Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/160623

Modelling the South African Covid-19 Induced Web Traffic Data Shift using Artificial Neural Networks

*Judah Soobramoney**, *Retius Chifurira and Temesgen Zewotir*

School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, South Africa

Received: 11 Jun. 2022, Revised: 3 Oct. 2022, Accepted: 22 Oct. 2022
Published online: 1 Nov. 2022

Abstract: In response to the coronavirus pandemic of the year 2019 (Covid-19), several governments around the world restricted commercial, economic and socio-economic activities. For emerging markets like South Africa, the pandemic and subsequent restrictions have negatively impacted the financial standing of many sectors, corporates and households. Several corporates that rely on digital technologies such as website marketing experienced a steep decline in traffic flow onto the website and a distinct change in online behaviour. Whilst such change in online behaviour directly impacted the revenue of corporates, such data shifts posed further challenges to the accuracy of machine learning models that were trained on online web behaviour prior to the data shift.
This paper aimed to explore the key features of behavioural change observed on a South African website during the covid-19 pandemic. This research utilises data obtained from the website of an engineering and training corporate in South Africa. Using artificial neural networks, the results indicated that the number of visitors, the sessions per visitor and seasonal period were important indicators of a data shift. However, whilst a drastic drop in volumes were noticed during the data shift, for those that did enter the website, the behaviour remained somewhat stable. The study also found that an artificial neural network was highly capable of detecting a data shift. Whilst the findings of the study are specific to the observed website, the methods applied could be adopted in various other applications.

Keywords: artificial neural network, covid-19, data shift, machine learning models

1 Introduction

To contain the spread of the coronavirus contagion of the year 2019 (Covid-19), the South African government attempted to balance economic activity and individual. In imposing harsh individual restriction, the result would imply unemployment whilst too loose individual restrictions would result in a higher spread of the contagion than the health care systems could support. The South African government approached the balance through a 'risk-adjusted' strategy by introducing lockdown levels ranging from lockdown level 1 to lockdown level 5. Lockdown level 5 allowed minimal economic activity and stronger isolation rules whilst lockdown level 1 permitted most industries to trade with the weakest individual isolation rules. During the peak covid-19 periods, global economic health has been severely impacted across many sectors of industry [1].

This paper sought to investigate the shift in online user behaviour during the peaks of the covid-19 pandemic. The objective of the study was to identify the key factors of change observed on an informative website of an engineering training and engineering service provider. Furthermore, the study proposed an artificial neural network model that can be employed to detect the presence of a future data shift based on the learnings of the Covid-19 induced data shift. In the event of another data shift occurring in the future, without being detected, the data driven business decisions could be misleading. Thereby, it is suggested, that the proposed model be run in parallel with other data-driven tools such as reports and machine learning models that guide business decisions to detect a data shift as soon as it occurs [2].

Online behaviour change has been of interest to many researchers. Rodda et al. (2018) studied online behaviour change to understand problematic gambling. Rodda et al. (2018) concluded that using change strategies to influence online problematic gambling behaviour was very complex and required further research with a broader population base [3].

Perski et al. (2017) studied user engagement with digital behaviour change interventions. Perski et al. (2017) proposed a conceptual framework where the

* Corresponding author e-mail: judahsoobramoney@gmail.com

digital behaviour change intervention was influenced by the intervention itself. Perski et al. (2017) claimed that the context and mechanisms used may moderate the influence on the user's response to the digital behaviour change intervention [4].

Richiello et al. (2022) conducted a study to understand the challenges influencing webchat counselling. Richiello et al. (2022) identified several possible methods to address the challenges and initiate a behaviour change. Richiello et al. (2022) made use of an online behaviour change wheel and embedded surveys to address challenges (for example to prompt online counselling, etc.). However, Richiello et al. (2022) found that low survey take-up rates may imply an impractical approach [5].

The literature below discusses research on shifts in data patterns (data shifts) across several other applications apart from online web data. Furthermore, some researches highlight the danger that data shifts impose on existing machine learning and analytical frameworks.

Kiley and Vaisey (2020) studied population-wide cultural change. The first aspect studied stated that people are actively updating their beliefs and behaviours as they process new information. Whilst the second aspect argued that following early socialization experiences, the dispositions are stable. The data used were sourced from the General Social Survey. The study revealed that whilst short-term change were noticed, persistent change occurred primarily amongst young people [6].

Stacke et al. (2021) have claimed that neural network machine learning models are very accurate within a stable data environment. However, in the presence of a data shift, and unseen data poses a challenge to the generalization of neural networks. The study focused on data shifts within image processing employed within histopathology. The study proposed the use of convolutional neural networks to identify data shifts thereby suggesting potential in-accuracies [7].

Taori et al. (2020) assessed the impact of natural distribution shifts in image data compared to current synthetic distribution variations on predictive model robustness. The findings of the study concluded that distribution shifts that arise from real world data remain an open research problem. This implied that such data shifts do impact machine learning accuracy [8].

In an application of machine learning, to identify new biomarkers, Dockes et al. (2021) have highlighted the impact of data shifts. The study conducted define the breaking point and manner in which machine learning extracted biomarkers fail in the presence of data-shifts [9].

Adams-Cohen (2020) used Twitter data and machine learning methods to study the shift in sentiment when the Supreme Court legalized "same-sex" marriages. The data shift indicated that the Supreme Court's decisions polarize public opinion in the short term [10].

Guo et al. (2020) proposed a method of enhancing neural networks model accuracy in the presence of a data shift. Guo et al (2020) developed a continuous kernel cut segmentation algorithm by factoring in normalized cuts and continuous regularization over the data. The outcome of the study proved that the method reduced segmentation variability and achieved excellent classification accuracy [11].

Xiong et al. (2020) emphasized on the hyper-parameter settings of machine learning models. In particular, Xiong et al. (2020) proposed a protocol for assessing the hyper-parameter sensitivity to data shifts. However, the results of the study indicated that no clear winner could be determined [2].

This study assesses the key features of the covid-19 induced data shift and further proposes an artificial neural network (ANN) model that can be employed to detect the occurrence of a future data shift based on learnings from the covid-19 induced data shift. At the time of the writing, no such literature existed in this context. Although online behaviour during the pandemic would be specific to the nature of the website itself, the approaches employed within this study can nonetheless be adopted on other websites to understand the significance of behavioural changes during the Covid-19 pandemic.

2 Methodology

This paper employed artificial neural network models to understand the data shift in online web behaviour experienced during the Covid-19 outbreak. The models were used to understand the key factors of change detected within the data. Furthermore, a predictive model has been proposed to detect the occurrence of data shift should one occur in the future.

2.1 Artificial Neural Networks

An artificial neural network (ANN) is statistical machine learning algorithm that imitates the process of the human brain. An ANN consists of artificial neurons that exist in several layers. These neurons are linked to each other through a network of connections (or nodes) [12]. The ANN structure processes the data from inputs to an output using biases allocated to the neurons, weights associated to the connectors and an activation function. Figure 1 illustrates a three-layer artificial neural network. As illustrated in Figure 1, the input layer neurons are connected to the hidden layer neurons via weights. These weights dictate the influence that each input layer neuron impose onto the hidden layer [12].

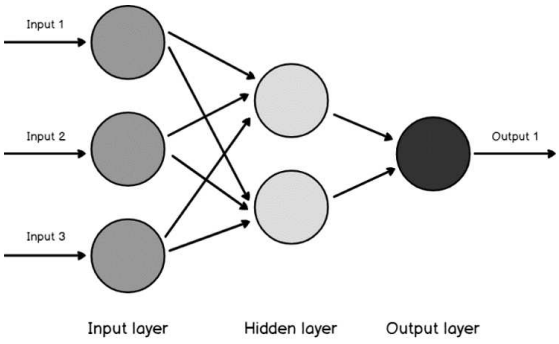

Fig. 1: Artificial neural network architecture.

An ANN is optimized through the number of hidden layers and nodes that the modeller must determine on a given set of data. The ANN is able to determine the nodes that are more important than others by assigning more influential weights to the more important nodes [12].

Whilst artificial neural networks are commonly employed in supervised problems and known to be high performance models, ANN may also be implemented in unsupervised applications [13].

There are several computational algorithms available to construct an ANN (such as multilayer feed forward, feed forward, feedback networks, etc.). However, multilayer feed forward models are considered the most frequently used architecture [13].

Following a general machine learning approach, modelling ANN consists of the collection of inputs and outputs (independent and dependent features), the selection of the model design (feedforward or feed backward), training the model and lastly, testing and validating the model. Depending on the outcome of the model, the process may be re-run by optimizing the hyper-parameters (number of hidden nodes and hidden layers) for optimal model accuracy [12]. Within the learning process of the model, the algorithm initially has a large prediction error when comparing the actual results to the predicted results. However, the system then minimizes the error by adjusting the nodal weights through an iterative process. This iterative process of optimizing the nodal weights is completed through back-propagation between the output nodes and hidden layers of the network [12]. During the training process of the artificial neural network, a validation step is applied to optimize the performance of the model by governing the number of epochs per training cycle. The epochs represent the iteration process used by the model to minimise prediction errors of the model [12].

The final step of the modelling entails testing the models performance on unseen data. By comparing the models prediction to the actual (or desired) results accuracy metrics are quantified [12].

3 Data

The study was conducted for a South African corporate within the engineering and engineering training sector (TEKmation). The underlying data was derived from the Google Analytics web tracking platform. Whilst the observed behaviour of online users were specific to such industry, the methods applied could be adopted in various other applications. The standard tracking platform provided information on the volume of visits to the website, the corresponding engagement whilst on the website, geographic and technology specific information (such as type of device, operating system, etc.) [14].

These online metrics supplied by the Google analytics platform were observed across the South African lockdown levels that influenced the observed data shift. The South African authorities regulated population and industrial behaviour through the implementation of "lockdown levels" at appropriate stages of the pandemic. The lockdown levels, as detailed in Table 1, sought to govern the trade-off between economic activity and the spread of the contagion.

Table 1: High-level description of the level of intensity of the lockdown restrictions.

SA Lockdown Level	**Description**
Level 0 (L_0)	No economic or personal restriction (the period prior to the covid-19 outbreak).
Level 1 (L_1)	Most normal activity can resume, with precautions and health guidelines followed at all times. Population prepared for an increase in alert levels if necessary.
Level 2 (L_2)	Physical distancing and restrictions on leisure and social activities to prevent a resurgence of the virus.
Level 3 (L_3)	High level of precaution and restrictions on personal activities.
Level 4 (L_4)	Extreme precautions to limit community transmission and outbreaks, while allowing some activity to resume.
Level 5 (L_5)	Drastic measures to contain the spread of the virus and save lives.

The South African covid-19 lockdown Level 5 represented the most strict economic and socio-economic restrictions on the country. Level 5 was activated during the early stages of the outbreak whilst emergency hospital facilities were in development and when the spread of the virus was outrageous. And accordingly, between the waves, when the number of active covid-19 cases were minimal, lockdown level 1 was activated at the lowest. It was found that during the harsh levels (4 and 5) of

lockdown, a distinct data shift was observed on the studied data. During each stage of the lockdown levels, relative to the period prior to the covid-19 pandemic, the following measures are observed as detailed in Table 2.

Table 2: Description of observed online variables.

Measure	Description
Bounce rate	The proportion of total visits that leave the website seconds after entering.
DataShift	Binary flag to indicate the covid-19 lockdown levels 4 and 5 when the data shift was experienced.
DesktopRate	The rate of visits from a desktop device.
Hits per session	The number of actions taken by the viewer whilst on the website.
LocalRate	The incidence of visits located within South Africa.
MobileRate	The rate of visits from a mobile device.
NewUserRate	On a given day, the incidence of new users whom there is no evidence of having visited the website before.
Pageviews per session	The number of pages viewed within a visit.
PublicHoliday	A binary flag to indicate South African public holidays.
Session duration	Time spent on the website during a visit.
Users per day	The number of unique viewers that visit the website per day.
Weekday	A binary flag to indicate week days.
YearEnd	A binary flag to indicate the year end period.

Of the observed features, several measures were directly sourced from the Google Analytics tracking tool, whilst some features were implicitly created (for instance, "YearEnd", which flags behaviour during the December holiday period to differentiate between seasonality and a data shift). The "datashift" feature flags the periods of lockdown level 4 and 5 when the data shift was experienced on the observed data.

4 Empirical Results

This section discusses the exploratory analysis of the studied online web behaviour to illustrate the data shift. Furthermore, the outputs of an artificial neural network model and subsequent feature importance are presented.

4.1 Online behaviour

This section discusses key aggregate online metrics between the various lockdown levels to assess the shifts in online behaviour. The analysis focuses on, firstly, traffic to the website and thereafter, the behaviour whilst on the website. Table 3, presents the average number of users visiting the website per day, the average session duration and the average number of pageview per session (by definition, a session refers to a user's visit to a particular website).

Table 3: Aggregate key online metrics at the various lockdown levels.

	Avg. Users Per Day	Avg. Session Duration (s)	Avg. Pageviews Per session
0. Level 0	5.39	387.80	3.32
1. Level 1	4.64	342.95	3.26
2. Level 2	4.97	331.76	3.52
3. Level 3	5.18	287.61*	3.21
4. Level 4	3.57*	298.23*	3.02
5. Level 5	2.83*	358.61*	3.10

From Table 3 for instance, on the observed website, an average of 5.4 users visited the website per day prior to lockdown and spent on average, 388 seconds during a typical session. Whilst during the harshest lockdown level 5, an average of 2.8 users visited the website per day and spent, on average, 359 seconds on the website. Using the Mann-Whitney U two-tailed test for significance (*), lockdown levels 4 and 5 record a significant difference in user visits per day relative to level 0 (at a 5% level of significance). The test for significance further informs that the session duration during levels 3, 4 and 5 are significantly different relative to the Pre-Covid period. Whilst the session duration show statistical significant difference, the extent of the change seems minimal. It is evident that during the harsh levels of lockdown (level 4 and 5), the studied website experienced a drastic drop in traffic flow into the website. However, whilst on the website, user behaviour remained fairly stable across the lockdown (as quantified by the page views and session duration).

4.2 Feature selection

The features initially considered to be explanatory of a data shift were further assessed for correlation between each other to identify redundancy and relevance in terms of distribution to identify features with little variance.

Figure 2 illustrates the correlation matrix of the feature set. According to Figure 2, the darker and larger blue bubbles represent a stronger positive correlation. And conversely, a darker and larger red bubble represents a stronger negative correlation.

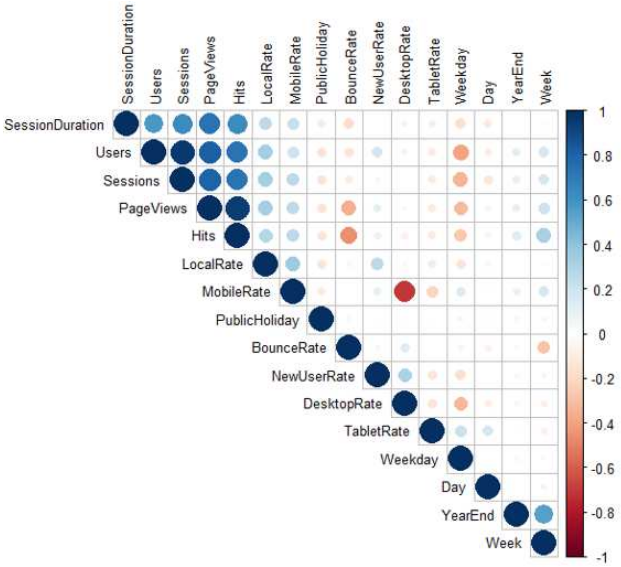

Fig. 2: Correlation matrix between features.

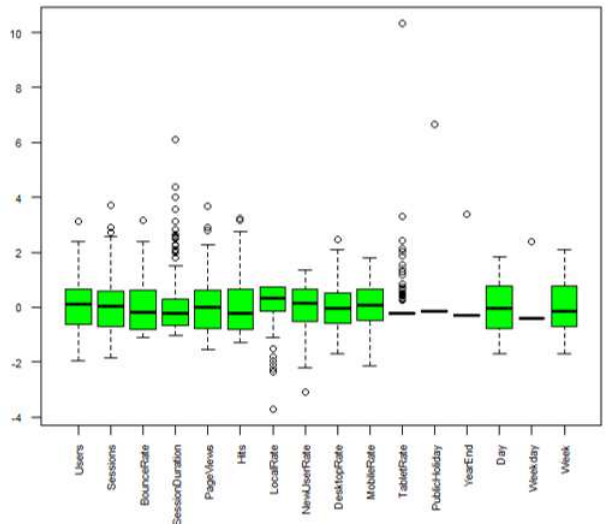

Fig. 3: Box plots of scaled features.

The correlation matrix indicated that the features "hits" and "pageviews" shared a strong positive correlation and hence one had to be omitted. Similarly, "users" and "sessions" share a strong positive relationship as naturally the "user" volume would influence the "session" volumes and due to the strength of the correlation, one ought to be omitted. Whilst "sessionduration" to "pageviews" and "sessionduration" to "hits" share a somewhat high positive correlation, it is likely that this may not always be the case. For instance, there may be users with a high volume of "pageviews", yet a short duration in cases where the users rushed through the website. Hence, to avoid possible loss of information, no omission would be implemented in this regard. A strong negative correlation was detected between "mobilerate" and "desktoprate". This was driven by devices used to access the website being primarily either mobile devices or desktop devices (tablet devices recorded a tiny incidence). Hence, to avoid possible loss of information from tablet device behaviour, both "mobilerate" and "desktoprate" were left within the model. Correlated features are known to hold redundant information and potentially influence model performance [15].

To identify features with no variance, Figure 3 depicts the box-and-whisker plots. Features with no variance ought to be omitted from classification models as such features would be irrelevant explanatory features [15].

Most scalded features included within the model do show variation. The features "Public Holiday", "YearEnd" and "Weekday" are binary flags and thus naturally would not hold much distribution. However, the "TabletRate" feature which quantifies the proportion of web visits that were from a tablet device on a given day shows minimal variation and thus was removed from the model.

Therefore, from the correlation and distribution assessments completed, the features "hits", "users" and "tabletrate" were omitted.

4.3 Artificial neural net

This section discusses the artificial neural network constructed to detect a data shift on the observed data. The proposed model could further be deployed to act as a sensor should a data shift occur on web traffic data in the future. Whilst the model has been trained on the covid-19 induced data shift, it could potentially alarm when data volumes and expected behaviour change drastically which would inherently affect business profitability.

The observed data was randomly split into two subsets, a training dataset (80%) and a test dataset (20%). The training dataset was used to build the artificial neural network whilst the test dataset was used to validate the model accuracy.

The architecture of the network where eleven input layers (implicit and explicit features sourced from Google Analytics Tracking). By design, there was one hidden layer with six nodes and the output layer, was the dependant variable that flagged the days that a data shift occurred (Figure 4).

The dependent variable ("datashift") naturally contained a class imbalanced majority of the data points

representing usual behaviour ("datashift = 0"). This was addressed through random under-sampling to maintain a proportion of 1:2 data shift events to regular events prior to training the ANN models [16]. The purpose of the under-sampling was to avoid a majority bias that would affect the prediction results. The model fitted is illustrated in equation 1:

$$DataShift \sim sessions + bouncerate + sessionduration + pageviews + publicholiday + yearend + localrate + newuserrate + desktoprate + mobilerate + weekday. \quad (1)$$

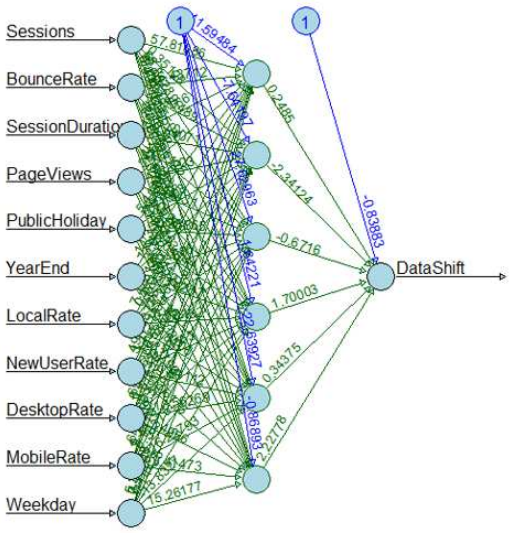

Fig. 4: Constructed artificial neural network.

During the validation step, the model generated from the training dataset was validated against unseen data (the test dataset). In doing so, the model has yielded a 91.07% classification accuracy.

4.4 Feature Importance

With the artificial neural network yielding satisfactory results, the feature importance's are discussed within this section and presented in Figure 5.

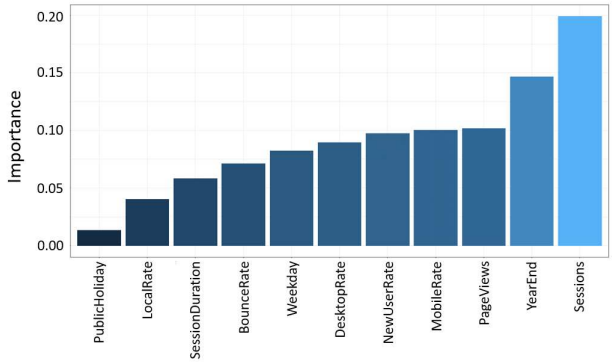

Fig. 5: Feature importance in identifying an online web data shift.

Figure 5 illustrates the feature importance's that emerged whilst building the model. Each feature is assigned a percentage contribution so that the sum of all features sum to one hundred percent. Thus, the higher a features importance percentage, the more important the feature has proven to be. According to Figure 5, the most important features (represented by the tallest bars) are "sessions", "yearend" flag, and "pageviews". The "sessions" and "pageview" features are volume based features that talk to the traffic flow levels on a particular day. The "yearend" feature is merely a flag to indicate if a data-point was during the December holiday period. The important features, thereby indicate that the model was able to accurately quantify the likelihood of a data shift occurring by assessing the volume based features ("sessions" and "pageviews") keeping in mind the time of year. The observed website, was prone to seasonal patterns, and traffic understandably would dip during the December period. Thereby, the "yearend" flag was used to inform the model on when data volumes are expected to drop and not being considered a data shift.

5 Conclusion

In this paper, artificial neural networks were used to model the covid-19 induced data shift in online web traffic on a South African informative website. The artificial neural network model has yielded a high detection accuracy on the event of a data shift even when deciphering between weekends and seasonal holidays (that would naturally result in a change in data patterns). The high detection accuracy yielded by the artificial neural network model, suggested that the model can be employed as a sensor to flag the occurrence of a future data shift. Although the model has been trained on a covid-19 induced data shift, the model may potentially be used to detect data shifts that are driven by other causes.

On the observed data, the website experienced a drastic drop in volume based readings but fairly stable

behaviour for those that have entered the website. Going into the harsh lockdown levels (when the data shift occurred), the observed number of visits and users has dropped significantly. However, for the portion of users that did access the website during the period of the data shift, the observed user behaviour showed somewhat stable behaviour relative to the period prior to the data shift (for instance, somewhat stable session duration and pageviews).

The artificial neural network highlighted that volume based measures (such "sessions" and "pageviews") were important indicators of a data shift occurring whilst bearing in mind the time year. The time of year was an important indicator since volumes would naturally drop during seasonal periods. This study and the corresponding findings were limited to one particular website, future work assessing the online behavioural change across several website spanning several different industries is recommended for future work. Although the techniques applied within this study could be employed elsewhere, such future work would further investigate the data patterns and artificial neural networks performance.

6 Abbreviations

Abbreviation	Definition
ANN	Artificial neural network
SA	South Africa

Acknowledgement

Acknowledgements go to Dr Retius Chifurira for his input on the study. Furthermore, to acknowledge Prof Temesgen Zewotir who served as the high-level supervisor. He was involved on the project at large, offering guidance and criticism. Both Dr Retius Chifurira and Prof Temesgen Zewotir along with Judah Soobramoney have met with the corporate whom own the data to discuss the findings.

References

[1] Ö. Açikgöz and A. Günay, The early impact of the Covid-19 pandemic on the global and Turkish economy, *Turkish Journal of Medical Sciences*, **50**, 520-526 (2020).

[2] Y. Xiong, X. Liu, L. Lan, Y. You, S. Si and C. Hsieh, How much progress have we made in neural network training? A new evaluation protocol for benchmarking optimizers, *ArXiv*, **1**, 1-16 (2020).

[3] S.N. Rodda, N. Hing, D.C. Hodgins, A. Cheetham, M. Dickins and D.I. Lubman, Behaviour change strategies for problem gambling: an analysis of online posts, *International Gambling Studies*, **18**, 420-438 (2018).

[4] O. Perski, A. Blandford, R. West and S. Michie, Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis, *Translational Behavioral Medicine*, **7**, 254-267 (2017).

[5] M.G. Richiello, G. Mawdsley and L.M. Gutman, Using the behaviour change wheel to identify barriers and enablers to the delivery of webchat counselling for young people, *Counselling and Psychotherapy Research*, **22**, 130-139 (2022).

[6] K. Kiley and S. Vaisey, Measuring stability and change in personal culture using panel data, *American Sociological Review*, **85**, 477-506 (2020).

[7] K. Stacke, G. Eilertsen, J. Unger and C. Lundstrom, Measuring domain shift for deep learning in histopathology, *IEEE J Biomed Health Inform*, **25**, 325-336 (2021).

[8] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht and L. Schmidt, Measuring robustness to natural distribution shifts in image classification, *ArXiv*, **2**, 1-69 (2020).

[9] J. Dockes, G. Varoquaux and J. Poline, Preventing dataset shift from breaking machine-learning biomarkers, *ArXiv*, **1**, 1-17 (2021).

[10] N.J. Adams-Cohen, Policy change and public opinion: measuring shifting political sentiment with social media data, *American Politics Research*, **48**, 612-621 (2020).

[11] F. Guo, M. Ng, M. Goubran, S. Petersen, S. Piechnik, S. Neubauer, and G. Wright, Improving cardiac MRI convolutional neural network segmentation on small training datasets and dataset shift: a continuous kernel cut approach, *Medical Image Analysis*, **61**, 101636 (2020).

[12] J.J. Hopfield, Artificial neural networks, *IEEE Circuits and Devices Magazine*, **4**, 3-10 (1988).

[13] A.F. Belhaj, K.A. Elraies, M.S. Alnarabiji, F.A. Abdul Kareem, J.A. Shuhli, S.M. Mahmood and H. Belhaj, Experimental investigation, binary modelling and artificial neural network prediction of surfactant adsorption for enhanced oil recovery application, *Chemical Engineering Journal*, **406**, 127081 (2021).

[14] A. Vecchione, D. Brown, E. Allen and A. Baschnagel, Tracking user behavior with google analytics events on an academic library web site, *Journal of Web Librarianship*, **10**, 1-14 (2016).

[15] T. Kavzoglu and P.A. Mather, The role of feature selection in artifial neural network applications, *International Journal of Remote Sensing*, **23**, 2919-2937 (2002).

[16] J. Johnson and T. Khoshgoftaar, Survey on deep learning with class imbalance, *Journal of Big Data*, **6**, 1-54 (2019).

Judah Soobramoney has more than 10 years of experience in data science and machine learning applications across several industries. At the time of writing, he was employed as a Senior Quantitative Analyst at Nedbank and a PhD candidate at the University of KwaZulu Natal, South Africa.

Retius Chifurira has more than 20 years of teaching experience and has emerged as a researcher in the field of applied statistics. He has been involved in the teaching, and designing of curriculum, at different levels of undergraduate and postgraduate modules. Dr Chifurira has successfully supervised many Masters students. He is a member of the South African Statistical Association.

Temesgen Zewotir is a pioneer and codirector of Data Science Unit for Business and Industry. He has a strong track record in research and postgraduate supervision, consulting and is a nationally- and internationally-recognised expert in statistics methods and applications. For his work in statistical methods and statistical education, Zewotir is an elected member of the International Statistical Institute (ISI), the highest international recognition accorded to established statisticians who have made significant contributions to their profession.