On Understanding the Influence of Controllable Factors with a Feature Attribution Algorithm: a Medical Case Study

Veera Raghava Reddy Kovvuri
Swansea University
Swansea,United Kingdom
v.r.r.kovvuri@swansea.ac.uk

Siyuan Liu
Nanyang Technological University
Singapore
syliu@ntu.edu.sg

Monika Seisenberger
Swansea University
Swansea,United Kingdom
m.seisenberger@swansea.ac.uk

Xiuyi Fan
Nanyang Technological University
Singapore
xyfan@ntu.edu.sg

Berndt Müller
Swansea University
Swansea,United Kingdom
berndt.muller@swansea.ac.uk

Hsuan Fu
Université Laval
Québec, Canada
hsuan.fu@fsa.ulaval.ca

***Abstract*—Feature attribution XAI algorithms enable their users to gain insight into the underlying patterns of large datasets through their feature importance calculation. Existing feature attribution algorithms treat all features in a dataset homogeneously, which may lead to misinterpretation of consequences of changing feature values. In this work, we consider partitioning features into *controllable* and *uncontrollable* parts and propose the *Controllable fActor Feature Attribution* (CAFA) approach to compute the relative importance of controllable features. We carried out experiments applying CAFA to two existing datasets and our own COVID-19 non-pharmaceutical control measures dataset. Experimental results show that with CAFA, we are able to exclude influences from uncontrollable features in our explanation while keeping the full dataset for prediction.**

***Index Terms*—Explainable AI, Feature Attribution, Medical Application**

I. Introduction

Feature attribution algorithms [15] are a popular class of Explainable AI (XAI) algorithms. Given a prediction instance, they tell the relative "importance" of each feature in the instance. In addition to "explaining" the prediction model, importance measures also reveal insight about the instance being explained, e.g., [4] shows that XAI can help "generating the hypothesis about causality" in developing decision support systems. In this sense, feature attribution algorithms are considered as a data mining tool for extracting and discovering patterns in large datasets. For instance, [6] uses feature attribution algorithms to understand important factors affecting cancer patient survivability; [7] employs feature attribution algorithms to study factors affecting the transmission of SARS-CoV-2; and [14] uses feature attribution to analyse factors affecting foreign exchange markets. However, existing feature attribution algorithms (see e.g., [2], [17], [22] for overviews) treat all features homogeneously when computing their relative importances. Such homogeneity may not always give desirable interpretations when feature attribution algorithms are used for data mining purposes. Consider the following hypothetical example.

> Suppose we want to estimate the chance for some individual having breast cancer, with features like *age*, *gender*, *weight*, *alcohol intake*, *smoking habits*, *family history*, etc. A predictive model estimates the likelihood of the person having breast cancer; and a feature attribution algorithm gives attributions like *age: 0.3*, *gender: 0.13*, *weight: 0.27*, *alcohol intake: 0.15*, *smoking: 0.3*, *family history: 0.36*, etc.

From these calculated values, we notice that certain features, such as *age*, *gender* and *family history*, while being influential to the prediction, are *uncontrollable risk factors* [5]. Knowing the relative importance of these features makes little contribution to clinical decision making. On the other hand, features representing *controllable risk factors* such as *weight*, *alcohol intake* and *smoking habits* are vital to clinical interventions [5]. Thus, from an intervention perspective, it is necessary to distinguish these two classes of factors and compute their influences accordingly. We raise the question:

What are the influences of controllable factors used in a prediction?

To answer this question, a naive approach would be to build another predictive model, which only considers controllable factors, and apply feature attribution algorithms to that model. However, as explained in [8] and [24], dropping features from models can negatively impact the model performance as we will show in Table I in experimental study. Thus, instead of building models with fewer features, we suggest creating algorithms that are able to treat controllable factors differently from uncontrollable ones.

In this paper, we present *Controllable fActor Feature Attribution (CAFA)*. Through *selective perturbation* and *global-for-*

local interpretation, CAFA computes the relative importance of controllable factors for individual instances using prediction models built from all features. We apply CAFA on lung cancer data in Simulacrum[1] and on the UCI breast cancer dataset[2] to study the influence of controllable factors on survival time or recurrence. In a second experiment, we apply CAFA to a COVID-19 virus transmission case study (Section V) to explore the effectiveness of non-pharmaceutical control measures.

II. Background

Given a prediction model $f \in \mathcal{F}$ where $\mathcal{F}$ is a set of models, let $\mathbf{y} = f(\mathbf{x})$ be the prediction made by f on the input $\mathbf{x} = \langle \mathbf{x}_1, \ldots, \mathbf{x}_m \rangle \in \mathbb{R}^m$, and a *feature attribution* algorithm give an explanation $\Phi_{\mathbf{x}} = \langle \phi_1, \ldots, \phi_m \rangle \in \mathbb{R}^m$, where ϕ_i can be viewed as the relative importance of $\mathbf{x}_i$ for the prediction $\mathbf{y} = f(\mathbf{x})$. We briefly review two feature attribution algorithms supporting this work as follows.

A. Local Interpretable Model-agnostic Explanations (LIME) [19]

To explain how a model f predicts a data instance $\mathbf{x}$, LIME generates a new dataset $D = \{(\mathbf{x}_1, f(\mathbf{x}_1)), \ldots, (\mathbf{x}_n, f(\mathbf{x}_n))\}$ consisting of n perturbed samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ within some proximity $\pi_{\mathbf{x}}$ of $\mathbf{x}$, and then fits an interpretable model g to D. The parameters of the new model are the explanation of $\mathbf{x}$. Formally, LIME computes explanations as:

$$\text{LIME}(\mathbf{x}) = \arg\min_{g \in \mathcal{G}} L(f, g, \pi_{\mathbf{x}}) + \Omega(g), \tag{1}$$

where L is a loss function comparing f and g, $\mathcal{G} \subseteq \mathcal{F}$ is a class of interpretable models, and $\Omega(g)$ the complexity of g.

B. SHapley Additive exPlanations (SHAP) [15]

SHAP is based on the coalitional game theory concept of a *Shapley value*, assigned to each feature of instance $\mathbf{x}$. The Shapley value of a feature is its marginal contribution to the prediction thus explains the prediction. Specifically, let g be the explanation model. For an instance $\mathbf{x}$ with m features, there is a corresponding $z \in \{0,1\}^m$ such that SHAP specifies g being a linear function of z:

$$g(z) = \phi_0 + \sum_{j=1}^{m} \phi_j z_j, \tag{2}$$

where $\phi_j (0 < j \leq m)$ is the Shapley value of feature j and ϕ_0 is the "average" prediction when none of the features in $\mathbf{x}$ is present. Both SHAP and LIME are local methods in the sense that they explain individual instances in a dataset. Global explanation, which describes the average behaviour of the dataset, can be simply obtained by taking the average of local explanations of instances in the dataset [17].

[1]Simulacrum is a dataset developed by Health Data Insight CiC derived from anonymous cancer data provided by the National Cancer Registration and Analysis Service, which is part of Public Health England.

[2]http://archive.ics.uci.edu/ml/datasets/Breast+Cancer

III. Our approach

CAFA computes feature importances for controllable factors through *selective perturbation* and *global-for-local interpretation*. Conceptually, CAFA is inspired by LIME such that a set of perturbed samples is generated to compute the feature importance. However, there are two main differences. Firstly, unlike LIME where the perturbation is carried out uniformly throughout all features, CAFA selectively perturbs features representing controllable factors. Secondly, with the dataset generated, instead of fitting a weak interpretable model for computing explanations, a strong model is chosen to fit the dataset. We then determine the feature importance of controllable factors by using an explainer to compute the global explanation on the dataset. Fig. 1 illustrates CAFA's selective perturbation strategy.

Given a prediction model f, for a data point $\mathbf{x}$ with m features partitioned into two sets F_c (*controllable*) and F_u (*uncontrollable*) such that $F_c \cap F_u = \{\}$, to compute feature importance for F_c, we construct a data set with n points

$$D_{\mathbf{x}} = \{(\mathbf{x}_1, f(\mathbf{x}_1)), \ldots, (\mathbf{x}_n, f(\mathbf{x}_n))\}$$

such that for all $(\mathbf{x}_i, f(\mathbf{x}_i)) \in D_{\mathbf{x}}$, the following two conditions hold:

1) $\delta(\mathbf{x}, \mathbf{x}_i) \leq \pi_{\mathbf{x}}$, where δ is a distance function and $\pi_{\mathbf{x}}$ is some proximity threshold, and
2) for $\mathbf{x} = \langle v_1, \ldots, v_m \rangle$, and $\mathbf{x}_i = \langle v_1^i, \ldots, v_m^i \rangle$, for all j $(1 \leq j \leq m)$, it is the case that if feature j is in F_u, then $v_j = v_j^i$.

For two instances $\mathbf{x}_1 = \langle v_1^1, \ldots, v_m^1 \rangle$ and $\mathbf{x}_2 = \langle v_1^2, \ldots, v_m^2 \rangle$, the distance function $\delta(\mathbf{x}_1, \mathbf{x}_2)$ is

$$\delta(\mathbf{x}_1, \mathbf{x}_2) = \frac{\sum_{i=1}^{m} \omega_i d(v_i^1, v_i^2)}{\sum_{i=1}^{m} \omega_i}, \tag{3}$$

where ω_i is the weight of feature i and $d(v_i^1, v_i^2)$ is defined by[3]:

- if feature i is categorical, then

$$d(v_i^1, v_i^2) = \begin{cases} 0 & \text{if } v_i^1 = v_i^2, \\ 1 & \text{otherwise;} \end{cases} \tag{4}$$

- if feature i is continuous, then

$$d(v_i^1, v_i^2) = |v_i^1 - v_i^2|. \tag{5}$$

We then build a strong prediction model g from $D_{\mathbf{x}}$ and calculate the global explanation $g(D_{\mathbf{x}})$ using SHAP by first computing local explanations for all instances in $D_{\mathbf{x}}$ and then averaging the results. Overall, for an instance $\mathbf{x}$ and explanations Φ_i computed over $D_{\mathbf{x}}$,

$$\text{CAFA}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \Phi_i. \tag{6}$$

[3]Note that we assume some standard normalization / scaling pre-processing is performed on the dataset so all continuous features take values in the range [0,1].

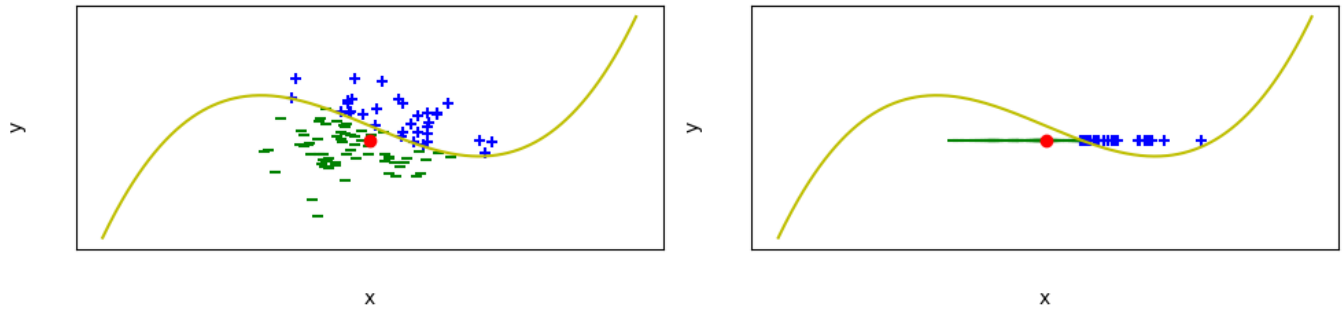


Fig. 1: Selective Perturbation in CAFA. The point of interest (explanation point) and the generated dataset are shown in the figures. The red dot denotes the point of interest in a 2D space. The yellow curve is the decision boundary. Blue "+" and green "-" denote generated positive and negative samples, respectively. The figure on the left illustrates the standard perturbation (LIME), where both features x and y are perturbed; the figure on the right illustrates the selective perturbation (CAFA), where only the x axis, representing the controllable factor, is perturbed.

Thus, we use the *global* explanation computed with a strong predictor on $D_{\mathbf{x}}$ as the *local* explanation for $\mathbf{x}$. This *global-for-local interpretation* is superior to LIME's local surrogate approach, as it has been shown that SHAP is more robust than LIME [10], [16], [21], [23].

Algorithm 1 describes the process in detail. Since all points in $D_{\mathbf{x}}$ have the same values for their uncontrollable features, these features have no correlation to class labels of points in $D_{\mathbf{x}}$. Thus, their feature importance will be assigned to 0, as they make no contribution to the prediction. By setting that each class contains K samples (Line 7), we ensure that $D_{\mathbf{x}}$ is balanced.

Algorithm 1 Selective Perturbation and Global-for-Local Interpretation.

Input: Data point $\mathbf{x}$, Prediction model f, Proximity threshold $\pi_{\mathbf{x}}$, Distance Function δ, Controllable features F_c, Sample class size K **Output:** Feature Importance Φ

1: Let $D'_{\mathbf{x}} = []$;
2: **do**
3: Randomly generate a data point $\mathbf{x}'$ such that for all features $v \in F_u$, $\mathbf{x}'$ contains the same value as $\mathbf{x}$ in v and $\delta(\mathbf{x}, \mathbf{x}') \leq \pi_{\mathbf{x}}$;
4: Append $(\mathbf{x}', f(\mathbf{x}'))$ to $D'_{\mathbf{x}}$;
5: Let r be the size of the smallest class in $D'_{\mathbf{x}}$;
6: **while** $r < K$;
7: Construct $D_{\mathbf{x}}$ from $D'_{\mathbf{x}}$ by sampling K elements from each class in $D'_{\mathbf{x}}$;
8: Let Φ be the global explanation for $g(D_{\mathbf{x}})$ with a strong predictor g;
9: **return** Φ;

IV. Experiments with Two Existing Medical Datasets

As an experiment, we apply CAFA to the lung cancer data in Simulacrum and the UCI breast cancer dataset. We predict 12-months survival on the lung cancer dataset, which contains 2,242 instances specified by 24 features:

- Four uncontrollable features: age, ethnicity, sex and height;
- 20 controllable features: morph, weight, dose administration, regimen outcome description, administration route, clinical trial, cycle number, regimen time delay, cancer plan, T best, N best, grade, CReg code, laterality, ACE, CNS, performance, chemo radiation, regimen stopped early, and M Best.[4]

The breast cancer dataset comprises 286 data instances, predicting cancer recurrence, each containing 9 features, which are:

- Two uncontrollable features: age and menopause;
- Seven numerical controllable features: tumor size, inv-nodes, node-caps, deg-malig, breast, breast-quad, and irradiate.

Random forest classifiers are used in both cases.

Firstly, we illustrate that simply dropping uncontrollable features will negatively impact the prediction accuracy. As shown in Table I, the accuracy drops across the three datasets, i.e., lung cancers, breast cancer, and covid19 (we will introduce covid19 dataset in the next section), suggesting features importances achieved from models from fewer features may be different from the ones achieved from using the original dataset.

	Original	Controllable features only
Lung Cancer	0.97	0.85
Breast Cancer	0.79	0.76
COVID19	0.94	0.88

TABLE I: The prediction for lung cancer, breast cancer, and COVID19 dataset by using the original dataset and the dataset with controllable features only.

We then explore the influence of controllable features on prediction results on individual instances (local explanations).

[4]Description of features used in this dataset can be found at the Cancer Registration Data Dictionary and the SACT Data Dictionary, with links available at:
https://simulacrum.healthdatainsight.org.uk/available-data/table-descriptions/.

To this end, we randomly sample an instance from each dataset, as follows:

- *Lung Cancer: age 71; ethnicity 5; sex 0; morph 8140; weight 49.8; height 1.83; dose administration 8; regimen outcome 1; administration route 1; clinical trial 2; cycle number 1; regimen time delay 0; cancer plan 0; T Best 3; N Best 0; grade 3; CReg Code 401; laterality 2; ACE 9; CNS 99; performance 0; chemo radiation 0; regimen stopped early 1; M Best 0.*
- *Breast Cancer: age 40; menopause 0; tumor-size 6; inv-nodes 0; node-caps 1; deg-malig 3; breast 0; breast-quad 3; irradiate 0.*

For each instance $\mathbf{x}$, we generate $D_{\mathbf{x}}$ containing 1,000 perturbed instances (binary classification, $K = 500$) and carry out the CAFA calculation as shown in Algorithm 1. We let $\pi_{\mathbf{x}}$ be the average distance between points and feature weights $\omega_i = 1$. Results from SHAP and CAFA are shown in Fig. 2. In this figure, the x-axis shows the features; y-axis shows feature importance. For each feature, the left (blue) bar shows the SHAP result of the feature, and the right (red) bar shows the importance calculated with CAFA. We observe that:

1) For uncontrollable features, i.e., "age", "ethnicity", "sex", and "height" from the lung cancer dataset as well as "age" and "menopause" from the breast cancer dataset, the assigned importance value is 0, as expected;
2) For controllable features, there is a strong correlation, 0.96 for lung cancer and 0.99 for breast cancer, between values represented by the blue and the red bars, suggesting that CAFA is agreeable with SHAP.

This suggests that CAFA successfully excludes influences of uncontrollable features with its calculation, while maintaining properties of standard feature attribution algorithms such as SHAP.

We further study the influence of uncontrollable features with CAFA for global explanations. We randomly sample 100 instances from each dataset and compute global explanations with SHAP and CAFA. We produce "violin plots" using the summary plot function from the SHAP library. Fig. 3 (a) and (b) illustrate global explanations for the lung and breast cancer datasets, respectively. There, the x-axis is the feature importance and the y-axis is the features. Color (red to blue) represents the value of a feature.

For Fig. 3 (a) and (b), the left-hand side figures show results from SHAP; and at the right-hand side figures show results from CAFA. We can see that: (1) as seen in local explanation cases (Fig. 2), all uncontrollable features are assigned an importance value 0; (2) similar patterns to SHAP on controllable features can be seen from CAFA, i.e., similar color patterns for a specific feature; (3) the orders of feature importance differ from SHAP to CAFA. We conclude that, for global explanations, CAFA precludes uncontrollable features from contributing to explanations, and CAFA produces distinct explanations to SHAP even if uncontrollable features are excluded.

V. UK COVID-19 Case Study

With the outbreak of the COVID-19 pandemic in December 2019, many countries have implemented some non-pharmaceutical control measures to contain the spread of the virus in the absence of effective vaccination and treatment. In this case study, we use CAFA to study the effectiveness of the non-pharmaceutical control measures implemented in the UK.

We formulate the effectiveness of control measures as an XAI modelling problem. We focus on studying the relationship between control measures and the daily reproduction rate R_t. R_t is one of the most important metrics used to measure the epidemic spread. A value greater than 1 suggests the epidemic being expanding; a value less than 1 indicates shrinking. We employ the approach presented in [9] for estimating R_t from daily infection cases. We then pose the following classification problem:

> *Given non-pharmaceutical control measures applied on a specific day, predict whether R_t is smaller or greater than 1 on that day.*

Having this prediction problem solved by a classifier, we use CAFA to identify control measures that make the greatest contribution to the prediction. Thus, by analysing the behaviour of the prediction model, we gain insight into the effectiveness of control measures.

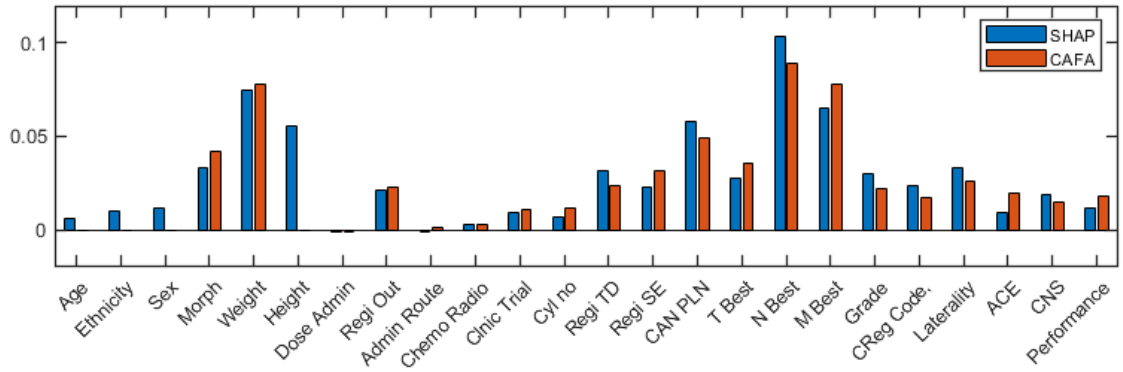
We have collected a dataset containing daily infection numbers and control measures from 04/January/2020 to 06/September/2021. Each instance consists of uncontrollable features (i.e., daily number of infections, cumulative cases, daily number of deaths and tests performed, temperature and humidity) and controllable features (i.e., implemented control measures). The numbers of daily cases, cumulative cases, deaths, and tests performed are collected from the Public Health England website[5]. Control measure information is retrieved from Wikipedia[6] and various news articles.

We have considered control measures *school closures (SC)*, restrictions on *meeting friends and family indoors (MInd), meeting friends and family outdoors (MOut), domestic travel (DT), international travel (IT), hospitals and nursing home visits (HV), opening of cafes and restaurants (CR), accessing pubs and bars (PB), sports and leisure venues (SL)*, and *non-essential shops (NS)*. The values for control measures are binary, e.g, for "school closure", the values are "open" and "closed"; for "restrictions on meeting indoors" the values are "High" (H) or "Moderate" (M). To accommodate the temporal effect of control measures, each feature is represented categorically. For instance, if they are open, then the "school closure" feature takes value 0; if the schools are closed for 0-5 days, then it takes value 1; etc.

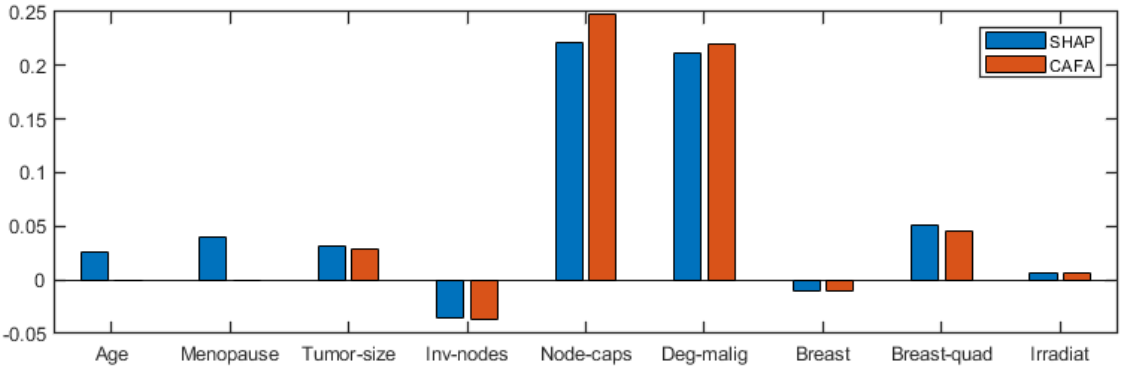
In total, we have collected 4,256 data points across 12 UK regions: East Midlands, East of England, London, North East, North West, South East, South West, West Midlands, Yorkshire

[5]COVID-19 Dashboard (UK): https://coronavirus.data.gov.uk

[6]For example, for Wales the control measure data has been collected from https://en.wikipedia.org/wiki/Timeline_of_the_COVID-19_pandemic_in_Wales



(a) A lung cancer instance randomly selected from the Simulacrum dataset. Uncontrollable features are: *Age*, *Ethnicity*, *Sex*, and *Height*.



(b) A breast cancer instance randomly selected from the UCI breast cancer dataset. Uncontrollable features are *Age* and *Menopause*.

Fig. 2: Illustration of CAFA vs. SHAP on two explanation instances selected from two medical datasets. We observe that (1) with CAFA, all uncontrollable features are assigned importance 0; (2) for controllable features, CAFA produces results that are agreeable with the ones given by SHAP.

and Humber, Northern Ireland, Scotland and Wales. To remove noise and achieve a more accurate R_t estimation, we drop data points with cumulative cases less than 20 for each region and keep 3,936 instances. A sliding-window mean filter of size 3 has been used to filter noise in daily cases.

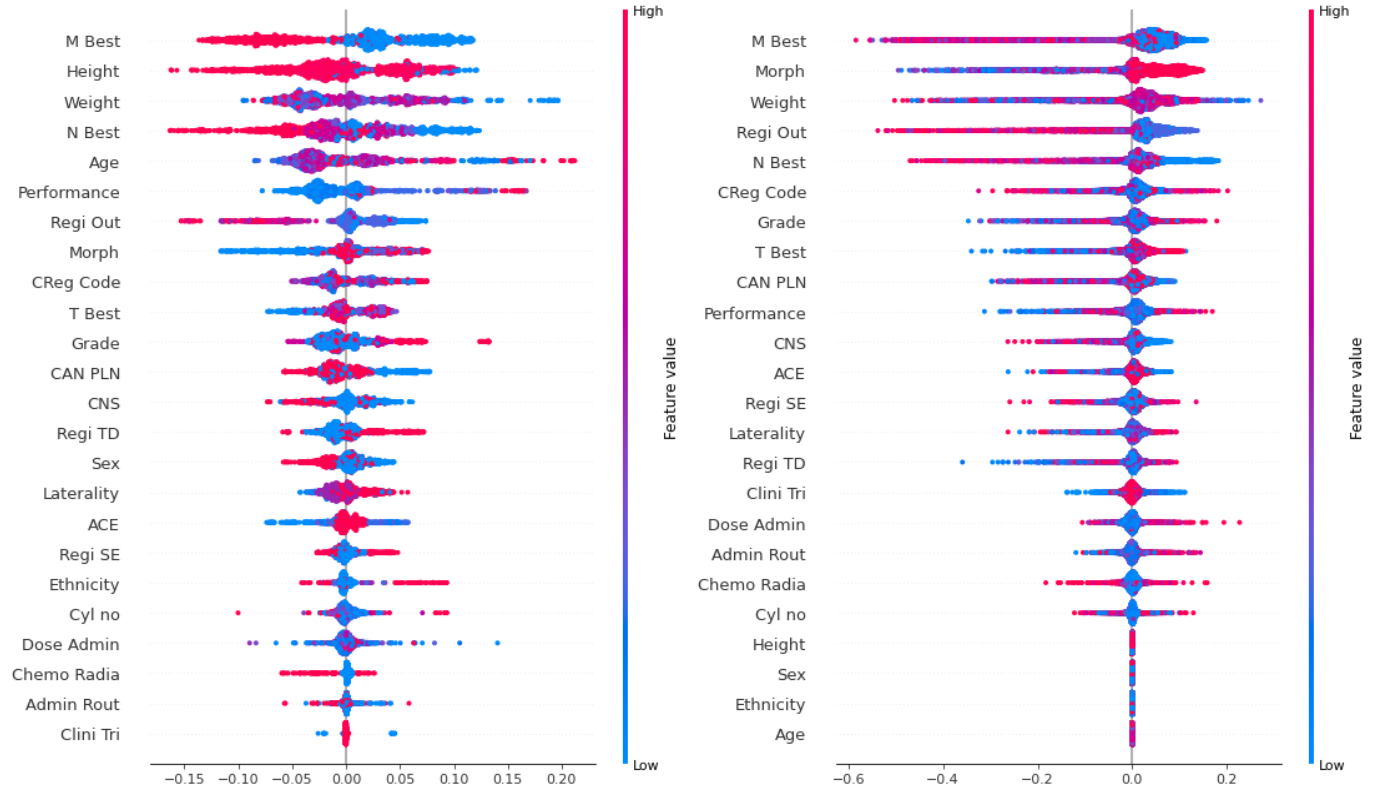
We split the dataset as 70% for training and 30% for testing, and use a random forest classifier. We achieve a high prediction accuracy of 94.4%. Since we aim to obtain a bird's-eye view of how control measures are affecting the disease, we focus on calculating global explanations. To this end, for each instance $\mathbf{x}$, we generate $D_{\mathbf{x}}$ with $K = 500$. π_x is the average distance between any two instances; $\omega_i = 1$. By following Algorithm 1, we obtain feature importance using CAFA. The global explanations are shown in Fig. 4, right-hand side, with SHAP results shown on the left.

The SHAP results at the left demonstrate that the number of daily cases and cumulative cases both have strong impact in predicting R_t. However, as both are uncontrollable, knowing that they have strong influence to the prediction does not help us understand the effectiveness of control measures. With CAFA (Fig. 4 right-hand side), the importance of all uncontrollable features are assigned to 0. Overall, we observe that:

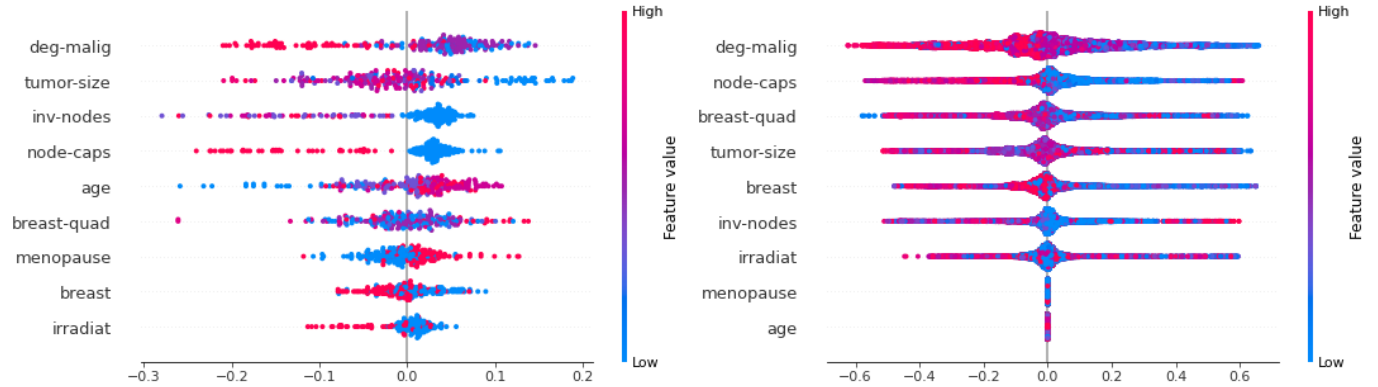
- SHAP considers *High Restriction on Cafes and Restaurants Access (CR_H)*, *High Restriction on Pubs and Bars Access (PB_H)*, *Number of Daily Infections (Cases)*, *Number of Daily Infections (Cases)*, *Medium Restriction on Pubs and Bars Access (PB_M)*, and *High Restriction Sport and Leisure Facilities (SL_H)* as the top five effective control measures; whereas
- CAFA considers *CR_H, PB_H, PB_M*, *Medium Restriction on Hospital and Nursing Home Visits (HV_M)* and *Medium Restriction on Cafes and Restaurants Access (CR_M)* as the top five effective control measures.

CAFA's results are in alignment with WHO's COVID-19 guideline stating the "Three C's" rule that the virus is more transmissible with (1) *Crowded places*; (2) *Close-contact settings*; and (3) *Confined and enclosed spaces with poor ventilation*.[7] Focusing on restricting access to cafes and restaurants as well as pubs and bars seem to be a very reasonable strategy in reducing the virus transmission, for the reason that these are the most prominent locations meeting the Three C's for most of the population.

[7]https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19-how-is-it-transmitted



(a) Global views of lung cancer cases in the Simulacrum (left: SHAP; right: CAFA). Uncontrollable features are: *Age*, *Ethnicity*, *Sex*, and *Height*.



(b) Global views of the UCI Breast Cancer dataset (left: SHAP; right: CAFA). Uncontrollable features are: *Age*, and *Menopause*.

Fig. 3: Global explanations calculated using SHAP and CAFA on the Simulacurm Lung Cancer dataset and the Breast Cancer dataset. Same as Fig. 2, we see that uncontrollable features in both datasets have importance 0; and CAFA produces similar results to SHAP for controllable features.

VI. Related Work

There has been some research conducted to extend feature attribution algorithms to achieve more meaningful explanations. For example, [1] extended the Kernel SHAP method to handle dependent features through different approaches to estimate the conditional distribution. Experiments over simulated datasets suggest that the dependencies between features are handled properly using proposed Shapley value approximations. An aggregation of the Shapley values of dependent features was also introduced to ease the interpretation and use of the Shapley values. In [20], a model-agnostic explanation approach 'anchors' was proposed based on if-then rules, which depends on input perturbation to approximate local explanations. Experimental results over classification, structured prediction, and text generation machine learning tasks demonstrated the usefulness of anchors. In [3], a variant of LIME for continuous data was proposed. Theoretical analysis was performed to derive explicit closed form expressions for the explanations output. It was also demonstrated that post hoc

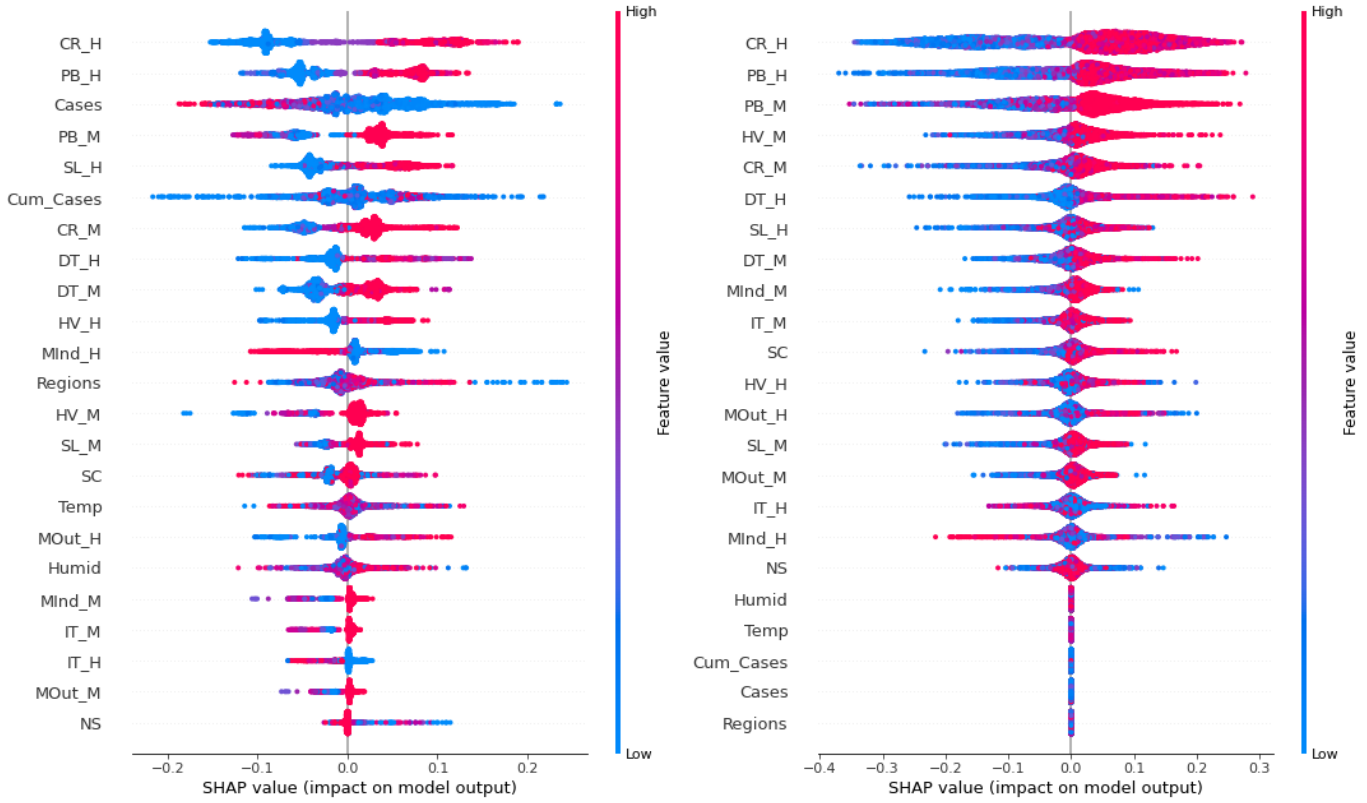


Fig. 4: Global views of the COVID dataset (SHAP Left; CAFA Right). Uncontrollable features are: *Humidity (Humid)*, *Temperature (Temp)*, *Cumulative Cases (Cum_cases)*, *Daily Infections (Cases)* and *Regions*.

explanation methods will converge to the same explanations when the number of perturbed samples used by these methods is large.

A large amount of effort has been put into studying the effectiveness of control measures for containing the COVID-19 pandemic. For example, in [13], the authors estimated the instantaneous reproduction number (R_t) of COVID-19 in four Chinese cities and ten provinces. They found that tough aggressive non-pharmaceutical interventions (e.g., city lockdown) had abated the first wave of COVID-19 outside of Hubei. The effect of physical distancing measures on the progression of the COVID-19 epidemic was explored in [18]. An extensive simulation based on an age-structured susceptible-exposed-infected-removed model [12] was carried out. The simulation results show that sustained physical distancing measures have a potential to reduce the magnitude of the epidemic peak of COVID-19. The impact of physical distancing measures in the UK was evaluated through comparing the contact patterns during the "lockdown" to patterns of social contact made before the epidemic [11]. It was found that the estimated change in reproduction number significantly decreased, suggesting that the physical distancing measures adopted by the UK public would probably lead to a decline in cases.

VII. Conclusion and Future Work

Feature attribution XAI algorithms tell users the relative contribution of a feature in a prediction, which can help users gain insight by shedding light onto the underlying patterns in large datasets. However, existing feature attribution algorithms treat controllable and uncontrollable features homogeneously, which may lead to incorrect estimation of the importance of controllable features. In this paper, we proposed CAFA to compute the relative importance of controllable factors through generating perturbed instances. Specifically, for each prediction instance, CAFA creates a dataset by selectively perturbing features representing controllable factors while leaving uncontrollable ones unchanged and then computing the global explanation on the generated dataset as the local explanation for the prediction instance.

We tested CAFA on two existing medical datasets, the lung cancer data from the Simulacrum dataset and the UCI breast cancer dataset. Experimental results show that with CAFA, although the prediction model is built over all features, the explanations of controllable features are not interfered with by the uncontrollable ones. We further applied CAFA in a case study on understanding the effectiveness of COVID-19 non-pharmaceutical control measures implemented in the UK during the period of January 2020 to February 2021. We found that restricting access to cafes and restaurants as well as pubs and bars are the most effective measures in containing the

disease, represented by reaching an R_t value smaller than 1.

Our work was carried out on classification tasks using a popular class of supervised machine learning techniques. In future, we want to further refine our CAFA technique. First, we plan to extend it to studies involving time series, and secondly, when sampling perturbed data points, we aim to take the density of the data into account and integrate a more fine-grained proximity measure. We expect this to extend the applicability of this technique and to increase its robustness.

REFERENCES

[1] Aas, K., Jullum, M., Loland, A.: Explaining individual predictions when features are dependent: More accurate approximations to shapley values. AI Journal 298 (2021).

[2] Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access 6 (2018).

[3] Agarwal, S., Jabbari, S., Agarwal, C., Upadhyay, S., Wu, S., Lakkaraju, H.: Towards the unification and robustness of perturbation and gradient based explanations. In: Proc. of ICML. vol. 139, pp. 110–119 (2021).

[4] Antoniadi, A. M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B. A., Mooney, C.: Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems : A systematic review. Applied Sciences (Switzerland) 11 (2021).

[5] Draper, L.: Breast cancer: Trends, risks, treatments, and effects. AAOHN Journal 54(10), 445–453 (2006).

[6] Duell, J., Fan, X., Burnett, B., Aarts, G., Zhou, S. M.: A comparison of explanations given by explainable artificial intelligence methods on analysing electronic health records. In: Proc. of BHI (2021).

[7] Fan, X., Liu, S., Chen, J., Williams, M., Henderson, C. T.: An investigation of covid-19 spreading factors with explainable ai techniques. International Journal of Information Technology 26 (2020).

[8] Fereshte, K and Percy., L.: Removing Spurious Features can Hurt Accuracy and Affect Groups Disproportionately. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, 196–205 (2021).

[9] Flaxman, S., Mishra, S., Gandy, A., Unwin, H. J. T., Mellan, T. A., Coupland, H., Whittaker, C., Zhu, H., Berah, T., Eaton, J. W., Monod, M., Perez-Guzman, P. N., Schmit, N., Cilloni, L., Ainslie, K. E. C., Baguelin, M., Boonyasiri, A., Boyd, O., Cattarino, L., Bhatt, S.: Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. Nature (2020).

[10] Honegger, M.: Shedding light on black box machine learning algorithms: Development of an axiomatic framework to assess the quality of methods that explain individual predictions. CoRR abs/1808.05054 (2018).

[11] Jarvis, C. I., van Zandvoort, K., Gimma, A., Prem, K., Auzenbergs, M., O'Reilly, K., Medley, G., Emery, J. C., Houben, R. M. G. J., Davies, N., Nightingale, E. S., Flasche, S., Jombart, T., Hellewell, J., Abbott, S., Munday, J. D., Bosse, N. I., Funk, S., Sun, F., Edmunds, W. J.: Quantifying the impact of physical distance measures on the transmission of covid-19 in the UK. BMC medicine 18, 1–10 (2020).

[12] Klepac, P., Caswell, H.: The stage-structured epidemic: linking disease and demography with a multi-state matrix approach model. T. Ecology 4(3), 301 (2011).

[13] Leung, K., Wu, J. T., Liu, D., Leung, G. M.: First-wave covid-19 transmissibility and severity in china outside hubei after control measures, and second-wave scenario planning: a modelling impact assessment. The Lancet (2020).

[14] Liu, S., Yalcin, O. M., Fu, H., Fan, X.: An investigation of the impact of covid-19 non-pharmaceutical interventions and economic support policies on foreign exchange markets with explainable AI techniques. In: Proc. of XAI-FIN21 (2021).

[15] Lundberg, S. M., Lee, S. I.: A unified approach to interpreting model predictions. In: Proc. of NIPS (2017).

[16] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S. I.: Explainable AI for trees:From local explanations to global understanding. CoRR abs/1905.04610 (2019).

[17] Molnar, C.: Interpretable Machine Learning, A Guide for Making Black Box Models Explainable (2019), https://christophm.github.io/interpretable-ml-book/.

[18] Prem, K., Liu, Y., Russell, T. W., Kucharski, A. J., Eggo, R. M., Davies, N., Flasche, S., Clifford, S., Pearson, C. A. B., Munday, J. D., Abbott, S., Gibbs, H., Rosello, A., Quilty, B. J., Jombart, T., Sun, F., Diamond, C., Gimma, A., Zandvoort, K., Klepac, P.: The effect of control strategies to reduce social mixing on outcomes of the covid-19 epidemic in Whuhan, CHINA: a modelling study. The Lancet Public Health (2020).

[19] Ribeiro, M. T., Singh, S., Guestrin, C.: "Why should I trust you?": Explaining the predictions of any classifier. In: Proc. of SIGKDD (2016).

[20] Ribeiro, M. T., Singh, S., Guestrin, C.: Anchors: High-precision model-agnostic explanations. In: Proc. of AAAI (2018).

[21] Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H.: Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In: Proc. of AIES. pp. 180–186. ACM (2020).

[22] Tjoa, E., Guan, C.: A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Trans. on Neural Networks and Learning Systems (2020).

[23] Yalcin, O., Fan, X., Liu, S.: Evaluating the correctness of explainable AI algorithms for classification. CoRR abs/2105.09740 (2021).

[24] Zliobaite, I., Custers, B.: Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. AI and Law 24 (2016).