explainable and black-box models on the security and medical applications.

INDEX TERMS Black-box, white-box, explainable artificial intelligence, deep learning.

I. INTRODUCTION

Three decades ago, a part of the international scientific community working on computer science was focused on creating machine learning models for solving theoretical challenges [1], [2]. Nowadays, there exist both theoretical and practical progress in computer science. However, the international scientific community working on machine learning has begun an important debate about the relevance of the black-box approach and the explainable approach (a.k.a white-box approach) from a practical point of view [3], [4].

On the one hand, the international scientific community of machine learning has labeled as black-box models all those proposals containing a complex mathematical function (like support-vector machine and neuronal networks) and all those needing a deep understanding of the distance function and the representation space (like k-nearest neighbors), which are very hard to explain and to be understood by experts in practical applications [4]–[6]. On the other hand, those models based on patterns, rules, or decision trees are labeled as white-box models; and they, usually, ca be understood by experts in practical applications due to they provide a model closer to the human language [7]–[9].

There has been a trend of moving away from black-box models towards white-box models, particularly for critical industries such as healthcare, finances, and military (e.g. battlefields). In other words, there has been a focus to obtain white-box models, as well as fusing white- and black-box models for explaining to both experts and an intelligent lay audience the results obtained by the applied model [3], [4].

This trend is due to experts who are needing both understandable and accurate models. Also, currently, in several practical problems, it is mandatory to have an explanation of the obtained results. For example, the Equal Credit Opportunity Act of the US labels as illegal those denied credits to a

The associate editor coordinating the review of this manuscript and approving it for publication was Kim-Kwang Raymond Choo.

customer where there are indefinite or vague reasons; hence, any classifier used by the financial institution should provide an explanatory model [10].

Both white- and black-box approaches have shown accurate results for different practical problems, but usually, when one is suitable for a problem obtaining accurate results, then, another one obtains poor results [4], [8], [11]–[14]. As a consequence, some questions arise:

- What approach (black- or white-box) should be used in a practical problem?
- Is it necessary to move from a black-box approach to a white-box approach?
- Are the white-box-based models easy to interpret, and the black-box-based models very hard to understand?
- Is it always necessary that experts in the application domain should understand the black-box models?
- Is it feasible to fuse white- and black-box models?
- How to measure the effectiveness of the applied machine learning model by using expert opinions jointly with statistical methods?

All these questions are responded throughout this paper by using a state-of-the-art review for both, white- and black-box, approaches. Hence, the main contributions of this paper are:

- A review of the most outstanding models, following the white- or black-box approach, from a practical point of view.
- A practical analysis of both approaches taking into account their advantages and weaknesses.
- Ideas for obtaining new machine learning models by fusing white- and black-box models.
- One way to measure the effectiveness of the applied machine learning models by using expert opinions jointly with statistical methods.
- To the best of our knowledge, it is the first paper discussing about white- and black-box approaches from a practical point of view.

It is important to highlight that this paper is different to the ones recently published by [3], [4], where the author outlines several key reasons why explainable black-box models should be avoided in high-stakes decisions. In this paper we provide another point of view for understanding that both, white- and black-box, approaches are suitable for solving practical problems, but experts in machine learning need to understand the input data, the problem to solve, and the best way for showing the output data before applying a machine learning model.

This paper is organized as follows: Section II shows a brief introduction to the black-box approach as well as those relevant models following this approach. Also, advantages and weaknesses of the black-box approach in practical scenarios are presented. Next, Section III shows a similar structure than Section II but for the white-box approach. After, Section IV presents some papers fusing white- and black-box approaches and our points of view about how to fuse both approaches. Next, Section V shows one way to measure the effectiveness of the applied machine learning model by using expert opinions jointly with statistical methods. Finally, Section VI presents the conclusion of this paper.

II. BLACK-BOX APPROACH

The term *black-box* is mainly used for labeling all those machine learning models that are (from a mathematical point of view) very hard to explain and to be understood by experts in practical domains [3], [4], [8]. These black-box-based models can be grouped into the following categories: based on hyperplanes, like used by the support-vector machines (SVMs) [15]; inspired on the biological neural networks that constitute animal brains [16], [17]; based on probabilistic and combinatory logic, like the probabilistic logic networks (PLNs) [18], [19]; and those based on instances (a.k.a lazy learning) where the function is only approximated locally, like the k-nearest neighbors [20], [21]. Below we will detail the most used models within each of these categories and why they are labeled as black-box models.

Hyperplane-based models try to find a subspace that allows separating the problem's classes. In this way, a SVM-based model builds a hyperplane or set of hyperplanes in a high-dimensional space, which can be used for classification, regression, or other tasks like outliers detection [15], [22], [23].

For a better understanding, Fig. 1 shows an example based on data points each belong to one of two classes (green dashes and red crosses). Any hyperplane can be defined as a set of points $\vec{p}$ satisfying $\vec{w}*\vec{p}-a=0$, where $\vec{w}$ is the normal vector to the hyperplane and a is an arbitrary constant. From Fig. 1, we can see that there are two hyperplanes: $\vec{w}*\vec{p}-a=1$, where any point on or above this boundary belongs to the green class; and $\vec{w}*\vec{p}-a=-1$, where any point on or below this boundary belongs to another class. The region bounded by these two hyperplanes is called the "margin" and its distance between them is $\frac{2}{\|\vec{w}\|}$. The parameter $\frac{a}{\|\vec{w}\|}$ determines the offset of the hyperplane from the origin along the normal vector $\vec{w}$. In this example, we can see that the problem's classes are linearly separable, but usually, the practical problems are not linearly separable; consequently, other functions (a.k.a kernel) for separating those non-linear problems have been proposed [22].

One of the most outstanding application areas for the SVMs is biometrics [13], [14], [24]–[28]. For example, the SVM-based model has been widely used for face verification and recognition by using Local Binary Patterns (LBP) as feature extractor [26]. In the same vein, SVM has widely been used as classifier for iris and fingerprint authentication/verification and identification/recognition [13], [14], [27], [28].

Another context where the SVM-based model has widely been applied is author profiling. In this context, the main idea is to identify the authorship based on the information provided by several documents. For doing this, all gathered information is converted in vector feature representations, and after, using an SVM-based model, the authorship of a query text can be identified [29]–[31].

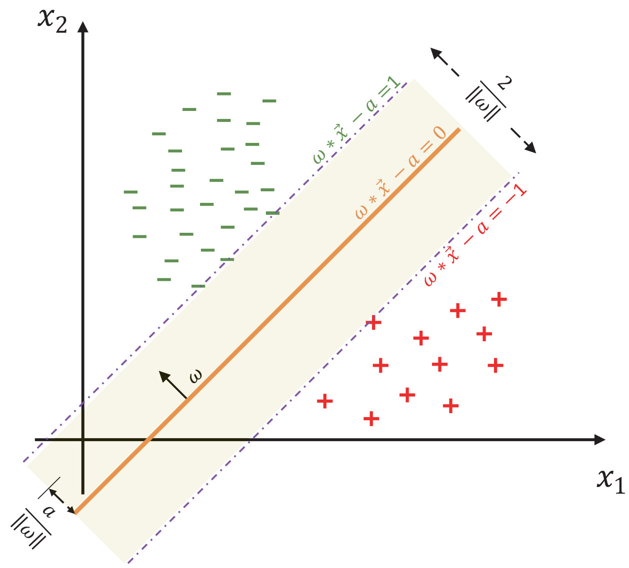

FIGURE 1. **Example of a two-class (green dashes and red crosses) problem, linearly separable by using hyperplanes. Notice that there are two hyperplanes (between the purple dashed line and the orange line). Any point on or above the upper purple dashed line belongs to the green class, and any point on or below the bottom purple dashed line belongs to another class. Hyperplane-based models are categorized as black-box models because it is complicated to understand this type of models applied to a more complex problem than shown in this figure, e.g., a problem with at least 10 classes, thousands of objects, and hundreds of features.**

From the above mentioned, we can say that SVM-based models have been used in several practical problems. However, the mathematical operations supporting this type of models are hard to understand for both machine learning experts and specialists in the application area. In this way, SVM-based models are categorized into the black-box approach.

It is important to highlight that nowadays, the computer science community continues creating new kernels and methodologies for improving the existing SVM-based models [15], [23], [32], [33].

Other of the models labeled as black box are those inspired on the biological neuronal networks. Artificial neural networks (ANNs) marked a milestone due their utilization of potential functional similarities between human and artificial information processing systems [16]. In the 80's decade, ANNs showed an important advantage in the supervised- and unsupervised-based classification using images as input. ANNs are based a set of artificial neurons, which are connected by using a weight that adjusts as learning proceeds [17].

From 1980 to nowadays, the approach behind the ANNs has been subject to continuous improvements; as a result, convolutional neuronal networks (CNNs) were proposed by [34], which marked a milestone. CNNs are regularized versions of multilayer perceptrons coming from ANNs, which are interconnected among them to create an input layer, at least one hidden layer, and an output layer [35]. Fig. 2 shows a typical CNN architecture. Using CNNs, the input image is processed by using several convolutional filters based on sliding dot products and cross-correlations; as a consequence, only spatially local correlations by enforcing sparse local connectivity patterns between neurons of adjacent layers are conduced [35].

In 2014, Goodfellow *et al.* proposed a new type of ANN labeled as Generative Adversarial Networks (GANs). A GAN-based model is trained using two ANNs. One is called *generator*, which learns to generate new plausible samples. The other is called *discriminator*, which learns to differentiate generated examples from real examples (see Fig. 3). Both ANNs are set up in a contested space (similar to the game-theory-based approach), where the generator network seeks to fool the discriminator network, and the latter should be able to detect generated samples from real ones. After training, the generator network can then be used to create new plausible samples on demand [36].

From our state-of-the-art review, we see that there exist a large number of interesting applications of GANs [37], such as generating examples for image datasets [36], [38], creating photographs of human faces, objects, and scenes [39]–[41], generating cartoon and Pokemon characters [42], image-to-image translation [43], [44], text-to-image translation [45]–[48], semantic-image-to-photo translation [49], face frontal view generation [50], generating new human poses [51], photos to emojis [52], photograph editing [53]–[55], face aging [56], [57], photo blending [58], super resolution [59]–[61], photo inpainting [62]–[64], video prediction [65], and 3D object generation [66], [67]; among others.

Recently, Razavi *et al.* proposed a Vector Quantized Variational AutoEncoder (VQ-VAE) model for large scale image generation. This model is able to scale and improve the previous VQ-VAE-based models [68], which allows generating synthetic samples of much higher coherence and fidelity than possible before. The proposal of Razavi *et al.* uses simple feed-forward encoder and decoder networks, which making this model an attractive candidate for applications where the encoding and/or decoding speed is critical.

From our review of GANs, we can say that nowadays, several researchers are moving to this type of networks. The main reason is that GANs are providing solutions to practical problems, which are revolutionizing the computer science; mainly, where the images are the inputs of the problem. Nevertheless, other approaches (like probabilistic networks) have shown good results in other contexts where images are not the inputs of the problem [19], [69].

It is important to highlight that ANN-based models (CNN and GANs families) are the most difficult to understand by both machine learning experts and specialist in the application area due to the several transformations made to the input data.

Probabilistic networks is one of the pioneering approaches used for solving practical problems, which continue showing good progress on practical problems [19], [70]. Markov networks [18], [70] and Bayesian networks [19], [71] are the

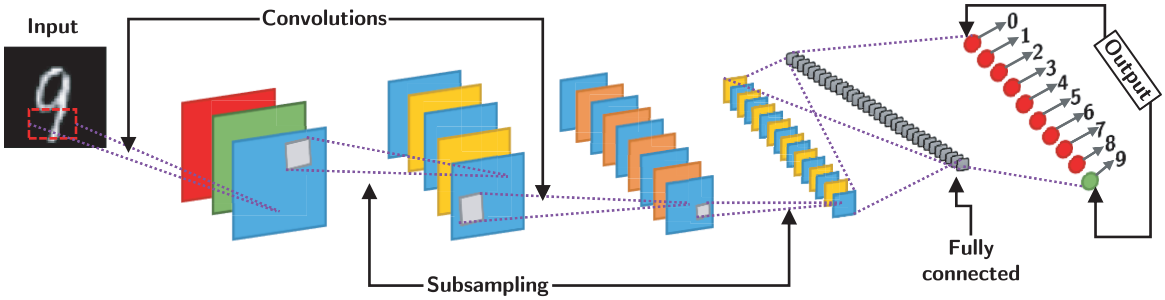

FIGURE 2. Example of a typical CNN architecture for hand-written digits recognition. The input is a handwritten digit, after, there are several hidden layers, and finally, using the fully connected layer the output is a label representing a number between zero and nine. Note that, for human, is hard to understand the number of mathematical operation used behind the convolutional filters applied in each network's layer.

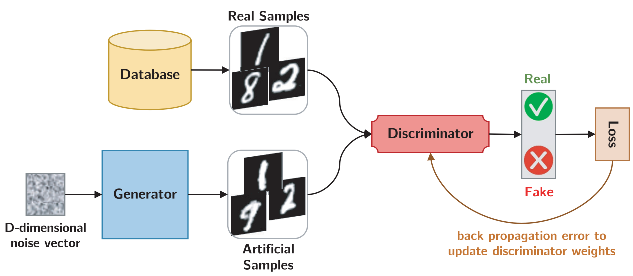

FIGURE 3. Example of a typical GANs architecture for both hand-written digits recognition and generation. The discriminator tries to identify between real samples and those artificial samples generated by other ANN. Both, generator and discriminator are trained by using a back propagation procedure.

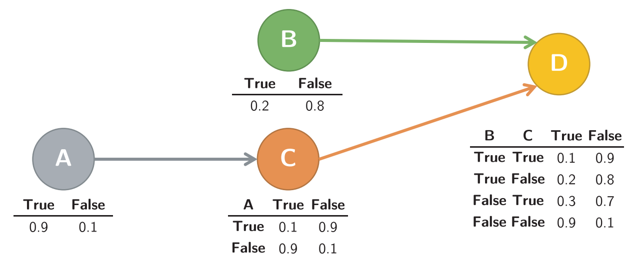

FIGURE 4. Example of a Bayesian Network architecture and its probabilities. Each circle represents a feature and its probabilities by using (or not) the probabilities of other features. Each circle has associated a table showing probabilities. Note that each row of each table should sum 1. Note that a graph similar to this image but containing hundreds of nodes and probability tables is more complicated to understand by a human than other structure like a decision tree, which is acyclic and it does not contain probability tables.

main models used for solving practical problems by using a probabilistic approach.

On the one hand, a Markov network [70] is based on the joint distribution of a set of features $F = \{f_1, f_2, \ldots f_n\}$, which is composed of an undirected graph G and a set of potential functions $\Phi = \{\phi_1, \phi_2, \ldots \phi_n\}$. The graph contains a node for each feature, and the model has a potential function for each clique in the graph. A potential function is a non-negative real-valued function of the state of the corresponding clique. The joint distribution represented by a Markov network is given by $P(F) = \frac{1}{Z}\prod_k \phi_k(f_{\{k\}})$, where $f_{\{k\}}$ is the state of the kth clique (i.e., the state of the variables that appear in that clique), and Z (known as the partition function) is computed by $Z = \sum_{f \in F}\prod_k \phi_k(f_{\{k\}})$.

Markov network-based models have been used in different applied context, such as financial engineering [72], image annotation [73], social network analysis tools in ecology [74], and tool-wear monitoring [75]; among others. For applying Markov network-based models is necessary that the problem can be represented as a graph, and the input data provides a timeline, which difficult its application in some contexts [69].

On the other hand, a Bayesian network (BN) is a probabilistic directed acyclic graphical model [19], [71]. BNs use nodes to represent features, arcs to signify direct dependencies between the linked nodes and conditional probabilities to quantify the dependencies (see Fig. 4). A set of features $F = \{f_1, f_2 \ldots f_n\}$ can be represented in a BN by using a directed acyclic graph with n nodes, where each node j ($1 \leq j \leq n$) is associated with each f_j feature. It can be represented by $P(f_1, f_2 \ldots f_n) = \prod_{j=1}^{n} P(f_j|\Psi(f_j))$; where $\Psi(f_j)$ denotes the set of features in the graph, which are connecting the node i with the node j [19].

BN-based models have been applied on several practical problems, such as diagnosis of Alzheimer's disease [76], fault location on distribution feeder [77], human cognition [78], educational testing [79], analysis of resistance pathways against HIV-1 protease inhibitors [80], ovarian cancer diagnosis [81], fault diagnostic system for proton exchange membrane fuel cells [82], information retrieval [83], inferring missing climate data for agricultural planning [84], and predicting protein-protein interactions from genomic data [85]; among others.

It is important to highlight that the main drawback of BN-based models is that their network structure depends on the order of all problem's features. If the order is chosen carelessly, the resulting network structure may fail to reveal many conditional non-dependencies in the domain. Consequently, BNs are challenging to apply for solving practical problems [86].

Some authors [85] can consider BN-based models as an interpretable model compared with other ANN-based models. However, in a practical problem, a BN-based model

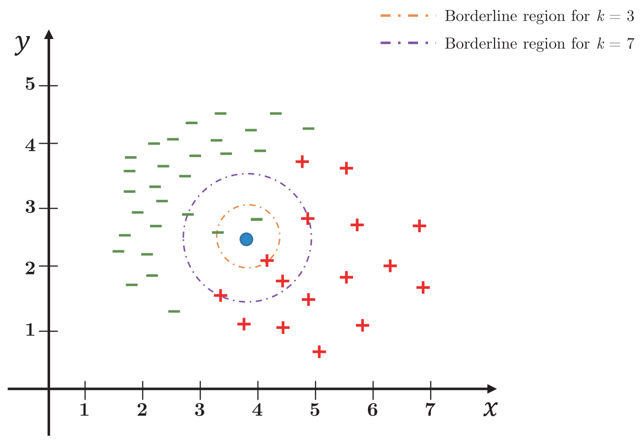

FIGURE 5. Example of kNN-based classification based on the Euclidean distance and the majority vote strategy. The query object (blue-fill circle) should be classified either to the green-dash class or to the red-cross class depending of the *k* value (3 or 7) and its borderline region. Note that for problems containing nominal features and which need to use other distance function (different than the Euclidean distance), it is no easy to understand this type of model.

produces a graph containing hundreds of nodes and probability tables. Consequently, this BN-based model is more complicated to understand by a human than other models like a decision tree, which is acyclic, and it does not contain probability tables. For that reason, BN-based models are considered as black-box models.

The last one labeled as a black-box approach is lazy learning. Models following this approach are based on a target function, which is approximate locally [20]. One of the most outstanding models following the lazy learning approach is the nearest neighbors-based model [21]. The k-nearest neighbors-based models (kNN) rely on a target function f for given a query object o and a training dataset T. Then, the aim is to find the k-nearest neighbors in T by using f that allow classifying o taking into account a classification strategy. For example, Fig. 5 shows an example of kNN-based classification based on the Euclidean distance and the majority vote strategy. The query object (blue-fill circle) should be classified either to the green-dash class or to the red-cross class. If $k = 3$ (dashed-line orange circle), the query object is assigned to the green-dash class because two objects are belonging to the green-dash class and only one object belonging to the red-cross class on or inside the inner orange circle. If $k = 7$ (dashed-line purple circle), the query object is assigned to the red-cross class because four objects are belonging to the red-cross class vs. three objects belonging to the green-dash class on or inside the purple circle.

Lazy learning-based approach has widely been used in several practical contexts, such as data streams [87], air quality planning [88], predictive toxicology based on a chemical ontology [89], prediction of customer demands [90], hydrologic forecasting [91], air quality prediction [92], magnetorheological damper [93], and spam filtering [94]; among others.

Lazy learning-based models, like kNN, suffer from other intrinsic problems coming from application domains, such as the class imbalance problem [95], [96], and the class confidence problem; where objects are weighted regarding their confidence to all classes of the problem [97].

It is important to highlight that lazy learning-based models mainly rely on a distance function, which biases the classification results. Ideally, a distance function should arise from the interaction between experts in the application domain and mathematics specialists due to experts have the known about the problem to solve and provide important information about the comparison between objects containing nominal values. However, in practice, experts in machine learning use different distance functions proposed for general purposes, and they select the best one according to previous results [98]. Also, it is known that learning-based models obtain different classification results when the distance function is changed [99]. The before stated comments are the main reasons for labeling the lazy learning-based models as black-box models.

A. DISCUSSION

After we have stated the main models following the black-box approach, it is important to expose why this approach could obtain good or bad results from a practical point of view. For that, firstly, we need to understand some important keys, which are related below.

Commonly, experts in the application domain have focused their knowledge on understanding the phenomenons of their expertise area instead of learning about machine learning [4]. As a consequence, for most of these experts it is complicate to understand models containing a complex mathematical function (like SVM and Neuronal Networks) or those needing a deep understanding of the distance function and the representation space (like kNN). For most of these experts, any application based on machine learning should provide help for decision making in a clear and precise way. Even, most of these experts claim that they do not need to understand the mathematical supporting the machine learning model applied in their expertise domain [6], [100], [101].

Based on the above, the following question arises: what is the reluctance of some experts of applying black-box models in their application domain? For answering this question, first we need to understand about how people learn and trust naturally.

Usually, people are learning from experts or teachers, which based on logical reasoning and images can transmit their knowledge to apprentices [102], [103]. This method must be respected when experts in machine learning try to apply an artificial intelligence-based model to a practical problem because if we break this method then, we can obtain reluctance from experts in practical problems.

For example, Fig. 6 shows a mammography image (left-hand side), and after using a Faster R-CNN model, an image (right-hand side) containing two red bounding-boxes for delimiting those zone of possible lesions is showed.

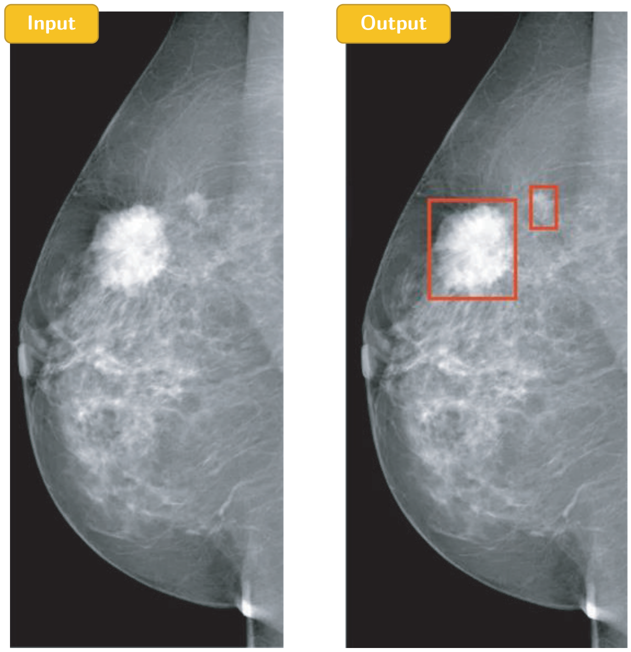

FIGURE 6. Example of a mammography image (left-hand side), and after using a Faster R-CNN model, the output image (right-hand side) containing two red bounding-boxes for delimiting those zone of possible lesions. Both images were taken from the paper published by [104]. Note that for an expert in the application area is easy to understand the output image and detect the affected zones, independently of the mathematical operations executed by the applied model.

These images were taken from [104], where the authors proposed a computer aided detection (CAD) system based on a VGG16 network, which is a 16 layer deep CNN [105]. The proposed model can detect two types of lesions: benign and malignant.

The CAD system, proposed in [104], outputs an image that is easy to understand by an expert (in this case an oncologist) because this system outputs red bounding-boxes for delimiting those zone of possible lesions, which facilitate to localize the possible affected zones from a mammography image. Notice that this model is based a CNN of 16 layers, which contains a strong mathematical foundation based on several convolutional filters. This model is hard to understand by the expert in the application domain but its output is straightforward to understand due to the visual form used. It is an excellent example of how to use a black-box-based model from a practical point of view because the output produced by the model is in correspondence with its input data.

On the other hand, there exist other practical contexts where it is hard to obtain an understandable output for experts in the application domain when a black-box approach is used. For example, in author profiling the main idea is to identify the authorship based on the information provided by several documents [29]. For doing this, usually, all collected information is converted in vector representations, and after, using an SVM-based model, the authorship of a query text can be identified [29]–[31]. Notice that, commonly, the input data in the author profiling problem are texts but after several

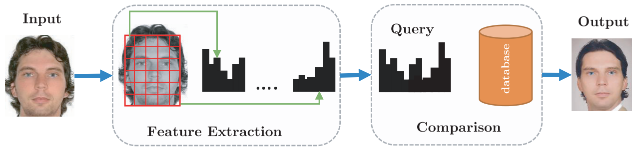

FIGURE 7. A face verification system diagram. First, the raw image is processed for extracting features, in this case, using a local binary pattern procedure. After that, a pairwise comparison between the extracted vector and each vector contained in the databases is executed. Finally, the system output the image corresponding to the vector more similar found in the database. Note that it is easy for experts in biometrics to compare the input and output images. The author of this paper has provided his face images to be used in this figure as illustrations of the face verification system diagram.

mathematical operations they are transformed to vector representations. The final result is hard to understand by experts in the application domain because the output result is very different to the input data. In other words, the input domain (which is known by the expert) was changed by a new domain non-understandable by the expert.

It is important to highlight that the difficulty of understanding the model's output in several practical problems is not attributed to the use of a black-box-based model but the transformation of the output data. For example, the SVM-based model has widely been used for face verification and recognition [24], [25] but although the input images suffers several transformations during the training, the output result provided by the model is an image understandable by experts in the domain application (see Fig. 7).

Based on the above mentioned, we can conclude that, on the one hand, experts in the application domain do not need to understand the mathematical transformation behind the applied black-box-based model, they only need a natural form of interpreting the model's output. For doing that, the easiest way is to provide to experts the output data in the same form that they provided the input data but highlighting the important keys discovered by the applied model.

On the other hand, it is mandatory for experts in machine learning to understand the black-box-based model to be applied due to, in most occasions, these models need to be tuned for obtaining accurate results. Consequently, experts in machine learning need to know where the model is failing or it needs to be tuned, and for doing that, they need to understand the black-box-based model in depth. This last it is hard to achieve because some black-box-based models, like CNN, are designed for applying several transformations over the input data and debugging these models, at any stage, is not a straightforward task.

III. WHITE-BOX APPROACH

The terms *white-box*, *understandable model,* and *explainable artificial intelligence* (XAI) are used for labeling all those machine learning models providing results associated to their models that are easy to understand by experts in the application domain. Usually, these models provide a good trade-off between accuracy and explainability [3], [4], [106]. Usually, the terms understandable and interpretable are used

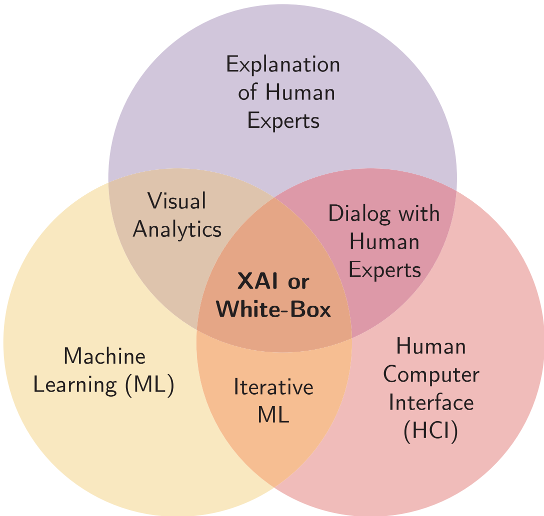

FIGURE 8. A graphic showing the interaction of the different areas forming the XAI model.

for referring to all those models providing an explanation to experts in the application area. However, based on the explanation provided by [4], an understandable model refers to those machine learning needing an additional model or other features for providing an explanation to experts in the application area. On the other hand, an interpretable model is able to provide explanations to experts without using any additional model. In this paper, we will refer to this type of models as white-box models.

As we showed in Fig. 8, an explainable model is possible by the interaction of different areas, such as machine learning, human computer interface, explanation of human experts, visual analytic, iterative machine learning, and dialog between machine learning experts and human experts in the application domain [4], [106].

There exist different families of algorithms following the white-box approach but we will describe the most used in the literature, such as decision trees [1], [2], [107], rule-based systems [108], [109], contrast patterns [7], [8], [110], and fuzzy patterns [5], [110].

Decision tree-based model was a pioneer model providing both accurate results and an understandable model for experts in the application domain [1], [2], [107]. A decision tree (DT) contains elements of tree structure, based on the graph theory, and decision support system. Consequently, a DT can be described as an directed graph in which any two vertices are connected by exactly one path [2], [107].

There are two approaches for inducing decision trees: top-down and bottom-up [111]. The top-down approach is the most used from them, which is based on the divide and conquer approach [107]. For inducing a top-down-based decision tree, it starts building a root node with all objects of the training database D. Then, it splits the root node into two disjoint subsets (left child D_l and right child D_r) and repeats this process recursively over the children nodes until certain stopping criterion is met [111].

Fig. 9 shows an example of a binary decision tree where the root node (grey rectangle) contains 30 objects belonging to the A class and other 100 belonging to the B class. After, the objects are recursively distributed into two disjoint subsets (green and blue ovals) by using test conditions. Finally, orange squares represent the leaf (or decision) nodes, which based on some classification strategy will define the class of a query object.

The above-explained procedure allows inducing just one decision tree. However, several authors have shown that using a model containing several and diverse decision trees attains significantly better classification results than using only one decision tree; even, than other popular state-of-the-art classifiers, which are not based on decision trees [112]–[114]. For inducing diverse decision trees, there exist three popular ways:

Random: it selects a subset of features randomly and, by using the selected features, generating as many binary splitting criteria as possible depending on the type of the feature [1], [2], [107]. This procedure is used at each level of the decision tree while it is being induced.

Bagging: it selects a subset of objects randomly from the training dataset for each decision tree to be induced [113].

Boosting: it selects a subset of object randomly from the training dataset and after that, it takes the remaining no selected objects for validation where each object is weighted. Finally, based on the weight assigned to each object, a new subset of objects randomly from the training dataset is selected to induce a new decision tree [113].

Decision tree-based models have been used for solving several practical problems, such as classification of cancer data by analyzing gene expression [115], diagnosis and drug codes for analyzing chronic patients [116], medical diagnosis [117], diagnosis of sport injuries [118], prediction and sensitivity analysis of bubble dissolution time in 3D selective laser sintering [119], and part-of-speech tagging [120]; among others.

The main advantages of the decision trees are [1], [2], [107], [111]:

- They are self-explanatory and when contain a reasonable number of leaves, they are also easy to follow.
- They can handle both nominal and numeric input features as well as missing values.
- Ensemble of decision trees can deal with noisy objects and outlier data.

On the other hand, as any machine learning model, decision trees have disadvantages [1], [2], [107], [111]; among the most important are:

- An ensemble of decision trees is complicated to understand by experts in the application domain due to the number of decision trees induced, e.g., some authors have stated that an ensemble containing 100 decision trees obtains the best classification results among the

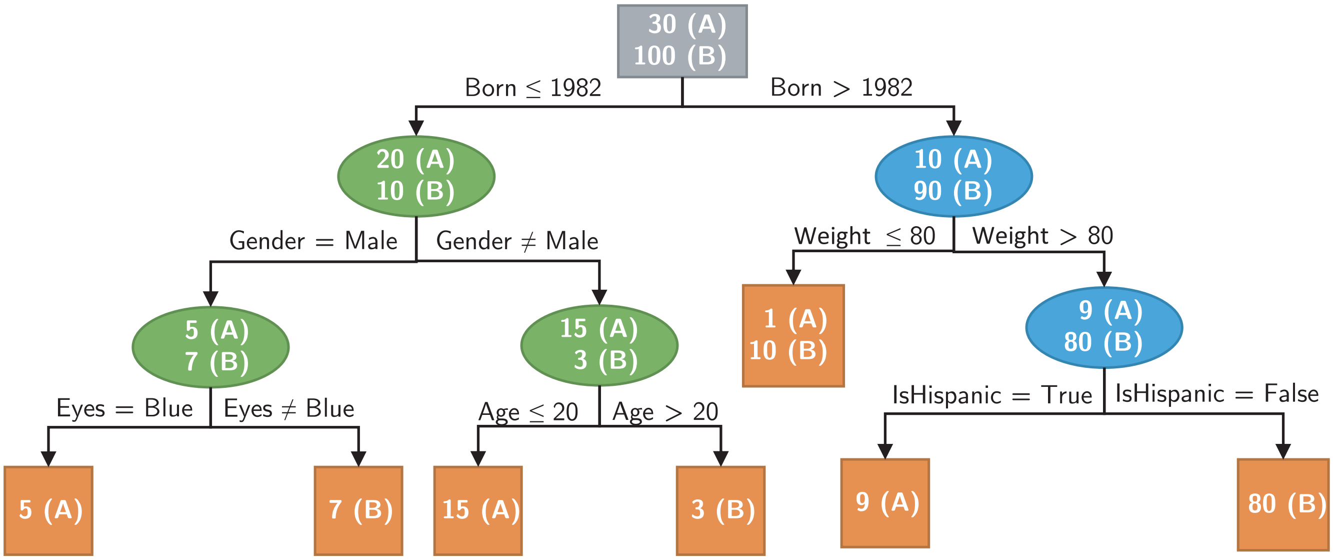

FIGURE 9. A example of a binary decision tree. The root node (grey rectangle) contains 30 objects belonging to the A class and other 100 belonging to the B class. The objects are recursively distributed into two disjoint subsets (green and blue ovals) by using test conditions. The orange squares represent the leaf (or decision) nodes.

different number of decision trees tested for creating a DT-based ensemble [1], [111]. Notice that for a human, it is complicate to understand 100 decision trees.

- They present over-sensitivity to the training dataset, which makes the model unstable. Small variations in the training dataset can cause a different selection of one split candidate near the root node, and consequently, all the subtree will change [107].
- An ensemble of decision trees can provide the same path (from the root node to a leaf node) in several of the induced decision trees, which can overwhelm the other important paths.

As was stated by [107], rules can be extracted from decision trees. Rule-based models are considered as understandable models. A rule is an expression describing a collection of objects. Usually, a rule is represented as a logic implication (IF-THEN) by using a conjunction of relational statements (a.k.a *antecedent*), and another part (a.k.a *consequent*) implicating one or more labels [108].

Agrawal *et al.* proposed an algorithm for mining rules, which is able to discover regularities between products in large-scale transaction data recorded by point-of-sale (POS) systems in supermarkets. For example, $\{milk, butter\} \Rightarrow \{bread\}$, which would indicate that if a customer buys milk and butter together, they are likely to also buy bread. Such rule can be used for decisions about marketing activities [108].

Following the definition proposed by [109], the problem of rule mining is defined as: Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of n binary features called items, and let $T = \{t_1, t_2, \ldots, t_m\}$ be a set of transactions (a.k.a database); where each transaction in T is unique and contains a subset of the items in I. Then, a rule is defined as an implication of the form: $X \Rightarrow Y$; where $X, Y \subseteq I$.

Usually, algorithms for mining rules are prone to extract several rules. Consequently, several quality measures have been proposed for evaluating the extracted rules. The best-known quality measures for rules are support and confidence [121], [122].

On the one hand, the support for a given set of items X regarding to a given database T is computed as the proportion of transactions $T' \subseteq T$ containing X. On the other hand, the confidence value for a given rule, $X \Rightarrow Y$ regarding to a given database T, is computed as the proportion of the transactions $T' \subseteq T$ that contains X which also contains Y; i.e., support of $X \cup Y$/ support of X .

Although, Agrawal *et al.* proposed their algorithm for mining association rules in contexts where there are not classes (a.k.a *unsupervised classification*), other authors, like [1], [107], introduced variations of the algorithm proposed by Agrawal *et al.* for mining rules from decision trees in those contexts containing classes (a.k.a *supervised classification*).

Rule-based algorithms have widely been applied to different practical contexts, such as software defect prediction [123], inferring causal gene regulatory networks [124], evaluating the efficiency of currency portfolios [125], ranking of text documents [126], relationship between student engagement and performance in E-learning environment [127], assessing web sites quality [128], and exploring shipping accident contributory factors [129]; among others.

Rule-based systems have shown good classification results in several practical problems. Also, these systems provide rules understandable to experts in the application domain. Nevertheless, rule-based models have some drawbacks, such as exponential complexity [130], they need an a-priori discretization for all numerical features [131], and rule mining strategies provides a large number of rules, which cannot be handled by experts effectively [132].

At the end of the 90s, Dong & Li proposed the contrast pattern-based model, which is similar to a rule-based model, but it has not the consequent as a rule has. A pattern is an expression defined in a certain language that describes a collection of objects [7], [8], [100], [110], [133], [134]. Usually, a pattern is represented by a conjunction of relational statements (a.k.a *items*), each with the form: $[f_i \ \# \ v_j]$, where v_j is a value in the domain of feature f_i, and # is a relational operator from the set $\{\in, \notin, =, \neq, \leq, >\}$ [7], [8], [135]. For example, $[Hour\_in\_Server \in [0, 5]] \wedge [Number\_of\_URL > 5] \wedge [Number\_of\_Follower \leq 10] \wedge [Mobile = \text{``}False\text{''}]$ is a pattern describing a collection of tweets issued from a botnet [9]. Let p be a pattern, and $C = \{C_1, C_2, C_3, \ldots, C_n\}$ a set of classes such that $C_1 \cup C_2 \cup C_3 \cup \ldots \cup C_n = U$; then, support for p is the fraction resulting from dividing the number of objects belonging to C_i described by p by the total number of objects belonging to C_i [7], [8], [110], [135], [136]. A contrast pattern (cp) for a class C_i is a pattern p whereby the support of p for C_i is significantly higher than any support of p for every class other than C_i [7], [8], [110], [135], [136].

For building a contrast pattern-based classifier, there are three phases [7], [8], [110]:

Mining: this phase is dedicated to finding patterns from a training dataset by an exploratory analysis using a search-space, which is defined by a set of inductive constraints provided by the user. Contrast pattern mining algorithms can be grouped into two groups: (i) exhaustive-search based algorithms, which perform an exhaustive search of combination of values for a set of features appearing significantly in a class regarding the remaining classes; and (ii) decision trees based algorithms, which extract cps from a collection of decision trees.

Filtering: as usually several patterns are extracted at the mining phase, filtering is dedicated to select a set of high-quality patterns, which allows obtaining equal or better results than using all patterns extracted at the first phase. Filtering algorithms are divided into two groups: (i) based on set theory, which are suitable for removing redundant items and duplicate patterns, as well as removing specific patterns (a.k.a maximal patterns); and (ii) based on quality measure, which allow generating a pattern ranking based on the discriminative power of the patterns.

Classification: it is responsible for searching the best strategy for combining the information provided by a collection of patterns, which allows building an accurate model based on patterns. Classification strategies are divided into two categories: (i) those providing an unweighted score, which are easy to compute and understand, but they can be affected by the nature of the problem; for example, on class imbalance problems; and (ii) those providing a weighted score, which are suitable for handling balanced and imbalanced problems.

Algorithms for mining contrast patterns can use an exhaustive search, like rule-based models or decision trees, for extracting contrast patterns from a training dataset. Some authors [107], [137] have shown that those algorithms based on decision trees have advantages regarding those approaches not based on trees for extracting contrast patterns. First, the local discretization performed by decision trees with numeric features avoids doing a priori global discretization, which might cause information loss. Second, with decision trees, there is a significant reduction in the search space of potential patterns.

Contrast pattern-based classifiers have been applied in several practical problems, such as bot detection on twitter [9], road safety [138], crime pattern [139], cerebrovascular examination [140], describing political figures [141], sales trends [142], detection of frequent alarm patterns [143], bot detection on weblog [6], complex activity recognition in smart homes [144], network traffic [145], and studying the patterns of expert gamers [146]; among others.

Contrast patterns were improved by using fuzzy sets [147], deriving a new type of pattern called fuzzy pattern [5], which shows a language closer to the human experts than provided by contrast patterns. A fuzzy pattern is a pattern containing conjunctions of selectors $[Feature \in FuzzySet]$, where $\in$ is the membership of the feature value to FuzzySet. This way an object satisfies a given pattern to a certain degree according to the degree the object feature values satisfy the item expressed in the pattern. For example, $[Temperature \in hot] \wedge [Humidity \in normal]$ is a fuzzy pattern describing the weather in a fuzzy domain. For mining fuzzy patterns, first, this procedure creates a fuzzification of all features. For non-numeric features, a collection of singleton fuzzy sets is created, i.e. for each different value, a fuzzy set having membership 1 for that value, and 0 for the remaining values is created. For numeric features, a traditional fuzzification method is applied. After that, [5] use a fuzzy variant of the ID3 method [2] for building a set of different fuzzy decision trees, from where several fuzzy patterns are extracted.

García-Borroto *et al.* proposed extending the fuzzy patterns by using linguistic hedges (e.g., "very", "often", and "somewhat"), which are commonly used for fixing the discretization of continuous features. These fuzzy patterns look more closely to the language used by experts than other types of patterns.

For a better understanding of the difference between the three types of patterns aforementioned, we provide an example of each one for the same domain.

CP. $[Temperature > 35] \wedge [Humidity \in [55, 70]$, it is a contrast pattern.

FP. $[Temperature \in hot] \wedge [Humidity \in normal]$, it is a fuzzy pattern using discretization of continuous features.

FPLH. [*Temperature is very*(*hot*)]$\wedge$ [*Humidity is somewhat*(*normal*)], it is a fuzzy pattern by using linguistic hedges.

Notice that the three items aforementioned are patterns describing the same domain but the fuzzy pattern by using linguistic hedges looks more closely to the language used by experts than other two types of patterns.

Models based on decision trees, rules, or patterns have their major advantage in the explanation provided by them, further their accuracy. However, like all classification model, they have important drawbacks, which we mentioned below:

- Their explanatory power is limited by the nature of input data and the feature representation. For example, they cannot provide a suitable explanation when the input data are images.
- Usually, they provide a considerable number of patterns, rules, or decision trees, which are hard to understand by experts in the application domain, and the proposed filtering methods do not provide a reasonable filtered collection of them.

A. DISCUSSION

As we stated in Section II and this one, black-box-based models are so good as white-box-based models are. Their performances depend on the application domain and the input data. In practical contexts, the results obtained by a white-box-based model where the input data are described in a matrix containing features and values issued by experts in the application area are easier to understand by that experts than other types of feature representation, which are transformations of input data. For example, a rule-based system using images as input data for document image segmentation provides rules like *if* $((H_{t,min} \leq height < H_{t,max})$ *AND* $((A_{t,min} \leq aspect\ ratio < A_{t,max}))CC \Rightarrow TEXT$ and *if* $(aspect\ ratio \leq A_{L,min})CC \Rightarrow NONTEXT$, which are hard to understand by experts in the application area and even, usually, these types of system output a huge number of rules; being more complicated for understanding [148]. On the other hand, there exist black-box-based models proving both cropped images and bounding boxes for document image segmentation [149], which are easier to understand by experts in the application area than white-box-based models, like rule-based system.

As we have stated on before sections, it is essential to understand the input data as well as the best way for showing the output data to experts before applying a machine learning model to a practical problem. For some contexts a white-box-based model is better than using a black-box-based models; for other contexts it is the other way around. However, there are machine learning experts fusing both white- and black-box approaches for obtaining the best performance in practical applications.

IV. FUSING BLACK- AND WHITE-BOX APPROACHES

On the one hand, in several contexts, black-box-based models have shown better classification results than white-box-based models. However, in most of these contexts, an accurate model is not the only desired characteristic by experts in the applied context, the used model should provide an explanation of their results, which should be easy to understand by experts in that context [7]–[9].

On the other hand, there exist several contexts where white-box-based models obtain a good explanation of the problem's classes for experts in the application context, but they obtain lower accuracy than other black-box-based models. Consequently, new models fusing white- and black-box approaches are necessary [150]–[152].

One of the pioneer ideas for fusing white- black-box approaches was unifying decision tree-based classification with the representation learning functionality known from deep convolutional networks by training them in an end-to-end way [150]–[155]. This approach starts with random initialization of the decision nodes parameters and iterates the learning procedure for a given number of epochs.[1] At each epoch, it initially obtains an estimation of the prediction node parameters given the actual value of by running an iterative scheme, starting from the uniform distribution in each leaf. Then, it splits the training set into a random sequence of mini-batches.[2] After each epoch, it could eventually change the learning rate according to predetermined schedules. In Fig. 10, we show how the hidden layer and the leaf nodes are connected into a Deep CNN for obtaining a random forest by using the representation learning functionality.

As we stated in Section III, from decision trees several rules can be extracted. Consequently, several authors [156]–[158] have proposed to extract rules from different deep convolutional networks.

For extracting rules from CNN, the main idea is processing every hidden layer in a descending order for obtaining a set of rules for each class (CNN output). This approach extracts rules for each hidden layer that describes its behaviour based on the preceding layer. At the end, all rules for one class are getting merged [156]–[158].

Although models fusing rule and CNN have been used in some applied context like aspect level sentiment analysis [158] and semantic image labeling [153], the extracted rules are not easy to understand by experts in the application domain due to these rules are extracted from several hidden layers into the CNN model, which contain connections inherent from the CNN. Also, the applied pruning strategies for rules do not provide a subset of rules easy to understand by experts [156]–[158]. Hence, other authors [150] have proposed to integrate both visualization techniques and rules for understanding the output provided by CNN. Samek *et al.* proposed to use a heatmap and a rule-based model for extracting how much does each pixel contribute to prediction and how much do changes in each pixel affect

[1] An epoch is a hyperparameter belonging to deep learning-based models. It is considered an epoch when an entire dataset is passed both forward and backward through the neural network only once [34].

[2] Batch is a hyperparameter used by deep learning-based models for representing a proportion of the total number of training objects [34].

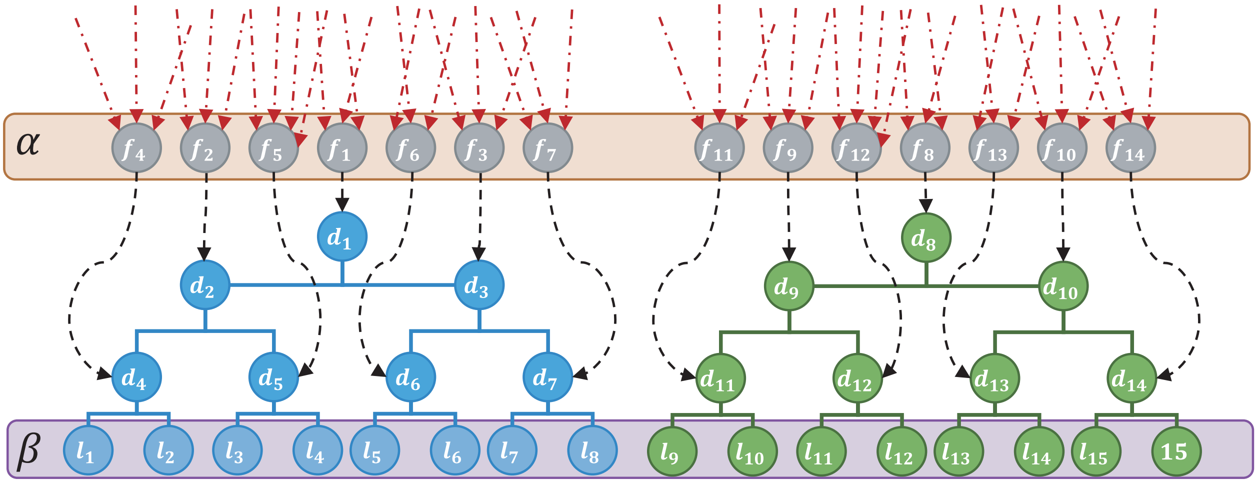

FIGURE 10. Red-dashed arrows are hidden layers coming from a Deep CNN. The α block represents the fully connected layer used to provide functions $F = \{f_1, f_2, f_3 \ldots f_n\}$. Each output of f_i is brought in correspondence with a split node in a tree, eventually producing the routing (split) decisions $d_i(x) = \sigma(f_i(x))$. The order of the assignments of output units to decision nodes can be arbitrary (the one we show allows a simple visualization). The block of circles at bottom (β) correspond to leaf nodes, which contain probability distributions. All blue circles represent a decision tree and the green circles another one.

the prediction. Nevertheless, the proposed explanatory model is yet hard to be understandable by experts in the application area [156].

Recently, Kanehira & Tatsuya proposed a framework to generate complemental explanations by using three different neural networks: predictor, linguistic explainer, and example selector. The aim is that given a query object o and a class c associated to o, the model projects them to the common-space and element-wise summation is applied. After one more projection, they are normalized by the the softmax[3] function. The output vector of the network $f(o, c)$ is the same as that of the vector representation, and each vector indicates the probability that each type of feature is selected as an explanation [159].

Fig. 11 shows an example for illustrating how to work the proposal of [159]. The main drawback of the Kanehira & Tatsuya's proposal is that the image databases should contain additional features describing the images, which is no common in the image databases.

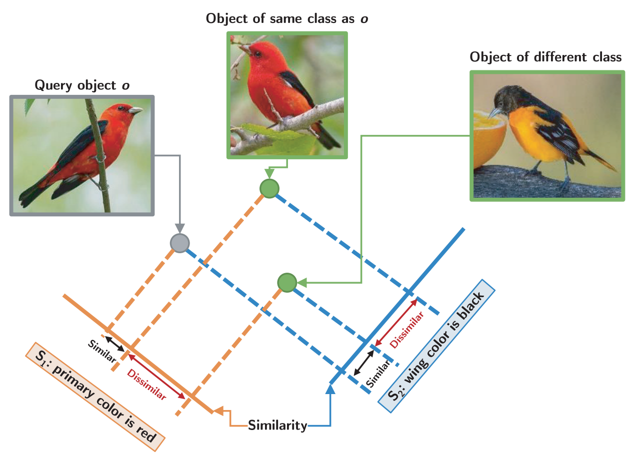

FIGURE 11. The model proposed by [159] predicts the query object o (gray circle) by referring other objects based on the similarity space (orange and blue non-dashed lines) corresponding to each linguistic explanation (S_1 and S_2).

A. IDEAS FOR FUSING WHITE- AND BLACK-BOX APPROACHES

Although models fusing white- and black-box approaches have recently been proposed, and they have shown good classification results, they need to improve their explanatory models to provide an explanation close to the human language [151], [152], [156]. Consequently, we will state a couple of ideas for doing that, which, as far as we know, have not been proposed before.

One idea is to extract information and a feature representation using one of the approaches (white- or black-box) and after, use another approach for extracting more information and other feature representation. After that, we need to prove that both feature representations contain a strong correlation, and finally, the result provided by an approach should be complemented by the result provided by another one. For example, see Fig. 12, imagine that we need to classify human brain cancer (malignant or benign) from an input image, where both white- and black-box approaches contribute together to the solution and understanding of the model's output. A possible solution is to use a CNN for both detecting and classifying the types human brain cancer and by other way, extract important information from both the image and the patient's clinical

[3]The softmax is a function taking as input a vector of n real numbers, and normalizes it into a probability distribution consisting of n probabilities proportional to the input numbers.

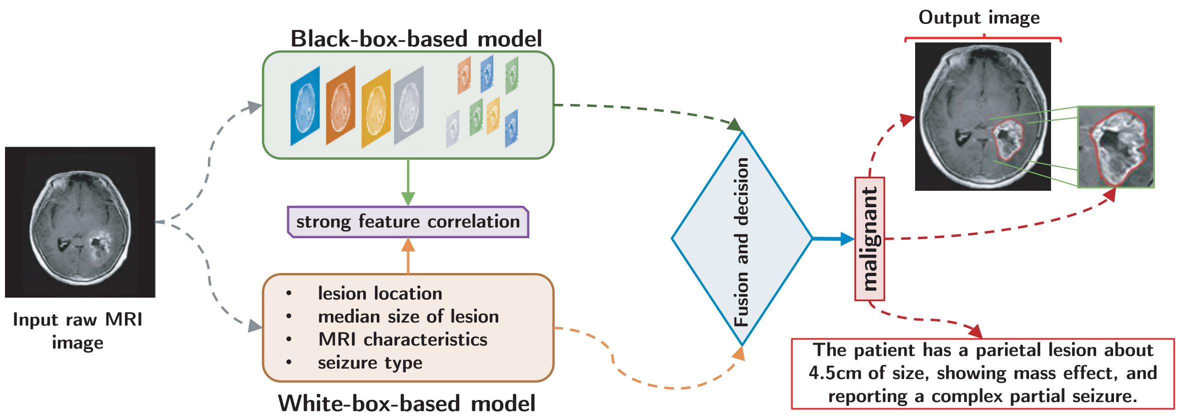

FIGURE 12. An example of fusing white- and black-box approaches for human brain cancer diagnosis. The brain cancer images were taken from [160] and the human brain cancer characteristics were extracted from [161]. Note that this idea of fusing white- and black-box approaches can provide both accurate and understandable results.

TABLE 1. A taxonomy for the white- and black-box approaches as well as the fusing of both, taking into account different characteristics.

Items		White-box	Black-box	Fusing approaches
Input Data*	Image		x	x
	No-Image	x	x	x
Feature Type	Numerical	x	x	x
	Non-Numerical	x		x
Interpretability of the model's output	Input Images		x	x
	Input Texts	x		x
Working with missing values	Modifying the input data	x	x	x
	Original data	x		x

*It refers to what type of input data is the best for the models following any of the reviewed approaches.

history, in a language close to the human expert. Finally, the model's output provide an image with a bonding box for detecting the affected area, a classification of the type of brain cancer, and a description in a language close to the human expert for explaining the classification.

Another idea is to design a stacking approach considering some white- and black-box models by using a structure similar to the one stated in Fig. 12. Using a stacking approach, the predictions from both models are used as inputs for each sequential layer, and combined to form a new set of predictions. A stacking approach can deal when the white-box-based model's output differs from the black-box-based model's output. As was stated by [162], stacking approach has shown better results than using ensembles of classifiers of the same nature.

One of the contributions of this paper is Table 1, where a taxonomy for the white- and black-box approaches is shown. In this taxonomy, we analyzed what type of input data and feature type are better for the models following any of the reviewed approaches. Also, we analyzed the interpretability of the model's output, taking into account the input data, and if the reviewed approach is able for working with the original dataset containing missing values. For example, if you select a black-box-based model, then, based on Table 1 you can infer that the selected model can work with different types of input data, only with numerical features (it needs a data transformation for working with other types of features). Also, it is feasible to interpret the model's output when the input data are images, and it needs to modify the input data for working with missing values.

Finally, based on the reviews stated in Sections II-III, it is important to show a way for measuring the effectiveness of the applied model, where statistical procedures and the expert opinions can be taken into account together.

V. MEASURING THE EFFECTIVENESS OF THE APPLIED MODEL

For measuring the effectiveness of a machine learning model, there are two main procedures: internal and external validation [163]. On the one side, internal validation estimates how accurately a predictive model will perform in practice by using a set of databases. Most of the published machine learning papers use an experimental setup based on the k-fold cross-validation (k-FCV) procedure as internal validation. For doing that, the original database is randomly partitioned into k equal-sized datasets, where a single dataset is retained as the testing dataset, and the remaining $k-1$ datasets are used as the training dataset. The cross-validation procedure is repeated k times, where each of the k datasets is used only once as the testing data. After, using some measure, the k results are averaged to produce a single estimation. Some machine learning areas, like biometrics, use one round of cross-validation where the database is partitioned into two complementary datasets [163].

On the other side, external validation uses at least two databases (d_1 and d_2) from the same nature, but they cannot share any objects. The idea is to use a database d_1 for training the model and use this database d_1 as internal validation, and after that, the second database d_2 is used as external validation. This type of validation is widely used in biometrics problems such as iris and fingerprint identification [13], [14], [24]–[28], [164].

The main weakness of using internal and external validation procedures for validating machine learning models, which will be applied in practical scenarios, is that expert opinions are not being taken into account. For applied machine learning models, the experts in the application area have the last word. However, the main question is: how to validate the suitability of an applied machine learning model by using the opinions of several experts in the application area?

A statistical method for validating the suitability of an applied machine learning model by using the opinions of several experts in the application area is the Delphi method [165], [166]. This method is an effective and systematic procedure for collecting expert opinions on a particular topic. The Delphi method achieves a consensus from all opinions issued by experts by using an evaluation questionnaire. The main novelty of the Delphi method is the use of a structured questionnaire to which the different opinions of the experts in the subsequent rounds are added or modified, until at least three rounds are completed [165], [166]. The Delphi method, as a validation instrument for questionnaires, has been widely used in numerous applied areas, such as sports, economics, marketing, medical sciences, and curriculum planning [167]–[169].

For applying the Delphi method, it is essential to have: (i) at least three experts in the application area, having ample experience in the area; and (ii) a suitable questionnaire for measuring the effectiveness of the applied machine learning model. The questionnaire should provide a set of suitable questions, which can extract from the experts their objective opinion for the applied machine learning model. Also, this questionnaire should provide a numerical scale for assigning a value, according to that scale, to each question [165], [166].

This questionnaire could include some of the following question for evaluating an applied machine learning model:

- Is it suitable the applied model?
- Is it accurate the model?
- Does the model's output is easy to understand?
- Does the model's output help to the end-user for taking a decision?
- Does the model provides an explanation that justifies its recommendation, decision, or action?

The Delphi method uses the evaluations issued by the experts using the applied questionnaire, the numerical scale, and the grade of expertise of each expert for outputting the suitability of the assessed model [165], [166].

The Delphi method can be described in the following items:

- The Delphi method is a methodology to arrive at a consensus or decision by surveying a panel of experts.
- Experts fulfill all the questionnaires after several rounds, and the responses are aggregated and shared with the group of experts after each round.
- Based on the interpretation of the *group response* provided by each expert in the before item, they can change their answers after each round.
- The final result is meant to be a true consensus of what the group thinks.

In most of the published papers about white-box models applied to practical problems, the authors claim that the proposed model is explainable because it follows a white-box approach [8], [9], [108]. However, it is essential that those white-box model applied to practical problems can be evaluated by using the Delphi method or other methods requiring the opinion of experts in the application area. In this way, the proposed white-box model can be validated by experts in the application area.

VI. CONCLUSION

In this paper, we provide a brief introduction for both white- and black-box approaches, where the most outstanding models based on each one were reviewed. Also, we presented advantages and weaknesses of both white- and black-box approaches from both theoretical and practical points of view; taking special interest in those security and biomedical applications. Additionally, one the one hand, we reviewed and provided idea about how to fuse white- and black-box approaches for creating better machine learning models than other proposed in the literature. On the other hand, we provide a way to measure the effectiveness of the applied models by using statistical procedures and experts opinions jointly.

From this paper, we can conclude the following:

(i) It is essential to understand the input data as well as the best way for showing the output data to experts before applying a machine learning model to a practical problem; showing a special interest in those problems related to security and medicine.
(ii) Experts in the application domain do not need to understand the mathematical transformation behind the applied black-box-based model, they only need a natural form of interpreting the model's output. For doing that, the output data should be provided to experts in a similar form that they provided the input data but highlighting the important keys discovered by the applied model.
(iii) White-box-based models are so good as black-box-based models are, and their performances depend on the application domain and the input data.
(iv) Experts in the application domain do not need to understand the inside of the applied model but it is mandatory

for experts in machine learning to understand this model due to, on most occasions, the model need to be tuned for obtaining accurate results.

(v) It is not necessary to move all black-box models towards white-box models, but it is mandatory to analyze the models for selecting the best one to be applied in the given problem and the best form to show the model's output.

(vi) New models fusing white- and black-box approaches are necessary for providing models more easy to interpret than those previously proposed.

ACKNOWLEDGMENT

The author thanks Prof. Miguel Angel Medina-Pérez, Ph.D. and Prof. Leonardo Chang, Ph.D.; for their considerable efforts to provide thoughtful comments and discussion about black- and white-box approaches. Besides, thanks to Prof. Leyanes Díaz-López, PhD; for her useful comments and discussion about black- and white-box approaches from her point of view as a biologist working on applied biotechnology problems. Also, all of them contributed to improving the grammar and style of this paper.

REFERENCES

[1] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. Wadsworth, OH, USA: CRC Press, 1984. [Online]. Available: http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0412048418

[2] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986. doi: 10.1007/BF00116251.

[3] C. Rudin, "Please stop explaining black box models for high stakes decisions," in *Proc. 32nd Conf. Neural Inf. Process. Syst. (NIPS), Workshop Critiquing Correcting Trends Mach. Learn.*, 2018, pp. 1–20. [Online]. Available: https://arxiv.org/abs/1811.10154

[4] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019. doi: 10.1038/s42256-019-0048-x.

[5] M. García-Borroto, J. Martínez-Trinidad, and J. Carrasco-Ochoa, "Fuzzy emerging patterns for classifying hard domains," *Knowl. Inf. Syst.*, vol. 28, no. 2, pp. 473–489, 2011. doi: 10.1007/s10115-010-0324-x.

[6] O. Loyola-González, R. Monroy, M. A. Medina-Pérez, B. Cervantes, and J. E. Grimaldo-Tijerina, "An approach based on contrast patterns for bot detection on Web log files," in *Advances in Soft Computing*, I. Batyrshin, M. Martínez-Villaseñor, and H. P. Espinosa, Eds. Cham, Switzerland: Springer, 2018, pp. 276–285.

[7] O. Loyola-González, M. A. Medina-Pérez, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, R. Monroy, and M. García-Borroto, "PBC4cip: A new contrast pattern-based classifier for class imbalance problems," *Knowl.-Based Syst.*, vol. 115, pp. 100–109, Jan. 2017.

[8] G. Dong, *Exploiting the Power of Group Differences: Using Patterns to Solve Data Analysis Problems* (Synthesis Lectures on Data Mining and Knowledge Discovery), vol. 11, no. 1. San Rafael, CA, USA: Morgan & Claypool, 2019, pp. 1–146. doi: 10.2200/S00897ED1V01Y201901DMK016.

[9] O. Loyola-González, R. Monroy, J. Rodríguez, A. López-Cuevas, and J. I. Mata-Sánchez, "Contrast pattern-based classification for bot detection on Twitter," *IEEE Access*, vol. 7, pp. 45800–45817, 2019.

[10] D. Martens, B. Baesens, T. Van Gestel, and J. Vanthienen, "Comprehensible credit scoring models using rule extraction from support vector machines," *Eur. J. Oper. Res.*, vol. 183, no. 3, pp. 1466–1476, 2007.

[11] D. Quick and K.-K. R. Choo, "Pervasive social networking forensics: Intelligence and evidence from mobile device extracts," *J. Netw. Comput. Appl.*, vol. 86, pp. 24–33, May 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1084804516302843

[12] D. Quick and K.-K. R. Choo, "IoT device forensics and data reduction," *IEEE Access*, vol. 6, pp. 47566–47574, 2018.

[13] A. Bansal, R. Agarwal, and R. K. Sharma, "Pre-diagnostic tool to predict obstructive lung diseases using iris recognition system," in *Smart Innovations in Communication and Computational Sciences*, B. K. Panigrahi, M. C. Trivedi, K. K. Mishra, S. Tiwari, and P. K. Singh, Eds. Singapore: Springer, 2019, pp. 71–79.

[14] B. Topcu and H. Erdogan, "Fixed-length asymmetric binary hashing for fingerprint verification through GMM-SVM based representations," *Pattern Recognit.*, vol. 88, pp. 409–420, Apr. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320318304187

[15] Y. Ma and G. Guo, *Support Vector Machines Applications*. New York, NY, USA: Springer, 2014.

[16] J. M. Zurada, *Introduction to Artificial Neural Systems*, 1st ed. Boston, MA, USA: PWS Publishing Co., 1999.

[17] I. A. Basheer and M. Hajmeer, "Artificial neural networks: Fundamentals, computing, design, and application," *J. Microbiol. Methods*, vol. 43, no. 1, pp. 3–31, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167701200002013

[18] M. Richardson and P. Domingos, "Markov logic networks," *Mach. Learn.*, vol. 62, no. 1, pp. 107–136, Feb. 2006. doi: 10.1007/s10994-006-5833-1.

[19] B. Cai, X. Kong, Y. Liu, J. Lin, X. Yuan, H. Xu, and R. Ji, "Application of Bayesian networks in reliability evaluation," *IEEE Trans. Ind. Informat.*, vol. 15, no. 4, pp. 2146–2157, Jul. 2019.

[20] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Mach. Learn.*, vol. 6, no. 1, pp. 37–66, Jan. 1991. doi: 10.1007/BF00153759.

[21] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *Amer. Statist.*, vol. 46, no. 3, pp. 175–185, 1992. doi: 10.1080/00031305.1992.10475879.

[22] B. Schlkopf and A. J. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.

[23] C.-F. Lin and S.-D. Wang, "Fuzzy support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 464–471, Mar. 2002.

[24] Y. Martínez-Díaz, N. Hernández, R. J. Biscay, L. Chang, H. Méndez-Vázquez, and L. E. Sucar, "On Fisher vector encoding of binary features for video face recognition," *J. Vis. Commun. Image Represent.*, vol. 51, pp. 155–161, Feb. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1047320318300245

[25] Y. Martínez-Díaz, H. Méndez-Vázquez, L. López-Avila, L. Chang, L. E. Sucar, and M. Tistarelli, "Toward more realistic face recognition evaluation protocols for the youtube faces database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2018, pp. 413–421.

[26] L. Liu, P. Fieguth, M. Pietikäinen, D. Hu, and G. Zhao, "Extended local binary patterns for face recognition," *Inf. Sci.*, vols. 358–359, pp. 56–72, Sep. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020025516302560

[27] H.-A. Park and K. R. Park, "Iris recognition based on score level fusion by using SVM," *Pattern Recognit. Lett.*, vol. 28, no. 15, pp. 2019–2028, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S016786550700178X

[28] Y. Yao, P. Frasconi, and M. Pontil, "Fingerprint classification with combinations of support vector machines," in *Audio- and Video-Based Biometric Person Authentication*, J. Bigun and F. Smeraldi, Eds. Berlin, Germany: Springer, 2001, pp. 253–258.

[29] A. P. López-Monroy, M. Montes-y-Gómez, H. J. Escalante, L. Villaseñor-Pineda, and E. Stamatatos, "Discriminative subprofile-specific representations for author profiling in social media," *Knowl.-Based Syst.*, vol. 89, pp. 134–147, Nov. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950705115002427

[30] H. J. Escalante, E. Villatoro-Tello, S. E. Garza, and A. P. López-Monroy, M. Montes-y-Gómez, and L. Villaseñor-Pineda, "Early detection of deception and aggressiveness using profile-based representations," *Expert Syst. Appl.*, vol. 89, pp. 99–111, Dec. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417417305171

[31] R. M. Ortega-Mendoza, A. P. López-Monroy, A. Franco-Arcega, and M. Montes-y-Gómez, "Emphasizing personal information for author profiling: New approaches for term selection and weighting," *Knowl.-Based Syst.*, vol. 145, pp. 169–181, Apr. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950705118300224

[32] J. Baek and E. Kim, "A new support vector machine with an optimal additive kernel," *Neurocomputing*, vol. 329, pp. 279–299, Feb. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231218312207

[33] C. Chen, X. Li, A. N. Belkacem, Z. Qiao, E. Dong, W. Tan, and D. Shin, "The mixed kernel function SVM-based point cloud classification," *Int. J. Precis. Eng. Manuf.*, vol. 20, no. 5, pp. 737–747, May 2019. doi: 10.1007/s12541-019-00102-3.

[34] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," in *Proc. Adv. Neural Inf. Process. Syst.*, 1990, pp. 396–404.

[35] S. Mallat, "Understanding deep convolutional networks," *Philos. Trans. A, Math., Phys., Eng. Sci.*, vol. 374, no. 2065, 2016, Art. no. 20150203. doi: 10.1098/rsta.2015.0203.

[36] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, Inc., 2014, pp. 2672–2680. [Online]. Available: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[37] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *CoRR*, vol. abs/1809.07294, pp. 1–24, Sep. 2018. [Online]. Available: https://arxiv.org/abs/1809.07294

[38] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: https://arxiv.org/abs/1511.06434

[39] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," *CoRR*, vol. abs/1710.10196, pp. 1–26, Oct. 2017. [Online]. Available: https://arxiv.org/abs/1710.10196

[40] M. Brundage *et al.*, "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation," *CoRR*, vol. abs/1802.07228, pp. 1–101, Feb. 2018. [Online]. Available: https://arxiv.org/abs/1802.07228

[41] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," *CoRR*, vol. abs/1809.11096, pp. 1–35, Sep. 2018. [Online]. Available: https://arxiv.org/abs/1809.11096

[42] Y. Jin, J. Zhang, M. Li, Y. Tian, H. Zhu, and Z. Fang, "Towards the automatic anime characters creation with generative adversarial networks," *CoRR*, vol. abs/1708.05509, pp. 1–16, Aug. 2017. [Online]. Available: https://arxiv.org/abs/1708.05509

[43] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CoRR*, vol. abs/1611.07004, pp. 1–17, Nov. 2016. [Online]. Available: https://arxiv.org/abs/1611.07004

[44] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *CoRR*, vol. abs/1703.10593, pp. 1–18, Mar. 2017. [Online]. Available: https://arxiv.org/abs/1703.10593

[45] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. N. Metaxas, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," *CoRR*, vol. abs/1612.03242, Dec. 2016. [Online]. Available: https://arxiv.org/abs/1612.03242

[46] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," *CoRR*, vol. abs/1605.05396, pp. 1–10, Jun. 2016. [Online]. Available: https://arxiv.org/abs/1605.05396

[47] S. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning what and where to draw," *CoRR*, vol. abs/1610.02454, pp. 1–9, Oct. 2016. [Online]. Available: https://arxiv.org/abs/1610.02454

[48] A. Dash, J. C. B. Gamboa, S. Ahmed, M. Liwicki, and M. Z. Afzal, "TAC-GAN—Text conditioned auxiliary classifier generative adversarial network," *CoRR*, vol. abs/1703.06412, pp. 1–9, Mar. 2017. [Online]. Available: https://arxiv.org/abs/1703.06412

[49] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," *CoRR*, vol. abs/1711.11585, pp. 1–14, Nov. 2017. [Online]. Available: https://arxiv.org/abs/1711.11585

[50] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2439–2448.

[51] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," *CoRR*, vol. abs/1705.09368, pp. 1–11, May 2017. [Online]. Available: https://arxiv.org/abs/1705.09368

[52] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," *CoRR*, vol. abs/1611.02200, pp. 1–14, Nov. 2016. [Online]. Available: https://arxiv.org/abs/1611.02200

[53] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," *CoRR*, vol. abs/1606.07536, pp. 1–32, Sep. 2016. [Online]. Available: https://arxiv.org/abs/1606.07536

[54] A. Brock, T. Lim, J. M. Ritchie, and N. Weston, "Neural photo editing with introspective adversarial networks," *CoRR*, vol. abs/1609.07093, pp. 1–15, Sep. 2016. [Online]. Available: https://arxiv.org/abs/1609.07093

[55] H. Zhang, V. Sindagi, and V. M. Patel, "Image de-raining using a conditional generative adversarial network," *CoRR*, vol. abs/1701.05957, pp. 1–14, Jan. 2017. [Online]. Available: https://arxiv.org/abs/1701.05957

[56] G. Antipov, M. Baccouche, and J.-L. Dugelay, "Face aging with conditional generative adversarial networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 2089–2093.

[57] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," *CoRR*, vol. abs/1702.08423, pp. 1–9, Feb. 2017. [Online]. Available: https://arxiv.org/abs/1702.08423

[58] H. Wu, S. Zheng, J. Zhang, and K. Huang, "GP-GAN: Towards realistic high-resolution image blending," *CoRR*, vol. abs/1703.07195, pp. 1–11, Mar. 2017. [Online]. Available: https://arxiv.org/abs/1703.07195

[59] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," *CoRR*, vol. abs/1609.04802, pp. 1–19, Sep. 2016. [Online]. Available: https://arxiv.org/abs/1609.04802

[60] H. Bin, C. Weihai, W. Xingming, and L. Chun-Liang, "High-quality face image SR using conditional generative adversarial networks," *CoRR*, vol. abs/1707.00737, pp. 1–6, Jul. 2017. [Online]. Available: https://arxiv.org/abs/1707.00737

[61] S. Vasu, N. T. Madam, and R. A. Narayanan, "Analyzing perception-distortion tradeoff using enhanced perceptual super-resolution network," *CoRR*, vol. abs/1811.00344, pp. 1–17, Nov. 2018. [Online]. Available: https://arxiv.org/abs/1811.00344

[62] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," *CoRR*, vol. abs/1604.07379, pp. 1–12, Apr. 2016. [Online]. Available: https://arxiv.org/abs/1604.07379

[63] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," *CoRR*, vol. abs/1607.07539, pp. 1–19, Jul. 2016. [Online]. Available: https://arxiv.org/abs/1607.07539

[64] Y. Li, S. Liu, J. Yang, and M.-Y. Yang, "Generative face completion," *CoRR*, vol. abs/1704.05838, pp. 1–9, Apr. 2017. [Online]. Available: https://arxiv.org/abs/1704.05838

[65] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," *CoRR*, vol. abs/1609.02612, pp. 1–10, Sep. 2016. [Online]. Available: https://arxiv.org/abs/1609.02612

[66] J. Wu, C. Zhang, T. Xue, W. T. Freeman, and J. B. Tenenbaum, "Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling," *CoRR*, vol. abs/1610.07584, pp. 1–11, Oct. 2016. [Online]. Available: https://arxiv.org/abs/1610.07584

[67] M. Gadelha, S. Maji, and R. Wang, "3D shape induction from 2D views of multiple objects," *CoRR*, vol. abs/1612.05872, pp. 1–10, Dec. 2016. [Online]. Available: https://arxiv.org/abs/1612.05872

[68] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Proc. Adv. Neural Inf. Process. Syst.*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, Inc., 2017, pp. 6306–6315.

[69] H. Shen, Y. Zhu, L. Zhang, and J. H. Park, "Extended dissipative state estimation for Markov jump neural networks with unreliable links," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 2, pp. 346–358, Feb. 2017.

[70] S. Dami, A. A. Barforoush, and H. Shirazi, "News events prediction using Markov logic networks," *J. Inf. Sci.*, vol. 44, no. 1, pp. 91–109, 2018. doi: 10.1177/0165551516673285.

[71] M. Henrion, "Propagating uncertainty in Bayesian networks by probabilistic logic sampling," in *Uncertainty in Artificial Intelligence* (Machine Intelligence & Pattern Recognition), vol. 5, J. F. Lemmer and L. N. Kanal, Eds. Amsterdam, The Netherlands: North Holland, 1988, pp. 149–163. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780444703965500194

[72] S. Shi and A. S. Weigend, "Taking time seriously: Hidden Markov experts applied to financial engineering," in *Proc. IEEE/IAFE Comput. Intell. Financial Eng. (CIFEr)*, Mar. 1997, pp. 244–252.

[73] A. Morales-González, E. García-Reyes, and L. E. Sucar, "Hierarchical Markov random fields with irregular pyramids for improving image annotation," in *Advances in Artificial Intelligence—IBERAMIA*, J. Pavón, N. D. Duque-Méndez, and R. Fuentes-Fernández, Eds. Berlin, Germany: Springer, 2012, pp. 521–530.

[74] J. C. Johnson, J. J. Luczkovich, S. P. Borgatti, and T. A. B. Snijders, "Using social network analysis tools in ecology: Markov process transition models applied to the seasonal trophic network dynamics of the chesapeake bay," *Ecol. Model.*, vol. 220, no. 22, pp. 3133–3140, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0304380009004414

[75] A. G. Vallejo, Jr., J. A. Nolazco-Flores, and R. Morales-Menéndez, L. E. Sucar, and C. A. Rodríguez, "Tool-wear monitoring based on continuous hidden Markov models," in *Progress in Pattern Recognition, Image Analysis and Applications*, A. Sanfeliu and M. L. Cortés, Eds. Berlin, Germany: Springer, 2005, pp. 880–890.

[76] P. R. Pinheiro, A. K. A. de Castro, and M. C. D. Pinheiro, "A multicriteria model applied in the diagnosis of Alzheimer's disease: A Bayesian network," in *Proc. 11th IEEE Int. Conf. Comput. Sci. Eng.*, Jul. 2008, pp. 15–22.

[77] C.-F. Chien, S.-L. Chen, and Y.-S. Lin, "Using Bayesian network for fault location on distribution feeder," *IEEE Trans. Power Del.*, vol. 17, no. 3, pp. 785–793, Jul. 2002.

[78] R. A. Jacobs and J. K. Kruschke, "Bayesian learning theory applied to human cognition," *Wiley Interdiscipl. Rev., Cognit. Sci.*, vol. 2, no. 1, pp. 8–21, 2011. doi: 10.1002/wcs.80.

[79] J. Vomlel, "Bayesian networks in educational testing," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 12, pp. 83–100, Jan. 2004. doi: 10.1142/S021848850400259X.

[80] K. Deforche *et al.*, "Bayesian network analysis of resistance pathways against HIV-1 protease inhibitors," *Infection, Genet. Evol.*, vol. 7, no. 3, pp. 382–390, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1567134806001043

[81] P. Antal, H. Verrelst, D. Timmerman, Y. Moreau, S. Van Huffel, B. De Moor, and I. Vergote, "Bayesian networks in ovarian cancer diagnosis: Potentials and limitations," in *Proc. 13th IEEE Symp. Comput.-Based Med. Syst. (CBMS)*, Jun. 2000, pp. 103–108.

[82] L. A. M. Riascos, M. G. Simoes, and P. E. Miyagi, "A Bayesian network fault diagnostic system for proton exchange membrane fuel cells," *J. Power Sources*, vol. 165, no. 1, pp. 267–278, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378775306025262

[83] B. Ribeiro-Neto, I. Silva, and R. Muntz, "Bayesian network models for information retrieval," in *Soft Computing in Information Retrieval*. Heidelberg, Germany: Physica-Verlag HD, 2000, pp. 259–29. doi: 10.1007/978-3-7908-1849-9_11.

[84] L. Lara-Estrada, L. Rasche, L. E. Sucar, and U. A. Schneider, "Inferring missing climate data for agricultural planning using Bayesian networks," *Land*, vol. 7, no. 1, p. 4, 2018. [Online]. Available: https://www.mdpi.com/2073-445X/7/1/4

[85] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein, "A Bayesian networks approach for predicting protein-protein interactions from genomic data," *Science*, vol. 302, no. 5644, pp. 449–453, 2003. [Online]. Available: https://science.sciencemag.org/content/302/5644/449

[86] D. Heckerman, "A tutorial on learning with Bayesian networks," in *Innovations in Bayesian Networks*. Dordrecht, The Netherlands: Springer, 1998, pp. 301–35. doi: 10.1007/978-94-011-5014-9_11.

[87] P. Zhang, B. J. Gao, X. Zhu, and L. Guo, "Enabling fast lazy learning for data streams," in *Proc. IEEE 11th Int. Conf. Data Mining*, Dec. 2011, pp. 932–941.

[88] C. Carnevale, G. Finzi, A. Pederzoli, E. Turrini, and M. Volta, "Lazy learning based surrogate models for air quality planning," *Environ. Model. Softw.*, vol. 83, pp. 47–57, Sep. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1364815216301177

[89] E. Armengol and E. Plaza, "Lazy learning for predictive toxicology based on a chemical ontology," in *Artificial Intelligence Methods And Tools For Systems Biology*. Dordrecht, The Netherlands: Springer, 2004, pp. 1–1. doi: 10.1007/978-1-4020-5811-0_1.

[90] M. Kück and B. Scholz-Reiter, "A genetic algorithm to optimize lazy learning parameters for the prediction of customer demands," in *Proc. 12th Int. Conf. Mach. Learn. Appl.*, vol. 2, Dec. 2013, pp. 160–165.

[91] D. P. Solomatine, M. Maskey, and D. L. Shrestha, "Eager and lazy learning methods in the context of hydrologic forecasting," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Jul. 2006, pp. 4847–4853.

[92] G. Corani, "Air quality prediction in Milan: Feed-forward neural networks, pruned neural networks and lazy learning," *Ecol. Model.*, vol. 185, nos. 2–4, pp. 513–529, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0304380005000165

[93] M. J. L. Boada, J. A. Calvo, B. L. Boada, and V. Díaz, "Modeling of a magnetorheological damper by recursive lazy learning," *Int. J. Non-Linear Mech.*, vol. 46, no. 3, pp. 479–485, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020746208002187

[94] F. Fdez-Riverola, E. L. Iglesias, and F. Díaz, J. R. Méndez, and J. M. Corchado, "Applying lazy learning algorithms to tackle concept drift in spam filtering," *Expert Syst. Appl.*, vol. 33, no. 1, pp. 36–48, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417406001175

[95] W. Liu and S. Chawla, "Class confidence weighted *k*NN algorithms for imbalanced data sets," in *Advances in Knowledge Discovery and Data Mining*, J. Z. Huang, L. Cao, and J. Srivastava, Eds. Berlin, Germany: Springer, 2011, pp. 345–356.

[96] G. Haixiang, L. Yijing, L. Yanan, L. Jinlinga, and L. Xiao, "BPSO-Adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification," *Eng. Appl. Artif. Intell.*, vol. 49, pp. 176–193, Mar. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0952197615002110

[97] Y. Gao and F. Gao, "Edited AdaBoost by weighted kNN," *Neurocomputing*, vol. 73, pp. 3079–3088, Apr. 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231210003358

[98] Y. Bao, N. Ishii, and X. Du, "Combining multiple k-nearest neighbor classifiers using different distance functions," in *Intelligent Data Engineering and Automated Learning—IDEAL*, Z. R. Yang, H. Yin, and R. M. Everson, Eds. Berlin, Germany: Springer, 2004, pp. 634–641.

[99] T. Yamada, K. Yamashita, N. Ishii, and K. Iwata, "Text Classification by Combining Different Distance Functions withWeights," in *Proc. 7th ACIS Int. Conf. Softw. Eng., Artif. Intell., Netw., Parallel/Distrib. Comput. (SNPD)*, Jun. 2006, pp. 85–90.

[100] O. Loyola-González, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, D. Hernández-Tamayo, and M. García-Borroto, "Detecting pneumatic failures on temporary immersion bioreactors," in *Pattern Recognition*, vol. 9703. Cham, Switzerland: Springer, 2016, pp. 293–30. doi: 10.1007/978-3-319-39393-3_29.

[101] O. Loyola-González, M. A. Medina-Pérez, D. Hernández-Tamayo, R. Monroy, J. A. Carrasco-Ochoa, and M. García-Borroto, "A pattern-based approach for detecting pneumatic failures on temporary immersion bioreactors," *Sensors*, vol. 19, no. 2, p. 414, 2019. [Online]. Available: http://www.mdpi.com/1424-8220/19/2/414

[102] C. A. Clement and R. J. Falmagne, "Logical reasoning, world knowledge, and mental imagery: Interconnections in cognitive processes," *Memory Cognition*, vol. 14, no. 4, pp. 299–307, Jul. 1986. doi: 10.3758/BF03202507.

[103] M. Knauff, "A neuro-cognitive theory of deductive relational reasoning with mental models and visual images," *Spatial Cognition Comput.*, vol. 9, no. 2, pp. 109–137, 2009. doi: 10.1080/13875860902887605.

[104] D. Ribli, A. Horváth, Z. Unger, P. Pollner, and I. Csabai, "Detecting and classifying lesions in mammograms with deep learning," *Sci. Rep.*, vol. 8, no. 1, 2018, Art. no. 4165. doi: 10.1038/s41598-018-22437-z.

[105] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[106] DARPA. (2016). *Broad Agency Announcement, Explainable Artificial Intelligence (XAI)*. [Online]. Available: https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf

[107] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA, USA: Morgan Kaufmann, 1993.

[108] R. Agrawal and T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," *SIGMOD Rec.*, vol. 22, no. 2, pp. 207–216, Jun. 1993. doi: 10.1145/170036.170072.

[109] R. Agrawal and T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proc. 1993 ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, New York, NY, USA, 1993, pp. 207–216. doi: 10.1145/170035.170072.

[110] G. Dong, "Preliminaries," in *Contrast Data Mining: Concepts, Algorithms, and Applications* (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series), G. Dong and J. Bailey, Eds. London, U.K.: Chapman & Hall, 2012, ch. 1, pp. 3–12.

[111] L. Rokach and O. Maimon, *Data Mining with Decision Trees: Theory and Applications* (Series in Machine Perception and Artificial Intelligence), 2nd ed. Singapore: World Scientific, 2014. [Online]. Available: http://www.worldscientific.com/worldscibooks/10.1142/9097

[112] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, 2006. doi: 10.1007/s10994-006-6226-1.

[113] S. B. Kotsiantis, "Decision trees: A recent overview," *Artif. Intell. Rev.*, vol. 39, no. 4, pp. 261–283, 2013. doi: 10.1007/s10462-011-9272-4.

[114] N. Cepero-Pèrez, L. A. Denis-Miranda, R. Hernández-Palacio, M. Moreno-Espino, and M. García-Borroto, "Proactive forest for supervised classification," in *Progress in Artificial Intelligence and Pattern Recognition*, Y. H. Heredia, V. M. Núñez, and J. R. Shulcloper, Eds. Cham, Switzerland: Springer, 2018, pp. 255–262.

[115] S. A. Ludwig, S. Picek, and D. Jakobovic, "Classification of cancer data: Analyzing gene expression data using a fuzzy decision tree algorithm," in *Classification Cancer Data: Analyzing Gene Expression Data Using a Fuzzy Decision Tree Algorithm*. Cham, Switzerland: Springer, 2018, pp. 327–34. doi: 10.1007/978-3-319-65455-3_13.

[116] C. Soguero-Ruiz, A. A. Díaz-Plaza, P. de Miguel Bohoyo, J. Ramos-López, M. Rubio-Sánchez, A. Sánchez, and I. Mora-Jiménez, "On the use of decision trees based on diagnosis and drug codes for analyzing chronic patients," in *Bioinformatics and Biomedical Engineering*, I. Rojas and F. Ortuño, Eds. Cham, Switzerland: Springer, 2018, pp. 135–148.

[117] A. Freitas, A. Costa-Pereira, and P. Brazdil, "Cost-sensitive decision trees applied to medical data," in *Data Warehousing and Knowledge Discovery*, I. Y. Song, J. Eder, and T. M. Nguyen, Eds. Berlin, Germany: Springer, 2007, pp. 303–312.

[118] I. Zelič, I. Kononenko, N. Lavrač, and V. Vuga, "Induction of decision trees and Bayesian classification applied to diagnosis of sport injuries," *J. Med. Syst.*, vol. 21, no. 6, pp. 429–444, Dec. 1997. doi: 10.1023/A:1022880431298.

[119] H.-B. Ly, E. Monteiro, T.-T. Le, V. M. Le, M. Dal, G. Regnier, and B. T. Pham, "Prediction and sensitivity analysis of bubble dissolution time in 3D selective laser sintering using ensemble decision trees," *Materials*, vol. 12, no. 9, p. 1544, 2019. [Online]. Available: https://www.mdpi.com/1996-1944/12/9/1544

[120] L. Márquez and H. Rodríguez, "Part-of-speech tagging using decision trees," in *Machine Learning: ECML-98*, C. Nédellec and C. Rouveirol, Eds. Berlin, Germany: Springer, 1998, pp. 25–36.

[121] L. Geng and H. J. Hamilton, "Interestingness measures for data mining: A survey," *ACM Comput. Surv.*, vol. 38, no. 3, pp. 1–32, Sep. 2006.

[122] L. Geng and H. J. Hamilton, "Choosing the right lens: Finding what is interesting in data mining," in *Quality Measures in Data Mining* (Studies in Computational Intelligence), F. J. Guillet and H. J. Hamilton, Eds. Berlin, Germany: Springer, 2007, vol. 43, pp. 3–24.

[123] D.-L. Miholca, G. Czibula, and I. G. Czibula, "A novel approach for software defect prediction through hybridizing gradual relational association rules with artificial neural networks," *Inf. Sci.*, vol. 441, pp. 152–170, May 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020025518301154

[124] S. S. Ahmed, S. Roy, and P. P. Choudhury, "Inferring causal gene regulatory networks using time-delay association rules," in *Advanced Informatics for Computing Research*, A. K. Luhach, D. Singh, P.-A. Hsiung, K. Hawari, P. Lingras, and P. K. Singh, Eds. Singapore: Springer, 2019, pp. 310–321.

[125] C.-P. Lai and J.-R. Lu, "Evaluating the efficiency of currency portfolios constructed by the mining association rules," *Asia–Pacific Manage. Rev.*, vol. 24, no. 1, pp. 11–20, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S102931321730026X

[126] S. Jabri, A. Dahbi, T. Gadi, and A. Bassir, "Ranking of text documents using TF-IDF weighting and association rules mining," in *Proc. 4th Int. Conf. Optim. Appl. (ICOA)*, Apr. 2018, pp. 1–6.

[127] A. Moubayed, M. Injadat, A. Shami, and H. Lutfiyya, "Relationship between student engagement and performance in e-learning environment using association rules," in *Proc. IEEE World Eng. Educ. Conf. (EDUNINE)*, Mar. 2018, pp. 1–6.

[128] R. Rekik, I. Kallel, J. Casillas, and A. M. Alimi, "Assessing Web sites quality: A systematic literature review by text and association rules mining," *Int. J. Inf. Manage.*, vol. 38, no. 1, pp. 201–216, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0268401215302875

[129] J. Weng and G. Li, "Exploring shipping accident contributory factors using association rules," *J. Transp. Saf. Secur.*, vol. 11, no. 1, pp. 36–57, 2019. doi: 10.1080/19439962.2017.1341440.

[130] R. J. Bayardo, Jr., "Efficiently mining long patterns from databases," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, New York, NY, USA, 1998, pp. 85–93. doi: 10.1145/276304.276313.

[131] U. Fayyad and K. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in *Proc. 13th Int. Joint Conf. Artif. Intell.*, San Francisco, CA, USA, 1993, pp. 1022–1027.

[132] G. I. Webb, "Efficient search for association rules," in *Proc. 6th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, 2000, pp. 99–10. doi: 10.1145/347090.347112.

[133] O. Loyola-González, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, and M. García-Borroto, "Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases," *Neurocomputing*, vol. 175, pp. 935–947, Jan. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231215015908

[134] O. Loyola-González, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, and M. García-Borroto, "Effect of class imbalance on quality measures for contrast patterns: An experimental study," *Inf. Sci.*, vol. 374, pp. 179–192, Dec. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020025516309379

[135] X. Zhang and G. Dong, "Overview and analysis of contrast pattern based classification," in *Contrast Data Mining: Concepts, Algorithms, and Applications* (Data Mining and Knowledge Discovery Series), G. Dong and J. Bailey, Eds. London, U.K.: Chapman & Hall, 2012, ch. 11, pp. 151–170.

[136] G. Dong and J. Li, "Efficient mining of emerging patterns: Discovering trends and differences," in *Proc. 5th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, 1999, pp. 43–52.

[137] M. García-Borroto, J. F. Martínez-Trinidad, and J. A. Carrasco-Ochoa, "Finding the best diversity generation procedures for mining contrast patterns," *Expert Syst. Appl.*, vol. 42, no. 11, pp. 4859–4866, 2015.

[138] F. Z. El Mazouri, M. C. Abounaima, and K. Zenkouar, "Data mining combined to the multicriteria decision analysis for the improvement of road safety: Case of France," *J. Big Data*, vol. 6, no. 1, Jan. 2019, Art. no. 5. doi: 10.1186/s40537-018-0165-0.

[139] D. Usha, K. Rameshkumar, and B. V. Baiju, "Association rule construction from crime pattern through novelty approach," in *Advances in Big Data and Cloud Computing*, J. D. Peter, A. H. Alavi, and B. Javadi, Eds. Singapore: Springer, 2019, pp. 241–250.

[140] C. P. Wulandari, C. Ou-Yang, and H.-C. Wang, "Applying mutual information for discretization to support the discovery of rare-unusual association rule in cerebrovascular examination dataset," *Expert Syst. Appl.*, vol. 118, pp. 52–64, Mar. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417418306213

[141] O. Loyola-González, A. López-Cuevas, M. A. Medina-Pérez, B. Camiña, J. E. Ramírez-Márquez, and R. Monroy, "Fusing pattern discovery and visual analytics approaches in tweet propagation," *Inf. Fusion*, vol. 46, pp. 91–101, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253517307716

[142] C.-H. Weng and C.-K. Huang, Tony, "Observation of sales trends by mining emerging patterns in dynamic markets," *Appl. Intell.*, vol. 48, no. 11, pp. 4515–4529, Nov. 2018. doi: 10.1007/s10489-018-1231-1.

[143] W. Hu, T. Chen, and S. L. Shah, "Detection of frequent alarm patterns in industrial alarm floods using itemset mining methods," *IEEE Trans. Ind. Electron.*, vol. 65, no. 9, pp. 7290–7300, Sep. 2018.

[144] H. T. Malazi and M. Davari, "Combining emerging patterns with random forest for complex activity recognition in smart homes," *Appl. Intell.*, vol. 48, no. 2, pp. 315–330, Feb. 2018. doi: 10.1007/s10489-017-0976-2.

[145] E. A. Chavary, S. M. Erfani, and C. Leckie, "Summarizing significant changes in network traffic using contrast pattern mining," in *Proc. ACM Conf. Inf. Knowl. Manage. (CIKM)*, New York, NY, USA, 2017, pp. 2015–2018. doi: 10.1145/3132847.3133111.

[146] O. Cavadenti, V. Codocedo, J.-F. Boulicaut, and M. Kaytoue, "What did I do wrong in my MOBA game? Mining patterns discriminating deviant behaviours," in *Proc. IEEE Int. Conf. Data Sci. Adv. Anal. (DSAA)*, Oct. 2016, pp. 662–671.

[147] L. A. Zadeh, "Fuzzy sets," *Inf. Control*, vol. 8, no. 3, pp. 338–353, Jun. 1965. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S001999586590241X

[148] J. L. Fisher, S. C. Hinds, and D. P. D'Amato, "A rule-based system for document image segmentation," in *Proc. 10th Int. Conf. Pattern Recognit.*, vol. 1, Jun. 1990, pp. 567–572.

[149] K. Chen, M. Seuret, J. Hennebert, and R. Ingold, "Convolutional neural networks for page segmentation of historical document images," in *Proc. 14th Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 1, Nov. 2017, pp. 965–970.

[150] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *CoRR*, vol. abs/1708.08296, pp. 1–8, Aug. 2017. [Online]. Available: https://arxiv.org/abs/1708.08296

[151] G. Biau, E. Scornet, and J. Welbl, "Neural random forests," *Sankhya A*, pp. 1–40, Jun. 2018. doi: 10.1007/s13171-018-0133-y.

[152] Y. Yang, I. G. Morillo, and T. M. Hospedales, "Deep neural decision trees," *CoRR*, vol. abs/1806.06988, pp. 1–7, Jun. 2018. [Online]. Available: https://arxiv.org/abs/1806.06988

[153] S. Bulò and P. Kontschieder, "Neural decision forests for semantic image labelling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 81–88.

[154] P. Kontschieder, M. Fiterau, A. Criminisi, and S. R. Bulò, "Deep neural decision forests," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1467–1475.

[155] A. Roy and S. Todorovic, "Monocular depth estimation using neural regression forest," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5506–5514.

[156] T. Hailesilassie, "Rule extraction algorithm for deep neural networks: A review," *CoRR*, vol. abs/1610.05267, pp. 1–6, Sep. 2016. [Online]. Available: https://arxiv.org/abs/1610.05267

[157] J. R. Zilke, E. L. Mencía, and F. Janssen, "DeepRED—Rule extraction from deep neural networks," in *Discovery Science*, T. Calders, M. Ceci, and D. Malerba, Eds. Cham, Switzerland: Springer, 2016, pp. 457–473.

[158] P. Ray and A. Chakrabarti, "A mixed approach of deep learning method and rule-based method to improve aspect level sentiment analysis," *Appl. Comput. Inform.*, pp. 1–9, Mar. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2210832718303156

[159] A. Kanehira and T. Harada, "Learning to explain with complemental examples," *CoRR*, vol. abs/1812.01280, pp. 1–9, May 2019. [Online]. Available: https://arxiv.org/abs/1812.01280

[160] J. Cheng, W. Huang, S. Cao, R. Yang, W. Yang, Z. Yun, Z. Wang, and Q. Feng, "Enhanced performance of brain tumor classification via tumor region augmentation and partition," *PLoS ONE*, vol. 10, no. 10, Oct. 2015, Art. no. e0140381. doi: 10.1371/journal.pone.0140381.

[161] E. F. Chang, M. B. Potts, G. E. Keles, K. R. Lamborn, S. M. Chang, N. M. Barbaro, and M. S. Berger, "Seizure characteristics and control following resection in 332 patients with low-grade gliomas," *J. Neurosurg.*, vol. 108, no. 2, pp. 227–235, 2008. [Online]. Available: https://thejns.org/view/journals/j-neurosurg/108/2/article-p227.xml

[162] S. Džeroski and B. Ženko, "Is combining classifiers with stacking better than selecting the best one?" *Mach. Learn.*, vol. 54, no. 3, pp. 255–273, 2004. doi: 10.1023/B:MACH.0000015881.36452.6e.

[163] P. Gramatica, "Principles of QSAR models validation: Internal and external," *QSAR Combinat. Sci.*, vol. 26, no. 5, pp. 694–701, 2007. doi: 10.1002/qsar.200610151.

[164] D. Valdes-Ramirez, M. A. Medina-Pérez, R. Monroy, O. Loyola-González, J. Rodríguez, A. Morales, and F. Herrera, "A review of fingerprint feature representations and their applications for latent fingerprint identification: Trends and evaluation," *IEEE Access*, vol. 7, pp. 48484–48499, 2019.

[165] N. Dalkey, O. Helmer, and O. Helmer, "An experimental application of the Delphi method to the use of experts," *Manage. Sci.*, vol. 9, no. 3, pp. 458–467, Apr. 1963.

[166] C.-C. Hsu and B. A. Sandford, "The Delphi technique: making sense of consensus," *Practical Assessment, Res. Eval.*, vol. 12, no. 10, pp. 1–8, 2007.

[167] B. Ludwig, "Predicting the future: Have you considered using the Delphi methodology," *J. Extension*, vol. 35, no. 5, pp. 1–4, 1997.

[168] B. S. Kramer, A. E. Walker, and J. M. Brill, "The underutilization of information and communication technology-assisted collaborative project-based learning among international educators: A Delphi study," *Educ. Technol. Res. Develop.*, vol. 55, no. 5, pp. 527–543, Oct. 2007. doi: 10.1007/s11423-007-9048-3.

[169] H.-L. Hung, J. W. Altschuld, and Y.-F. Lee, "Methodological and conceptual issues confronting a cross-country Delphi study of educational program evaluation," *Eval. Program Planning*, vol. 31, no. 2, pp. 191–198, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S014971890800013X

OCTAVIO LOYOLA-GONZÁLEZ received the B.Eng. degree in informatics engineering and the M.Sc. degree in applied informatics from the University of Ciego de Ávila, Cuba, in 2010 and 2012, respectively, and the Ph.D. degree in computer science from the National Institute of Astrophysics, Optics and Electronics, Mexico, in 2017. He is currently a Research Professor with the Tecnologico de Monterrey, where he is also a member of the GIEE-ML (machine learning) Research Group. His research interests include contrast pattern-based classification, data mining, class imbalance problems, bot detection, cyber security, and biometrics. He has been involved in many research projects about pattern recognition, which have been applied on biotechnology and security problems. He is also a member of the Mexican Researchers System (Rank 1).