A Framework for Data-Driven Solutions with COVID-19 Illustrations

MWITONDI, Kassim S. and SAID, Raed A.

SPECIAL COLLECTION: MULTIDISCIPLINARY DATA ACTIVITIES BRIDGING THE RESEARCH COMMUNITY AND SOCIETY

RESEARCH PAPER

A Framework for Data-Driven Solutions with COVID-19 Illustrations

KASSIM S. MWITONDI

RAED A. SAID

*Author affiliations can be found in the back matter of this article

ABSTRACT

Data-driven solutions have long been keenly sought after as tools for driving the world's fast changing business environment, with business leaders seeking to enhance decision making processes within their organisations. In the current era of Big Data, applications of data tools in addressing global, regional and national challenges have steadily grown in almost all fields across the globe. However, working in silos has continued to impede research progress, creating knowledge gaps and challenges across geographical borders, legislations, sectors and fields. There are many examples of the challenges the world faces in tackling global issues, including the complex interactions of the 17 Sustainable Development Goals (SDG) and the spatio-temporal variations of the impact of the on-going COVID-19 pandemic. Both challenges can be seen as non-orthogonal, strongly correlated and requiring an interdisciplinary approach to address. We present a generic framework for filling such gaps, based on two data-driven algorithms that combine data, machine learning and interdisciplinarity to bridge societal knowledge gaps. The novelty of the algorithms derives from their robust built-in mechanics for handling data randomness. Animation applications on structured COVID-19 related data obtained from the European Centre for Disease Prevention and Control (ECDC) and the UK Office of National Statistics exhibit great potentials for decision-support systems. Predictive findings are based on unstructured data-a large COVID-19 X-Ray data, 3181 image files, obtained from GitHub and Kaggle. Our results exhibit consistent performance across samples, resonating with cross-disciplinary discussions on novel paths for data-driven interdisciplinary research.

CORRESPONDING AUTHOR:
Kassim S. Mwitondi
Sheffield Hallam University, College of Business, Technology & Engineering, GB
k.mwitondi@shu.ac.uk

KEYWORDS:
Animation; Big Data; Convolutional Neural Networks; COVID-19; Data Science; Sustainable Development Goals (SDG)

TO CITE THIS ARTICLE:
Mwitondi, KS and Said, RA. 2021. A Framework for Data-Driven Solutions with COVID-19 Illustrations. *Data Science Journal*, 20: 36, pp. 1–21. DOI: *https://doi.org/10.5334/dsj-2021-036*

1 INTRODUCTION

Drawn to address global challenges–poverty, health, inequality, climate change, innovation, environmental degradation, peace and justice, the 17 United Nations Sustainable Development Goals (SDG) have, since their inception in 2015, remained at the centre of development strategies for central and local governments, businesses and institutions across the world (United-Nations 2015). The impact of COVID-19 is felt across the sectors and areas that describe the SDG (Rothan & Byrareddy 2020). While tackling global challenges of this nature naturally entails efforts from across disciplines, sectors, borders and legislations; silo working continues to dominate research initiatives in many fields, constantly creating knowledge gaps. The COVID–19 pandemic–a typical example of a global challenge, has reminded us of such gaps in our knowledge, requiring even stronger collaborative and interdisciplinary data-driven initiatives to fill. Despite its devastating impact on our ways of life, it has been argued that COVID–19 has presented us with an excellent opportunity for accelerating attainment of the SDG through data–driven technologies (Pan & Zhang 2020), particularly because COVID–19 is happening amidst data deluge and growing capabilities in handling Big Data (Wang et al. 2020). Different countries have been dealing with the pandemic using different strategies, and the need for sharing data across geographical borders has never been greater.

One of the main issues researchers face and will continue to face in the future is spatio–temporal variations and their impact on the conclusions we reach on data-driven solutions. Despite the devastating effects, the spatio-temporal variations of COVID–19 present an excellent opportunity for the research community to bridge knowledge gaps in addressing societal challenges through interdisciplinary data modelling. As data-driven solutions are dependent on the stability of the underlying data modelling assumptions, knowledge gaps inevitably arise when the assumptions are violated. We present a generic framework for filling such gaps, based on two data-driven algorithms that combine data, machine learning and interdisciplinarity to bridge societal knowledge gaps by highlighting timing and conditions for interventions. Using structured COVID–19 data obtained from the European Centre for Disease Prevention and Control (ECDC); data on its impact, obtained from the UK Office of National Statistics, and unstructured imagery COVID–19 X–obtained from GitHub and Kaggle, we present two algorithms–one for animation and visualisation and the other for enhanced classification based on an adaptive Convolutional Neural Network (CNN) model.

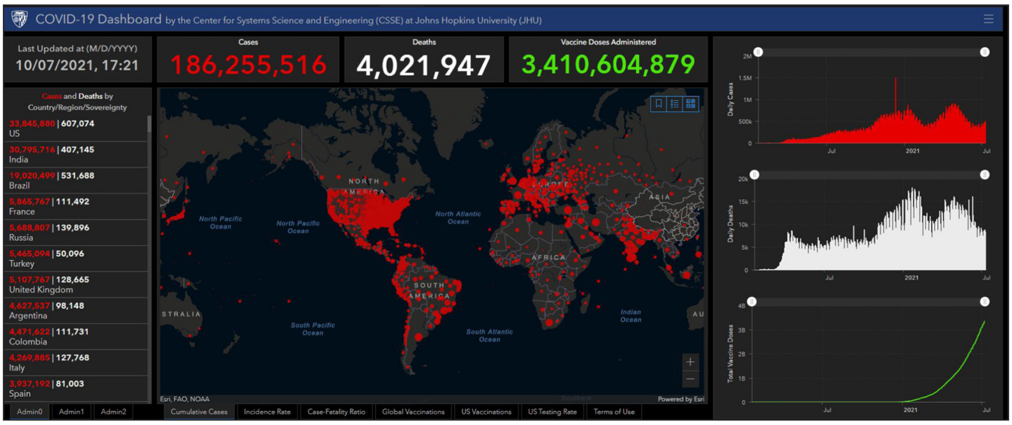
Novelty of the paper is embedded in the two algorithms–both adapted from the Sampling-Measuring-Assessing (SMA) algorithm for addressing data randomness, originally developed by Mwitondi et al. (2018*a*, *b*, 2020) for modelling structured data based on statistical model fitting and evaluation. The adaptation in Section 2, resonates with cross-disciplinary research discussions in tackling global challenges. The paper is organised as follows. Section 1 provides an introduction, motivation, research question and objectives. Section 2 presents the methods–framework, data sources and modelling techniques. Section 3 presents the analyses and Section 4 concludes the work and highlights new directional paths for research.

1.1 RELATED WORK

As noted above, this work was motivated by Big Data Modelling of SDG (BDMSDG) (Mwitondi et al. 2020, 2018*a, b*) and, particularly, by the way COVID–19 has impacted our ways of life (Zambrano-Monserrate et al. 2020, Bartik et al. 2020). The complex interactions of the SDG, the magnitude and dynamics of their data attributes as well as the deep and wide socio–economic and cultural variations across the globe present both a challenge and an opportunity to the SDG project. These attributes impinge on data–driven solutions as they contribute to not only data randomness but also to variations in underlying data relationships and definitions over time, commonly known as concept drift (Zenisek et al. 2019). It is, therefore, reasonable to align the spatio–temporal variations of the impact of COVID–19 with the potential to bridge societal knowledge gaps and gain a better understanding of the challenges we face through data–driven solutions. Data variations have been extensively studied and this work draws from existing modelling techniques such as the standard variants of cross-validation (Bo et al. 2006, Xu & Goodacre 2018) and permutation feature importance (Galkin et al. 2018). The work derives from statistical models like bagging and bootstrapping, which either rely on aggregation of classifiers or sample representativeness (Mwitondi et al. 2019). The SMA algorithm’s superiority lies in its built–in mechanics for efficiently handling data randomness (Mwitondi et al. 2019, 2020).

Since the onset of the COVID–19 pandemic, data visualisation tools have become increasingly common across the world. Many pre-existing dashboards like Our World in Data (Roser et al. 2018), the World Bank Group (WBGroup 2018), Johns Hopkins University Coronavirus Resource Center (CRC 2021) and the Millennium Institute (MI 2021) have developed tools for mapping the pandemic across the globe, some in near real-time. ***Figure 1***, captured from the Johns Hopkins University COVID–19 dashboard on 7th July 2021 at 17:21 hrs, displays cases, deaths and vaccine doses administered by country as well as other data attributes via the menu items. While this kind of pattern visualisation is informative of the direction the pandemic is taking, just like the aforementioned tools, it is essentially an enhanced descriptive statistics generator. Its reliance on country–specific data accuracy leaves many unanswered questions. For instance, does it truly reflect data collection and reporting in all countries displayed? Does it provide a better understanding of the challenges we face? For answers to these and many other general questions, Zhang et al. (2014) recommend a bottom–up approach. Accuracy, completeness, consistency and other aspects of data quality have been widely studied and they remain a focal point in many fields (Cai & Zhu 2015, Zhang et al. 2017).

The advent of dashboards has prompted further thinking on how they can be used for enhancing decision making processes by increasing transparency, accountability, stakeholders' engagement, governance and institutional arrangements Matheus et al. (2020). Our work focuses on how to complement accessible descriptive data, through dashboards or otherwise, by modelling techniques to support decision making processes. It is guided by spatio-temporal variations in gaining insights into how different societies have been impacted by the pandemic.



COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)

Last Updated at (M/D/YYYY)
10/07/2021, 17:21

Cases
186,255,516

Deaths
4,021,947

Vaccine Doses Administered
3,410,604,879

Cases and Deaths by Country/Region/Sovereignty

33,845,880 | 607,074
US

30,795,716 | 407,145
India

19,020,499 | 531,688
Brazil

5,865,767 | 111,492
France

5,688,807 | 139,896
Russia

5,465,094 | 50,096
Turkey

5,107,767 | 128,665
United Kingdom

4,627,537 | 98,148
Argentina

4,471,622 | 111,731
Colombia

4,269,885 | 127,768
Italy

3,937,192 | 81,003
Spain

Figure 1 A Johns Hopkins COVID–19 visualisation dashboard.

(Source: *https://coronavirus.jhu.edu/map.html*)

This paper proposes an interdisciplinary approach for addressing the foregoing general questions based on structured and unstructured data modelling methods. The former is an interactive data animation and visualisation tool with a built-in ability to fire warning alerts, while the latter provides a predictive power using imagery data.

1.2 RESEARCH QUESTION AND OBJECTIVES

Spatio–temporal variations and data randomness are some of the main factors known to impinge on the conclusions we draw from data–driven solutions (Mwitondi & Said 2013). This work combines the power of Big Data, machine learning and interdisciplinarity to address those issues. Using real–life examples based on COVID–19 pandemic data, we examine how country–specific approaches to global challenges fit in the global prism of data–driven solutions. We seek to answer the question: **How do spatio–temporal variations resonate with interdisciplinary tackling of global challenges?** To answer the foregoing question, we set the following objectives.

1. To illustrate the efficacy of national level multi–dimensional visualisation of COVID–19 impact on societies.
2. To demonstrate the efficacy of combining data, techniques and skills in an interdisciplinary context.
3. To use multi-dimensional data visualisation in a two-dimensional space for timely decisions on the impact of the pandemic and other societal challenges.

4. To provide practical implementations of a robust machine learning algorithm, with built-in capabilities for accommodating interdisciplinary skills.
5. To highlight a roadmap for aligning national strategies to the global prism of data–driven solutions.

1.3 CONTRIBUTION TO KNOWLEDGE

The paper's novelty derives from applied mechanics of ***Algorithms 1*** and ***2***, within the context of the data-driven framework in ***Figure 2*** and the implementation flow in ***Figure 3***. Its main idea hinges on data randomness that characterises all learning models as a major cause of spatio-temporal variations, as described in Mwitondi & Said (2013). Using COVID-19 illustrations, the application highlights paths for combining domain knowledge, data, tools and skills in addressing global challenges across the SDG spectrum. Based on evidence from literature, we highlight the following aspects of contribution to knowledge.

1. **Addressing Data Randomness:** Leading to enhanced modelling techniques for decision support systems.
   (a) *Animation*: Multi-dimensional visualisation of data attributes in 2-D space, allowing for manual or automated intervention via ***Algorithm 1***, is an enhanced data-driven decision support system that is not provided by any of the tools discussed under related work in Section 1.1. ***Algorithm 1*** is adaptable to a wide range of applications and COVID-19 is used as a special case to illustrate its mechanics.
   (b) *Model Optimisation*: Rather than just averaging an ensemble of models to reduce variance (as in the bagging case) or evaluating surrogate models, ***Algorithm 2*** combines cross-validation, bagging and step-wise assessment based on updatable parameters (model weights, in this case), exhibiting a robust performance of the algorithm. Application of CNN on COVID-19 data to illustrate robust data-driven solutions for global challenges, ***Algorithm 2*** is adaptive to a wide range of techniques–unsupervised and supervised.
2. **Applications:** A novel approach towards application in the context of SDG initiatives.
   (a) It complements dashboard descriptive data, in an interdisciplinary context as shown in ***Figure 2***.
   (b) Spatio-temporal variations provide insights into how different societies have been impacted by the pandemic. Researchers focusing on other SDG-related challenges can easily adapt the mechanics of the two algorithms and the data-driven generic framework to their specific needs in search of robust performance.

The study methodology is based on structured and unstructured data. Its basic ideas are in the Sample-Measure-Assess (SMA) algorithm originally developed for structured data (Mwitondi et al. 2020, Said & Mwitondi 2021).

2 METHODOLOGY

The section hinges on addressing data randomness that characterises all learning models and it is organised as follows. Section 2.1 provides a data–driven generic framework for addressing societal challenges from an interdisciplinary perspective, using sophisticated data modelling tools. It is followed by a data description in Section 2.2 and an outline of the implementation strategy in Section 2.3.

2.1 A DATA-DRIVEN GENERIC FRAMEWORK

Figure 2 highlights the overlap of global challenges, data and relevant skills, from which the motivation of this work derives. It constitutes a logical relationship between the three categories that are fundamental in addressing cross–sectoral or global challenges, with its overlapping components forming the basis for addressing data randomness, as presented in Section 2.3. The intersections #1 through #4 are crucial as they resonate with the interdisciplinary approach to problem solving. For example, #1 and #4 relate to aspects of data science, while #2 and #4 may relate to specific knowledge domains. Similar interpretations can be made for #1, #2 and #4 or the other tripartites.

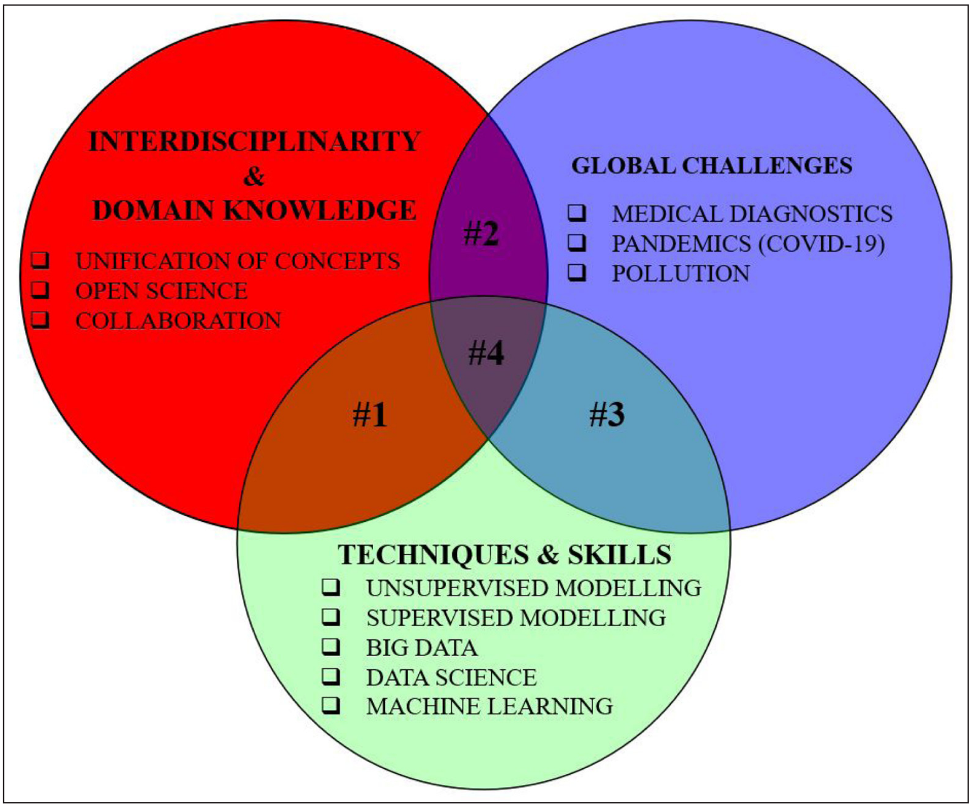


Figure 2 A diagrammatical illustration of the interaction of challenges, data and skills.

Interdisciplinary approaches to tackling global challenges, combining domain knowledge, data, tools and skills are well-documented. In recent years researchers have focused on integrating different sources of knowledge across the broad spectrum of SDG, with poverty, food security, gender equality, health, education, innovation and climate change standing out (Mwitondi et al. 2018*b*). One good example would be the ongoing debates on the role of disparate knowledge sets and expertise in managing the impact of climate change which expose cross-sectoral gaps in learning about data, national policies and various aspects of science as outlined in Pearce et al. (2018). COVID-19 delivers an even better example of the need for interdisciplinarity in tackling global challenges. Evidence of direct correlation between environmental pollution and contagion dynamics imply that interdisciplinarity is required in understanding the pandemic's contagion diffusion patterns in relation to multiplicity of environmental, socio–economic as well as its geographical diversity (Bontempi et al. 2020). This is particularly important, since COVID-19 has generated arguments and counter-arguments on how it should be managed–from balancing societal health and economic aspects to vaccine uptakes and their ramifications on social interactions. Apparently, detaching the categories creates knowledge gaps and the more they overlap, the more cohesive knowledge is attained. These dynamics inevitably lead to data randomness, inherently affecting modelling results and hence the conclusions drawn from them. The setup naturally appeals to developing robust solutions for SDG challenges such as COVID–19, in which not only data variability abounds (Mwitondi et al. 2013), but also definitions and interpretations tend to vary over time, a phenomenon commonly referred to as concept drift (Tsymbal et al. 2008, Žliobaitė et al. 2016). Data randomness and concept drift present natural challenges to algorithmic learning, on which this paper focuses (Mwitondi & Said 2021).

2.2 DATA SOURCES AND VISUALISATION

In the light of the impact of COVID–19 on SDG, data deluge and computing power, each SDG can reasonably be seen as a source of Big Data (Kharrazi 2017, Kruse et al. 2016, Yan et al. 2015, Mwitondi et al. 2018*a*, *b*). For the purpose of this work, structured data came from the European Centre for Disease Prevention and Control (ECDC) (ECDC 2020) and the UK Office of National Statistics website (ONS 2020). The former provided daily updates on cases and deaths per country based on a 14-day notification rate of new COVID-19 cases and deaths while the latter provided multiple data files on business, industry and trade as well as on the general economy and on the dynamics on the labour market before and during the pandemic. Preparation of structured data for animation and visualisation required re-arranging the data points in such a way that the adapted ***Algorithm 1***, described below, could iterate across attributes. ***Table 1*** lists a typical choice of variables of interest. Notice that while this list satisfies the requirements for the illustrations in this paper, it is by no means exhaustive. Its elements are dependent on the problem at hand and must carefully be selected based on the data-driven generic framework in ***Figure 2***. In other words, variable selection is problem-dependent and it should be guided by expert knowledge in both the underlying domain and data analytics. Identifying the necessary skills and modelling techniques is also a function of the problem space. It is that multi-dimensional joint decision that defines the functionality of the framework in ***Figure 2***.

Table 1 Typical variables of interest for animation and visualisation.

VARIABLES	NOTATION	DESCRIPTION AND RELEVANCE
Population	δ	Population affected by a phenomenon: This may be a national, regional or city population from which other variables are obtained
GDP	γ	Gross Domestic Product of a country: Vital for comparative purposes
Unemployment	ξ	Unemployment rate: Global, national, regional or city level
Location	λ	Where a phenomenon happens: Useful for spatio–temporal comparisons
Time	τ	Year, month, week, day etc: Useful for spatio–temporal comparisons
COVID–19	κ	Deaths, infections, hospitalisation rates, variants
PPE	π	Personal Protective Equipment: Associated with COVID–19 etc.

The unstructured dataset is a large COVID–19 X–Ray collection of 1840 image files, downloaded from GitHub (Cohen et al. 2020) and 1341 normal X–Ray image files obtained from from Kaggle (Kaggle 2020). Sources of both structured and unstructured data used in this research are regularly updated, which makes it possible for the paper's modelling results to be reproduced and updated. The adopted implementation strategy is based on a two–fold adaptation of the SMA algorithm as outlined below.

2.3 IMPLEMENTATION STRATEGY

Adaptation of the SMA algorithm is two–fold. The first modification is for animation and visualisation, in search of informative COVID–19 patterns from multiple attributes in a two-dimensional space. The second is for the classification of unstructured data using the Convolutional Neural Network (CNN) model as described in (LeCun, Jackel, Boser, Denker, Graf, Guyon, Henderson, Howard & Hubbard 1989), (LeCun, Boser, Denker, Henderson, Howard, Hubbard & Jackel 1989) and (Fukushima 1980), which is also used to carry out multiple sampling of COVID–19 imagery data. Both adaptations have the potential for providing crucial information to decision makers.

2.3.1 SMA Adaptation for Animation and Visualisation

This adaptation is designed for carrying out animation and graphical data visualisation of selected variables, to reflect the multi-dimensional impact of COVID–19 in a two–dimensional space. Its specific applications will vary and must typically be guided by the framework in ***Figure 2***. For example, the choice of attributes to be animated and/or visualised will depend on the intended purpose of the study. Which variables to display and which cut-off points to trigger which alarms are decisions that require underlying domain knowledge and not a purely data science problem. ***Algorithm 1*** represents a simple variant of the SMA algorithm. It is designed to display multiple variables in a two–dimensional space, comparing relevant parameters and firing a message on meeting pre-specified criteria. Its mechanics are illustrated below, using the notation in ***Table 1***, collectively featuring a super set of data sources Γ.

```
procedure VISUALISATION(Access Data Repository or Superset of Variables)
    Data Access $\Gamma = \{\delta, \gamma, \xi, \lambda, \tau, \kappa, \pi\}$ Accessible Data Source
    Select Variables to plot: $\phi \subseteq \Gamma$
    Define alarm triggering parameters as a set: $\psi$
    Activate: install.packages("gapminder")
    Activate: library(gapminder)
    Activate: library(ggplot2)
    while V $\neq \emptyset$ do
        $\alpha$ :=$ggplot(\mathrm{V}, aes(.))$
        Print $\alpha$
        if $\xi \geq$ a $\in \psi$ and $\kappa \geq$ b $\in \psi$ then
            Trigger Alarm X
        else
            if $\frac{\gamma}{\pi} \geq$ m $\in \psi$ then
                Trigger Alarm Y
            end if
        end if
    end while
end procedure
```

Algorithm 1 Adaptation of the SMA Algorithm (Mwitondi et al. 2020) for Animation & Visualisation.

The subset $\phi \subseteq \Gamma$ contains variables of interest, based on which the algorithm iteratively displays multi-dimensional data in a two-dimensional space, triggering alarms in accordance with pre-set conditions. For example, as the unemployment rate in a particular borough in England reaches a specific level, e.g. $\xi \geq 3.5\%$ while death rates are above 1000 per day, at time $\tau = t^*$, the Chancellor of the Exchequer may need to consider taking action on the furlough scheme, say. Presenting structured data in both visual static and animated forms, provides clear insights to stakeholders in addressing societal challenges such as COVID–19. The main focus is on both Γ and ϕ which will always need to be adapted to handle new cases. Practical illustrations of the algorithm's mechanics are given in Section 3.1.

2.3.2 SMA Adaptation for Convolutional Neural Networks

Figure 3 provides a graphical illustration of our adaptation of the SMA algorithm in addressing data randomness via multiple model training, testing and assessing. The data repository is a large data source from which multiple training and testing samples are drawn, with or without replacement. Its data contents can be either structured or unstructured.

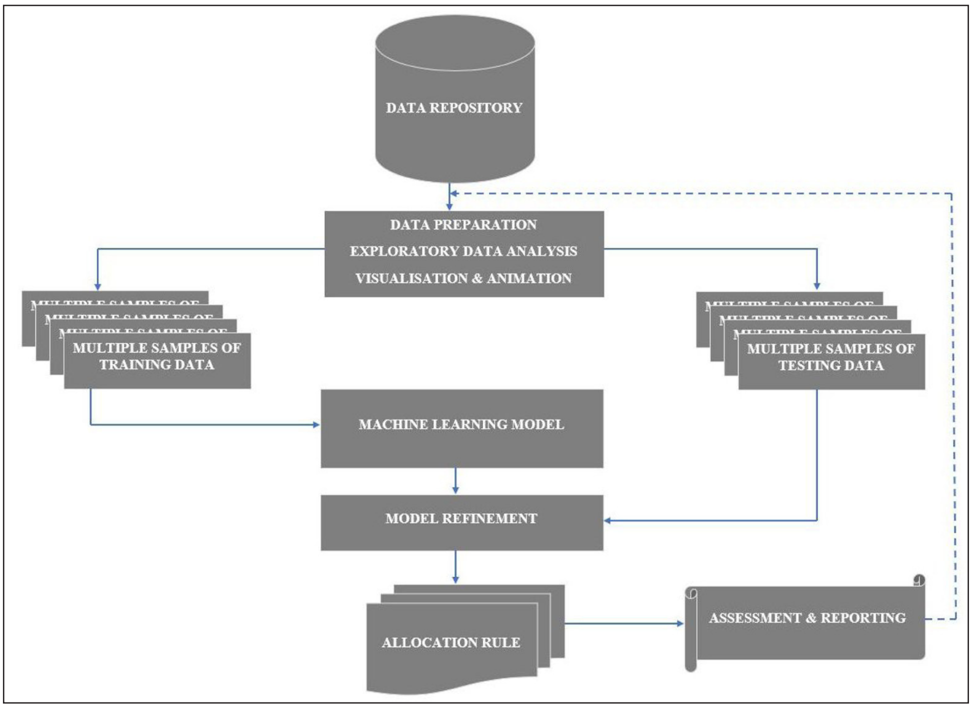

DATA REPOSITORY
DATA PREPARATION
EXPLORATORY DATA ANALYSIS
VISUALISATION & ANIMATION
MULTIPLE SAMPLES OF TRAINING DATA
MULTIPLE SAMPLES OF TESTING DATA
MACHINE LEARNING MODEL
MODEL REFINEMENT
ALLOCATION RULE
ASSESSMENT & REPORTING

Figure 3 Graphical illustration of the CNN classification and assessment process.

At the preparatory level, the investigator examines the overall behaviour of the data through visualisation, animation or other methods of inspection, such as outlier detection and missing values, in order to ascertain its validity for applying the adopted modelling technique. A machine learning model trained and tested on different training samples will typically yield different outcomes. Performance assessment is made on the basis of specific metrics generated and assessed via the two algorithms, as described in Section 2.3. Given a dataset with class labels, *y*, the SMA algorithm applies a learning model which, without loss of generality, we can define as in Equation 1

$$F(\phi) = \underset{x,y\sim\mathcal{D}}{P}\left[\phi(x) \neq y\right] \tag{1}$$

where $\mathcal{D}$ is the underlying distribution and $P[\phi(x) \neq y]$ is the probability of disparity between the predicted and actual values. By repeatedly sampling from the provided data source, modelling and carrying out a comparative assessment of the results, the SMA algorithm provides a unifying environment with the potential to yield consistent results across samples. For classification problems, it proceeds by training and validating the model in Equation 1 on random samples, keeping the samples stateless across all iterations. Thus, multiple machine learning models are fitted, compared and updated over several iterations, finally selecting the best performing model based on the probability

$$P\left(\Psi_{D,POP} \geq \Psi_{B,POP}\right) = 1 \Leftrightarrow \mathbb{E}\left[\Psi_{D,POP} - \Psi_{B,POP}\right] = \mathbb{E}[\Delta] \geq 0 \tag{2}$$

where $\mathbb{E}[\Delta]$ is the estimated difference between the population error $\psi_{D,POP}$ and the validation error $\psi_{B,POP}$. Adaptation of the SMA algorithm is illustrated via Convolutional Neural Network (CNN)–a machine learning technique, typically used for classifying image data such as the X-Ray data, in this case. Its original ideas derive from the work of a Japanese Scientist, Kunihiko Fukushima (Fukushima 1980), on neocognitron–a basic image recognition neural network and developed through the work of LeCun, Jackel, Boser, Denker, Graf, Guyon, Henderson, Howard & Hubbard (1989), LeCun, Boser, Denker, Henderson, Howard, Hubbard & Jackel (1989) into the modern day CNN via the ImageNet data challenge Krizhevsky et al. (2012).

A CNN model performs classification based on image inputs and a target variable of known classes of the images. It is typically composed of multiple layers of artificial neurons, imitating biological neurons, as graphically illustrated in ***Figure 4***. It processes the convolution computing for the input multichannel extracting features on its plane.

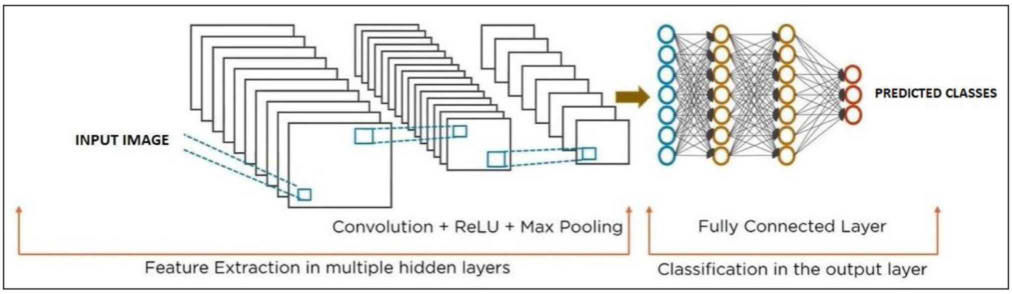

INPUT IMAGE
PREDICTED CLASSES
Convolution + ReLU + Max Pooling
Fully Connected Layer
Feature Extraction in multiple hidden layers
Classification in the output layer

Figure 4 An architecture of a CNN model.

Each convolutional kernel is convolved across the width and height of 2D input volumes from the previous layer, computing the dot product between the kernel and the input. If we let **X** be an $n \times n$ data matrix and **W** a $k \times k$ matrix of weights, which is a 2-dimensional filter with $k \leq n$ (see ***Figure 5***), then

$$X_{k(i,j)} = \begin{bmatrix} x_{i,j} & x_{i,j+1} & x_{i,j+2} \cdots x_{i,j+k-1} \\ x_{i+1,j} & x_{i+1,j+1} & x_{i+1,j+2} \cdots x_{i+1,j+k-1} \\ x_{i+2,j} & x_{i+2,j+1} & x_{i+2,j+2} \cdots x_{i+2,j+k-1} \\ \cdots & \cdots & \cdots\cdots\cdots \\ x_{i+k-1,j} & x_{i+k-1,j+1} & x_{i+k-1,j+2} \cdots x_{i+k-1,j+k-1} \end{bmatrix} \tag{3}$$

where $\mathbf{X}_{k(i,j)}$ is the $k \times k$ submatrix of $\mathbf{X}$ and $1 \le i, j \le n - k + 1$. Now, given a $k \times k$ matrix $\Lambda \in \mathbb{R}^{k \times k}$

$$\sum(\Lambda) = \sum_{i=1}^{k} \sum_{j=1}^{k} \lambda_{i,j} \tag{4}$$

We can then express the 2-dimensional convolution of **X** and **W** using the sums of the element-wise products as

$$\mathrm{X} \star \mathrm{W} = \begin{bmatrix} \sum(\mathrm{X}_k(1,1) \odot \mathrm{W}) \ldots & \sum(\mathrm{X}_k(1, n-k+1) \odot \mathrm{W}) \\ \sum(\mathrm{X}_k(2,1) \odot \mathrm{W}) \ldots & \sum(\mathrm{X}_k(2, n-k+1) \odot \mathrm{W}) \\ \ldots & \ldots \\ \ldots & \ldots \\ \sum(\mathrm{X}_k(n-k+1,1) \odot \mathrm{W}) \ldots & \ldots \sum(\mathrm{X}_k(n-k+1, n-k+11) \odot \mathrm{W}) \end{bmatrix} \tag{5}$$

such that $\sum(\mathrm{X}_k(i,j) \odot \mathrm{W}) = \sum_{\alpha=1}^{k} \sum_{\beta=1}^{k} x_{i+\alpha-1, j+\beta-1} \cdot w$, for $i, j = 1,2,3, \ldots n - k + 1$. It can be shown that the convolution of $\mathbf{X} \in \mathbb{R}^{k \times k}$ is an $(n - k + 1) \times (n - k + 1)$ matrix (Zaki & Mera 2020). A CNN is driven by mathematical functions that calculate the weighted sum of multiple inputs to generate an output based on an activation value function.

We can envision a CNN output, $y_{i,j,k}$ as denoting the neuron output in the i^{th} row and the j^{th} column of feature map k of the l^{th} convolutional layer. To get the convolutional values, we slide the kernel over the input data, multiplying the corresponding values and summing up and fill the matrix of the same dimension as the kernel, as shown in *Figure 5*. Note that the filter has reduced the input matrix to a smaller dimension of its size.

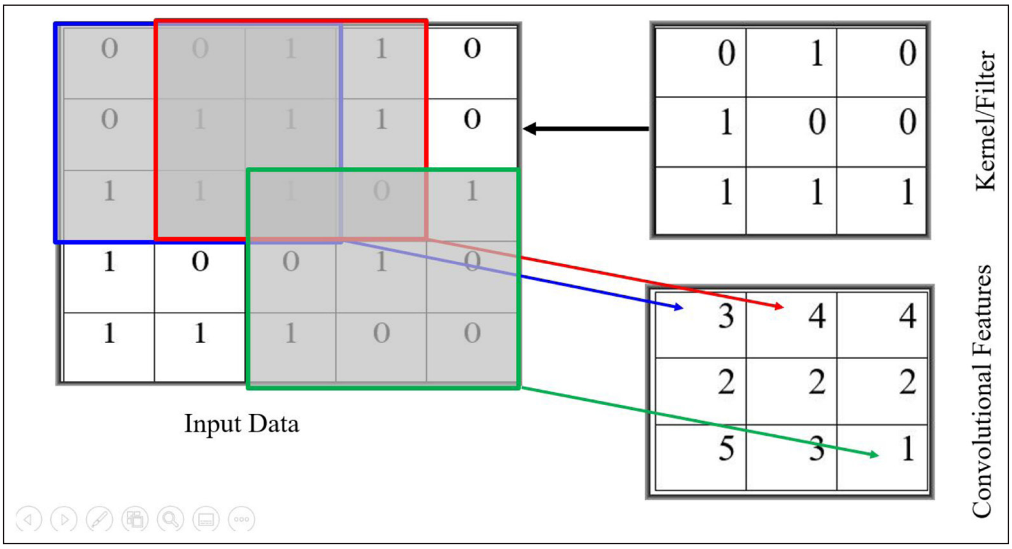


Figure 5 Convolutional values are obtained by sliding the kernel over data.

It is also important to consider the vertical and horizontal strides as they impinge on the model's capability of feature capturing. The pooling layer reduces the dimensionality of the rectified feature map, using different filters to identify different parts of the image – like edges, corners, curves etc. Flattening converts the 2-D arrays from the pooled layer into a one-dimensional vector. The Fully Connected layer then receives this as input, for classifying the image. A 2 × 2 pooling layer, filtering with a sliding of 2 downsamples at every depth of the input discards 75% of the activations. The Rectified Linear Unit (ReLU) applies the activation function in Equation 6.

$$f(x) = \begin{cases} 0, & \text{if} \quad x < 0 \\ x, & \text{otherwise} \end{cases} \tag{6}$$

which basically replaces negative values from the activation map by zero and adopting the actual values otherwise. Other activation functions like the hyperbolic tangent and the sigmoid, in Equation 7, are also commonly used.

$$f(x)=\frac{1}{\left(1+e^{-}x\right)} \tag{7}$$

Central to the performance of the CNN is its architecture. ***Figure 4*** exhibits a typical structure of a CNN model, consisting of the input, convolutional, pooling and the fully connected layers. Each input is weighted in a similar way as are the coefficients in a linear regression model. CNN models are trained using an optimization process, driven by a loss function that calculates the classification error. The maximum likelihood methods is a framework that describes the loss function choice. Other common methods include the cross-entropy and mean squared error. Typically, a CNN model is trained using the stochastic gradient descent optimization algorithm, via which the weights are updated by back-propagating the error. That is, the model with specific weights performs predictions and the resulting allocation error is calculated. At every epoch, the weights are changed to improve performance at the next stage. Equation 8 shows how each weight expresses the rate of change in the total loss (L), as the weight (w) changes by one unit.

$$\frac{\partial L}{\partial w}=\lim_{dw\to 0}\left[\frac{L(w+dw)-L(w)}{dw}\right] \tag{8}$$

As with all models that learn rules from data, performance of the CNN is associated with variations due to data randomness (Mwitondi et al. 2013, Mwitondi & Said 2013). Hence, our CNN implementation will draw multiple samples from the data sources in Section 2.2 in order to attain a generalised performance and attain model robustness.

A number of factors are known to affect the accuracy of CNN–they include the network's number of layers, number of neurons and the learning rate–see, for instance, Géron (2019), Rawat et al. (2020), Wang et al. (2019). These parameters impinge on the model's accuracy and loss–two crucial parameters to the performance of CNN–accuracy and loss. The former is the number of correctly predicted data points as a proportion of the total number of predictions. Loss is the quantitative measure of deviation or difference between the predicted and actual values–it measures the mistakes the CNN makes in applying Equation 1 to any given dataset. Due to the random nature of the sampled data, the accuracies and mistakes the model makes will vary. We shall be seeking to stabilise these variations across samples, using our adaptation of the SMA algorithm as shown below. Thus, the second adaptation of the SMA algorithm, described below, conditions these random samples to the foregoing parameters and varying proportions of training, validation and testing samples for CNN classification. Different strategies to reduce the learning rate during training are known, including those outlined in ***Table 2***.

Table 2 Strategies for Reducing Learning Rate.

STRATEGY	FORMULATION	DESCRIPTION
Power Scheduling	$\eta(t)=\frac{\eta_0}{(1+\frac{t}{k})^{\nu}}$	The learning rate η_0, the steps k and the power ν are typically set to 1 at the beginning. The learning rate will keep dropping at each step, much faster in the early stages than later on. Fine tuning η(t) is one of the functions of the algorithm.
Exponential Scheduling	$\eta(t)=\eta_0\times 0.1^{\frac{t}{k}}$	A much faster option for reducing η_0, which drops by a factor of 10 every k steps. The researcher can fine tune the constant 0.1 to suit their needs
Piecewise Constant Scheduling	$\eta(t)$ = 0 for 10 epochs	A constant η_0 for a number of epochs (e.g. η_0 = 0.2 for k = 10 then η_0 = 0.1 for k = 30 etc).
Performance scheduling	ε_v	Validation Error: Measuring it helps decide on reducing η_0 by a specified factor when δ_v stops dropping.

Algorithm 2 experiments with as many learning rates as possible in search of an optimal model and it only stops once it is evident that the changes have no much impact on the parameters δ_t and δ_t, i.e., changes in the training and validation errors respectively. To arrive at the best model at step #36, the CNN model runs with a check point that monitors both training and validation accuracy, saving the best weights and reporting each time performance improves. In Python *checkpointer = ModelCheckpoint(filepath=”best_weights, monitor = ‘accuracy’, save_best_only=True)* is called by the best CNN fit as an argument alongside other training and validation parameters and number of epochs.

```
procedure CNN Classification of Imagery Data
    Set $\Omega$ : Large data source of imagery data
    Initialise $\mathcal{R} := \{n_1(t), n_2(t), n_3(t)\}$ CNN Model Learning Rates
    Initialise: $\mathcal{M} := \{\mu_1, \mu_2, \ldots \mu_k\}$ Model Architecture
    Initialise: $S_t \subset \Omega$ Training sample without replacement
    Initialise: $S_v \subset \Omega$ Validating sample without replacement
    Initialise: $\epsilon_t := \emptyset$ Training sample error
    Initialise: $\epsilon_v := \emptyset$ Validation sample error
    Initialise: $l_t := \emptyset$ Training loss
    Initialise: $l_v := \emptyset$ Validation loss
    Initialise: $\psi^* := \emptyset$ Model weights
    Initialise: $\delta_t^* := \emptyset$ Training error changes vector across epochs
    Initialise: $\delta_v^* := \emptyset$ Validation error changes vector across epochs
    while $i \leq m$ do
        for $j := 1 \rightarrow n$ do
            Fit $\mathcal{M}[i]$ CNN models in set $\mathcal{M}$
            $\epsilon_t^* \leftarrow \epsilon_t[ij]^*$ Update training error vector
            $\epsilon_v^* \leftarrow \epsilon_v[ij]^*$ Update validating error vector
            $\delta_t^* \leftarrow \delta_t[ij]^*$ Update training error changes vector
            $\delta_v^* \leftarrow \delta_v[ij]^*$ Update validating error changes vector
            Plot $\delta_t^*$ and $\delta_v^*$
            Plot $\epsilon_t$ and $\epsilon_v$
            if $l_t \gg l_v$ or $\epsilon_t \downarrow$ while $\epsilon_v \uparrow$ then
                Over-fitting
            else
                if $l_t \ll l_v$ or $\epsilon_t \uparrow$ while $\epsilon_v \downarrow$ then
                    Under-fitting
                    if $l_t \approx l_v$ and $\epsilon_t \approx \epsilon_v$ then
                        Retain $\psi^*$
                    end if
                end if
            end if
        end for
        Update $\psi^*$ with weights from $\mathcal{M}[i]$
    end while
    Output the Best Model $\underbrace{P}_{x,y\sim\mathscr{D}} [\phi(x) \neq y]$ with $\psi^*$
end procedure
```

Algorithm 2 Adaptation of the SMA Algorithm (Mwitondi et al. 2020) for CNN Classification.

While the implementations in Section 3 were carried out by a combination of R and Python libraries, the two algorithms are amenable to any data analytics tool. Our animation and visualisation in Section 3.1 were driven by the *gapminder* package in R, hence their explicit inclusion in ***Algorithm 1*** but, again, these steps are transferable to other packages and libraries. The same applies to the implementation in Section 3.2, which was carried out in Python's Keras.

3 ANALYSES

Analyses in this section are two–fold. Section 3.1 presents graphical images captured from animated patterns for the structured data and Section 3.2 presents unstructured data results, based on a CNN model. The section also provides discussions on the impact of the digital divide (Ramsetty & Adams 2020) in the fight against COVID–19.

3.1 DATA VISUALISATION

The UK Office for National Statistics (ONS 2020) data repository has datasets going back many years, but we examine data on employment and the Gross Domestic Product (GDP) for the last 2 years before and through the pandemic. The plots in ***Figure 6*** are selected animation patterns from the period 2008 to 2020. They exhibit GDP and labour market patterns for the first and second quarters over the period–i.e., before and during the pandemic. Because of the furlough scheme introduced by the UK Government at the beginning of the pandemic, unemployment between the two quarters of 2020 appears to be at the same level, but there is a huge variation between the two GDP figures.

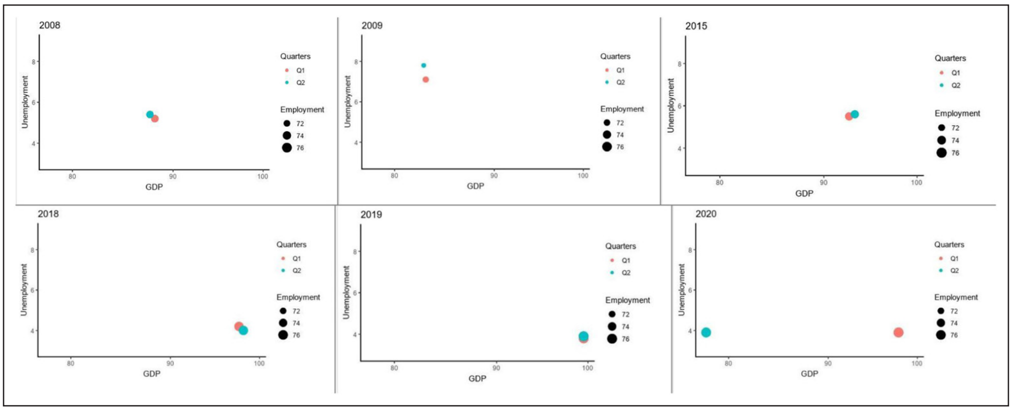


Figure 6 GDP and unemployment patterns for selected 1st and 2nd quarters over the period 2008–2020.

Figure 7 shows the number of deaths involving and not involving the coronavirus (COVID-19) in Wales and selected regions of England, occurring between 1 March and 31 July 2020. The data, obtained from the UK Office of National Statistics show that the highest deaths occurred in the South East–with 18.7% of the total 39,154 deaths being COVID–19 related. The lowest number of deaths occurred in the North East, but with 20.9% of the total 13,507 deaths being COVID–19 related. London had the highest proportion of COVID–19 related deaths–i.e., 30.6% of the total 27,908. Due to the effect of the lockdown and other measures, the number of deaths went down in July, but the South East and the North East maintained their respective statuses–highest and lowest. Typically, a COVID–19 related death will be one that has COVID–19 appearing on the death certificate. It is therefore important to note the inherent randomness in interpreting these statistics, as it has an impact on the patterns.

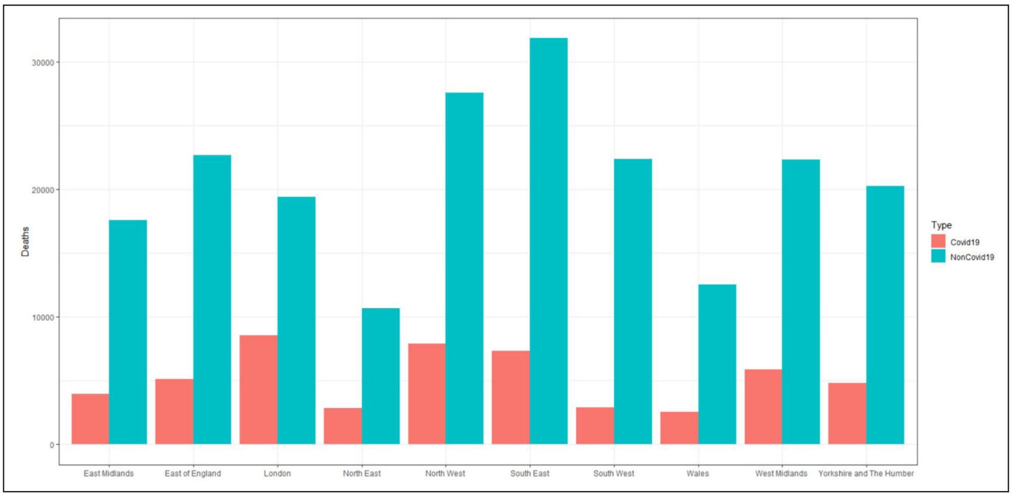


Figure 7 Recorded deaths in parts of the UK between March and July 2020.

The six panels in ***Figure 8*** show the number of COVID–19 related cases and deaths in Brazil, France, Italy, Japan, South Africa, US and the UK. They are captured from an animation model run on the data in Section 2.2. This kind of data visualisation enables researchers to view up to five data attributes in a 2–D plot, with the option to interrupt animation to capture a desired part of the data. As noted earlier, by inserting a conditional check in ***Algorithm 1***, the animation can be used to trigger a warning alarm or raise an alert about some good news.

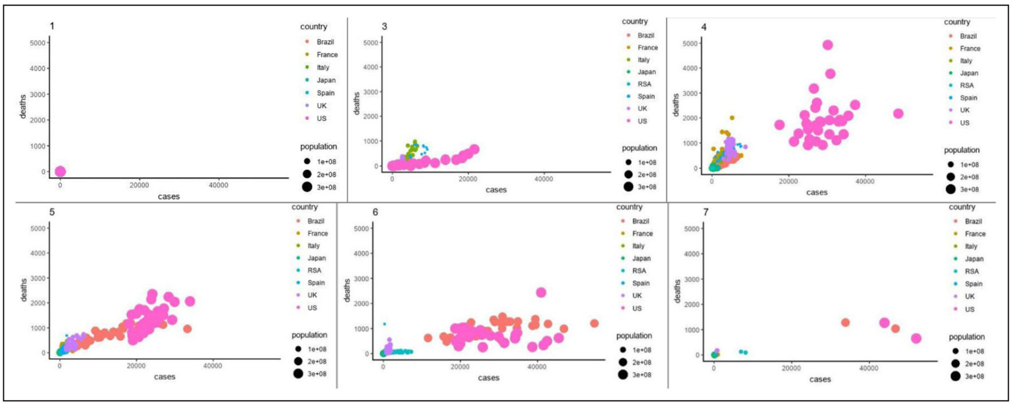


Figure 8 Images captured from animated plots for the first 7 months of 2020.

All three data visualisation examples in ***Figures 6*** through ***8*** are prone to data randomness, which analysts need to pay attention to. For example, the ECDC acknowledges that the data might not be very accurate, as the calculations by the ECDC Epidemic Intelligence are affected by variations in national testing strategies, laboratory capacities effectiveness of surveillance systems. This implies that reporting and hence monitoring and control of the pandemic will vary across regions, which underlines the need for collaborative work in managing global challenges.

3.2 CONVOLUTIONAL NEURAL NETWORKS

Under pandemic conditions, doctors and radiologists are under pressure to distinguish COVID-19 X-Ray data described in Section 2.2 images. The output arrows in ***Figure 9*** are class predictions from the input data. At different levels of convolutions, extracted features provide useful data for predicting the class of an image, which underlines differences in imagery representation under different machine learning conditions. While the application of CNN in modelling imagery data is not new, model optimisation challenges remain a focal point for research. It is in this context that we emphasise the need to adopt the interdisciplinary framework in ***Figure 2***, for guiding data-driven solutions.

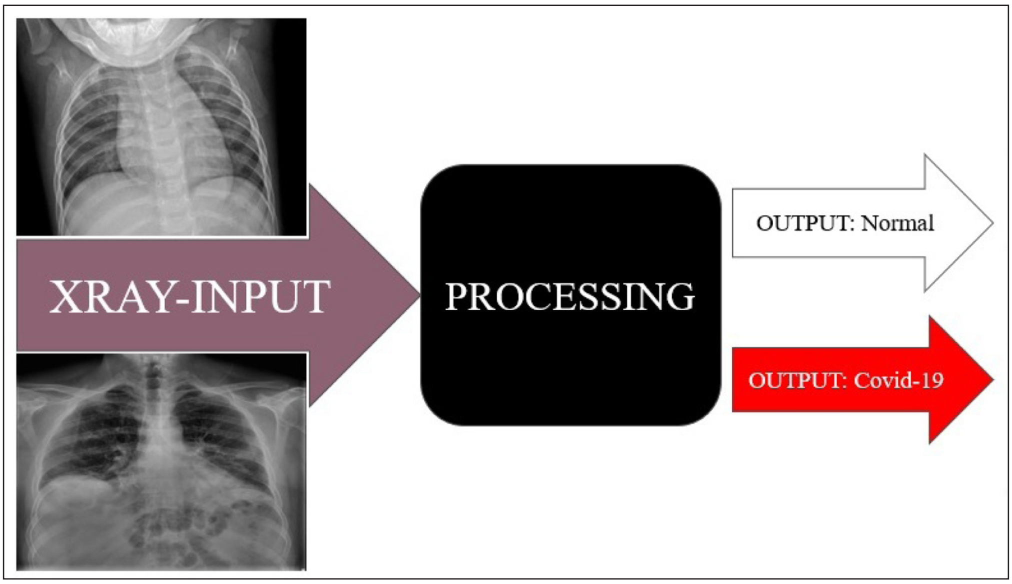


Figure 9 A CNN model is trained on imagery data to perform classification based on known classes.

3.2.1 Training and Validation

Implementation of the adapted ***Algorithm 2*** was driven mainly by Keras–a deep learning Application Programming Interface (API) running on top of TensorFlow–an open-source machine learning platform (Géron 2019, Grattarola & Alippi 2020). The open-source models were chosen in consideration of the interdisciplinary nature of the proposed methods, as they provide easy access to a wide range of stakeholders–beginners and experts alike (Zhang et al. 2021). Further discussions on the relevance of interdisciplinarity to modelling mechanics of various machine learning models are in Section 3.3.2. As already noted, learning rules from data is inevitably associated with variations due to data randomness (Mwitondi & Said 2013, Mwitondi et al. 2013), which can negatively affect model performance. For generalisation and robustness, we adopted the SMA algorithm (Mwitondi et al. 2018a,b, 2020), taking multiple samples and conditioning each on the training and validation proportions as in ***Table 3***.

SAMPLE #	TRAIN %	VALID %	TRAIN-START	TRAIN-CONVERGE	VALID-START	VALID-CONVERGE
1	80%	20%	87.98%	99.57%	20.99%	98.95%
2	70%	30%	88.79%	99.71%	95.99%	100.00%
3	60%	40%	90.68%	100.00%	83.99%	99.00%
4	50%	50%	87.00%	99.71%	94.99%	100.00%

Table 3 Selected training and validation model accuracy based on 50 CNN epochs.

Training and validation data for the first run was split into 80%–20% respectively, running 50 epochs on two classes with 744 training and 186 validation images. Other samples were of 70% (training) and 30% (validation), corresponding to 651 images and 279 images respectively, 60%–40% (558 training and 372 validation) and 50% for training and validation on 2 classes. All but the first sample started and converged at high training and validation accuracy.

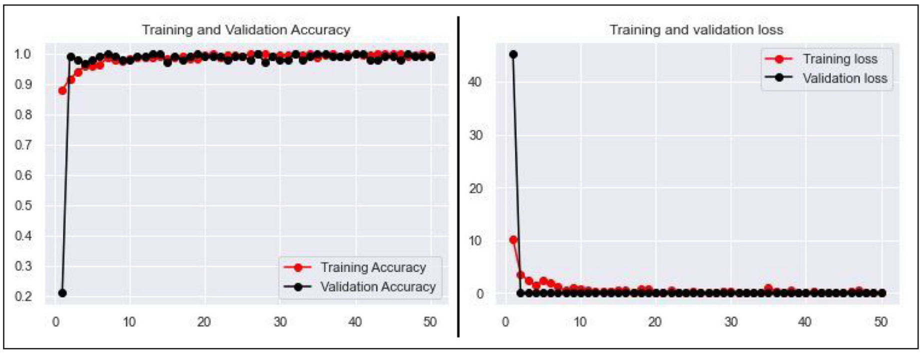


Figure 10 Training and validation accuracy and loss patterns based on the 80%–20% split.

The panels in ***Figure 10*** correspond to model accuracy (left-hand side) and model loss (right-hand panel). They are based on the training-validation split of 80%–20% respectively. In this case, both start with very high accuracy and quickly converge, after only ten epochs. The panel to the right shows the model loss–a quantitative measure of deviation between the predicted and actual values. This is the measure of the mistakes the CNN model makes in predicting the output. As the loss is approximately equal to validation loss, the model is perfectly fitting on both training and validation data, as can be seen on the left-hand side panel.

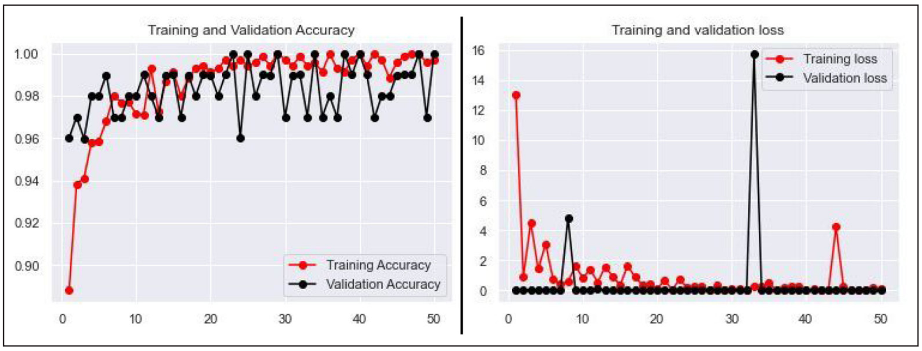


Figure 11 Training and validation accuracy and loss patterns based on the 70%–30% split.

The two panels in ***Figure 11*** are based on the training-validation split of 70%–30% respectively. In this case, while they start with high accuracy, it isn't until about 25 epochs that the training accuracy converges while the validation accuracy continues to oscillate around 97%. In the panel to the right, the training loss exceeds validation loss up until 30 epochs, an indication of underfitting, but the model fits well at higher epochs, except for the two spikes.

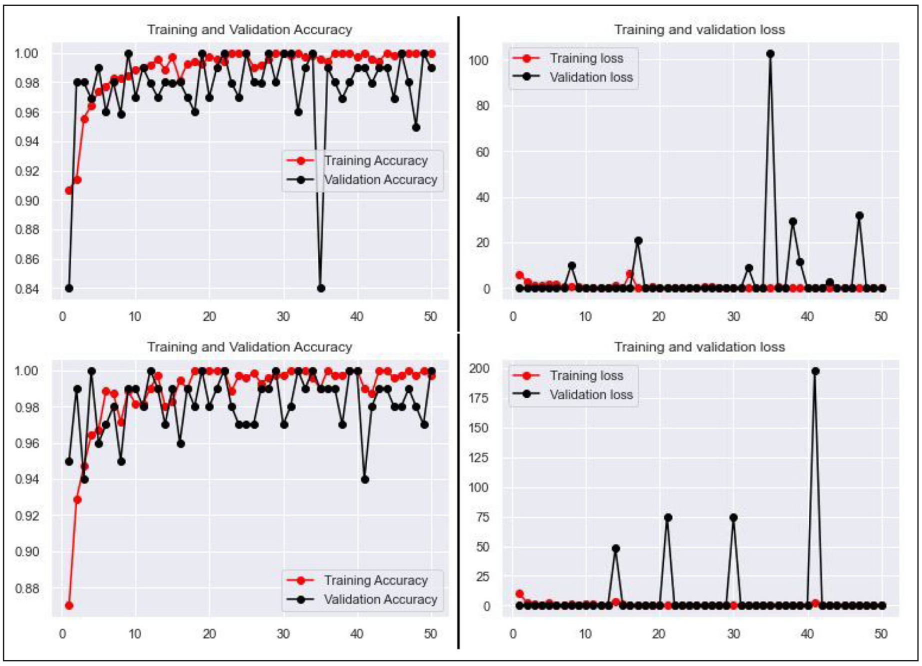


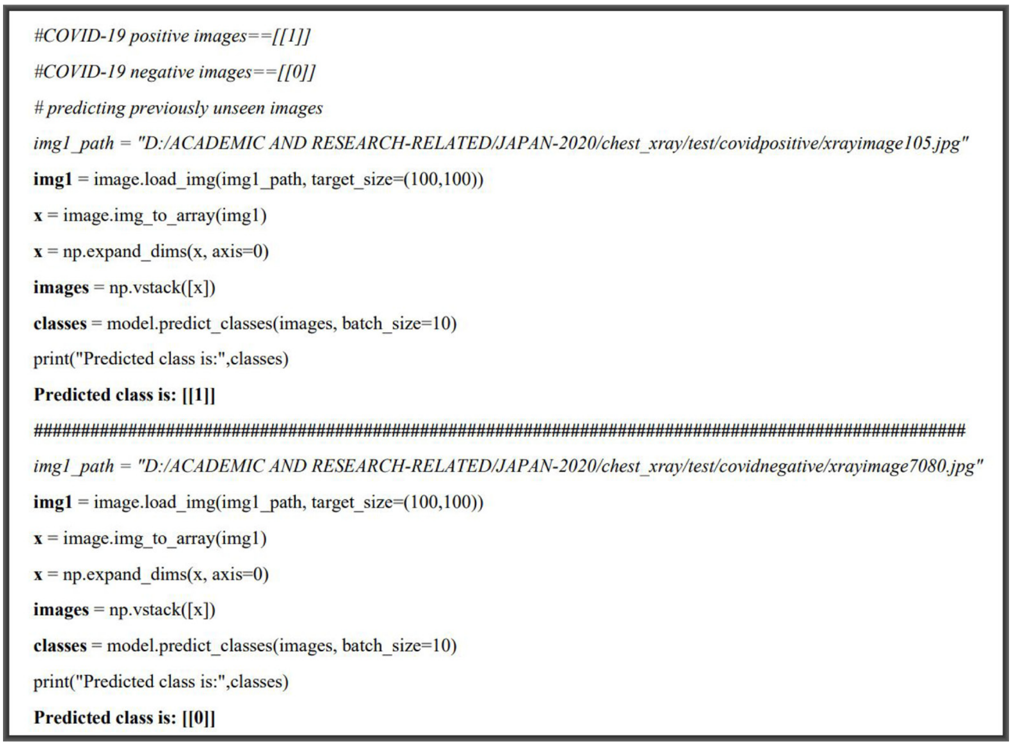
Figure 12 Accuracy and loss patterns based on the 60%–40% (top) and 50%–50% (bottom) splits.

The two panels at the top of ***Figure 12*** correspond to the 60%–40% split, while the bottom panels represent the 50%–50% split. In both cases, the training and loss rates are stable above 25 epochs, but the validation rates appear to be consistently oscillating, an indication of huge variations attributed to randomness in unseen data. In all plots in ***Figures 10*** through ***12*** attention is on the model's consistency, i.e., whether the model is predicting the classes well. The panels to the right show the model loss–a measure of the mistakes the CNN model makes in predicting the output.

When training loss exceeds validation loss, we have the case of underfitting, a rarity. The most common scenario is that of over-fitting–i.e., when training loss is significantly less than validation loss, which implies that the model is adapting so well to the training data that it considers random noise as meaningful data. In other words, the model fails to generalize well to previously unseen data. The ideal scenario is when training loss is approximately equal to validation loss, as that would mean that the model is perfectly fitting on both training and validation data. It is important to note that while these technical issues are fundamental, interpreting data visualisation and modelling findings must always be considered in problem–specific and interdisciplinary context.

3.2.2 Testing and Assessment

Ensuring that the CNN model performs well after training is crucial before its deployment on previously unseen data. The model was tested on new data and yielded convincingly high accuracy and consistent loss patterns. We did this by repeatedly running the CNN model with a "check point" via ***Algorithm 2***, monitoring both training and validation accuracy, saving the best weights and reporting each time performance improves. The saved best weights (the model) were then used to predict the class of any previously unseen X–ray images as illustrated in ***Figure 13***.



```
#COVID-19 positive images==[[1]]
#COVID-19 negative images==[[0]]
# predicting previously unseen images
img1_path = "D:/ACADEMIC AND RESEARCH-RELATED/JAPAN-2020/chest_xray/test/covidpositive/xrayimage105.jpg"
img1 = image.load_img(img1_path, target_size=(100,100))
x = image.img_to_array(img1)
x = np.expand_dims(x, axis=0)
images = np.vstack([x])
classes = model.predict_classes(images, batch_size=10)
print("Predicted class is:",classes)
Predicted class is: [[1]]
########################################################################################################
img1_path = "D:/ACADEMIC AND RESEARCH-RELATED/JAPAN-2020/chest_xray/test/covidnegative/xrayimage7080.jpg"
img1 = image.load_img(img1_path, target_size=(100,100))
x = image.img_to_array(img1)
x = np.expand_dims(x, axis=0)
images = np.vstack([x])
classes = model.predict_classes(images, batch_size=10)
print("Predicted class is:",classes)
Predicted class is: [[0]]
```

Figure 13 Accurate predictions of unlabelled new data for both positive and negative COVID–19 cases.

We assessed model performance based on the metrics inside ***Algorithm 2***, measuring loss as the distance between the predicted and true values. Minimising this loss means making fewer errors on the data. In our binary classification, application we had access to probabilities of class membership and we computed the loss as the sum of the difference between the predicted probability of the real class of the test image and 1. Parameter tuning is necessary to achieve optimal results and different applications may require different tunings. However, this can generally be monitored at the model refinement stage in ***Figure 3***. Adapting Equation 2 to a loss function informs how the model is performing. In a binary classification, anything above 0.5 will allocate to one class and to another class, otherwise. We used the loss function to evaluating how well the CNN model functioned through the algorithm in modelling our dataset. ***Figure 12*** exhibits a very low loss output, which indicates a good performance of the algorithm.

3.3 DISCUSSIONS

Addressing global challenges is conditional on capturing relevant data attributes across areas of interest and making that data readily and equitably available to the international scientific and research community. For example, by admitting that some of the COVID–19 data might not be accurate, as it is conditional on regional and technical variations, the ECDC acknowledges that there are significant potential consequences in the decisions we take. In dealing with COVID–19 related data, the findings in Section 3.1 suggest that any comparisons should be made with care, possibly in combination with other factors like "...testing policies, number of tests performed, test positivity, excess mortality and rates of hospital and Intensive Care Unit (ICU) admissions." In particular, such comparisons must be done by teams of data scientists, epidemiologists, and other medical and social experts.

Sustainability of our livelihood and natural habitat requires an adaptive understanding of the triggers of known and potential positive and negative phenomena we face. Thus, SDG monitoring in post COVID–19 conditions should reflect realities in a spatio-temporal context, focusing on, *inter-alia*, citizen science data, machine learning, IoT and mobile applications. We will need an interdisciplinary approach to respond to new challenges and exploit new opportunities in sectors like manufacturing, agriculture, business, health and education. Tracking global variations in recovery strategies in various sectors and addressing real-life issues like food security, innovation, productivity and many others, will be crucial. In the end, we look at the most important challenges and opportunities that researchers face when working with COVID databases (repositories). The success stories relate to interoperability, interdisciplinarity and free access.

3.3.1 Potential Extensions of Algorithm 1 Applications

Table 4 provides selected examples of the role of interdisciplinarity (***Figure 2***) in addressing SDG. The complex interactions of SDG present an ideal case for the mechanics of ***Algorithm 1***. Given established interactions, the algorithm can be applied to monitor SDG at all levels–national, regional or global. ***Algorithm 1*** provides scope for interventions based on automated multi-dimensional animation for a wide range of applications. More specifically, describing "what is interesting" (the basis for problem identification), is based on ***Figure 2***–i.e., underlying domain knowledge, problem space and modelling expertise. In ***Algorithm 1***, this amounts to identifying Γ and $\phi \subseteq \Gamma$.

Table 4 Selected scenarios of interest for intervention through Algorithm 1.

SDG APPLICATION	RELATED ASPECTS OF DEVELOPMENT	INTERDISCIPLINARITY
SDG #1 (Poverty)	1. Sustainable livelihoods 2. Access to basic social services 3. International cooperation	Various attributes describe poverty eradication & empowerment: The impact of poverty on women requires gender specialist intervention (SDG #5). Co-ordinated efforts between donors & recipients (SDG #17). Good health (SDG #3) and education (SDG #4) lead to productivity (SDG #9), improved income and reduced inequality (SDG #10)
SDG #9 (Innovation)	1. Resilient infrastructure 2. Supporting economic development and human well-being 3. Research and development 4. Industrialisation	To deliver sustainable and resilient infrastructure countries need enhanced financial, technological and technical co-operation (SDG #17). Enhanced productivity in manufacturing, agriculture & services sectors requires quality education (SDG #4).
SDG #13 (Climate Action)	1. Disaster risk reduction 2. Sustainable transport 3. Sustainable human settlement 4. National strategies	Climate action spans across SDG from multi-disciplinary angles. Its key aspects include national strategies, disaster risk reduction, sustainable transport, sustainable cities & human settlement (SDG #11).

The impact of COVID-19 on SDG has recently been widely studied, particularly during the first 18 months of the pandemic. In one recent publication, the pandemic is reported to have led to an unprecedented rise in poverty, in a generation, in parts of the world. For example, the Government of Bangladesh is said to have struggled to provide social safety net packages for marginalised groups, leading to a huge socio-economic inequality and exclusion (SDG #10) (IISD 2021). The three examples in ***Table 4*** underline not only the SDG overlaps but also the need for interdisciplinary consensus in adapting and executing ***Algorithm 1***.

3.3.2 Potential Extensions of Algorithm 2 Applications

For ***Algorithm 2***, variability derives from data, deployed models and model parameters. Like in all other applications, validity of the results is hugely influenced by Ω and model-specific parameters, which implies that interdisciplinarity plays a crucial role in identifying "what is interesting". ***Table 5*** highlights two examples for the rationale of the framework in ***Figure 2*** in searching for optimal machine learning models–unsupervised or supervised.

Table 5 Selected examples of interdisciplinary involvement for machine learning.

MODELLING TECHNIQUE	PERFORMANCE INFLUENTIAL FACTORS	INTERDISCIPLINARY INVOLVEMENT
K-Means	1. Data distributional behavior 2. Initial centroids 3. Distance function adopted	Data choice is problem-driven but it is vital to have thorough considerations as to "what is interesting" before, during and after clustering.
CNN	1. Topology/Architecture 2. Initial weights 3. Updating rule 4. Learning rate 5. Epochs 6. Data/Data augmentation	Data choice is problem-driven and while the decision on the architecture may initially be by a data scientist, underlying domain knowledge is crucial in interpreting the results. Parameter tuning image data augmentation, handling of over-fitting/under-fitting require interdisciplinarity.

For all learning models, the choice and/or tuning of parameters is inherently interdisciplinary. For instance, pre-specifying the number of clusters in the data hugely impinges on the performance of the K-Means algorithm, implying that this decision has to be made based on some level of prior knowledge of the phenomenon. In applying ***Algorithm 2*** for K-Means clustering, these considerations must be made. For example, instead of using a single predefined number of centroids, multiple sets might be considered. For the second example, in ***Table 5***, the convergence of the back propagation network in neural computing is a function of factors such as initial weights, learning rate, updating rule as well as the quality and size of training and validation data. The multi-parameter dependence yields different results, the interpretations of which determines whether the underlying problem is addressed or not.

Attaining model optimisation through training and validation is crucial for the performance of all learning algorithms, yet data randomness remains a major challenge to researchers. The plots in ***Figures 10*** through ***12*** exhibit one common challenge in predictive modelling–attaining generalisation for which we need to avoid both underfitting and overfitting. They reflect the challenges of model optimisation and while they provide guidelines in selecting the best performing model, we can attain unified understanding of the concepts and work towards scientific consensus if we work collaboratively across regions and disciplines, openly sharing resources. Some of the general commonalities for addressing global challenges in problem–specific and interdisciplinary contexts are summarised in ***Table 6***.

Table 6 Basic considerations for data–driven approaches to addressing global challenges.

COMMONALITIES	FOCAL POINTS	DESCRIPTION
Data	1. Data owners 2. Data managers 3. National Statistics Offices 4. Open access repositories	Making relevant data available to those who need it, when they need it
Computing Resources	1. High Performance Computing 2. Security 3. Internet of Things (IoT)	Providing robust, secure and versatile computing resources for users by both the public and private sectors
Skills	1. Data Science 2. Domain–specific knowledge 3. Interdisciplinarity	Adopting interdisciplinary approaches for the purpose of attaining unified solutions to global challenges
Strategies	1. Research collaboration 2. Students Exchange Programmes 3. Apprenticeships & Internships 4. Knowledge Transfer Partnerships	Devising institutional frameworks for sharing resources and knowledge through educational, vocational and research institutions
Legislations	1. Privacy (e.g., GDPR) 2. Cross–border data sharing 3. Access to computing resources 4. Patents and copy rights	Working towards operating open systems that talk to each other

For our sustainability and that of species around us, we are required to make right decisions at the right time. Co-ordinated initiatives are required in responding to global challenges that defy geographical boundaries and national or regional legislations. While the foregoing geo-political variations may not disappear overnight, the scientific community is duty bound to engage in co-ordinated studies for addressing the current and potential future global challenges.

4 CONCLUDING REMARKS

This paper focused on addressing global challenges from a data modelling perspective, illustrating use cases based on the data-driven generic framework in Section 2.1 and the two adaptations of the SMA algorithm in Section 2.3. The adaptive nature of the two algorithms highlights the paper's contribution to knowledge as outlined in an interdisciplinary context, highlighting where errors could occur in the process of knowledge extraction from data. The algorithms and the framework form a system with which actors–any users, addressing SDG-related challenges interact to reach desired outcomes. ***Tables 4*** and ***5*** present some of the preconditions which must hold for the use case to run. Identifying the triggers of the events for which data-driven solutions are entailed cannot be confined to a single discipline. The current circumstances entailed the illustrations based on COVID–19 related data.

Based on the objectives outlined in Section 1.2, the paper highlighted the potentials of combining underlying domain knowledge, on the one hand, and data science–technical skills and soft skills, on the other. It underlined the role of interdisciplinarity in addressing global challenges, and these were viewed in the context of SDG. There are many lessons from the COVID-19 pandemic, not least how we generate and share data. Generally, the five objectives in Section 1.2 were met. The X–Ray examples used in this paper present only very basics of deep and machine learning methods for biomedical imaging and related clinical data, which academic, biomedical and industry will need to explore further as a way of decreasing diagnostic errors and developing and scaling novel phenotypes to enhance precision in the medical research and related fields. We emphasised interdisciplinarity and data randomness because even though CNN models can detect patterns that might go unnoticed to the human eye, for all their power and complexity, they do not provide thorough interpretations of the imagery data. Further, they may perform poorly on previously unseen data. We observed that lessons derived from COVID-19 can help enhance our understanding of the mutual impact–positive and negative, resulting from our interaction with our environment.

There can be no better way to view the bigger picture than through the SDG initiative. Aspects of SDG like species facing extinction, hunger and poverty, low productivity, land degradation, gender inequality or gaps in health and education quality as well as technological achievements span across sectors and regions. These geo-political variations of SGD metrics reflect the inverted COVID-19 patterns in terms of data access and mitigation. The two algorithms–both relating to objectives 3 through 5, provide a range of opportunities in addressing societal challenges of the COVID–19 nature and others. This paper was prepared using open source data and tools. It is expected that it will stimulate novel discussions into the way the scientific community interact based on the elements in ***Figure 2*** and ***Table 6***.

ADDITIONAL FILE

The additional file for this article can be found as follows:

- **Japan.** BIB file for references. DOI: *https://doi.org/10.5334/dsj-2021-036.s1*

ACKNOWLEDGEMENTS

We would like to thank the Joint Support-Center for Data Science Research (DS), through the Japanese Polar Environment Data Science Center (PEDSC), who invited us to Tokyo to present the original ideas that lead to this work. We would also like to acknowledge the role played by the Data Intensive Research Initiative of South Africa (DIRISA), through the South African Council for Scientific and Industrial Research (CSIR), in the run up to this work.

FUNDING INFORMATION

This research was not supported by any external funding, but we would like to express our gratitude to our respective institutions–Sheffield Hallam University and Canadian University Dubai, for allowing us time to complete it and for agreeing to meet our Article Processing Charge (APC) costs.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTION

The original ideas of this work derive from previous collaborative work of both authors, part of which has involved other colleagues. They both contributed equally to this work, with the problem definition being initiated by the second author and the first author focusing on the methodology. The second author also made significant contributions in developing the framework and related literature, while the second author carried out most of the data cleaning and developing the algorithms. Coding, writing up and proof-reading were equally shared across several iterations, from the initial submission through all the reviews and resubmissions.

AUTHOR AFFILIATIONS

Kassim S. Mwitondi *orcid.org/0000-0003-1134-547X*
Sheffield Hallam University, College of Business, Technology & Engineering, GB
Raed A. Said *orcid.org/0000-0003-4378-7029*
Canadian University Dubai, Faculty of Management, UAE

REFERENCES

Bartik, AW, Bertrand, M, Cullen, Z, Glaeser, EL, Luca, M and **Stanton, C.** 2020. *The impact of covid-19 on small business outcomes and expectations*, 117(30): 17656–17666. DOI: *https://doi.org/10.1073/pnas.2006991117*

Bo, L, Wang, L and **Jiao, L.** 2006. Feature scaling for kernel fisher discriminant analysis using leave-one-out cross validation. *Neural Computation*, 18(4): 961–978. DOI: *https://doi.org/10.1162/neco.2006.18.4.961*

Bontempi, E, Vergalli, S and **Squazzoni, F.** 2020. Understanding covid-19 diffusion requires an interdisciplinary, multi-dimensional approach. *Environmental Research*, 188: 109814. URL: *https://www.sciencedirect.com/science/article/pii/S001393512030709X*. DOI: *https://doi.org/10.1016/j.envres.2020.109814*

Cai, L and **Zhu, Y.** 2015. The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, 14(2): 1–10. DOI: *https://doi.org/10.5334/dsj-2015-002*

Cohen, JP, Morrison, P and **Dao, L.** 2020. Covid-19 image data collection. *arXiv 2003.11597*. URL: *https://github.com/ieee8023/covid-chestxray-dataset*.

CRC. 2021. Coronavirus Resource Center. URL: *https://coronavirus.jhu.edu/*.

ECDC. 2020. Covid-19 data. URL: *https://www.ecdc.europa.eu/en/publications-data*.

Fukushima, K. 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36: 193–202. DOI: *https://doi.org/10.1007/BF00344251*

Galkin, F, Aliper, A, Putin, E, Kuznetsov, I, Gladyshev, VN and **Zhavoronkov, A.** 2018. Human microbiome aging clocks based on deep learning and tandem of permutation feature importance and accumulated local effects. DOI: *https://doi.org/10.1101/507780*

Géron, A. 2019. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, O'Reilly Media, Inc.

Grattarola, D and **Alippi, C.** 2020. Graph Neural Networks in Tensorflow and Keras with Spektral. DOI: *https://doi.org/10.1109/MCI.2020.3039072*

IISD. 2021. Covid-19 Wreaking Havoc on Bangladesh's Poor: A Story of Food, Cash, and Health Crises. URL: *https://sdg.iisd.org*.

Kaggle. 2020. Chest x-ray images (pneumonia). URL: *https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia*.

Kharrazi, A. 2017. Challenges and opportunities of urban big-data for sustainable development. *Asia-Pacific Tech Monitor*, 34(4): 17–21.

Krizhevsky, A, Sutskever, I and **Hinton, GE.** 2012. Imagenet classification with deep convolutional neural networks. In: Pereira, F, Burges, CJC, Bottou, L and Weinberger, KQ (eds.), *Advances in Neural Information Processing Systems 25*, 1097–1105. Curran Associates, Inc. URL: *http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf*.

Kruse, CS, Goswamy, R, Raval, Y and **Marawi, S.** 2016. Challenges and opportunities of big data in health care: A systematic review. *JMIR Medical Informatics*, 4(4): e38. DOI: *https://doi.org/10.2196/medinform.5359*

LeCun, Y, Boser, B, Denker, JS, Henderson, D, Howard, RE, Hubbard, W and **Jackel, LD.** 1989. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4): 541–551. DOI: *https://doi.org/10.1162/neco.1989.1.4.541*

LeCun, Y, Jackel, LD, Boser, B, Denker, JS, Graf, HP, Guyon, I, Henderson, D, Howard, RE and **Hubbard, W.** 1989. Handwritten digit recognition: Applications of neural net chips and automatic learning. *IEEE Communication*, 41–46. Invited paper. DOI: *https://doi.org/10.1109/35.41400*

Matheus, R, Janssen, M and **Maheshwari, D.** 2020. Data science empowering the public: Data-driven dashboards for transparent and accountable decision-making in smart cities. *Government Information Quarterly*, 37(3): 101–284. DOI: *https://doi.org/10.1016/j.giq.2018.01.006*

MI. 2021. isdg: Integrayed Simulation Tool. URL: *https://www.millennium-institute.org/isdg*.

Mwitondi, K, Munyakazi, I and **Gatsheni, B.** 2018a. Amenability of the united nations sustainable development goals to big data modelling. *International Workshop on Data Science-Present and Future of Open Data and Open Science*, 12–15 Nov 2018, Joint Support Centre for Data Science Research, Mishima Citizens Cultural Hall, Mishima, Shizuoka, Japan.

Mwitondi, K, Munyakazi, I and **Gatsheni, B.** 2018b. An interdisciplinary data-driven framework for development science. *DIRISA National Research Data Workshop, CSIR ICC*, 19–21 June 2018, Pretoria, RSA.

Mwitondi, K, Munyakazi, I and **Gatsheni, B.** 2020. A robust machine learning approach to sdg data segmentation. *Journal of Big Data*, 7(97). DOI: *https://doi.org/10.1186/s40537-020-00373-y*

Mwitondi, KS, Moustafa, RE and **Hadi, AS.** 2013. A data-driven method for selecting optimal models based on graphical visualisation of differences in sequentially fitted roc model parameters. *Data Science Journal*, 12: WDS247–WDS253. DOI: *https://doi.org/10.2481/dsj.WDS-045*

Mwitondi, KS and **Said, RA.** 2013. A data-based method for harmonising heterogeneous data modelling techniques across data mining applications. *Journal of Statistics Applications & Probability*, 2(3): 293–305. DOI: *https://doi.org/10.12785/jsap/020312*

Mwitondi, KS and **Said, RA.** 2021. Dealing with Randomness and Concept Drift in Large Datasets. *Data*, 6(7). URL: *https://www.mdpi.com/2306-5729/6/7/77*. DOI: *https://doi.org/10.3390/data6070077*

Mwitondi, KS, Said, RA and **Zargari, SA.** 2019. A robust domain partitioning intrusion detection method. *Journal of Information Security and Applications*, 48: 102360. URL: *http://www.sciencedirect.com/science/article/pii/S2214212617305823*. DOI: *https://doi.org/10.1016/j.jisa.2019.102360*

ONS. 2020. Office for national statistics. URL: *https://www.ons.gov.uk/*.

Pan, SL and **Zhang, S.** 2020. From fighting covid-19 pandemic to tackling sustainable development goals: An opportunity for responsible information systems research. *International Journal of Information Management*, 102196. URL: *http://www.sciencedirect.com/science/article/pii/S0268401220311154*. DOI: *https://doi.org/10.1016/j.ijinfomgt.2020.102196*

Pearce, W, Mahony, M and **Raman, S.** 2018. Science advice for global challenges: Learning from trade-offs in the ipcc. *Environmental Science & Policy*, 80: 125–131. URL: *https://www.sciencedirect.com/science/article/pii/S1462901117310298*. DOI: *https://doi.org/10.1016/j.envsci.2017.11.017*

Ramsetty, A and **Adams, C.** 2020. Impact of the digital divide in the age of covid-19. *Journal of the American Medical Informatics Association*, 27: 1147–1148. DOI: *https://doi.org/10.1093/jamia/ocaa078*

Rawat, J, Logofătu, D and **Chiramel, S.** 2020. Factors affecting accuracy of convolutional neural network using vgg-16. In: Iliadis, L, Angelov, PP, Jayne, C and Pimenidis, E (eds.), *Proceedings of the 21st EANN (Engineering Applications of Neural Networks) 2020 Conference*, 251–260. Cham: Springer International Publishing. DOI: *https://doi.org/10.1007/978-3-030-48791-1_19*

Roser, M, Ortiz-Ospina, E, Ritchie, H, Hasell, J and **Gavrilov, D.** 2018. Our world in data: Research and interactive data visualizations to understand the world's largest problems.

Rothan, HA and **Byrareddy, SN.** 2020. The epidemiology and pathogenesis of coronavirus disease (covid-19) outbreak. *Journal of Autoimmunity*, 109: 102433. URL: *http://www.sciencedirect.com/science/article/pii/S0896841120300469*. DOI: *https://doi.org/10.1016/j.jaut.2020.102433*

Said, RA and **Mwitondi, KS.** 2021. An Integrated Clustering Method for Pedagogical Performance. *Array*, 11: 100064. URL: *https://www.sciencedirect.com/science/article/pii/S2590005621000126*. DOI: *https://doi.org/10.1016/j.array.2021.100064*

Tsymbal, A, Pechenizkiy, M, Cunningham, P and **Puuronen, S.** 2008. Dynamic integration of classifiers for handling concept drift. *Information Fusion*, 9(1): 56–68. Special Issue on Applications of Ensemble Methods. URL: *http://www.sciencedirect.com/science/article/pii/S1566253506001138*. DOI: *https://doi.org/10.1016/j.inffus.2006.11.002*

United-Nations. 2015. Sustainable development goals. URL: *https://www.un.org/sustainabledevelopment/sustainable-development-goals/*.

Wang, CJ, Ng, CY and **Brook, RH.** 2020. *Response to covid-19 in Taiwan: Big Data Analytics, new technology, and proactive testing*, 323(14): 1341–1342. DOI: *https://doi.org/10.1001/jama.2020.3151*

Wang, H, Chong, D, Huang, D and **Zou, Y.** 2019. What affects the performance of convolutional neural networks for audio event classification. In: *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 140–146. DOI: *https://doi.org/10.1109/ACIIW.2019.8925277*

WBGroup. 2018. Atlas of sustainable development goals from world development indicators.

Xu, X and **Goodacre, R.** 2018. On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing*, 2(3): 249–262. DOI: *https://doi.org/10.1007/s41664-018-0068-2*

Yan, M, Haiping, W, Lizhe, W, Bormin, H, Ranjan, R, Zomaya, A and **Wei, J.** 2015. Remote sensing big data computing: Challenges and opportunities. *Future Generation Computer Systems*, 51: 47–60. DOI: *https://doi.org/10.1016/j.future.2014.10.029*

Zaki, M and **Mera, W.** 2020. Data Mining and Machine Learning Fundamental Concepts and Algorithms, second edn. Cambridge University Press. DOI: *https://doi.org/10.1017/9781108564175*

Zambrano-Monserrate, MA, Ruano, MA and **Sanchez-Alcalde, L.** 2020. Indirect effects of covid-19 on the environment. *Science of The Total Environment*, 728: 138813. URL: *http://www.sciencedirect.com/science/article/pii/S0048969720323305*. DOI: *https://doi.org/10.1016/j.scitotenv.2020.138813*

Zenisek, J, Holzinger, F and **Affenzeller, M.** 2019. Machine learning based concept drift detection for predictive maintenance. *Computers & Industrial Engineering*, 137: 106031. URL: *https://www.sciencedirect.com/science/article/pii/S0360835219304905*. DOI: *https://doi.org/10.1016/j.cie.2019.106031*

Zhang, P, Xiong, F, Gao, J and **Wang, J.** 2017. Data quality in big data processing: Issues, solutions and open problems. In: *2017 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (Smart-World/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, 1–7. DOI: *https://doi.org/10.1109/UIC-ATC.2017.8397554*

Zhang, R, Jayawardene, V, Indulska, M, Sadiq, S and **Zhou, X.** 2014. A data driven approach for discovering data quality requirements. In: *Proceedings of ICIS – Decision Analytics, Big Data and Visualisation*. URL: *https://aisel.aisnet.org/icis2014/proceedings/DecisionAnalytics/13*.

Zhang, Z, Yang, Y, Xia, X, Lo, D, Ren, X and **Grundy, J.** 2021. Unveiling the Mystery of API Evolution in Deep Learning Frameworks: A Case Study of Tensorflow 2. In: *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, 238–247. DOI: *https://doi.org/10.1109/ICSE-SEIP52600.2021.00033*

Žliobaitė, I, Pechenizkiy, M and **Gama, J.** 2016. An Overview of Concept Drift Applications, Springer International Publishing, Cham, 91–114. DOI: *https://doi.org/10.1007/978-3-319-26989-4_4*

TO CITE THIS ARTICLE:
Mwitondi, KS and Said, RA. 2021. A Framework for Data-Driven Solutions with COVID-19 Illustrations. *Data Science Journal*, 20: 36, pp. 1–21. DOI: *https://doi.org/10.5334/dsj-2021-036*

Submitted: 08 April 2021
Accepted: 26 October 2021
Published: 18 November 2021

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.