ExMed: An AI Tool for Experimenting Explainable AI Techniques on Medical Data Analytics

Marcin Kapcia*, Hassan Eshkiki*, Jamie Duell*, Xiuyi Fan*, Shangming Zhou†, Benjamin Mora*
*Department of Computer Science, Swansea University, Swansea, United Kingdom
† Centre for Health Technology, Faculty of Health, University of Plymouth, Plymouth, United Kingdom

***Abstract*—The recent explosion of demand for Explainable AI (XAI) techniques has encouraged the development of various algorithms such as the Local Interpretable Model-Agnostic Explanations (LIME) and the SHapley Additive exPlanations ones (SHAP). Although these algorithms have been widely discussed by the AI community, their applications to wider domains are rare, potentially due to the lack of easy-to-use tools built around these methods. In this paper, we present ExMed, a tool that enables XAI data analytics for domain experts without requiring explicit programming skills. In particular, it supports data analytics with multiple feature attribution algorithms for explaining machine learning classifications and regressions. We illustrate its domain of applications on two real world medical case studies, with the first one analysing COVID-19 control measure effectiveness and the second one estimating lung cancer patient life expectancy from the artificial Simulacrum health dataset. We conclude that ExMed can provide researchers and domain experts with a tool that both concatenates flexibility and transferability of medical sub-domains and reveal deep insights from data.**

***Index Terms*—Explainable AI, Medical Data Analytics, Explainability, Interpretability, COVID-19, Cancer.**

I. Introduction

Explainable AI (XAI) has drawn tremendous attention in the recent years [1]. XAI systems not only aim to make intelligent decisions or accurate predictions, but also provide an insight into the process of AI decison making [2]. A goal of enabling explainability in AI systems "is to ensure algorithmic predictions and any input data triggering those predictions can be explained" [3]. In the context of Machine Learning (ML), XAI focuses on developing human-understandable prediction models producing explanations, along with predictions and model agnostic techniques that generate explanations to existing ML models.

From a data science perspective, equipped with its "explanation power", XAI is not only useful in bringing trust to AI models [1], but is also about providing deeper insights from data. By explaining why an AI model makes a certain prediction, one also gains knowledge about the underlying data used to build the model. Explanations thus can reveal previously unknown patterns in the data and may facilitate new discoveries. As a relative new field in AI, current development in XAI software is fragmented in the sense that various algorithm implementations are scattered in multiple libraries written in different programming languages. As various levels of coding are often required for using such libraries, they are predominately intended for data science developers rather than domain experts. The lack of easy-to-use XAI tools hinders further development of XAI and its applications in the wider context.

In this paper, we present ExMed, a self-contained XAI toolkit for domain experts. ExMed performs XAI analysis for prediction models. With its simple user interface, it supports both *global* explanations presenting patterns of the entire dataset and *instance* explanations that are local to individual predictions, for both classification and regression tasks. Although various XAI techniques have been proposed in recent years – e.g., a good overview of these techniques is presented in [4] – we focus on feature attribution explanation techniques [5] due to the transparency of their explanations, their computational effectiveness and general popularity.

To better illustrate our work, we present two real world case studies that demonstrate ExMed's functionalities. In case study I, a COVID-19 transmission study reveals how different COVID-19 control measures were used and impacted the virus transmission rates. In case study II, we examine lung cancer patient life expectancy using the Simulacrum dataset.[1] Through the two case studies, we illustrate how ExMed can be used for making predictions and generating explanations.

The rest of this paper is organised as follows. Section II reviews some research in the area of XAI and medicine. We present background on feature attribution XAI algorithms in Section III before introducing ExMed in Section IV. Section V and VI discuss the two application cases studies, respectively.

II. Related Work

The use of Machine Learning (ML) has become more prominent in several areas of healthcare, such as diabetes, arthritis, cancer [6]–[8], with varying input formats ranging from tabular data in stored in relational databases to large scale image datasets [9]. Stemming from the involvement of data sensitivity in the medical domain is the necessity of gaining human trust towards ML application [10]. Thus, we see a recent surge in the production of interpretable results using

Jamie Duell is supported by the UKRI AIMLAC CDT, funded by grant EP/S023992/1. Xiuyi Fan is supported with a funding contribution from the Welsh Government Office for Science, Ser Cymru III programme – Tackling Covid-19.

[1] https://simulacrum.healthdatainsight.org.uk/

state-of-the-art models such as Local Interpretable Model-Agnostic (LIME) and SHapley Additive exPlanations (SHAP) to supplement the outputs provided by black-box algorithms, with much work showing the intent of XAI expansion through new prediction model architectures [11], [12].

XAI aims to improve the usability of AI by providing justifications behind a given ML prediction [13]. It is a field of exploration by creating a framework to cater for usability adhering to each step of an ML pipeline, including data manipulation, pre-processing methods and explainability. Whilst yielding the same prediction performance, the explainable aspect helps reduce bias and promote fairness of the prediction model. This interpretable aspect in junction with human input can then create an optimal pipeline of fairness whilst obtaining best results by providing various reasons that can provide an insight into the decision making process.

Currently, many ML pipelines are especially designed for an identified problem with a high degree of specificity. In other words, there is the necessity of alteration when it comes to rebuilding a pipeline for new data. Therefore, without a basic knowledge of ML, accessibility to ML applications can be limited if data interpretation is not outsourced to ML domain experts. In addition, there is a demand for more human-input and interaction within the AI model to support explainability [14], [15], with the exploration of explainable architectures having received recently some important development from web-based interfaces supporting image segmentation with a readily available interface [16].

A few open-source applications have been created to ease the application of AI to datasets, e.g., [17]–[20]. Much data in biology is stored as images and Fiji [17] is an example of an open-source tool designed from biological-image analyses that aims to prototype algorithms for image-processing. WEKA [18], [19] is another workbench designed to combine different ML libraries for supporting various analysis through a graphic user interface. WEKA creates fast access to the information within the datasets, allowing selection of areas of interest. WEKA however does not provide both merging and concatenation of datasets, which is often required when introducing new medical data. None of these tools has focused on explanations.

III. Feature Attribution Algorithms

Before introducing ExMed, this section present first the two main feature attribution XAI algorithms used in this paper. Feature attribution algorithms compute explanations for predictions of data instances in the form of "feature weights". For some prediction model $f \in \mathcal{F}$ with $\mathcal{F}$ a set of models that takes input $\mathbf{x} \in \mathbb{R}^n$ and produces output $\mathbf{y} \in \mathbb{R}^n$, a feature attribution algorithm is a function $\Pi : \mathcal{F} \times \mathbb{R}^n \mapsto \mathbb{R}^n$. In other words, given a prediction model f, for each input $\mathbf{x} = \langle x_1, \ldots, x_n \rangle$, a feature attribution algorithm computes an explanation $\boldsymbol{\phi} = \langle \phi_1, \ldots, \phi_n \rangle$ for $\mathbf{y} = f(\mathbf{x})$. Each ϕ_i in $\boldsymbol{\phi}$ represents the "weight" of x_i in the prediction.

Local Interpretable Model-Agnostic Explanations (LIME) is a feature attribution algorithm that explains individual predictions of black-box machine learning models. LIME is model-agnostic, so it is applicable to any classifier [21]. LIME tests how predictions change when a user perturbs the input data. Given a black-box model f and a data instance $\mathbf{x}$, LIME generates a set of perturbed instances around $\mathbf{x}$ and compute their corresponding predictions with f in order to explain the prediction of $\mathbf{x}$ made with f. LIME then creates a linear model g from some interpretable model class $\mathcal{G}$ as a local surrogate to f based on generated data, such that:

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

where L is an error function with respective inputs f, g, π_x as the size of the locality around x and $\Omega(g)$ as the complexity of g. Using parameters of g, LIME returns an explanation as a list of feature contributions to the prediction $f(\mathbf{x})$.

SHapley Additive exPlanations (SHAP) [5] is based on the coalitional game theory concept *Shapley value*, assigned to each feature of a data instance. A Shapley value is defined to answer the question: "What is the fairest way for a coalition to divide its payout among players"? It assumes that payouts should be assigned to players in a game depending on their contribution towards total payout. In machine learning terms, the features are the "player"'s characteristics and the "total payout" is the value that needs to be predicted. In this setting, the Shapley value of a feature represents its contribution to the prediction and thus explains the prediction. For a data instance $\mathbf{x}$, SHAP computes the marginal contribution of each feature to the prediction of $\mathbf{x}$ as its feature weight.

Specifically, given an explanation model g, for an input $\mathbf{x}$ with n features, there is a corresponding $z \in \{0, 1\}^n$ 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 $\phi_j (j > 0)$ is the Shapley value of feature j and ϕ_0 is the "average" prediction when none of the feature in $\mathbf{x}$ is present, both computed with the original model f. The idea is that if $z_j = 0$, the corresponding feature value is absent in $\mathbf{x}$. Otherwise, the corresponding feature value is present in $\mathbf{x}$.

In this work, we use the tree-based model, TreeSHAP, for estimating Shapley values of features introduced in [22], as which is shown to be a superior method than the KernelSHAP introduced in [5].

IV. ExMed Workflow

When sending medical data to an ML model, complications unhinging the ubiquity between data and the applied model can arise from issues with the records such as noise and human-induced errors. Therefore, providing clean datasets would provide the best possible results. Moreover, as one of the leading challenges in medical data analysis is to aggregate data from multiple data sources for performing joint analysis [23], it is crucial for medical data analytic platforms to support

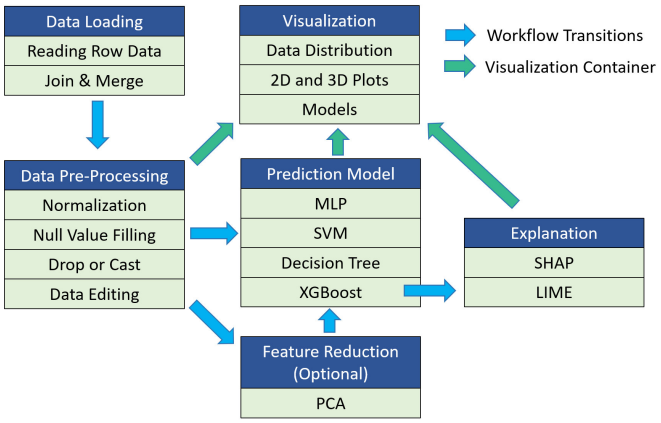


Fig. 1: ExMed Activities. ExMed provides the user with a sequence of simple actions, including loading, merging and editing data, and creating prediction as well as explanation models. Various visualisation techniques are supported in several stages of this pipeline.

such functionalities. Our new application ExMed [24] addresses both challenges and makes the integration of data preprocessing tools easy in order to minimise error and increase the baseline ML performance of the model. Overall, ExMed provides explanations through both data exploration and data processing using state of the art methods to unravel the black-box models applied to medical data that are potentially collected from multiple sources.

ExMed's main functionalities, architecture and selected interface illustrations are shown in Figure 1, 2 and 3, respectively. ExMed implements a wide set of tools to load, process, predict, interpret and explain data. Its back end design is modular so that more tools can be easily added at a later stage. ExMed can accept most common data files as input (e.g. Excel, CSV, or SAS, and XPT files) with the possibility for easy integration of new file types. Input data can be then combined through classic database join operators, whether or not a common key exists. This gives users the potential to create larger datasets from different file types - potentially collected from different sources - rapidly. Cells, rows, columns and data types can be edited by the user directly within ExMed, allowing greater freedom for data manipulation and quality checks. Data validation is supported by various visualisation tools included with the interface. These tools can represent data trends in many ways (see Fig 4 for a few examples) to provide fast data insight to users and can be applied to either the entire dataset or just part of it.

Creating a model can easily be done by selecting a target label (i.e., column) in the interface. Non-categorical columns selected from inference can be edited and transformed into a category (e.g., using a thresholding operator, Fig. 3) prior to creating the model. Once data has been finalised and validated, and a target label has been created, a range of machine models can then be applied, including SVM, Random Forest Classifier, MLP Regression and XGBoost. There is also an option to apply dimensionality reduction by preprocessing data with an automated Principal Component Analysis (PCA) process.

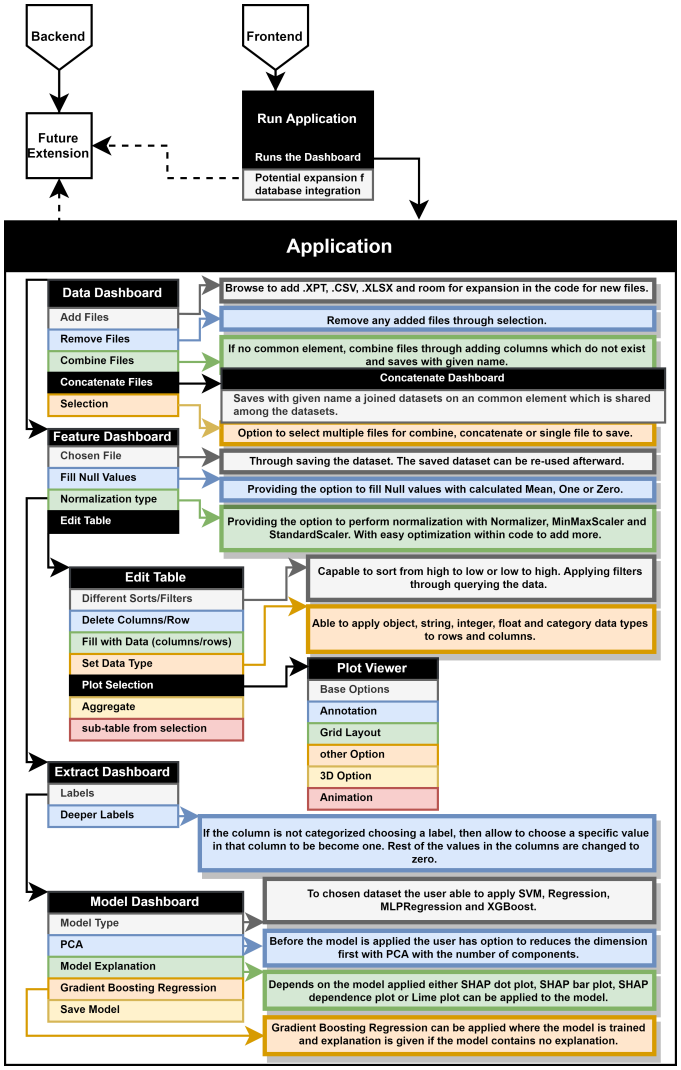
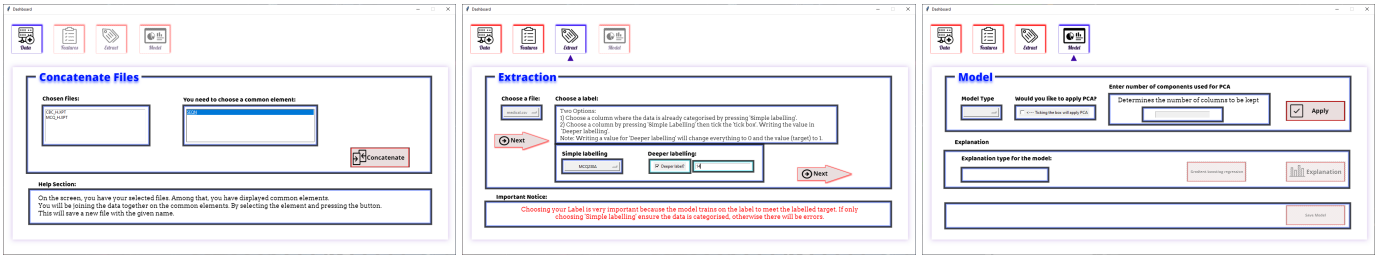


Fig. 2: ExMed Operation Overview. This figure shows the flow of ExMed operations, along with the key features available in the interface. Once the "Application" is running, the first window "Data Dashboard" is shown. Black arrows denote event-driven actions that take the user to the next window in chronological order. Colours highlight the Key features for each window, along with a short description provided for each feature. The indentation of boxes represents a dependency between windows. For instance, the 'Feature Dashboard' window can lead to the 'Edit Table' window, which subsequently can open the 'Plot Viewer' window. The dotted line represents a database extension that is to be added in the future.

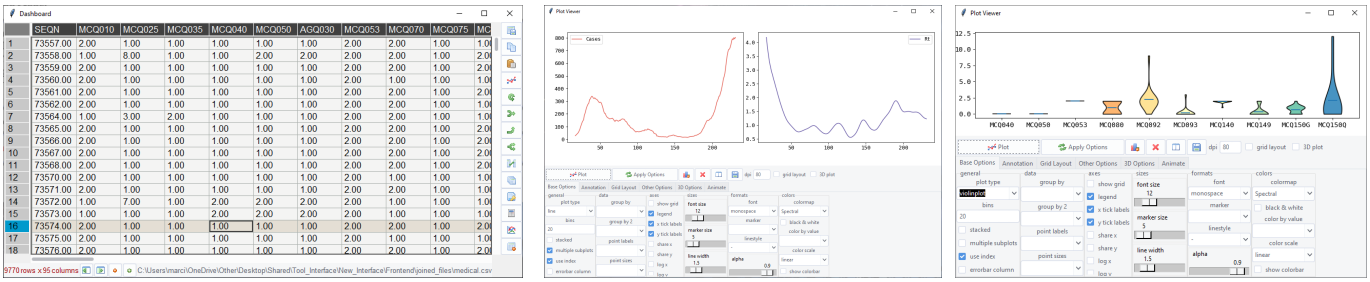
Moreover, the result of the PCA can be visualised in 2D or 3D from the two or three largest eigen vectors respectively.

To interpret data, individual models have their own functions to offer specific explanations. SHAP dot plots, SHAP bar plots, SHAP dependence plots and LIME plots can be used for this purpose. This will show different ways of explaining the reasoning behind the results. We explore explanations and ExMed capabilities on two case studies in sections V and VI. LIME and SHAP – introduced in section III – adhere to



(a) File Joining and Concatenation (b) Data Pre-Processing (c) Model Selection and Construction

Fig. 3: ExMed interface for some of the main activities as describe in Fig. 1. (a) Data from various supported file types is loaded, with the option to combine this data with other datasets. (b) Data is optionally pre-processed with some of the plugins available. (c) A model is created, with the option of reducing the number of features with a PCA algorithm and explanation generation with SHAP and LIME.



(a) Data Editor (b) Data Visualiser - Line Plot (c) Data Visualiser - Violin Plot

Fig. 4: Data exploration tools in ExMed. (a) is the Data Editor that supports standard data editing functions. (b) and (c) are the Data Visualiser that supports different plots types such as Line, Scatter, Bar, Histogram, Violin Plots and Pie chart. For each plot type, various customisation options are implemented, including changing the axes, layout, and adding texts.

ML local interpretability requirements for patient instances; expressed as a necessity from clinicians [10], whilst also producing global explanations. To invoke trust, we provide explanations from both LIME and SHAP as both models see a lack of ubiquity in feature priority, but may still provide valuable insight into the data as these methods still often see the same trend in feature attribution [25]. Also, feature attribution algorithms allow for a better understanding of data, as we are able to visualise bias, error, and gain insight into patient instances.

V. Case Study I: COVID-19 Control Measures

In this case study, we demonstrate how ExMed can be used in investigating relative effectiveness of COVID control measures used in the UK.

From the Public Health England website[2], we collect daily infection numbers reported across 9 regions in England *East Midlands, East of England, London, North East, North West, South East, South West, West Midlands*, and *Yorkshire and the Humber*, as well as and the other three nations in the UK: *Wales, Scotland and Northern Ireland*. Non-pharmaceutical control measure data were collected based on UK's COVID policies as summarised in Table I. Data are collected from from various sources including the Wikipedia and major news agencies such as BBC. Control Measures are coded based on

TABLE I: Non-pharmaceutical COVID Control Measures.

Control Measures	Type
Meeting Friends / Family (Indoor)	Categorical
Meeting Friends / Family (Outdoor)	Categorical
Domestic Travel Control	Categorical
International Travel Control	Categorical
Cafes and Restaurants Control	Categorical
Pubs and Bars Control	Categorical
Sports and Leisure Closure	Categorical
Hospitals / Care and Nursing Home Visits	Categorical
Non-Essential Shops Closure	Binary
School Closure	Binary

the level of severity ("High", "Moderate" or "Low") for all control measures excluding Non-essential shops and School closures, which are coded as binary choices ("Open" and "Closed"). Temperature and humidity data obtained from the weather website Raspisaniye Pogodi Ltd[3] were also included. This represents a total of 4,257 data points that were collected between February 2020 and February 2021.

We study the effectiveness of control measures by observing their impacts to the virus transmission rate R_t. Specifically, from daily infection numbers, we estimate R_t using the method reported in [26], [27]. R_t is one of the most important quantities used to measure the epidemic spread. If $R_t > 1$, then the epidemic is expanding at time t, whereas if $R_t < 1$,

[2]https://www.gov.uk/government/organisations/public-health-england

[3]https://rp5.ru/Weather_in_the_world

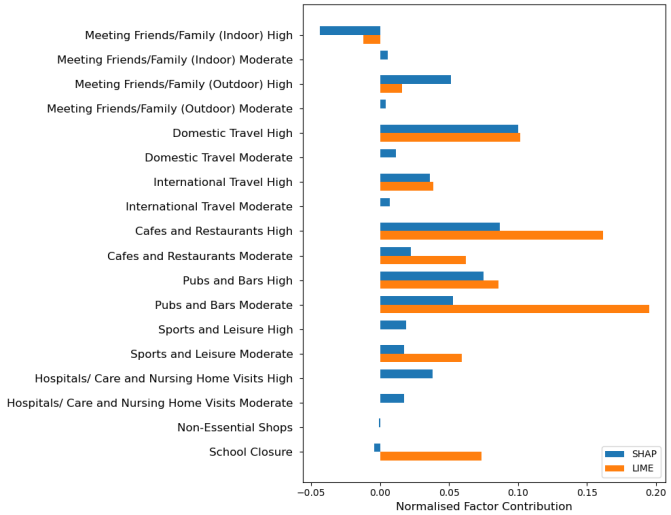


Fig. 5: Example of an Explanation computed with SHAP and LIME. For this instance, both explainers consider top measures contributing to this prediction being *Domestic Travel*, *Cafes and Restaurants Closure* and *Pubs and Bars Closure*.

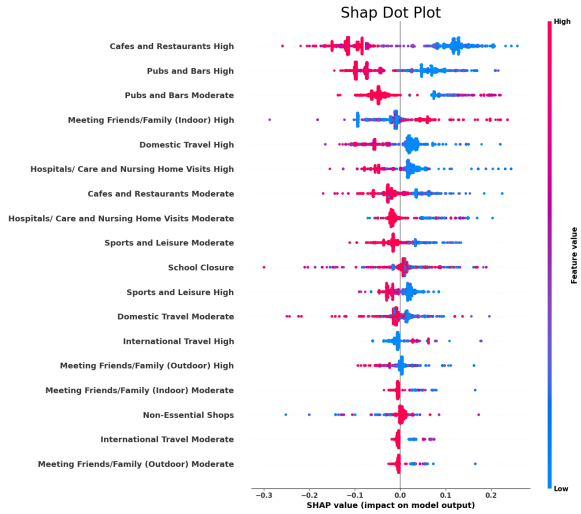


Fig. 6: Global explanations generated using SHAP on our COVID dataset for the prediction whether $R_t \geq 1$. We see that closing down cafes and restaurants as well as pubs and bars are the most effective control measures. When their feature values are high (red), they have string negative impact to the prediction; whereas when their feature values are low (blue), they have strong negative impact to the prediction.

then it is shrinking at time t. A *serial interval distribution*, which is a Gamma distribution $\mathrm{g}(\tau)$ with mean 7 and standard deviation 4.5, is used to model the time between a person getting infected and he/she subsequently infecting another person on day τ. The number of new infections c_t on a day t is computed as:

$$c_t = R_t \sum_{\tau=0}^{t-1} c_\tau g_{t-\tau}, \tag{3}$$

where c_τ is the number of new infections on day τ,

$$g_1 = \int_{\tau=0}^{1.5} \mathrm{g}(\tau) d\tau,$$

and for $s = 2, 3, \ldots$,

$$g_s = \int_{\tau=s-0.5}^{s+0.5} \mathrm{g}(\tau) d\tau.$$

From Equation 3, we have:

$$R_t = \frac{c_t}{\sum_{\tau=0}^{t-1} c_\tau g_{t-\tau}} \tag{4}$$

For $x = t$ and τ, c_x is the difference between the confirmed case on day x and the confirmed case on day $x-1$, which is available from the dataset directly.

Using this data, we pose a simple classification question:

> *Given the infection number and control measures implemented on a day t, can we predict* $R_t \geq 1$?

As control measures take time to affect the infection rate, we expand the dataset to include the duration of control measure implementation for all control measures. For example, *"Meeting Indoors (High) = 2"* means that *"it is the second week that meeting indoors has been banned completely"*. Similarly, *"International Travel (Low) = 0"* means that *"there is no restriction implemented on international travel"*. We also drop instances before March 15, 2020 across all 12 regions and nations in our dataset due to the low number of infections.[4] In this way, we form a data file with 18 features and 3,937 instances with 1,550 positive ones.

TABLE II: Prediction performance on the COVID dataset with four different classifiers.

Classifier	MLP	Random Forest	SVM	XGBoost
Precision	0.87	0.90	0.87	0.87
Recall	0.79	0.84	0.78	0.79
F1-score	0.83	0.87	0.83	0.84

The classification results are summarised in Table V. We can see that all four classifiers are able to achieve good performance on this dataset with a 70/30 training/testing split. As an illustration, for a prediction query instance such that:

- all control measures shown in Table I except *International Travel (IT)* and *Hospital / Care and Nursing Home Visits (HCNHV)* are implemented for more than 35 days at the level *High*;
- *IT* has been implemented for more than 35 days at the level *Moderate*; and
- *HCNHV* implemented for 20-25 days at the level *High*.

Using Random Forest as our prediction model, it correctly predicts that $R_t < 1$; and SHAP and LIME explanations are shown in Figure 5. We see that SHAP and LIME produce similar explanations for the instance. In addition to local explanations, ExMed can also use SHAP to compute global explanations for the entire dataset - describing the "trend" of all instances - as illustrated in Figure 6. We observe that control measures *Cafes and Restaurants Control* and *Pubs and*

[4]As can be seen from Equation 4, when c_x is small, R_t can flatten in a unrealistically large range and generate noises in the dataset.

TABLE III: Each patient is described with 20 features.

Feature	Value	Feature	Value
ACE	2.0	T Best	0.0
Sex	M	M Best	3.0
CNS	9.0	N Best	4.0
Age	68	Cycle Number	0.0
Grade	0.0	Ethnicity	1.0
Height	1.6	Cancer Plan	1.0
Weight	75.6	CReg Code	4.0
Morph	8041.0	Chemo Radiation	N
Laterality	901.0	Regimen Time Delay	N
Performance	1.0	Regimen Stopped Early	N

Bars Control have the most influence to predictions made with this dataset, this can be interpreted as:

> *From February 2020 to February 2021, the most effective non-pharmaceutical COVID control measures implemented in the UK are closing cafes and restaurants as well as pubs and bars.*

VI. Case Study II: Lung Cancer Life Expectancy

Our second case study investigates the application of XAI to electronic patient records for cancer research instead of using public health epidemiology data in order to emphasise the transferability provided by ExMed. Especially, we use artificial data from the Simulacrum[5], a synthetic dataset developed by Health Data Insight CiC and derived from anonymous cancer data provided by the National Cancer Registration and Analysis Service[6], which is part of Public Health England. This dataset contains 1,322,100 cancer patient instances.

We first isolate a cohort of interest, opting for lung cancer patients as they represent a large portion of cancer-based deaths [28]. With lung cancer patients, we define the medical question as a prediction of patient survival time, and pose the following multi-class classification question:

> *Given a set of features for a patient, what will be the predicted survival time for the patient? Under six months, six to twelve months, or more than twelve months?*

To study this, we first identify the subset of lung cancer patients in the Simulacrum with an ICD-10 code "C34" *Malignant neoplasm of bronchus and lung* and a deceased status, and includes 108,282 patients in total. We removed records from the original dataset with obvious errors and included only patients with a vital status date posterior to the diagnosis date.

A major challenge in medical data analytic, as exemplified in the Simulacrum, is missing or incomplete patient records. This results in a large number of "null" entries in the dataset. To address this, we identify a smaller cohort of patients such that each patient contains 20 features, with each patient instance only able to contain a maximum of one "null" value. This explicit filtering isolates a cohort of 2,260 patients. This also provides a well-balanced dataset with each group containing a similar amount of patients as shown in Table IV.

[5]https://simulacrum.healthdatainsight.org.uk/

[6]http://www.ncin.org.uk/

TABLE IV: Survival Time Feature Value Count

Survival Time	**Value Count**
Greater than 1 Year	842
Between 6 Months and 1 Year	748
Less than 6 Months	670

TABLE V: Prediction performance on the Lung Cancer dataset with four different classifiers.

Classifier	MLP	Random Forest	SVM	XGBoost
Precision	0.86	0.90	0.77	0.69
Recall	0.76	0.90	0.98	0.66
F1-score	0.81	0.90	0.86	0.67

We first provide a local explanation example using both SHAP and LIME for a patient instance as shown in Table III. We observe that both explainers give similar explanations as shown in Fig 7. Using the entire dataset, we produce a global explanation determining feature importance towards each output class in Fig 8 (a). We then provide granularity to feature value importance towards each class with Fig 8 (b) - (d). We interpret these results as:

> *Cancer grades, BMI, age, patient performance and the absence of distant metastatic spread are key indicators for estimating patients survival time.*

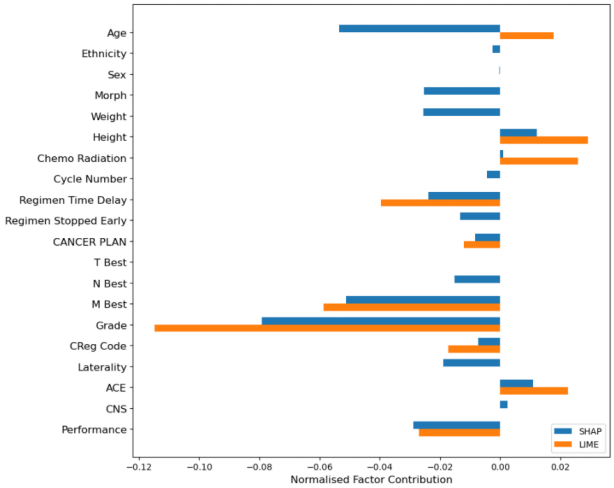
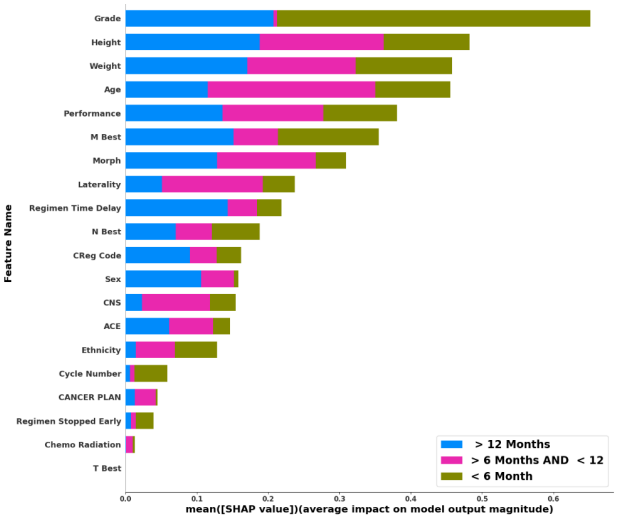


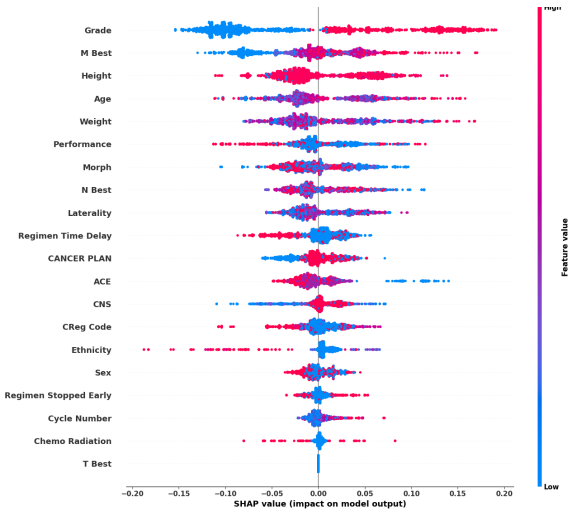
Fig. 7: Local explanation on the Lung Cancer life expectancy data set for a patient instance. We see that the most impactful features amongst SHAP and LIME are the same: "Grade" *How the cancer cells act; the higher the grade the less normality the cell resembles and it may act more aggressive* and "M Best" *Presence or Absence of Distant Metastatic Spread*, followed by a disagreement on age attribution.

VII. Conclusion

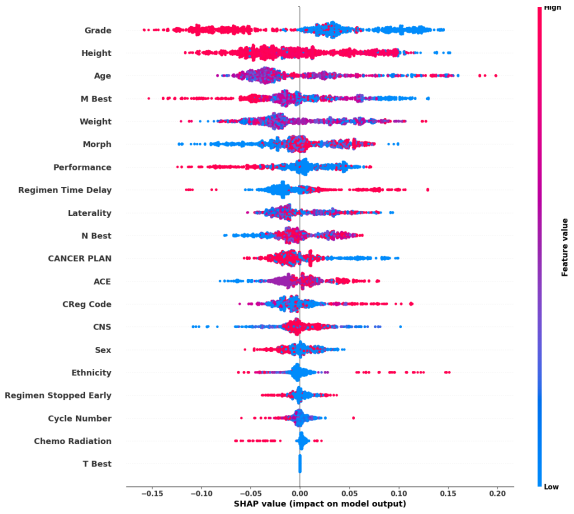
In this paper, we present ExMed, a self-contained software package that enables Explainable AI data analysis for medical domain experts without the need for explicit programming. With the development of ExMed, we aim to provide a



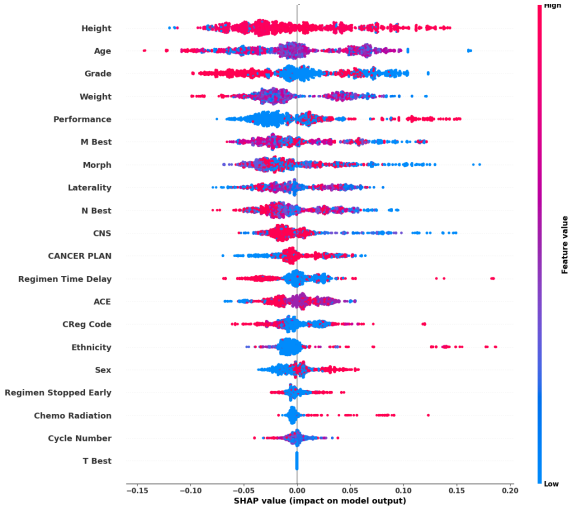
(a) We see that the largest impact towards the survival boundaries *greater than* 1 year and *less than* 6 months is the cancer grading - having direct impact on the longest and least time survived. This, followed by an associative relationship between height, weight and the patient age determinants of body mass index (BMI), having high attribution towards each class. This, then followed by cancer specific traits such as "M Best" and laterality of the tumour.



(b) Global explanation for feature attribution measured against the class *Survival time of less than 6 months*, where we see the cancer grade of higher value - indicative of cell abnormality and more aggressive, followed by "M Best" *Presence or Absence of Distant Metastatic Spread*, with the associative BMI attributes "height", "age" and "weight" following this.



(c) Global explanation for feature attribution measured against the class *Survival time of greater than 12 months*, we see an inverse plot of cancer grade to that shown in Fig.8 (a), such that a lower grade and what seems to be a better controlled BMI and a lower "M Best" contributing to a longer survival time.



(d) Global explanation for feature attribution measured against the class *Survival time between 6 and 12 months*, we see that a controlled BMI and lower cancer grade are attributive to this survival boundary, whilst the distributive "M Best", performance and cancer grade containing high values in both positive and negative impacts on the model are likely the reason for the central survival boundary.

Fig. 8: Global explanations on the lung cancer life expectancy data set.

tool that both concatenates the flexibility of medical sub-domain transferability and obtain an essence of trust through explainability using XAI methods. ExMed accepts multiple data input types and supports several standard pre-processing operations. It employs a number of different prediction models and visualisation techniques, while implementing two popular feature attribution XAI algorithms.

We have experimented ExMed with two real-world case studies in the domains of epidemiology in public health and cancer research with electronic patient records. In particular, we have studied effectiveness of COVID control measures in the UK using data from March 2020 to January 2021 and the life expectancy of lung cancer patients using the Simulacrum dataset. From the COVID case study, we observed that closing down cafes and restaurants as well as pubs and bars had the most impact in reducing the virus transmission rate. From the cancer case study, we saw that cancer grades, BMI, age and M Best are amongst the most influential factors for survival.

In the future, we plan to (1) experiment ExMed with healthcare professionals and conduct user studies to evaluate effectiveness of various XAI approaches; (2) further expand the functionality of ExMed and explore features such as parameter tuning; (3) incoporating additional missing value imputation techniques such as MICE [29] and SICE [30]; and (4) introducing additional XAI techniques such as Anchors [31] in ExMed.

REFERENCES

[1] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, 2019.

[2] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[3] D. Carvalho, E. Pereira, and J. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, p. 832, 07 2019.

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

[5] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 4765–4774.

[6] J. Li, J. Huang, L. Zheng, and X. Li, "Application of artificial intelligence in diabetes education and management: Present status and promising prospect," *Frontiers in Public Health*, vol. 8, p. 173, 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpubh.2020.00173

[7] N. Bhargava, R. Purohit, S. Sharma, and A. Kumar, "Prediction of arthritis using classification and regression tree algorithm," in *2017 2nd International Conference on Communication and Electronics Systems (ICCES)*, 2017, pp. 606–610.

[8] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," *Procedia Computer Science*, vol. 83, pp. 1064–1069, 2016, the 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016) / The 6th International Conference on Sustainable Energy Information Technology (SEIT-2016) / Affiliated Workshops. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050916302575

[9] Y. Shu, J. Zhang, B. Xiao, X. Luan, L. Liu, and C. Hu, "Aft-net: Active fusion-transduction for multi-stream medical image segmentation," in *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, 2020, pp. 753–760.

[10] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What clinicians want: Contextualizing explainable machine learning for clinical end use," in *Proceedings of the 4th Machine Learning for Healthcare Conference*, ser. Proceedings of Machine Learning Research, F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, and J. Wiens, Eds., vol. 106. Ann Arbor, Michigan: PMLR, 09–10 Aug 2019, pp. 359–380. [Online]. Available: http://proceedings.mlr.press/v106/tonekaboni19a.html

[11] H. F. da Cruz, B. Pfahringer, T. Martensen, F. Schneider, A. Meyer, E. Böttinger, and M.-P. Schapranow, "Using interpretability approaches to update "black-box" clinical prediction models: an external validation study in nephrology," *Artificial Intelligence in Medicine*, vol. 111, p. 101982, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365720312471

[12] P. Mróz, A. Quemy, M. Ślażyński, K. Kluza, and P. Jemioło, "Gbex - towards graph-based explanations," in *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, 2020, pp. 112–117.

[13] T. Chen, E. Keravnou-Papailiou, and G. Antoniou, "Medical analytics for healthcare intelligence – recent advances and future directions," *Artificial Intelligence in Medicine*, vol. 112, pp. 1–5, Feb. 2021.

[14] A. Holzinger, "Interactive machine learning for health informatics: when do we need the human-in-the-loop?" *Brain Informatics*, vol. 3, pp. 119 – 131, 2016.

[15] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable AI systems for the medical domain?" *CoRR*, vol. abs/1712.09923, 2017.

[16] M. Pennisi, I. Kavasidis, C. Spampinato, V. Schinina, S. Palazzo, F. P. Salanitri, G. Bellitto, F. Rundo, M. Aldinucci, M. Cristofaro, P. Campioni, E. Pianura, F. Di Stefano, A. Petrone, F. Albarello, G. Ippolito, S. Cuzzocrea, and S. Conoci, "An explainable ai system for automated covid-19 assessment and lesion categorization from ct-scans," *Artificial Intelligence in Medicine*, p. 102114, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S093336572100107X

[17] J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.-Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, and A. Cardona, "Fiji: an open-source platform for biological-image analysis," *Nature Methods*, vol. 9, no. 7, pp. 676–682, 2012.

[18] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[19] M. S. Hammoodi, H. A. A. Essa, and W. A. Hanon, "The Waikato Open Source Frameworks (WEKA and MOA) for Machine Learning Techniques," *Journal of Physics: Conference Series*, vol. 1804, no. 1, p. 012133, 2021.

[20] T. Meinl, "What's new in KNIME?" *Journal of Cheminformatics*, vol. 4, no. S1, 2012.

[21] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. ACM, 2016, pp. 1135–1144.

[22] S. M. Lundberg, G. G. Erion, and S. Lee, "Consistent individualized feature attribution for tree ensembles," *CoRR*, vol. abs/1802.03888, 2018.

[23] S. S. Dhruva, J. S. Ross, J. G. Akar, B. Caldwell, K. Childers, W. Chow, L. Ciaccio, P. Coplan, J. Dong, H. J. Dykhoff, S. Johnston, T. Kellogg, C. Long, P. A. Noseworthy, K. Roberts, A. Saha, A. Yoo, and N. D. Shah, "Aggregating multiple real-world data sources using a patient-centered health-data-sharing platform," *NPJ Digit Med*, vol. 3, p. 60, 2020.

[24] M. Kapcia, "ExMed on Github," https://github.com/983046/ExMed, 2021, [Online; accessed 19-July-2021].

[25] J. A. Duell, X. Fan, B. Burnett, G. Aarts, and S. Zhou, "A comparison of explanations given by explainable artificial intelligence methods on analysing electronic health records," in *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI) (IEEE BHI 2021)*, Athens, Greece, Jul. 2021.

[26] S. Flaxman, S. Mishra, A. Gandy, H. Unwin, H. Coupland, T. Mellan, H. Zhu, T. Berah, J. Eaton, P. Perez Guzman *et al.*, "Report 13: Estimating the number of infections and the impact of non-pharmaceutical interventions on covid-19 in 11 european countries," Imperial College London, Tech. Rep., 2020.

[27] J. T. Wu, K. Leung, M. Bushman, N. Kishore, R. Niehus, P. M. de Salazar, B. J. Cowling, M. Lipsitch, and G. M. Leung, "Estimating clinical severity of covid-19 from the transmission dynamics in wuhan, china," *Nature Medicine*, pp. 1–5, 2020.

[28] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, pp. 209–249, 2021.

[29] S. V. Buuren, J. P. Brand, C. G. Groothuis-Oudshoorn, and D. B. Rubin, "Fully conditional specification in multivariate imputation," *Journal of Statistical Computation and Simulation*, vol. 76, no. 12, pp. 1049–1064, 2006.

[30] S. I. Khan and A. S. M. L. Hoque, "SICE: an improved missing data imputation technique," *J. Big Data*, vol. 7, no. 1, p. 37, 2020. [Online]. Available: https://doi.org/10.1186/s40537-020-00313-w

[31] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proc. of AAAI*. AAAI Press, 2018, pp. 1527–1535.