Explainable AI for Enhanced Medical Diagnostics

Explainable AI for Enhanced Medical Diagnostics

This is the subtitle of the paper, this document both explains and embodies the submission format for authors using Word

First Author's Name, Initials, and Last name*

First author's affiliation, an Institution with a very long name, xxxx@gmail.com

Second Author's Name, Initials, and Last Name

Second author's affiliation, possibly the same institution, xxxx@gmail.com

Third Author's Name, Initials, and Last Name

Third author's affiliation, possibly the same institution, xxxx@gmail.com

In recent years, the integration of artificial intelligence (AI) into medical diagnostics has ushered in a new era of promise, where accuracy and efficiency in disease detection and prognosis have achieved remarkable heights. AI's capacity to process vast datasets and extract intricate patterns has showcased its potential to revolutionize healthcare outcomes. However, as the reliance on AI-enhanced diagnostic tools grows, an essential concern has emerged: the opacity of AI algorithms in decision-making. This is a pressing concern, particularly in critical medical scenarios where timely and accurate decisions hold the key to saving lives. Numerous studies have illuminated AI's proficiency in diagnosing a diverse range of medical conditions, from radiological image analysis to genomic profiling. Yet, the "black-box" nature of many AI models has impeded their seamless integration into clinical practice. This opacity, where models generate predictions without offering insights into the reasoning process, has led to a trust gap between AI recommendations and medical practitioners. The challenge lies in ensuring that AI's diagnostic prowess is augmented by transparency and interpretability, fostering a harmonious collaboration between machine intelligence and human expertise.

CCS CONCEPTS • Insert your first CCS term here • Insert your second CCS term here • Insert your third CCS term here

Additional Keywords and Phrases: Insert comma delimited author-supplied keyword list, Keyword number 2, Keyword number 3, Keyword number 4

ACM Reference Format:
First Author’s Name, Initials, and Last Name, Second Author’s Name, Initials, and Last Name, and Third Author’s Name, Initials, and Last Name. 2018. The Title of the Paper: ACM Conference Proceedings Manuscript Submission Template: This is the subtitle of the paper, this document both explains and embodies the submission format for authors using Word. In Woodstock ’18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY. ACM, New York, NY, USA, 10 pages. NOTE: This block will be automatically generated when manuscripts are processed after acceptance.

* Place the footnote text for the author (if applicable) here.

1 INTRODUCTION

The term "explaining a prediction" refers to the act of delivering textual or visual evidence that offers a qualitative comprehension of the connection between the many elements of an instance (such as words in text or patches in an image) and the prediction made by the model. We contend that the provision of explanations for predictions is a crucial element in fostering trust and promoting the effective utilization of machine learning by humans, provided that these explanations are both accurate and comprehensible. The process of elucidating individual predictions is depicted in Figure 1. It is evident that a physician is significantly more capable of making informed decisions when supplied with coherent explanations in conjunction with a model. In this particular scenario, an explanation refers to a concise compilation of symptoms accompanied by their respective weights. These symptoms may either contribute to the forecast, denoted by the color green, or serve as evidence against it, denoted by the color red. Typically, individuals possess prior information pertaining to the specific field of application, which they can employ to either accept (trust) or reject a forecast, contingent upon their comprehension of the underlying rationale. Previous studies have indicated that the provision of explanations has the potential to enhance the acceptability of movie recommendations [11] and other automated systems [12]. Each machine learning application necessitates a certain level of faith in the model. The process of developing and assessing a classification model often involves the acquisition of annotated data, from which a portion is put aside for automated evaluation. While the pipeline described here is valuable for several applications, it is important to note that evaluating its performance on validation data may not accurately reflect its performance in real-world scenarios. This is because practitioners typically have a tendency to overestimate the correctness of their models [13]. Therefore, it is not advisable to simply rely on validation data for establishing trust in the model. Examining instances provides an alternate approach to evaluating the veracity of the model, particularly when the examples are accompanied by thorough explanations. Therefore, we propose elucidating a selection of exemplary individual predictions generated by a model as a means of offering a comprehensive comprehension. There exist multiple potential sources of error or shortcomings in both the construction of a model and its subsequent evaluation. Data leakage, also known as the inadvertent release of signal into the training (and validation) data that would not be present during deployment, has the potential to enhance accuracy [14]. Kaufman et al. [14] present a notable instance that poses a challenge, wherein the patient identification (ID) exhibits a strong correlation with the target class in both the training and validation datasets. Identifying this issue just through the observation of forecasts and raw data would pose a considerable challenge. However, the task becomes significantly more manageable with the provision of explanations, as exemplified in Figure 1, where patient ID is included as an explanatory factor for predictions. Another challenging issue that can be difficult to identify is known as dataset shift [15], which occurs when the training data differs from the test data (an example of this will be shown later using the well-known 20 newsgroups dataset). The elucidations provided by explanations are especially valuable in discerning the necessary steps to transform an unreliable model into a reliable one, such as eliminating compromised data or modifying the training data to mitigate dataset shift. Machine learning practitioners frequently encounter the task of model selection, which necessitates the evaluation of the comparative reliability of multiple models. Figure 3 illustrates the utilization of individual prediction explanations in conjunction with accuracy for the purpose of model selection. In this particular scenario, it is observed that the algorithm exhibiting greater accuracy on the validation set is, in reality, significantly inferior. This observation becomes apparent when explanations are supplied, leveraging human previous knowledge, but remains challenging otherwise. Moreover, it is common to observe a discrepancy between the metrics that can be calculated and improved upon (e.g., accuracy) and the metrics that truly matter, such as user engagement and retention. Although the quantification of these indicators may provide challenges, our understanding of how specific model behaviors can impact them is well-established. Hence, a professional in the field may opt for a model with lower precision in content suggestion, deliberately disregarding

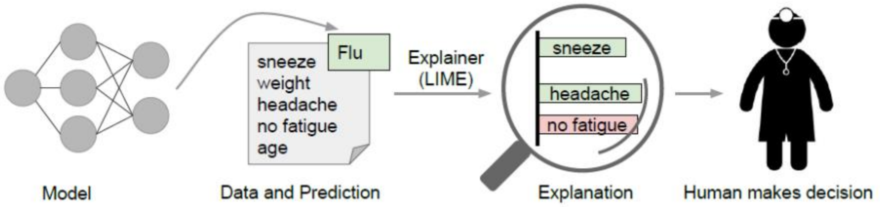


Figure 1: Elucidating the rationale behind individual predictions. According to the model's forecast, the patient exhibits the condition "u," and the LIME technique effectively identifies the specific symptoms within the patient's medical history that contributed to this prediction. The symptoms of sneezing and headache are depicted as factors that support the hypothesis; however, the absence of exhaustion serves as contradictory evidence. Based on these factors, a medical professional can make a well-informed determination regarding the reliability of the model's prognosis.[10]

attributes associated with "clickbait" articles (which could negatively impact user retention). This decision may be made despite the potential improvement in model accuracy during cross-validation by utilizing those features. It is worth noting that explanations are especially valuable in these situations, as they allow for the comparison of various models when a method is capable of generating explanations for any given model.

2 DATA PREPROCESSING

The current scholarly articles demonstrate the methodology of enhancing interpretability and transparency in the use of models, as depicted in Figure 2. Despite the detailed specification of models, datasets, criteria, and outcomes in numerous medical domain publications, it remains necessary to provide explanations and justifications for each individual case. In the coming years, there will be an increased demand for interactive artificial intelligence (AI) systems that offer explainability and facilitate engagement with domain experts. This desire originates from the need to continually enhance outcomes in response to numerous circumstances, including changes in human behavior, weather patterns, and medical problems. Tables 1 to 4 represent potential strategies for managing the corresponding infections or diseases, and are considered suitable for predicting recovery outcomes in a hospital setting. In this section, we will examine the preprocessing techniques employed in the recent study, the algorithms implemented in their respective models, and the resulting outcomes.

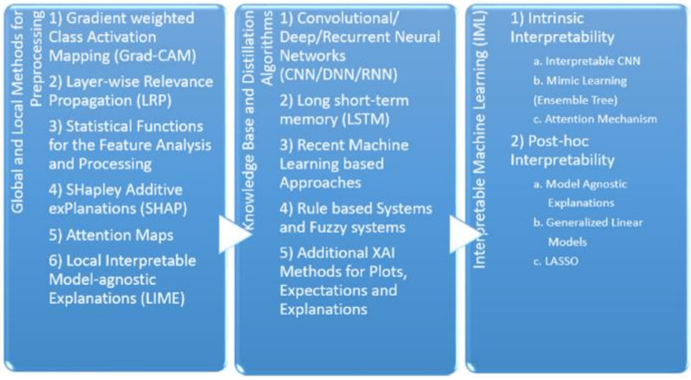

Global and Local Methods for Preprocessing
1) Gradient weighted Class Activation Mapping (Grad-CAM)
2) Layer-wise Relevance Propagation (LRP)
3) Statistical Functions for the Feature Analysis and Processing
4) SHapley Additive exPlanations (SHAP)
5) Attention Maps
6) Local Interpretable Model-agnostic Explanations (LIME)
Knowledge Base and Distillation Algorithms
1) Convolutional/ Deep/Recurrent Neural Networks (CNN/DNN/RNN)
2) Long short-term memory (LSTM)
3) Recent Machine Learning based Approaches
4) Rule based Systems and Fuzzy systems
5) Additional XAI Methods for Plots, Expectations and Explanations
Interpretable Machine Learning (IML)
1) Intrinsic Interpretability
a. Interpretable CNN
b. Mimic Learning (Ensemble Tree)
c. Attention Mechanism
2) Post-hoc Interpretability
a. Model Agnostic Explanations
b. Generalized Linear Models
c. LASSO

Figure 2: XAI Methods Categorization [32].

2.1 Gradient Weighted Class Activation Mapping (Grad-CAM)

The Grad-CAM technique [19] is employed to forecast the corresponding notion by leveraging the gradients of the target, which are propagated to the final convolutional layer. The coarse localization mapping is utilized to identify and emphasize the significant places. The aforementioned technique is recognized as a variation of a heat map, commonly employed in image registration to discern various image sizes and scales for predictive purposes. The Grad-CAM method is a propagation technique that offers a straightforward means of visualization and delivers explanations that are accessible to users. Object detection is a widely employed technique that has gained significant popularity, particularly in the medical field, for the purpose of identifying various diseases and affected regions in patients. The application is capable of effectively identifying chest X-rays (CXR), CT scans, brain tumors, and fractures in various human and animal anatomical regions. Given the potential limitations in accuracy when dealing with sensitive domains, several variants of CAM-supported analysis have been proposed. These include Guided Grad-CAM [16], Respond-CAM [17], Multi-layer CAM [18], among others. The utilization of Guided Grad-CAM involves the assessment of model predictions through the identification of visually prominent characteristics. Therefore, the saliency maps emphasize the relevant properties of the interest class. The technique of combining Grad-CAM and guided backpropagation by pointwise multiplication is commonly referred to as saliency maps in academic literature. The Guided Grad-CAM technique is recognized for its ability to provide maps that are specific to each class. These maps are generated by taking the dot product of the feature map from the last convolutional layers and the neurons, which are then combined to form a projected class score using partial derivatives. The Respond CAM is employed for the manipulation of three-dimensional images characterized by intricate macromolecular structures obtained from cellular electron cryo-tomography (CECT). The Respond-CAM algorithm possesses a "sum to score" characteristic that yields superior outcomes compared to the Grad-CAM method. It is specifically designed to emphasize the class-discriminative regions of three-dimensional pictures by employing weighted feature maps. The sum-to-score property of the Respond-CAM model can be represented as $y^{(c)}$, where $y^{(c)}$ is the class score. Additionally, $b^{(c)}$ represents the last layer CNN parameter, and $\sum_{i,j,k}(L_A^{(c)})$ represents the sum of class c for Grad-CAM/Respond-CAM. Furthermore, C denotes the number of classes as provided in Equation (1). The Multi-layer Grad-CAM method is employed to calculate the conditional probability of the chosen feature using a single maxout hidden layer. The architecture relies on the utilization of maxout units, which are employed in a solitary hidden layer alongside a softmax function that serves to normalize the output probability.

$$y^{(c)} = b^{(c)} + \sum_{i,j,k} \left(L_A^{(c)}\right)_{i,j,k} \tag{1}$$

2.2 Layer-Wise Relevance Propagation (LRP)

It is also one of the popularly used propagation methods, which operates by using the propagation rules for propagating the prediction backward in the neural network. The LRP can flexibly operate on input such as images, videos, and texts. The relevance scores can be recorded in each layer by applying different rules. The LRP is based and justified using a deep taylor decomposition (DTD). It can be set on a single or set of layers in the neural network and can be scaled in the complex DNN by providing high explanation quality. It is also popularly used in the medical domain consisting of CXR, axial brain slices, brain relevance maps, and abnormalities, etc. The versions available in LRP are LRP CNN, LRP DNN, LRP BiLRP, LRP DeepLight for the heatmap visualizations. The LRP relevance is higher as compared to other visualization/sensitivity analysis. The input representations are forward-propagated using CNN until the output is reached and back-propagated by the LRP until the input is reached. Thus, the relevance scores for the categories are yielded in LRP CNN [20]. For the LRP

DNN [21], the CNN is tuned with initial weights for the activity recognition with pixel intensity. In LRP BiLRP [22], the input features pairs having similarity scores are systematically decomposed by this method. The high nonlinear functions are scaled and explained by using composition of LRP. Thus, the BiLRP provides a similarity model for the specific problem by verifiability and robustness. The BiLRP is presented as a multiple LRP combined procedure and recombined on input layer. Here, x and x' are input which are to be compared for similarity, $\emptyset_x$ as a group of network layer with $\{\emptyset_1 \, to \, \emptyset_L\}$, and $y(x, x')$ as the combined output given in Equation (2). The DeepLight LRP [23] performs decoding decision decomposition, which is used to analyze the dependencies between multiple factors on multiple levels of granularity. It is used to study the fine-grained temporo-spatial variability of the high dimension and low sample size structures.

$$BiLRP(y, x, x') = \sum_{m=1}^{h} LRP([\emptyset_L \circ \ldots \circ \emptyset_1]_m, x) \otimes LRP([\emptyset_L \circ \ldots \circ \emptyset_1]_m, x^x) \tag{2}$$

2.3 Statistical Functions for the Feature Analysis and Processing

The comparison of survivors and non-survivors in terms of categorical variables was subjected to statistical analysis [24]. This analysis was conducted using the chi-square test or Fisher's exact test, and the results were presented in terms of interquartile range (IQR) and standard deviation or medians. The continuous variables were analyzed using either the Mann-Whitney U test or Student's t-test, and the results were reported as frequencies. The Kaplan-Meier method is commonly employed in academic research to visually analyze the association between two variables, accompanied by a log rank test to determine the statistical significance of this relationship. The multivariate Cox proportional hazards model is utilized to assess the impact of risk factors on the result. This model is further examined through the use of a log-log prediction plot. In instances of statistical analysis, a noteworthy p-value is considered to be less than 0.05 for univariate analysis and 0.10 for bivariate analysis. The utilization of the generalized estimating equation (GEE) [25] is employed to illustrate the associations among the sets of features that have been matched. The disparity in occurrence between feature inheritance with GEE matching lies in the changes made to the pre and post data. The Charlson comorbidity index score (Charlson et al., 1987) is employed to assess the impact of comorbidities on the one-year mortality risk of hospitalized patients. This scoring system assigns weights to different comorbid conditions in order to calculate an overall index score. The process of multivariate imputation involves utilizing multiple imputation for post-hoc sensitivity analysis on discrete and continuous data through the implementation of chained equations. The LMS approach, as described in reference [26], is employed for the computation of z-scores representing the usual lower limits of spirometric values. The kappa statistic is a measure of chance agreement, where a value of 1.0 indicates perfect agreement and a value of 0 indicates no agreement. The least absolute shrinkage and selection operator (LASSO) is a technique utilized in regression analysis to enhance prediction accuracy through variable selection and regularization [27]. The issue of imbalanced categorization is commonly addressed by the utilization of Synthetic Minority Oversampling Technique (SMOTE) [28]. Imbalance in datasets is commonly attributed to the presence of minority classes, which are subsequently replicated within the training set prior to model fitting. The act of duplicating class material serves to address the issue of class duplication, although it does not contribute any more knowledge.

2.4 Shapely Additive exPlanations (SHAP)

The SHAP [29] uses ranking based algorithms for feature selection. The best feature is listed in the descending values by using SHAP scores. It is based on the features attribution magnitude and is an additive feature attribution method. SHAP is a framework that uses shapley values to explain any model's output. This idea is a part of game theoretic approach which is known for its usability in optimal credit allocation. SHAP can compute well on the black box models as well as tree ensemble models. It is efficient to calculate SHAP values on optimized model classes but can suffer in equivalent settings of model-agnostic settings. Individual aggregated local SHAP values can also be used for global explanations due to their additive property. For deeper ML analysis such as fairness, model monitoring, and cohort analysis, SHAP can provide a better foundation.

2.5 Attention Maps

The LSTM RNN model is commonly utilized for its capacity to emphasize the precise instances in which predictions are primarily influenced by the input variables. This model also offers a high degree of interpretability for users [30]. In summary, the predicted accuracy, illness state analysis, performance breakdown, and interpretability of the RNN are enhanced. The attention vector is responsible for learning feature weights that establish a connection between the subsequent layer of the model and the most frequently utilized features. This vector is commonly employed in conjunction with LSTM to propagate attention weights towards the conclusion of the network. In this context, the weights obtained through the process of learning, denoted as W^k, are utilized to compute the value of a^k for each individual feature, denoted as x_k. In Equation (4), the value of y^k is determined by the learned attention vector, which assigns weights to the feature x_k at each time step.

$$a_k = softmax(W_k x_k) \tag{3}$$

DeepSOFA [31] showcases the imperative nature of capturing individual physiological data in a time-sensitive manner inside an ICU setting. The utilization of the attention mechanism is employed to emphasize the factors within time series data that play a critical role in predicting mortality outcomes. Subsequently, the time step is allocated with increasing weights, believed to possess greater influence on the final result.

$$y_k = a_k \odot x_k \tag{4}$$

3 MODEL SELECTION

3.1 Convolutional/Deep/Recurrent Neural Networks (CNN/DNN/RNN)

CNN, also known as Convolutional Neural Networks, is a prominent deep learning technique employed to model the intricate workings of the human brain. Its primary objective is to enhance performance and effectively address intricate problem-solving tasks. The process involves taking an input data or image and assigning weights and biases to its distinct elements, followed by differentiation between these factors. The filters employed in this context serve as a pertinent mechanism for transforming spatial and temporal interdependencies. Convolutional Neural Networks (CNNs) that have been specifically developed to generate structured output are commonly employed in the task of picture captioning [19]. The CNN + LSTM models have been observed to yield superior results in identifying local discriminative image regions, therefore enhancing the quality of captioning. The CNN scoring method (CNN stands for Convolutional Neural Network) offers accurate localization, as indicated by reference [16]. Subsequently, the scores are computed utilizing specific categories and predetermined criteria. The deep neural network (DNN) [33] is characterized by its architecture, which includes numerous hidden layers within the network. Once the deep neural network (DNN) has undergone training, it has the capability to deliver enhanced performance in detecting suspicious picture findings. This improved performance may be effectively utilized for the purpose of fault identification and status determination. Recurrent Neural Networks (RNNs) are predominantly employed in the domain of natural language processing due to their ability to effectively handle sequential input. The internal memory structure of a system is typically favored for the purpose of retaining its input, making it particularly well-suited for machine learning techniques that include sequential data. The bi-directional recurrent neural network (RNN) [18] has been specifically constructed to serve as both an encoder and a decoder, effectively simulating the process of scanning through sequences during decoding. Hence, it is possible to obtain the sequences of forward and backward concealed states.

3.2 Long-Short-Term Memory (LSTM)

The utilization of Long Short-Term Memory (LSTM) has facilitated progress in the areas of processing, categorizing, and predicting time series data. The issue of the vanishing gradient is commonly addressed through the utilization of Long Short-Term Memory (LSTM) networks. The utilization of the bi-directional Long Short-Term Memory (LSTM) [23] is employed for the purpose of modeling both the within and across numerous structures, taking into account the spatial dependencies. The Deeplight model has a bi-directional Long Short-Term Memory (LSTM) architecture, consisting of two separate LSTM units that operate in opposite directions. The outputs of these LSTM units are subsequently fed into a fully-connected softmax output layer. The Long Short-Term Memory (LSTM) encoder processes embedded sequences of size n using a dual-layer architecture with n cells, and generates dense layers as output. The second Long Short-Term Memory (LSTM) model is designed with a reverse architecture, sometimes referred to as a decoder, which aims to recover the input data. The inclusion of a dropout layer between the encoder and decoder can be employed as a means to mitigate the issue of overfitting. This study employs the linear/non-linear classifier f to analyze the input variable a, which has a dimension of d. The classifier's positive prediction, $f(a) > 0$, is considered. Additionally, the relevance of the single dimension R_d is taken into account.

$$f(a) \approx \sum_{d=1}^{D} R_d \tag{5}$$

In this context, $R_j^{(l)}$ represents a neural network layer with a single neuron at layer l. $R_{i\leftarrow j}^{(l-1,l)}$ refers to the deep light definition of the connection between neuron i at layer $l-1$ and neuron j at layer l. This connection is represented by Z_{ij}, which is calculated as the product of the input $a_i^{(l-1)}$ and the weight coefficient $w_{ij}^{(l-1,l)}$. Additionally, the stabilizer ϵ is included in Equation (6) to ensure stability.

$$\begin{gathered} R_j^{(l)} = \sum_{i\in(l)} R_{i\leftarrow j}^{(l-1,l)} \\ R_{i\leftarrow j}^{(l-1,l)} = \frac{Z_{ij}}{Z_j + \in . sign(Z_j)} R_j^{(l)} \end{gathered} \tag{6}$$

3.3 Recent Machine Learning-Based Approaches

Support Vector Machines (SVMs) are a type of supervised learning algorithms that are utilized for regression, classification, and outlier detection tasks. High-dimensional spaces are commonly preferred for its usage, often exceeding the size of the sample. The linear Support Vector Machine (SVM) [26] is commonly employed in the analysis of extremely large datasets to address multiclass classification tasks. Specifically, it utilizes the cutting plane technique as its underlying framework. The polynomial Support Vector Machine (SVM), also referred to as the polynomial kernel, is a mathematical model that represents polynomials in a feature space. This model is designed to analyze a training set by emphasizing the similarity between vectors. The degree parameter regulates the level of flexibility exhibited by the decision boundary. Therefore, the decision boundary has the potential to expand as a result of utilizing a kernel with a larger degree. The Support Vector Machine (SVM) also incorporates an additional kernel function referred to as the Gaussian Radial Basis Function (RBF). The RBF kernel is a value that is computed based on the distance from a certain point or origin. The term "deep belief network" (DBN) refers to a class or generative graphical model within the field of machine learning [34]. The construction of the model involves the incorporation of latent variables organized in several layers, wherein the layers are interrelated with the exception of the units within each layer. The deep rule forest (DRF) is a type of multilayer tree model that leverages rules to represent the combination of attributes and their interaction with outcomes [35]. The Discriminative Random Forest (DRF) is a method that utilizes techniques derived from random forest and deep learning to detect and analyze interactions. The reduction of validation errors can be achieved by the process of fine-tuning the hyperparameters of deep reinforcement learning frameworks (DRFs). The Dynamic Bayesian Network (DBN) is comprised of a sequence of transformations applied to a Restricted Boltzmann Machine (RBM), where each node in the RBM has a posterior probability that can take on values of either 1 or 0 [36].

$$P(h_i = 1|v) = f(b_i = W_i v) \tag{7}$$

$$P(h_i = 1|h) = f(a_i = W_i h) \tag{8}$$

Here, the $f(x) = 1\ /\ (1 + e^{-x})$, which has energy and distribution function as:

$$E(v,h) = -\sum_{i\in v} a_i v_i - \sum_{j\in h} b_j h_j - \sum_{i,j} v_i h_j w_{ij} \tag{9}$$

$$P(v,h) = -\frac{1}{z} e^{-E(v,h)} \tag{10}$$

The Restricted Boltzmann Machine (RBM) employs unsupervised learning techniques, utilizing a probability density function pdf $p(v)$ and a likelihood function θ that is parameterized by W, a, and b. The input vector v is given as $p(v,\theta)$. The gradient method is used to optimize the likelihood function $logp(v,\theta)$, and improved learning can be obtained by updating the gradient parameters using the partial derivative of $p(v,\theta)$ with respect to θ, denoted as $\frac{\partial p(v,\theta)}{\partial \theta}$.

$$\theta(n+1) = \theta(n) + a \times \left(-\frac{\partial p(v,\theta)}{\partial \theta}\right), \theta \in \{W, a, b\}$$

$$-\frac{\partial logp\left(v, w_{ij}\right)}{\partial w_{ij}} = E_v\left[p(h_i|v) \times v_j\right] - v_j^{(i)} \times f(W_i \times v^{(i)} + b_i)$$

$$-\frac{\partial logp(v, b_i)}{\partial b_i} = E_v\big[p(h_i|v) \times v_j\big] - f(W_i \times v^{(i)})$$

$$-\frac{\partial logp\big(v, a_j\big)}{\partial a_i} = E_v\big[p(h_i|v) \times v_j\big] - v_j \tag{11}$$

3.4 Rule-Based Systems and Fuzzy Systems

A rule-based system utilizes knowledge representation rules to acquire the knowledge encoded inside systems. The reliance on expert systems is absolute, as these systems employ reasoning methods akin to those used by human experts to address knowledge-intensive problems. Interpretable classifiers employing Bayesian analysis have been utilized in stroke prediction models [37]. The process of interpreting decision statements is made easier by discretizing if-then conditions in a high-dimensional and multivariate feature space. The posterior distribution of the decision list is obtained through the application of the Bayesian rule. The employed framework in this context, which is designed to promote sparsity, incorporates a medical grading system that exhibits a high level of accuracy. The utilization of gradient boosting trees in interpretable mimic learning has been found to yield strong prediction performance, making it an effective knowledge distillation strategy [38]. The approach of mimic learning involves the utilization of a model consisting of both a teacher and a student. In this framework, the teacher model serves the purpose of reducing noise and error present in the training data. Additionally, soft labels are employed as a kind of regularization to prevent overfitting in the student model. The application of this approach is observed within the medical field, namely in the context of acute lung injury, where it has demonstrated notable efficacy in generating accurate predictions. Furthermore, it has been observed that this approach can be effectively utilized in the domains of voice processing, multitask learning, and reinforcement learning. Fuzzy rules can be characterized as a type of conditional statement, specifically if-then sentences, which provide a degree of truth rather than a binary true/false outcome. The prediction of ICU patient mortality is facilitated by a sophisticated rule-based fuzzy system that incorporates a diverse dataset comprising both categorical and numeric attributes organized in a hierarchical structure [39]. The model contains interpretable fuzzy rules that are located in each unit of the hidden layer. In order to enhance interpretability, a guided random attribute shift is incorporated into the stack technique. Supervised clustering involves the utilization of a fuzzy partition matrix and cluster centers. In Equation (12), the output weight vectors, denoted as β_{dp}, correspond to a building unit indexed by dp. The partition matrix is represented by U_{dp}, and the output set is denoted as T.

$$\beta_{dp} = \left(\frac{1}{Const} I + U_{dp}^T U_{dp}\right)^{-1} U_{dp} T \tag{12}$$

The interpretability of the layer's prediction can be enhanced by using random projections to increase linear separability. In this context, α' represents the sub constants, Z_{dp} represents the random projection matrix, and Y_{dp} represents the output vector of the last unit.

$$X_{dp} = X + \alpha' Y_{dp} Z_{dp}$$

$$Y_{dp} = U_{dp} \beta_{dp} \tag{13}$$

3.5 Additional XAI Methods for Plots, Expectations, and Explanations

The partial dependence plot (PDP) in the field of machine learning illustrates the marginal impact of one or several input features on the ultimate prediction. Typically, this relationship exhibits a partial dependency. The PDP algorithm calculates the mean value of all input variables, excluding the PDP computed variable n [40]. The variable n is thereafter examined in relation to the alteration in the target variable for the intention of documenting and graphing. When comparing the PDP to individual conditional expectancies, the latter specifically examine particular cases that reveal differences in the recovery of subgroups within the patient population [41]. The optimal approach for explaining the classifier prediction using eXplainable Artificial Intelligence (XAI) is through the utilization of Local Interpretable Model-agnostic Explanations (LIME). LIME serves as an interpretable model that approximates the behavior of a black box model specifically for the instance being analyzed [42]. The artifacts refer to modules that are created by the user and can be interpreted. These modules are subsequently utilized to create local black boxes, specifically for neighboring instances. Semantic LIME (S-LIME) effectively addresses the constraints imposed by user intervention and artifact limitations. This is achieved through the utilization of independently generated semantic characteristics, which are obtained utilizing unsupervised learning techniques. The fidelity function is defined as follows: it involves a model g, an instance x and y, and a feature that measures agreement. The function π is used, which employs an exponential kernel with weighted σ and a distance D.

$$\mathcal{F}(x, f, g, \pi) = \sum_{y \epsilon X} \pi(x, y). \big(f(y) - g(y)\big)^2 \qquad (14)$$

$$D(x, y) = \sum_{x_1 = 1} |x_i - y_i| \qquad (15)$$

LIME is a widely utilized method for emphasizing significant aspects and offering explanations based on its coefficient. However, its utility is hindered by the presence of randomness in the sample step, rendering it unsuitable for implementation in medical contexts. In order to establish confidence, protect interests, and mitigate legal concerns, a proposed method called optimized LIME explanations (OptiLIME) is recommended for diagnostic purposes [43]. The mathematical aspects of OptiLIME are prominently emphasized and maintained consistently during multiple iterations to effectively explore the optimal kernel width in an automated manner. According to the formula provided in Equation (16), the diminishing R^2 is transformed into $l(kw, \tilde{R}^2)$, which represents a global maximum, in order to determine the optimal width. The $\widetilde{R}^2$ represents the anticipated level of adherence when random kw values are considered.

$$l\big(kw, \tilde{R}^2\big) = \begin{cases} R^2(kw), if\ R^2(kw) \leq \tilde{R}^2 \\ 2\tilde{R}^2 - \tilde{R}^2(kw), if R^2(kw) > \tilde{R}^2 \end{cases} \qquad (16)$$

The conventional receiver operating characteristic (ROC) plot and area under the curve (AUC) are influenced by the adjustable threshold, which in turn affects the occurrence of false positive and false negative errors [44]. The utilization of partial ROC and AUC measures in the context of unbalanced data is valuable. Additional approaches, such as partial AUC and the area under the precision-recall (PR) curve, have been proposed as optional solutions. However, it is important to note that these methods alone may not provide a comprehensive solution and should be used with caution. Hence, a novel approach referred to as partial area under the curve $pAUC$ and c statistics of receiver operating characteristic (ROC) have been introduced, preserving the continuous and discrete properties of the area under the curve (AUC), respectively. In the context of evaluating the performance of a binary classification model, the horizontal partial Area Under the Curve (AUC) is computed by considering $x = 1$ as the integration border for the AUC calculation, while designating the other regions as true negatives. When considering the integration with the baseline as the x-axis, it is important to note that the baseline x-value is set to 0 when swapping the x and y axes. Therefore, by converting the variable x (false positive rate) to $1 - x$ (true negative rate), the desired true negative rate (TNR) may be obtained, and when $x = 0$, it becomes 1.

$$pAUC_x \triangleq \int_{y_1}^{y_2} 1 - r^{-1}(y)dy \tag{17}$$

The normalized form of the partial c statistic (c_Δ) for ROC data is expressed in Equation (18). The c_Δ can be represented as a ratio of J from the set of positive elements P, while k can be considered as a subset of negative elements N.

$$\hat{C} \triangleq \frac{2PN.c_\Delta}{J.N + K.P} \tag{18}$$

The partial c statistic can by summed up as shown by the whole curve having q disjoint partial curves.

$$c = \sum_{i=1}^{q} (c_\Delta)_i \tag{19}$$

4 ATTENTION MECHANISMS

Integrate attention mechanisms to highlight regions in medical images that contribute significantly to the model's decision. Self-attention mechanisms, inspired by the transformer architecture, have shown promise in medical image analysis [1]. These mechanisms allow the model to focus on relevant areas of an image.

5 FEATURE IMPORTANCE ANALYTICS

Apply feature importance techniques such as SHAP (SHapley Additive exPlanations) [2]. SHAP values attribute the contribution of each feature to the prediction, offering insights into the decision-making process. This can be especially useful in cases where features are not visually interpretable, such as lab values or genomic data.

6 LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS (LIME)

Employ LIME to generate locally faithful explanations for individual predictions [3]. LIME creates surrogate interpretable models that approximate the behavior of the complex AI model in the vicinity of a specific prediction.

6.1 Interpretable Data Representations

It is crucial to make a distinction between features and interpretable data representations before introducing the explanation system. Regardless of the actual features that the model employs, interpretable explanations must use a representation that is understandable to people. While the classifier may employ more intricate (and incomprehensible) features like word embeddings, one potential interpretable form for text classification is a binary vector denoting the presence or absence of a word. The interpretable representation for image classification may be a binary vector that indicates the "presence" or "absence" of a contiguous patch of similar pixels (a super-pixel), whereas the classifier may represent the image as a tensor with three color channels per pixel. We denote $x \in R^d$ be the original representation of an instance being explained, and we use $x' \in \{0, 1\}^{d'}$ to denote a binary vector for its interpretable representation.

6.2 Fidelity-Interpretability Trade-off

In a formal manner, an explanation is defined as a model g that belongs to the class G. The class G consists of models that have the potential to be interpreted, such as linear models, decision trees, or falling rule lists [7]. In other words, a model g that belongs to G can be easily given to the user using visual or textual means. The domain of function g is defined as the set $\{0, 1\}^{d'}$, indicating that g operates based on the existence or absence of interpretable components. Not all elements $g \in G$ may possess a level of simplicity that allows for easy interpretation. Therefore, we define Ω(g) as a metric of complexity, rather than interpretability, for the explanation $g \in G$. As an illustration, in the case of decision trees, Ω(g) might represent the depth of the tree, whereas in the context of linear models, Ω(g) could denote the count of non-zero weights.

Let us represent the model being explained as $f: R^d \to R$. In the context of classification, the function $f(x)$ represents the probability or binary indicator denoting the membership of x within a specific class.

The proximity measure π_x(z) is employed to determine the distance between an instance z and x, hence establishing the concept of locality surrounding x. Let us denote $L(f, g, \pi_x)$ as a metric that quantifies the degree of discrepancy between the approximation of function g and the true representation of function x inside the specified locality indicated by π_x. To achieve both interpretability and local faithfulness, it is necessary to minimize the function $L(f, g, \pi_x)$ while ensuring that the complexity measure Ω(g) remains sufficiently low for human interpretability. The explanation generated by LIME is acquired using the following process:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} L(f, g, \pi_x) + \Omega(g) \tag{1}$$

This formulation has the potential to be applied with various families G, fidelity functions L, and complexity measures Ω. This study primarily centers on sparse linear models as a means of providing explanations, with a particular emphasis on conducting the search process through perturbations.

6.3 Sampling for Local Exploration

The objective is to reduce the locality-aware loss, denoted as $L(f, g, \pi_x)$, without imposing any assumptions on the function f. This is desired in order to ensure that the explanation remains independent of the specific model being used. Therefore, in order to understand the regional characteristics of function f when the comprehensible inputs change, we estimate the value of $L(f, g, \pi_x)$ by randomly selecting samples, with their weights determined by π_x. Instances around x' are sampled by randomly selecting nonzero items from x' in a uniform manner. The number of such selections is similarly uniformly sampled. In this study, we are provided with a disturbed sample z', where z' belongs to the set $\{0, 1\}$ and represents a percentage of the nonzero components of x'. Our objective is to restore the sample to its original representation z in R^d. Once we have obtained z, we calculate $f(z)$, which serves as the label for the explanation model. The dataset Z consists of altered samples together with their corresponding labels. We aim to optimize Equation (1) in order to obtain an explanation ξ(x). The fundamental concept underlying LIME is illustrated in Figure 2, where we select examples that are

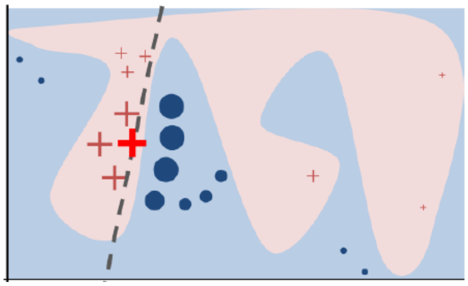


Figure 2: Illustrative example to demonstrate the conceptual understanding of LIME. The decision function f of the black-box model, which is not known to LIME, is of a complex nature. It is visually depicted by the blue/pink background, indicating that it cannot be accurately approximated by a linear model. The bold red cross represents the specific example that is being elucidated. LIME employs a sampling technique to choose instances, using a function f to generate predictions, and assigns weights based on the proximity of these instances to the instance under explanation, as depicted by their respective sizes. The dashed line represents the acquired explanation that exhibits local fidelity, but not global fidelity.[10]

both close to x (which have a higher weight due to π_x) and far from x (which have a lower weight from π_x). While the original model may include excessive complexity for a comprehensive explanation, LIME offers a locally faithful explanation that adheres to linearity in this particular instance. The concept of locality is effectively encapsulated by the variable π_x. It is important to acknowledge that our method demonstrates a considerable level of resilience to sampling noise due to the incorporation of sample weights based on π_x as described in Equation (1). In this study, we give a specific example that exemplifies the broader concept discussed.

6.4 Sparse Linear Explanations

In the subsequent sections of this study, we shall denote G as the set of linear models, where $g(z') = w_g * z'$. The locally weighted square loss, denoted as L, is employed in our study, as specified in Equation (2). In this context, we define $\pi_x(z)$ as an exponential kernel, which is mathematically represented as $exp(-D(x,z)^2\ /\sigma^2\ )$. This kernel is defined based on a distance function D, such as cosine distance for text or $L2$ distance for images, and it is characterized by a width parameter σ.

$$L(f, g, \pi_x) = \sum_{x,x' \in z} \pi_x(z)\big(f(z) - g(z')\big)^2 \tag{2}$$

In the context of text classification, it is imperative to guarantee that the provided explanation is easily understandable. This is achieved by employing an interpretable representation known as a bag of words. Additionally, a constraint is imposed by placing a limit, denoted as K, on the number of words. Mathematically, this constraint can be expressed as $\Omega(g)\ =\ \infty 1[\left|w_g\right|_0\ >\ K]$. It is possible to change the value of K to accommodate the user's capacity, or alternatively, employ varying values of K for different cases. In this study, a fixed value for the parameter K is employed, deferring the investigation of alternative values to subsequent research endeavors. In the context of image classification, a common approach involves utilizing "super-pixels" instead of words, which are obtained by the application of a standard algorithm. Consequently, the interpretable representation of an image is represented by a binary vector, where the value of 1 denotes the presence of the original super-pixel, while 0 signifies a grayed out super-pixel. The specific selection of Ω in Eq. (1)

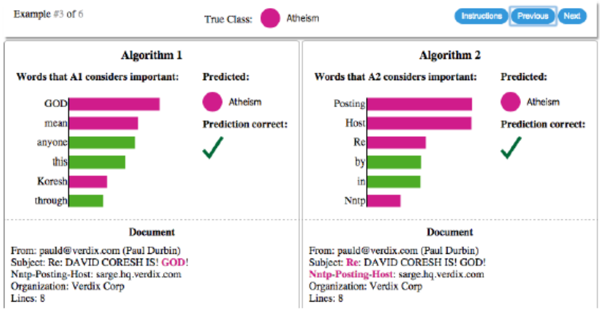
poses challenges for direct solution. However, we address this issue by employing an approximation method. Firstly, we employ Lasso with the regularization route [8] to pick K features. Subsequently, we estimate the weights through the least squares method. This technique is referred to as K-LASSO and is outlined in Algorithm 1. The difficulty of Algorithm 1 is independent of the dataset size, but rather relies on the computational time required to compute $f(x)$ and the number of samples N. In practical applications, the process of elucidating the concept of random forests, consisting of 1000 trees, is efficiently executed using the scikit-learn library (http://scikit-learn.org) on a laptop computer. Specifically, when the dataset size, denoted as N, is set to 5000, the aforementioned task may be completed in less than 3 seconds. It is important to note that this timeframe does not incorporate any optimization techniques such as utilizing graphics processing units (GPUs) or parallelization methods. The process of elucidating every forecast made by the Inception network [9] in the context of image categorization necessitates approximately 10 minutes. Every selection of interpretable representations and G will inevitably possess certain intrinsic limitations. Initially, it should be noted that although the fundamental model can be regarded as an opaque entity, there exist certain interpretable representations that may lack the capacity to elucidate specific behaviors. An instance can be illustrated by a model that predicts the retro nature of sepia-toned photographs, which cannot be elucidated only by the existence or non-existence of super pixels. Furthermore, the selection of G, specifically sparse linear models, may result in the absence of a reliable explanation if the underlying model exhibits significant non-linearity, even within the vicinity of the prediction. Nevertheless, it is possible to make an approximation of the accuracy of the explanation regarding Z and thereafter provide this data to the user. The measure of faithfulness described here can also be employed to choose a suitable set of explanations from a collection of interpretable model classes, thereby accommodating the specific dataset and classifier being utilized. The examination of this topic will be deferred to future research, as our trials have demonstrated that linear explanations are effective for many black-box models.

ALGORITHM 1: Sparse Linear Explanations using LIME[10]

```
Require: Classifier $f$, Number of samples $N$
Require: Instance $x$, and its interpretable version $x^2$
Require: Similarity kernel $\pi_x$, Length of explanation $K$
  $Z \leftarrow \{\}$
  for $i \epsilon \{1, 2, 3, \dots, N$ do
      $z'_i \leftarrow sample\ around(x')$
      $Z \leftarrow Z \cup (z'_i, f(z_i), \pi_x(z_i))$
  end for
  $w \leftarrow K - Lasso(Z, K)$ with $z'_i$ as features, $f(z)$ as target return $w$
```

6.5 Example 1: Text classification with SVMs

In the right side of Figure 3, we provide an explanation of the predictions made by a support vector machine with a radial basis function (RBF) kernel that was trained on unigrams. The purpose of this training was to distinguish between the topics of "Christianity" and "Atheism" using a subset of the 20-newsgroup dataset. Despite achieving a held-out accuracy of 94%, it is important to approach the classifier's results with caution. The explanation for an instance reveals that predictions are generated based on seemingly arbitrary factors, as phrases such as "Posting," "Host," and "Re" have no discernible link to either Christianity or Atheism. The term "Posting" is observed in 22% of instances within the training dataset, with 99% of these occurrences belonging to the category labeled as "Atheism". The classifier is capable of identifying the appropriate names of individuals who frequently contribute to the original newsgroups, even in the absence of headers. However, it should be noted that this ability does not extend to generalization. Upon gaining profound insights from the provided explanations, it becomes evident that the dataset in question exhibits significant flaws that are not readily apparent through a mere examination of the raw data or predictions. Consequently, it is imperative to exercise caution when relying on the classifier or held-out evaluation in this context. The identification of the problems and the formulation of appropriate measures to address these issues and enhance the reliability of the classifier are evident.



Document

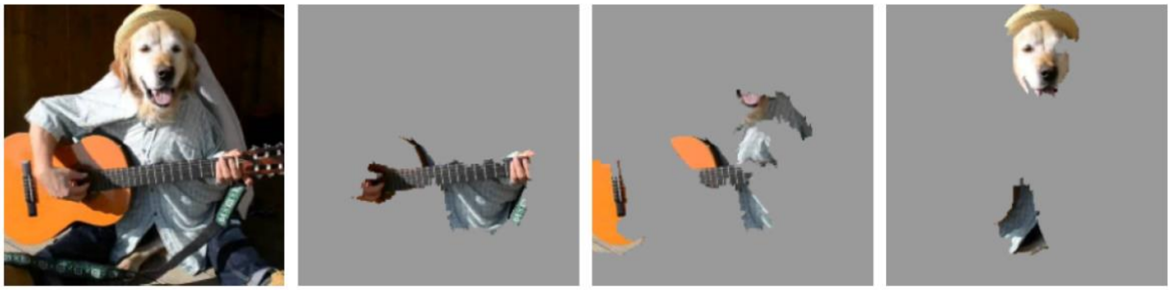
From: pauld@verdix.com (Paul Durbin)
Subject: Re: DAVID CORESH IS! GOD!
Nntp-Posting-Host: sarge.hq.verdix.com
Organization: Verdix Corp
Lines: 8

Document

From: pauld@verdix.com (Paul Durbin)
Subject: Re: DAVID CORESH IS! GOD!
Nntp-Posting-Host: sarge.hq.verdix.com
Organization: Verdix Corp
Lines: 8

Figure 3: This image aims to elucidate the rationale behind individual predictions made by competing classifiers in the task of discerning the thematic focus of a given document, namely whether it pertains to "Christianity" or "Atheism". The bar chart depicts the significance attributed to the most pertinent terms, which are also emphasized in the accompanying text. The color scheme employed in this context serves to denote the specific class to which a given term pertains. The color green is utilized to represent the class of words associated with "Christianity," whereas the color magenta is employed to signify the class of words associated with "Atheism."[10]

6.6 Example 2: Deep networks for images



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 3: This response aims to elucidate the prediction process of Google's Inception neural network for picture classification. The three classes that have been predicted with the highest probabilities are "Electric Guitar" (p = 0.32), "Acoustic Guitar" (p = 0.24), and "Labrador" (p = 0.21) [10].

When employing sparse linear explanations for image classifiers, it may be desirable to solely emphasize the super-pixels that exhibit positive weight towards a particular class. This approach provides insight into the reasoning behind the model's belief that the given class is likely to be present.

In this manner, we elucidate the process of predicting using Google's pre-trained Inception neural network [25] on a randomly selected image (Figure 3a). Figures 3b, 3c, and 3d depict the superpixels explanations for the three highest predicted classes, with the remaining portions of the image rendered in gray. The value of K was set to 10. The neural network demonstrates a natural ability to identify distinguishing features for each class, which aligns with human perception. Specifically, Figure 4b sheds light on the prediction of an acoustic guitar as electric, attributing it to the presence of the fretboard. This type of explanation serves to increase confidence in the classifier, even in cases when the highest predicted class is incorrect, as it demonstrates that the classifier is not behaving in an irrational manner.

7 ETHICAL CONSIDERATIONS

Integrate a module that analyzes potential ethical considerations in the diagnosis process. Ensure the AI system adheres to fairness, transparency, and privacy guidelines to avoid biased or harmful decisions [4].

8 USER INTERFACE

Develop a user-friendly interface that displays both diagnostic outcomes and the generated explanations. The interface should be intuitive for medical professionals to interact with, fostering trust and understanding [5].

9 EVALUATION AND VALIDATION

Perform rigorous evaluation and validation of the explainable AI system using appropriate metrics. Common metrics include accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). Additionally, consider domain-specific evaluation criteria that align with medical standards.

10 DOCUMENTATION

Document the system's architecture, algorithms, and design decisions thoroughly. Ensure that the documentation is accessible to both technical and non-technical audiences, including medical practitioners.

11 COMPLIANCE WITH REGULATIONS

Ensure that the AI system complies with relevant medical regulations and standards, such as HIPAA in the United States or GDPR in the European Union, to protect patient data and privacy [6].

ACKNOWLEDGMENTS

Acknowledgments are placed before the references. Add information about grants, awards, or other types of funding that you have received to support your research. Author can capture the **grant sponsor information**, by selecting the grant sponsor text and apply style ‘GrantSponsor’. After this, select grant no and apply ‘GrantNumber’ from style panel. Example of Grant sponsor: Competitive Research Programme and example of Grant no: CRP 10-2012-03.

HISTORY DATES

In case of submissions being prepared for Journals or PACMs, please add history dates after References as (*please note revised date is optional*):

Received November 2019; revised August 2020; accepted December 2020

REFERENCES

[1] Vaswani, A. et al. (2017). Attention Is All You Need. In Proceedings of NeurIPS.

[2] Lundberg, S. M. and Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. In Proceedings of NeurIPS.

[3] Ribeiro, M. T. et al. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of KDD.

[4] Obermeyer, Z. and Emanuel, E. J. (2016). Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. New England Journal of Medicine.

[5] Car, J. and Sheikh, A. (2003). Integrating health informatics and medical education. BMJ.

[6] General Data Protection Regulation (GDPR). (2018). Official Journal of the European Union.

[7] F. Wang and C. Rudin. Falling rule lists. In Artificial Intelligence and Statistics (AISTATS), 2015.

[8] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. Annals of Statistics, 32:407-499, 2004.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In Computer Vision and Pattern Recognition (CVPR), 2015.

[10] Marco Tulio Ribeiro, Sameer Singh and Carlos Guestrin. 2016. “Why Should I Trust You?” Explaining the Predictions of Any Classifier arXiv:1602.04938v3. University of Washington.

[11] J. L. Herlocker, J. A. Konstan, and J. Riedl. Explaining collaborative altering recommendations. In Conference on Computer Supported Cooperative Work (CSCW), 2000.

[12] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck. The role of trust in automation reliance. Int. J. Hum.-Comput. Stud., 58(6), 2003.

[13] K. Patel, J. Fogarty, J. A. Landay, and B. Harrison. Investigating statistical machine learning as a tool for software development. In Human Factors in Computing Systems (CHI), 2008.

[14] S. Kaufman, S. Rosset, and C. Perlich. Leakage in data mining: Formulation, detection, and avoidance. In Knowledge Discovery and Data Mining (KDD), 2011.

[15] J. Q. Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. Dataset Shift in Machine Learning. MIT, 2009.

[16] Tang, Z.; Chuang, K.V.; DeCarli, C.; Jin, L.W.; Beckett, L.; Keiser, M.J.; Dugger, B.N. Interpretable classification of Alzheimer’s disease pathologies with a convolutional neural network pipeline. Nat. Commun. 2019, 10, 2173.

[17] Zhao, G.; Zhou, B.; Wang, K.; Jiang, R.; Xu, M. RespondCAM: Analyzing deep models for 3D imaging data by visualizations. In Medical Image Computing and Computer Assisted Intervention—MICCAI 2018; Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G., Eds.; Springer: Cham, Switzerland, 2018; pp. 485–492.

[18] Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. arXiv 2014, arXiv:1409.0473.

[19] Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision, Venice, Italy, 22–29 October 2017; pp. 618–626.

[20] Arras, L.; Horn, F.; Montavon, G.; Müller, K.; Samek,W. ‘What is relevant in a text document?’: An interpretable machine learning approach. arXiv 2016, arXiv:1612.07843.

[21] Hiley, L.; Preece, A.; Hicks, Y.; Chakraborty, S.; Gurram, P.; Tomsett, R. Explaining motion relevance for activity recognition in video deep learning models. arXiv 2020, arXiv:2003.14285.

[22] Eberle, O.; Buttner, J.; Krautli, F.; Mueller, K.-R.; Valleriani, M.; Montavon, G. Building and interpreting deep similarity models. IEEE Trans. Pattern Anal. Mach. Intell. 2020, 44, 1149–1161.

[23] Thomas, A.W.; Heekeren, H.R.; Müller, K.-R.; Samek, W. Analyzing neuroimaging data through recurrent deep learning models. Front. Neurosci. 2019, 13, 1321.

[24] Burnham, J.P.; Rojek, R.P.; Kollef, M.H. Catheter removal and outcomes of multidrug-resistant central-line-associated bloodstream infection. Medicine 2018, 97, e12782.

[25] Fiala, J.; Palraj, B.R.; Sohail, M.R.; Lahr, B.; Baddour, L.M. Is a single set of negative blood cultures sufcient to ensure clearance of bloodstream infection in patients with Staphylococcus aureus bacteremia? The skip phenomenon. Infection 2019, 47, 1047–1053.

[26] Oonsivilai, M.; Mo, Y.; Luangasanatip, N.; Lubell, Y.; Miliya, T.; Tan, P.; Loeuk, L.; Turner, P.; Cooper, B.S. Using machine learning to guide targeted and locally-tailored empiric antibiotic prescribing in a children's hospital in Cambodia. Open Res. 2018, 3, 131.

[27] Hsu, C.N.; Liu, C.L.; Tain, Y.L.; Kuo, C.Y.; Lin, Y.C. Machine Learning Model for Risk Prediction of Community-Acquired Acute Kidney Injury Hospitalization From Electronic Health Records: Development and Validation Study. J. Med. Internet Res. 2020, 22, e16903.

[28] Greco, M.; Angelotti, G.; Caruso, P.F.; Zanella, A.; Stomeo, N.; Costantini, E.; Protti, A.; Pesenti, A.; Grasselli, G.; Cecconi, M. Artificial Intelligence to Predict Mortality in Critically ill COVID-19 Patients Using Data from the First 24h: A Case Study from Lombardy Outbreak. Res. Sq. 2021.

[29] Kim, K.; Yang, H.; Yi, J.; Son, H.E.; Ryu, J.Y.; Kim, Y.C.; Jeong, J.C.; Chin, H.J.; Na, K.Y.; Chae, D.W.; et al. Real-Time Clinical Decision Support Based on Recurrent Neural Networks for In-Hospital Acute Kidney Injury: External Validation and Model Interpretation. J. Med. Internet Res. 2021, 23, e24120.

[30] Kaji, D.A.; Zech, J.R.; Kim, J.S.; Cho, S.K.; Dangayach, N.S.; Costa, A.B.; Oermann, E.K. An attention based deep learning model of clinical events in the intensive care unit. PLoS ONE 2019, 14, e0211057.

[31] Shickel, B.; Loftus, T.J.; Adhikari, L.; Ozrazgat-Baslanti, T.; Bihorac, A.; Rashidi, P. DeepSOFA: A Continuous Acuity Score for Critically Ill Patients using Clinically Interpretable Deep Learning. Sci. Rep. 2019, 9, 1–12.

[32] Ruey-Kai Sheu and Mayuresh Sunil Pardeshi. 2022. A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System, Sensors 2022, 22(20), 8068, 21 October 2022, https://doi.org/10.3390/s22208068.

[33] Rueckel, J.; Kunz, W.G.; Hoppe, B.F.; Patzig, M.; Notohamiprodjo, M.; Meinel, F.G.; Cyran, C.C.; Ingrisch, M.; Ricke, J.; Sabel, B.O. Artificial intelligence algorithm detecting lung infection in supine chest radiographs of critically ill patients with a diagnostic accuracy similar to board-certified radiologists. Crit. Care Med. 2020, 48, e574–e583.

[34] Lee, H.-C.; Yoon, S.B.; Yang, S.-M.; Kim,W.H.; Ryu, H.-G.; Jung, C.-W.; Suh, K.-S.; Lee, K.H. Prediction of Acute Kidney Injury after Liver Transplantation: Machine Learning Approaches vs. Logistic Regression Model. J. Clin. Med. 2018, 7, 428.

[35] Kang, Y.; Huang, S.T.;Wu, P.H. Detection of Drug–Drug and Drug–Disease Interactions Inducing Acute Kidney Injury Using Deep Rule Forests. SN Comput. Sci. 2021, 2, 1–14.

[36] Hua, Y.; Guo, J.; Zhao, H. Deep Belief Networks and deep learning. In Proceedings of the 2015 International Conference on Intelligent Computing and Internet of Things, Harbin, China, 17–18 January 2015; pp. 1–4.

[37] Letham, B.; Rudin, C.; McCormick, T.H.; Madigan, D. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. Ann. Appl. Stat. 2015, 9, 1350–1371.

[38] Che, Z.; Purushotham, S.; Khemani, R.; Liu, Y. Interpretable Deep Models for ICU Outcome Prediction. AMIA Annu. Symp. Proc. 2017, 2016, 371–380.

[39] Davoodi, R.; Moradi, M.H. Mortality prediction in intensive care units (ICUs) using a deep rule-based fuzzy classifier. J. Biomed. Inform. 2018, 79, 48–59.

[40] Johnson, M.; Albizri, A.; Harfouche, A. Responsible artificial intelligence in healthcare: Predicting and preventing insurance claim denials for economic and social wellbeing. Inf. Syst. Front. 2021, 1–17.

[41] Xu, Z.; Tang, Y.; Huang, Q.; Fu, S.; Li, X.; Lin, B.; Xu, A.; Chen, J. Systematic review and subgroup analysis of the incidence of acute kidney injury (AKI) in patients with COVID-19. BMC Nephrol. 2021, 22, 52.

[42] Angiulli, F.; Fassetti, F.; Nisticò, S. Local Interpretable Classifier Explanations with Self-generated Semantic Features. In Proceedings of the International Conference on Discovery Science, Halifax, NS, Canada, 11–13 October 2021; Springer: Cham, Switzerland, 2021; pp. 401–410.

[43] Visani, G.; Bagli, E.; Chesani, F. OptiLIME: Optimized LIME explanations for diagnostic computer algorithms. arXiv 2020, arXiv:2006.05714.

[44] Carrington, A.M.; Fieguth, P.W.; Qazi, H.; Holzinger, A.; Chen, H.H.; Mayr, F.; Manuel, D.G. A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. BMC Med. Inform. Decis. Mak. 2020, 20, 1–12..

A APPENDICES

In the appendix section, three levels of Appendix headings are available.

A.1 General Guidelines (AppendixH2)

1. Save as you go and backup your file regularly.
2. Do not work on files that are saved in a cloud directory. To avoid problems such as MS Word crashing, please only work on files that are saved locally on your machine.
3. Equations should be created with the built-in Microsoft® Equation Editor included with your version of Word. (Please check the compatibility at http://tinyurl.com/lzny753 for using MathType.)
4. Please save all files in DOCX format, as the DOC format is only supported for the Mac 2011 version.
5. Tables should be created with Word’s “Insert Table” tool and placed within your document. (Tables created with spaces or tabs will have problems being properly typeset. To ensure your table is published correctly, Word’s table tool must be used.)
6. Do not copy-and-paste elements into the submission document from Excel such as charts and tables.
7. Footnotes should be inserted using Word’s “Insert Footnote” feature.
8. Do not use Word’s “Insert Shape” function to create diagrams, etc.
9. Do not have references appear in a table/cells format as it will produce an error during the layout generation process.
10. MS Word does not consistently allow the original formatting to be modified in the text. In these cases, it is best to copy all the document’s text from the specific file and paste into a new MS Word document and then save it.
11. At times there are font problems such as “odd” stuff/junk characters that appear in the text, usually in the references. This can be caused by a variety of reasons such as copying-and-pasting from another file, file transfers, etc. Please review your text prior to submission to make sure it reads correctly.

A.1.1 Preparing Graphics (AppendixH3)

1. Accepted image file formats: TIFF (.tif), JPEG (.jpg).
2. Scalable vector formats (i.e., SVG, EPS and PS) are greatly preferred.
3. Application files (e.g., Corel Draw, MS Word, MS Excel, PPT, etc.) are NOT recommended.
4. Images created in Microsoft Word using text-box, shapes, clip-art are NOT recommended.
5. IMPORTANT: All fonts must be embedded in your figure files.

6. Set the correct orientation for each graphics file.

A.2 Placeholder Text

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Vulputate sapien nec sagittis aliquam. Malesuada fames ac turpis egestas sed tempus urna. Posuere sollicitudin aliquam ultrices sagittis orci. Consequat id porta nibh venenatis cras sed felis eget. Pellentesque eu tincidunt tortor aliquam nulla facilisi cras fermentum odio. Tincidunt nunc pulvinar sapien et ligula ullamcorper malesuada proin. Tincidunt lobortis feugiat vivamus at augue. Eget nunc lobortis mattis aliquam faucibus. Egestas diam in arcu cursus euismod quis.

Erat pellentesque adipiscing commodo elit at imperdiet. In hac habitasse platea dictumst quisque sagittis purus. At lectus urna duis convallis. Eu mi bibendum neque egestas congue. Est ullamcorper eget nulla facilisi etiam dignissim diam. Sed ullamcorper morbi tincidunt ornare massa eget. Aenean vel elit scelerisque mauris pellentesque. Ullamcorper dignissim cras tincidunt lobortis feugiat vivamus. Cras fermentum odio eu feugiat pretium nibh. Congue eu consequat ac felis donec et odio pellentesque diam. Velit sed ullamcorper morbi tincidunt ornare massa eget egestas. In metus vulputate eu scelerisque felis imperdiet proin fermentum leo. Nulla malesuada pellentesque elit eget gravida cum.

Nullam ac tortor vitae purus faucibus ornare suspendisse. Libero enim sed faucibus turpis in eu mi bibendum neque. Sodales ut etiam sit amet nisl purus. Egestas diam in arcu cursus. Aliquet porttitor lacus luctus accumsan tortor. Pharetra magna ac placerat vestibulum lectus. Sit amet mauris commodo quis imperdiet massa tincidunt. In nisl nisi scelerisque eu ultrices vitae auctor. Nisi vitae suscipit tellus mauris a diam. Dui vivamus arcu felis bibendum ut tristique. Laoreet suspendisse interdum consectetur libero id.

Enim eu turpis egestas pretium. Nulla aliquet enim tortor at auctor urna. Id aliquet risus feugiat in. Non enim praesent elementum facilisis leo. Integer feugiat scelerisque varius morbi enim nunc faucibus. Egestas dui id ornare arcu odio ut sem nulla pharetra. Montes nascetur ridiculus mus mauris. Orci dapibus ultrices in iaculis. Enim sed faucibus turpis in eu mi bibendum neque. Faucibus pulvinar elementum integer enim neque volutpat ac tincidunt vitae. Et ultrices neque ornare aenean euismod elementum. Et pharetra pharetra massa massa ultricies mi quis hendrerit dolor. Tempus iaculis urna id volutpat lacus laoreet non curabitur gravida. Est velit egestas dui id ornare arcu odio. Eu facilisis sed odio morbi quis commodo odio. Lectus vestibulum mattis ullamcorper velit sed ullamcorper morbi tincidunt.

Eu non diam phasellus vestibulum lorem sed risus ultricies. Convallis aenean et tortor at risus viverra adipiscing at. Mauris pellentesque pulvinar pellentesque habitant morbi. Elementum sagittis vitae et leo duis. Massa enim nec dui nunc. Nisl tincidunt eget nullam non nisi est sit amet. Amet nisl purus in mollis nunc sed id semper. Fermentum leo vel orci porta non pulvinar neque laoreet suspendisse. Diam vel quam elementum pulvinar etiam non quam. Sagittis orci a scelerisque purus semper eget. Aliquet porttitor lacus luctus accumsan tortor. Integer vitae justo eget magna fermentum iaculis eu non diam. Egestas pretium aenean pharetra magna ac. Cursus metus aliquam eleifend mi in nulla. Cursus mattis molestie a iaculis at erat pellentesque adipiscing. Pulvinar pellentesque habitant morbi tristique senectus. Gravida cum sociis natoque penatibus et magnis dis parturient montes. In aliquam sem fringilla ut. Ut consequat semper viverra nam libero justo laoreet. Pellentesque diam volutpat commodo sed egestas.

Ornare arcu odio ut sem nulla pharetra diam. Ut enim blandit volutpat maecenas volutpat blandit aliquam. Tempus iaculis urna id volutpat lacus. Nascetur ridiculus mus mauris vitae. Venenatis cras sed felis eget velit aliquet sagittis id. Laoreet non curabitur gravida arcu ac tortor dignissim convallis aenean. Maecenas ultricies mi eget mauris pharetra et ultrices neque ornare. Egestas purus viverra accumsan in nisl nisi scelerisque eu ultrices. Tempus urna et pharetra pharetra massa massa. Pulvinar neque laoreet suspendisse interdum consectetur libero id.

Nisl rhoncus mattis rhoncus urna neque viverra justo nec ultrices. Morbi quis commodo odio aenean sed adipiscing diam donec. Neque gravida in fermentum et. Scelerisque purus semper eget duis at tellus. Volutpat blandit aliquam etiam erat velit scelerisque in dictum non. Odio ut sem nulla pharetra diam sit. Sed pulvinar proin gravida hendrerit lectus a. Diam ut venenatis tellus in metus vulputate eu scelerisque. Id semper risus in hendrerit. Vel quam elementum pulvinar etiam. Amet aliquam id diam maecenas ultricies mi. Auctor elit sed vulputate mi sit amet. Orci dapibus ultrices in iaculis nunc. Sed vulputate odio ut enim blandit volutpat maecenas volutpat. Auctor urna nunc id cursus metus. Integer enim neque volutpat ac tincidunt vitae.

Scelerisque in dictum non consectetur a erat. Vel risus commodo viverra maecenas accumsan lacus vel facilisis volutpat. Dignissim sodales ut eu sem integer vitae justo eget magna. Nunc non blandit massa enim nec dui nunc mattis enim. Sed vulputate odio ut enim blandit volutpat maecenas. Ante in nibh mauris cursus. Donec pretium vulputate sapien nec sagittis aliquam malesuada. Eu volutpat odio facilisis mauris sit amet massa. Blandit turpis cursus in hac habitasse platea dictumst quisque. Donec enim diam vulputate ut pharetra sit.

Magna fringilla urna porttitor rhoncus dolor purus non. Fames ac turpis egestas integer eget. Mattis rhoncus urna neque viverra. Laoreet sit amet cursus sit amet dictum sit amet. Vel pretium lectus quam id leo in vitae turpis massa. Euismod lacinia at quis risus sed vulputate odio ut. Lorem dolor sed viverra ipsum. Viverra justo nec ultrices dui sapien. Aliquam nulla facilisi cras fermentum odio eu feugiat pretium. Adipiscing commodo elit at imperdiet dui accumsan sit amet nulla. Morbi leo urna molestie at elementum eu facilisis sed. Habitant morbi tristique senectus et netus et malesuada. Viverra ipsum nunc aliquet bibendum enim. Integer vitae justo eget magna fermentum. Tincidunt id aliquet risus feugiat. Mauris ultrices eros in cursus turpis. Amet venenatis urna cursus eget nunc. Nisl nisi scelerisque eu ultrices vitae.

Non pulvinar neque laoreet suspendisse interdum consectetur libero. Facilisis leo vel fringilla est ullamcorper eget nulla facilisi. Ipsum dolor sit amet consectetur adipiscing elit pellentesque. Risus quis varius quam quisque id. Bibendum arcu vitae elementum curabitur vitae. Vitae et leo duis ut diam quam nulla. Orci eu lobortis elementum nibh tellus molestie nunc non blandit. Arcu odio ut sem nulla pharetra diam sit amet. Quis vel eros donec ac odio. Est lorem ipsum dolor sit amet consectetur adipiscing