MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220144

Augmentation of Electronic Medical Record Data for Deep Learning

Georgina Kennedy[a], **Mark Dras**[b], **Blanca Gallego**[a]

[a] *Centre for Big Data Research in Health, University of New South Wales, Sydney, NSW, Australia*
[b] *Department of Computing, Macquarie University, Sydney, NSW, Australia*

Abstract

Data imbalance is a well-known challenge in the development of machine learning models. This is particularly relevant when the minority class is the class of interest, which is frequently the case in models that predict mortality, specific diagnoses or other important clinical end-points. Typical methods of dealing with this include over- or under-sampling training data, or weighting the loss function in order to boost the signal from the minority class. Data augmentation is another frequently employed method — particularly for models that use images as input data. For discrete time-series data, however, there is no consensus method of data augmentation. We propose a simple data augmentation strategy that can be applied to discrete time-series data from the EMR. This strategy is then demonstrated using a publicly available data-set, in order to provide proof of concept for the work undertaken in [1], where data is unable to be made open.

Keywords:

Electronic Health Records, Deep Learning.

Introduction

Clinical prediction models frequently target rare endpoints such as mortality within a specific time-frame or other adverse events. This is a known challenge when developing machine learning models [2], as it is easy to over-train to the majority set, producing a classifier of high accuracy, but low utility. In machine learning by gradient descent, the weights of a model are updated based on the overall distance of the model output from the target state using the gradients of a predefined differentiable function. If each training data point contributes equally to this cost function, in a data set with a large imbalance between the majority and minority class the calculation quickly favors accuracy in the majority and will err on the side of under-classifying the class of interest. Under extreme levels of imbalance, this is true even in the instance where there is a strong signal from the minority class.

Data augmentation is an alternative to oversampling, where instead of repeating the same samples exactly, synthetic samples are created and used to expand the dataset more richly than repetition alone. Data augmentation in this context has two goals. If samples belonging to each class are augmented at a rate that is inversely proportional to their imbalance, this has an effect equivalent to an oversampling strategy as described above. In addition to this, it is possible to introduce an element of spatial or temporal invariance that improves the ability of the model to recognise patterns in unseen samples [3]. In an image classification task for instance, one would not want the model to rely on the precise orientation or positioning of the input to be able to detect the presence of the target class. Thus, by repeating each input image with random rotation, scale and skew factors, the model becomes robust in the face of input images that were captured in different contexts. More recently, data augmentation strategies using generative adversarial networks (GANs) have been applied to data from the electronic medical record (EMR) with some success [4], although this brings with it some additional challenges due to the complexity of the implementation and cost of significant additional model training. A GAN uses two models with opposing (adversarial) goals to produce realistic data samples —a generator network that creates synthetic data and a discriminator network that tries to differentiate these synthetic samples from the real data. As the discriminator becomes unable to differentiate between real and generated data samples, these samples are deemed sufficiently realistic, and treated as though they were part of the original dataset. This has been applied with success in medical image analysis [5], which are atypical images in their uniformity of scale and aspect. It is less common in other image domains, likely due to the availability of other more straightforward methods such as applying transformational filters that are not applicable to medical images (a skewed or scaled chest x-ray, for example, loses information that is relevant to the prediction task). It is possible to augment continuous time-series data in an analogous way, where noise can be added and filters applied in order to generate additional training samples that can improve model generalisability [6]. Discrete data, however, are more challenging to modify in this manner, as noise and multiplication factors are meaningless. The problem of finding a generalised solution for discrete, ordered tokens (as found in text or EMR data) is a known challenge [7]. We propose instead, a domain-specific method of augmentation, which makes clinically relevant assumptions about the way data is entered into the source system.

Methods

The source data for this work is an excerpt of the MIMIC-III Clinical Database [8]. This dataset was accessed using the Amazon Web Services Athena Cloud Formation scripts provided by MIT-LCP [9]. Code that builds on these scripts to produce the results in this paper can be found at https://github.com/CBDRH/PaTMan. These models were built using the TensorFlow library [10], version 2.0.

Input data

This dataset contains 61500 ICU admissions across 57773 hospital admissions, belonging to 46646 patients. Hospital admissions without an 'admit' record in the transfers table are excluded, as these represent either newborns, or incomplete records.

As input, we generate predictions only for the first ICU admission in each hospitalisation. Hospitalisations where the patient was discharged to general wards within 6 hours of their index ICU admission and where the patient died within the first 6 hours of their index ICU admission were excluded, leaving a final total of 52770 included index ICU admissions.

Endpoint targets

In order to demonstrate this technique, we selected three prediction targets, each having a differing level of endpoint imbalance (Table 1).

Table 1– Endpoint prevalence

Endpoint	True	False	Imbalance
Death in ICU	3346	49424	0.94
Death in admission	5304	47466	0.90
ICU duration > 7 days	7915	44855	0.85

Tokenisation

Discrete clinical events were gathered for patient demographics, historical admissions, historical diagnoses and historical ICU admissions. Pathology results were converted to discrete tokens according to their decile within all input data i.e. [test type]-[decile], or by [test type] alone for non-numeric results. These tokens were concatenated in as a temporally ordered list, which described the patient trajectory over time, e.g. [Admission, Female, 75, BUN-9, GFR-2, Ultrasound-Kidney, ..., Discharge, N17.8, ..., Admission, Female, 77, ..., ICUAdmission, ...] describes a patient with one prior admission and a history of kidney failure. Each trajectory contains the most recent 500 events that occur prior to prediction time. Diagnoses from the current admission are not included, as coded diagnostic information is not available in real-time. This description of patient trajectories in tokenised form is equivalent to the preprocessing described in [1].

Model architecture

A simple model architecture was implemented, with a small set of hyperparameters tested for each prediction task. Two versions of the model network were implemented with either LSTM or GRU bidirectional recurrent layers of 5, 10 or 15 nodes, in order to observe the robustness of the technique across simple architecture changes. This set of piloted architectures was held the same across all prediction tasks, as the purpose of this work is to demonstrate the effect of the augmentation strategy, rather than to produce the most precisely accurate classifier for each endpoint.

Augmentation strategies

We make a number of assumptions about data within the electronic medical record that allow the creation of augmented samples that can be used to improve model accuracy.

Temporal ordering is of course significant when determining whether or not the patient trajectory is trending towards recovery or deterioration, however it is unlikely to matter at a resolution shorter than one hour in duration. The data entry workflow is not instantaneous, and can be modulated by systems that are outside of the scope of the target patient's condition, e.g. the precise time that a pathology result is returned or manual data entry is completed may be heavily affected by the overall workload of the hospital on a given day. We therefore bucket data into time windows and randomly shuffle events in each of these buckets before reassembling the trajectory in order to increase the number of available samples.

We also assume that the length of available patient history is only somewhat related to patient outcomes. A more complex history of interactions with the healthcare system can be expected to indicate a more severely ill patient, however this dataset was not generated within a closed system of care, and therefore the lack of available history data does not strictly indicate that it does not exist, as patients may have interacted with numerous other providers prior to this admission. Thus, after bucketing and shuffling of data, we randomly truncate patient trajectories by dropping up to one third of the oldest events in each sample.

Finally, clinical data entry is a noisy process, affected by many external forces, and therefore we assume that up to half of each patient trajectory could be randomly masked without changing the clinical interpretation.

By combining these strategies multiple times, we generate additional samples proportional to the input distributions to train each model.

Time to event weighting:

The closer a patient is to time of death when a prediction is made, the more extreme their deterioration risk. Similarly, the longer the eventual ICU admission, the higher impact that early intervention may have on their overall trajectory.

We expect that amplifying the signal for subjects with the strongest evidence of deterioration risk will improve the overall calibration of our models.

For death endpoints, time to event was set to the number of days until death at prediction time and the weighting was inversely proportional to this value (i.e. more repetition of data for subjects with lower time to death). For the long ICU admission endpoint, time to event used was eventual ICU admission duration in weeks, and the weighting was directly proportional (higher repetition for the longest overall ICU admissions).

Data balancing

Five balancing strategies were tested -

1. None: Input data was fed to the model according to the original distribution.
2. Oversampling (simple): Minority class samples were randomly repeated at a rate required to approximately balance the input data.
3. Oversampling (time to event): As per oversampling strategy, except the rate of repetition is instead calculated based on the time to event for the minority class. The total repetition rate is equivalent to the rate for simple oversampling.
4. Augmentation (simple): Minority class samples were randomly augmented (first shuffling, then either truncating or masking). For data augmentation, we augment both majority and minority class samples, holding the ratio of these rates equivalent to the same rate as per simple oversampling.
5. Augmentation (time to event): As per augmentation strategy, weighted based on the time to event for minority class.

Evaluation Framework

Given the rare targets of these prediction models, we follow our previous work in [1] in reporting additional metrics to provide the necessary context that can be obscured by reporting the AUROC in isolation [11]. Specifically we focus on the effect of different training strategies on the workup to detection ratio

(WDR) versus sensitivity, as this gives a concrete measure of the excess workload on clinicians (i.e. how many patients they must assess for each one correctly targeted intervention) as compared to the potential benefit to the patient (i.e. what proportion of truly at-risk patients are correctly highlighted by the model).

In order to combat the known issue of poor calibration of deep learning models [12], we follow the same calibration process demonstrated in [1]. This strategy uses the distribution of predictions generated for a held-out calibration set to establish reference cut-off thresholds that reflect the expected distribution of the target event.

These quantiles are set using a stick-breaking process, which generates 10 thresholds that are then transformed to produce a risk score of between 1 and 10 for each input trajectory. The stick breaking process is defined such that approximately the same proportion of inputs are classified 'high risk' (risk score of 5 or more) as the observed proportion in the calibration set. In practice for the most rare events this makes the high-risk bands very narrow and the low-risk bands quite wide, reflecting the expectation that many more patients will be at low risk of experiencing these rare target events than will be at high risk.

Results

Predictive performance

Figure 1 summarises performance statistics for each model architecture as applied to each of the target endpoints. The AUROC metric (row 1) shows that the original data without any up-sampling applied rapidly fits to the majority class, struggling to capture much of the data signal at all, plateauing with an AUROC of close to 0.5 (where 0.5 is the AUROC for random classification, seen as a diagonal line). Reviewing the precision-recall curve (row 2) in combination with the workup to detection ratio (row 3) shows that for such imbalanced targets, all of the up-sampling techniques improve the performance somewhat, with the augmentation strategies generally outperforming the basic oversampling strategies across all metrics. The alerts per 100 patients versus sensitivity (bottom row) shows that in order to achieve 50% sensitivity, models trained using the original data distribution have to generate alerts for between 30 and 40% of patients, where the augmented data can achieve the same sensitivity while generating alerts for 10% of patients or fewer. For the prediction of death in ICU, time to event augmentation (AUROC=0.83) and basic data augmentation (AUROC=0.82) outperform time to event oversampling (AUROC=0.80) and basic oversampling (AUROC=0.73). Likewise for prediction of in-patient death, time to event and basic augmentation (AUROC=0.82, 0.81 respectively) outperform time to event and basic oversampling (AUROC=0.79, 0.80 respectively).

For the less severely imbalanced prediction task of long ICU admissions this also holds, with time to event (AUROC=0.80) and basic (AUROC=0.81) augmentation showing significant improvement over time to event (AUROC=0.74) and basic (AUROC=0.77) oversampling.

Model calibration

In Figure 2, raw model output from the time-weighted augmentation strategy is compared with predictions that have been calibrated according to the expected target distribution and a more traditional isotonic recalibration technique [13]. In all cases, the distribution-based strategy is much closer to the line showing correctly calibrated risk, however the very low number of positive cases in the calibration set limits its utility for predicting death in ICU across the whole range of probabilities. It does, however, retain its qualities of improved calibration, despite being unable to reach higher levels of confidence.

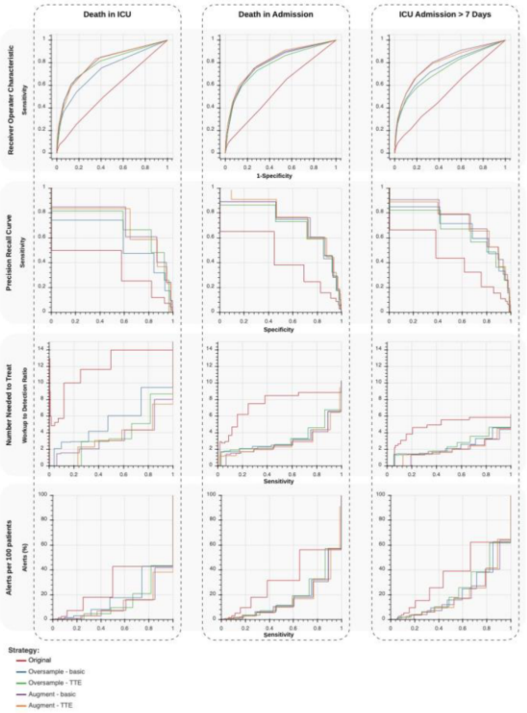


Figure 1 - Comparing model statistics across endpoints and sampling strategies

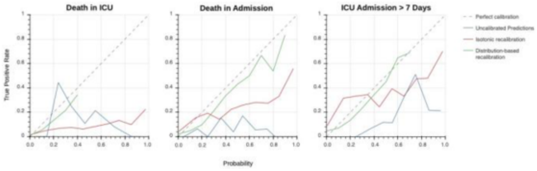
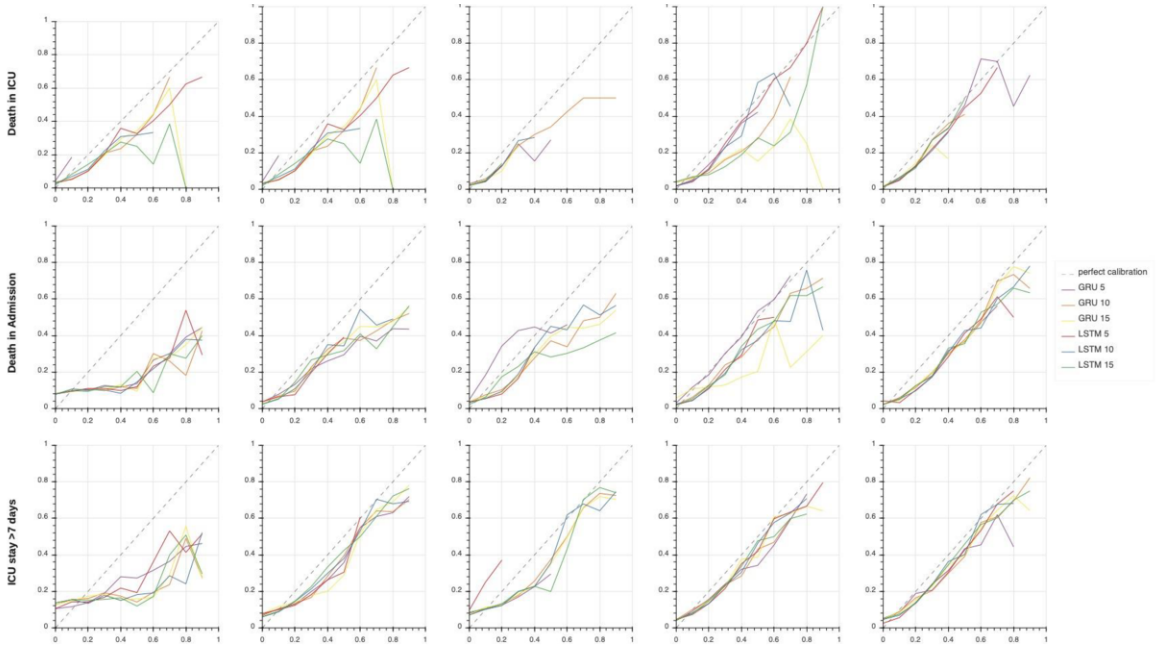


Figure 2 - Effect of different model calibration strategies

Figure 3 compares the calibration curves for each of the piloted architectures. Ideal calibration is shown as a diagonal line. All of the original data distribution training strategies fall significantly below this ideal line, as they fit to the majority class and predict very few patients to be at high risk. The discrimination is poor, as there are a similar number of positive samples within those predicted to be at low risk as those predicted at high risk. Particularly of note are the almost horizontal portions of the graphs below 50% risk for both death in admission and long ICU stay. The combination of time to event sampling and the data augmentation strategy has the most consistently acceptable calibration curves across all endpoints and architectures, meaning that there is less dependence on the model architecture itself, and the signal within the data is captured in a robust fashion.

Figure 3 - Calibration metrics across endpoints and architectures



The LSTM architecture with width of 10 units had the most stable calibration across sampling strategies and end-points, so for the rest of the results section where architectures are not being compared, these are the results reported.

The relative stability of the calibration of models trained on augmented data versus over-sampled data provides evidence that the augmentation strategy described in this work does indeed achieve the stated goal of introducing temporal invariance through the modulation of bucketed event windows.

Time to event weighting strategies

When reviewing models produced under time-to-event weighting, this appears to have a different effect under the over-sampled and augmentation strategies. Applied to augmented data the improved stability of model calibration is quite clear, although the performance across other statistics is similar. This suggests that increasing attention to the most high-risk samples does indeed improve discrimination of patients at most imminent risk of deterioration from those at moderately elevated risk, and is likely to be a better decision with respect to clinical outcomes, rather than attending only to improvements in AUROC.

For over-sampled data under time-to-event weighting, although there is some improvement in discrimination for the high-risk categories, this improvement is less consistent and comes at the expense of a jump in the workup to detection ratio due to an increase in false-positive predictions. This difference may be due to the fact that an augmentation strategy also acts as a sort of model ensembling, as all samples are augmented and therefore repeated multiple times, including those of the negative class. In the test set this means that all samples are repeated the same number of times with the prediction averaged, which can improve model performance in and of itself [14]. In addition, in the training set, those at extreme risk are augmented more frequently than those at elevated risk, but patients with elevated risk will still have significantly more samples than members of the negative class. If we aim to keep the overall distribution steady between oversampling under basic and time-to-event weighting in order to avoid overtraining to the minority class, an increase in the oversampling rate at the extremities will have the effect of decreasing the rate for positive class samples that are at less imminent risk, until they are only very slightly more prevalent than the negative class, and thus their signal is harder to capture.

Discussion

In this work, we do not implement a fully tuned architecture that is targeted to each specific endpoint of interest, as was demonstrated in [1], instead building a very simple, shallow network that can make explicit the effect of manipulating the data sampling strategy alone. In particular, this technique is specific to recurrent data, and thus we do not include the key component of the densely connected sub-model that ingests patient demographic factors. This fact notwithstanding, we still manage to produce a model that can predict half of inpatient deaths as high risk with a workup to detection ratio of 2.4 (for every 5 patients highlighted by the model, on average 2 will in fact die before discharge). Importantly, model calibration is greatly improved through the application of this sampling strategy, in a manner that is robust across different model architectures.

Traditional oversampling methods allow one to boost the signal of the minority class only, with a straightforward copy of each minority class sample. Using an augmentation strategy instead allows for more flexibility, where both the minority and majority class data may be strengthened by resampling each individual patient trajectory in a knowledge-driven fashion in order to create a much richer dataset for both classes. This strategy is common in imaging and continuous timeseries datasets, but the results presented here show that by making certain assumptions about the data collection methodology, it is possible to implement an equivalent strategy in discrete time-series data. This strategy has been designed around assumptions that are relevant to data entry in the electronic health record and proven against that data, however there are many equivalent input token-based datasets that may benefit from such treatment, for example consumer behaviour on websites that can be used to drive recommender systems.

Although generative models have been proposed for the purpose of creating augmented datasets for training models based in EMR data, they typically focus on generating aggregate data [15]. SMOTE is another alternative for adding synthetic data samples of the minority class [16], however this takes as its input tabular data, which limits its applicability to time-series data. Other methods of generating synthetic EMR data are knowledge-based and therefore restricted to specific disease domains [17, 18]. This is the only method to the authors' knowledge that is driven by known factors of the data entry paradigm as opposed to the data itself, and therefore generalizable across all patient classes and robust to unseen combinations of patient characteristics. In addition, this method is computationally and logically inexpensive in comparison to other generative methods. This factor not only reduces the cost of creating the input data (both time and financial), but also increases the applicability of model introspection techniques such as LIME [19] or SHAP [20]. These algorithms for model explainability output the factors of highest importance with respect to a specific prediction, which may be obfuscated by the use of truly synthetic data. By weighting model input according to the time-to-event parameter, we can ensure that risk immanency is captured and thereby robustly improve model calibration.

Conclusions

The pattern of improvement seen from applying the data augmentation strategy described in this work is conclusive – improving prediction results across the board for three distinct end-points, each with a different level of data imbalance. Time to event sampling improves model calibration for all endpoints, although its effect on other metrics is less consistent.

References

[1] G. Kennedy, J. Rihari-Thomas, M. Dras, and B. Gallego, "Developing a deep learning system to drive the work of the critical care outreach team," medRxiv, 2020.

[2] P. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modeling on imbalanced domains," ACM Computing Surveys (CSUR), vol. 49, no. 2, p. 31, 2016.

[3] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," Journal of Big Data, vol. 6, no. 1, p. 60, 2019.

[4] C. Elisa et al., "Generative adversarial networks applied to observational health data," arXiv preprint arXiv:2005.13510, 2020.

[5] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," Medical image analysis, vol. 58, p. 101552, 2019.

[6] A. Le Guennec, S. Malinowski, and R. Tavenard, "Data augmentation for time series classification using convolutional neural networks," 2016.

[7] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," arXiv preprint arXiv:1901.11196, 2019.

[8] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," Scientific data, vol. 3, p. 160035, 2016.

[9] MIT Laboratory for Computational Physiology, "MIMIC source code." https://github.com/MIT-LCP/mimic-code/tree/master/buildmimic/aws-athena. [accessed 1-Feb-2020].

[10] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," arXiv preprint arXiv:1603.04467, 2016.

[11] T. Raeder, G. Forman, and N. V. Chawla, "Learning from imbalanced data: Evaluation matters," in Data mining: Foundations and intelligent paradigms (D. E. Holmes and L. C. Jain, eds.), pp. 315–331, Berlin, Heidelberg: Springer, 2012.

[12] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proceedings of the 34th International Conference on Machine Learning- Volume 70, pp. 1321–1330, JMLR. org, 2017.

[13] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 694–699, 2002.

[14] C. Ju, A. Bibaut, and M. van der Laan, "The relative performance of ensemble methods with deep convolutional neural networks for image classification," Journal of Applied Statistics , vol. 45, no. 15, pp. 2800–2818, 2018.

[15] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, and J. Sun, "Generating multi-label discrete patient records using generative adversarial networks," arXiv preprint arXiv:1703.06490 , 2017.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of artificial intelligence research, vol. 16, pp. 321–357, 2002.

[17] J. Walonoski, M. Kramer, J. Nichols, A. Quina, C. Moesel, D. Hall, C. Duffett, K. Dube, T. Gallagher, and S. McLachlan, "Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record," Journal of the American Medical Informatics Association, vol. 25, no. 3, pp. 230–238, 2018.

[18] S. McLachlan, K. Dube, and T. Gallagher, "Using the caremap with health incidents statistics for generating the realistic synthetic electronic healthcare record," in 2016 IEEE International Conference on Healthcare Informatics (ICHI), pp. 439–448, IEEE, 2016.

[19] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier," CoRR , vol. abs/1602.04938, 2016.

[20] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in neural information processing systems, pp. 4765–4774, 2017.

Acknowledgements

This research is supported by an Australian Government Research Training Program (RTP) Scholarship.

The authors declare that they have no competing interests.

Address for correspondence

Corresponding author: georgina.kennedy@unsw.edu.au