DeepClean -- self-supervised artefact rejection for intensive care waveform data using deep generative learning

Tom Edinburgh[1]*

Peter Smielewski[2]

Marek Czosnyka[2]

Manuel Cabeleira[2]

Stephen J. Eglen[1]

Ari Ercole[3]

*Corresponding author, email: te269@cam.ac.uk

[1] Department of Applied Mathematics and Theoretical Physics, University of Cambridge, United Kingdom
[2] Brain Physics, Department of Clinical Neurosciences, University of Cambridge, United Kingdom
[3] Division of Anaesthesia, Department of Medicine, University of Cambridge, United Kingdom

Summary

Waveform physiological data is important in the treatment of critically ill patients in the intensive care unit. Such recordings are susceptible to artefacts, which must be removed before the data can be re-used for alerting or reprocessed for other clinical or research purposes. Accurate removal of artefacts reduces bias and uncertainty in clinical assessment, as well as the false positive rate of intensive care unit alarms, and is therefore a key component in providing optimal clinical care. In this work, we present DeepClean; a prototype self-supervised artefact detection system using a convolutional variational autoencoder deep neural network that avoids costly and painstaking manual annotation, requiring only easily-obtained `good' data for training. For a test case with invasive arterial blood pressure, we demonstrate that our algorithm can detect the presence of an artefact within a 10-second sample of data with sensitivity and specificity around 90%. Furthermore, DeepClean was able to identify regions of artefact within such samples with high accuracy and we show that it significantly outperforms a baseline principle component analysis approach in both signal reconstruction and artefact detection. DeepClean learns a generative model and therefore may also be used for imputation of missing data.

Introduction

Critically ill patients in the intensive care unit (ICU) may experience life-threatening deterioration sometimes over minutes or even seconds. Care of these patients is therefore highly dependent on data [1], which, by necessity, may be voluminous and complex. An extreme example of this are high frequency physiological waveforms, such as invasive arterial blood pressure (ABP), the electrocardiogram and other vascular or intracranial pressures, which provide a wealth of complex, heterogeneous yet highly structured data at optimal sampling frequencies of 100Hz or even more. Continuous monitoring of these waveforms form the basis of clinical research and care, in particular alerting clinicians to changes in the patient state in real time, and they may be further processed to derive other useful parameters; for example, measures for tracking cerebral autoregulation [2] and heart-rate variability [3], which has repeatedly been shown to be a determinant of physiological integrity. Further, more sophisticated metrics of these waveforms, based on non-linear dynamics [4] or information theory [5], may form novel digital biomarkers for precision care in the future (for example, [6]).

Waveform artefacts arise from a variety of internal and external sources, such as sensor noise, patient movement and clinical intervention. Such artefacts may reduce reliability in estimation of derived indices, confound analysis and create uncertainty in clinical decision making. Closely related, in the same manner, is the handling of missing data, often treated using simple methods that are biased and underestimate variability, such as linear interpolation of observed data. Imputation methods that maintain some statistical properties or features of the data offer a useful alternative [7]. Artefact detection is also important for ICU alerting systems [8] and reduces the likelihood of false positive alarm incidents. A high rate of false

alarms carries a significant risk as alarm fatigue often leaves clinical staff perceiving the alarms to be generally unhelpful and can therefore potentially result in delays to the appropriate clinical intervention [9].

Artefact detection has traditionally been a difficult and costly task, requiring time-consuming human annotation or thresholding based on signal-specific feature engineering [10]. Due to the complex morphologies of waveforms, annotation by experienced clinicians remains a gold standard for ICU multimodality monitoring, despite inherent biases and issues with replicability. Many standard supervised learning methods require samples that are annotated in this way in order to learn a model, though a recent study uses active learning to query and propose samples for annotation in an efficient manner [11]. An alternative unsupervised approach has foundations in spectral anomaly detection [12]. These methods seek an embedding of the data in an `information bottleneck' lower-dimensional space where the anomaly or artefact is separated from the normal data, and then form a reconstruction of the input from its lower-dimensional representation back to the original input space. Embedding in the `bottleneck' latent space is a lossy transformation and the aim is to capture salient features of the data whilst disregarding anomalous features. Subsequently, the reconstruction should restore the underlying `true' behaviour and the error in the reconstruction, compared to the input, can then be used to discriminate artefacts.

Generative modelling describes a class of models in which we want to learn a distribution $p_\theta(x)$ to approximate the true distribution of data X in a dataset. We can sample directly from this distribution to generate new example data that captures some statistical properties and features of the data. A subset of this are latent variable models, which condition on unobserved variables or features z, with a family of deterministic functions $f(z;\theta)$ mapping z to the data space via a decoder network, which is parameterised by θ. The aim is maximum likelihood estimation of the marginal (model evidence) $p(x) = \int p(x,z)dz = \int p_\theta(x|z)p(z)dz$ with respect to θ, but this is often intractable. Variational autoencoders (VAEs) [13] resolve this using variational Bayesian inference, reproducing this `bottleneck' architecture (Figure 1) via coupled but distinct encoder and decoder modules, where the encoder describes a variational distribution $q_\phi(z|x)$ that approximates the true posterior $p(z|x) = p(x|z)p(z)/p(x)$. During training, it learns both to encode some information about each data point in its latent representation $z \sim q_\phi(z|x)$ and stochastic sampling from this distribution allows the model to abstract and generate new candidate data $x \sim p_\theta(x|z)$ that accurately reconstructs inputs from the assumed underlying random process. This enables it to discriminate artefacts despite never being explicitly exposed to any, since we assume different latent mechanisms govern the waveform behaviour of any artefacts, and so model prediction over an input containing such an artefact will have low reconstruction probability and high reconstruction error [14]. For further detail, comprehensive summaries of the VAE can be found in the literature e.g. [15], [16], [17].

In this work, we have demonstrated the potential of deep generative learning, in particular VAEs, in the detection of artefacts in physiological waveforms, developing a VAE-based model and pipeline, which we name DeepClean. We

investigated whether, by training DeepClean on substantially clean ABP recordings as an example, regions of artefactual data could be identified in a self-supervised way without explicitly showing DeepClean any artefact exemplars. Furthermore, we aimed to see whether the DeepClean generative model could impute missing data over such regions. This work is a shortened version of a preprint we archived recently [18]; please refer to this for a fuller discussion.

Methods

Fully anonymous arterial blood pressure (ABP) data from a standard indwelling arterial line connected to a pressure transducer (Baxter Healthcare Corp. CardioVascular Group, Irvine, CA) was obtained as part of routine ICU clinical care. The data used in this study is ABP waveform from a single anonymised adult patient monitored almost continuously throughout a stay of several days in the ICU. The signal was sampled at frequency of 125Hz using the ICM+ software (Cambridge Enterprise Ltd., Cambridge, UK, http://icmplus.neurosurg.cam.ac.uk). Under UK regulations, ethical approval was not required for the re-use of anonymous data obtained as part of routine clinical practice for research.

DeepClean was trained and tested on 10-second windows of preprocessed data. In order to learn a generative model describing `clean' physiological waveforms, we first removed large, grossly abnormal sections by applying basic thresholding heuristics (for further detail, see [18]), to obtain substantially `clean' training data without human mark-up at a beat-to-beat level. This is a far easier task than manual markup of the waveform but inevitably leaves artefacts in the training set. Removing this preprocessing step doesn't prevent the learning of a generative model but without it training can be more difficult and the final generative model sub-optimal, as such training samples will have a larger contribution to the overall loss function. Though DeepClean is self-supervised, we required a labelled test set to determine model performance against manual expert mark-up. We selected this test set randomly from the unprocessed data, to avoid potential sources of selection bias, but with higher probability of selecting regions marked as abnormal in preprocessing, as we assumed waveform artefacts were rare compared to normal beats and we required a balanced test set, with a similar proportion of artefacts and valid signals (again, see [18]). The test set and preprocessing-marked regions were then removed from the dataset and the remaining data split into 10-second samples, shuffled, and divided into training and validation sets with ratio 9:1. Training, validation and test sets were then standardised by the training set mean and standard deviation. Each 10-second test sample was labelled as either containing an artefact or being artefact free by expert review (TE, AE). Furthermore, for samples judged to contain an artefact, sections of the waveform *within* each 10-second window were also annotated. We also assessed performance by comparison to principal component analysis (PCA), which performs a similar dimensionality reduction and reconstruction. The number of principal components is comparable to the latent dimension of a VAE, so we refer to both as the latent dimension for brevity.

We built a variational autoencoder with deep convolutional neural networks (CNNs) for both encoder and decoder modules. CNNs allow the network to learn translation-invariant local patterns, with successive layers building a spatial

hierarchy of increasing scale, a sensible approach in this context as physiological signal data is quasi-periodic and highly-structured. We fixed an encoder network architecture with three relatively small convolutional layers, alternated with two pooling layers to increase the receptive field, and finally two dense layers that split the network graph into separate branches for the two parameters that define the latent space. The decoder architecture mirrored this in reverse, with pooling replaced by up-sampling. Both encoder and decoder contained approximately 20,000 trainable parameters. We then trained separate models with increasing latent space dimension. For each latent dimension, we repeated training five times, presenting the model that minimised the validation loss. To train DeepClean, we used an NVidia Pascal P100 GPU, with code written in python using the Keras library [19] with TensorFlow backend [20]. We optimised the evidence lower bound (ELBO) objective [13], [15], [17] and made typical assumptions about the distributional families of the prior, variational and conditional.

We assessed the performance of our approach in detecting whether a 10-second sample of data contained an artefact or not via the mean squared error (MSE) between the sample and its reconstruction, with a suitable threshold. The goal of this work was to develop a `blind' self-supervised classification procedure, and so any threshold on a metric applied to the reconstruction error must therefore be chosen independently of the test data. Figure 2 illustrates that values of metrics are generally not independent of hyperparameter choice. Therefore, without prior knowledge of a suitable threshold, a pragmatic approach was to set a threshold based on the 90^{th} percentile of the same metric calculated instead on the training data and their corresponding reconstructions. Note, by decreasing the threshold on a metric, we classify more samples as artefacts, regardless of the correctness of this classification, and will therefore increase the specificity at the cost of decreasing the sensitivity (Figure 2). We assessed the model performance by comparing the DeepClean classification to our annotation, using measures widely employed in clinical settings: accuracy, sensitivity and specificity. With samples identified as artefacts, we can then use a metric on subsets of the sample (such as a sliding window approach) to identify artefacts more precisely within these samples [18].

Results

We work with 486,984 seconds of ABP waveform data from a single anonymised adult patient. We marked 11,082 seconds of 486,984 seconds (2.28%) as grossly abnormal in preprocessing, leaving training and validation sets of 37,821 and 4,728 10-second samples. Training required under 10 minutes of computation time on average, although this increased with latent dimension. Subsequent prediction, on test data or on new data, is inexpensive and requires only milliseconds. Our annotation marked 130 test samples out of a test set of 200 as containing an artefact.

DeepClean was able to much more accurately reconstruct the waveform inputs than PCA. In particular, DeepClean was able to encode sub-pulse components, such as the dicrotic notch, even for minimal latent dimensions. The PCA reconstruction is particularly poor for a small number of principal components but is much better for larger latent dimensions

(as seen in training set log-MSE values in Figure 2). DeepClean substantially outperformed the baseline PCA in sample-wide artefact detection in model sensitivity (Table 1), using our training-set defined threshold. Whilst we give a heuristic for identifying this threshold, DeepClean also had a significantly higher receiver operating characteristic area under the curve (ROC AUC) than PCA. There is in general a clear distinction between the log-MSE of artefacts and valid data for the DeepClean reconstruction, since DeepClean is able to distinguish samples similar to and unlike those belonging to the underlying generative process of the waveform from the training set data. In contrast, the range of log-MSE values for PCA reconstructions were similar for both training set and test set artefacts, so the latter cannot be easily identified in this way. In both cases, the non-artefact test samples followed a similar distribution to the training data, so the specificity was generally be close to the training set 90^{th} percentile threshold. As PCA reconstructions have higher error than DeepClean unless the number of principal components is very large, the threshold was different for each method and so an identical reconstruction may be identified differently by DeepClean and PCA. In particular, a poor DeepClean reconstruction of a sample containing an artefact may be identified as such when an identical reconstruction from PCA may not.

Discussion

We have demonstrated the use of a VAE to clean ABP signals in a self-supervised manner, suggesting a clear potential role for deep generative learning in this clinical application. DeepClean requires only two basic steps alongside jointly training encoder and decoder neural networks: straightforward thresholding heuristics for preprocessing, which may even be omitted at the cost of a slightly weaker model, and a metric or decision rule for discriminating artefacts based on the reconstruction error, which may include learning an acceptance threshold from the training data based on an approximate target false negative rate. Unlike a recent study [21], which employed a stacked convolutional autoencoder (SCAE) combined with a CNN, our approach does not require pulse pre-segmentation (which is in itself a difficult task requiring heuristics) or a supervised final classification network. Furthermore, since DeepClean functions in the time domain, it avoids the quadratic complexity of first mapping the pulse morphology to a 2-dimensional space.

We decided to split the data into 10-second samples for training and prediction, since such a segment will typically contain a small number of beats and we do not expect physiology to vary grossly over this period from clinical experience. A sample of this length should include a sufficient number of beats that the model can learn and generalise the structure of the waveform, and therefore distinguish artefacts that occur over similar or longer timescales. It is important that the model recognises clinical events as part of the same generative process, and so does not classify these as artefacts. Failure to flag a clinical event that has been incorrectly identified as a signal artefact risks delayed intervention and treatment. Whilst such events are characterised by sudden changes across multiple frequency bands concurrently, the structure of the waveform during a clinical event is largely retained within a sample of this length, so it is a suitable choice.

Artefact detection is often only the first of two parts involved in handling invalid data, as simply removing artefacts creates missing data that may also bias further analysis. One major advantage of a generative model is that we can generate realistic, synthetic data by sampling directly from the latent distribution. An obvious solution then is to replace an artefact sample with its reconstruction, and further, we can also set missing data to a fixed non-viable value and treat this as an artefact similarly. This may only be suitable for a given sample if, for any artefacts within that sample, the reconstruction can approximate (with large enough reconstruction error such that it is classified as an artefact) the underlying waveform behaviour that would be expected in the absence of the artefact mechanism, whilst simultaneously maintaining a small error for any non-artefact regions in the same sample (e.g. Figure 3, left and third from top). The latent representation, described by the variational distribution, $q_\phi(z|x)$, of an artefact gives some hint as to when DeepClean is successful at this. In imposing the generative model on new data that does not come from the same process or mechanism as the training data, the probability that an artefactual sample is generated by latent variables that are similar to those of valid data is very low and, as a result, the artefactual sample is forced to have a latent representation with vanishingly small probability mass in the average encoding distribution or the prior. As the model has spent little or no time training this region of the latent space, the reconstruction is therefore very unlike that of any valid data, and we cannot impute using it. Density-based anomaly detection methods, which identify anomalies by the sparsity of the region of space in which they occupy, may be useful to recognise artefacts for which this is the case. An alternative is to track the trajectory of a patient in the latent space and sample at points close to this trajectory when an artefact has been identified, but further work is needed to understand the structure of the latent space. For example, β-VAEs [22], [23] can improve the ability of the network to disentangle meaningful features in the data via constrained optimisation.

We have considered expert manual mark-up for `ground truth' artefact assignment. Some artefacts are clearly identifiable to clinicians because their signal profile is so extreme. For example, blood sampling or subsequent flushing of the arterial line, which may be very variable in profile but will clearly contain signal excursions which are unphysiological. There are other signals that are not so clear-cut to categorically assign as artefact. Patient movement may introduce vibration, which in the extreme may render the signal unusable but in less severe cases often simply decreases the signal-to-noise ratio; whether such signals are identified as artefacts is to some extent arbitrary. Changes in the resonant properties of the ABP transduction system due to blood clots or bubbles represent another difficulty. This may occur as a result of `over-damping' (mean pressure preserved) or attenuation [24] (mean pressure not preserved): in either case the pulse amplitude is reduced and high-frequency features are lost. Since the presence of high frequency features varies with cardiac output, it is impossible to absolutely ascribe such regions as valid or otherwise. As a result of these considerations, a good gold standard is lacking and therefore imposes limitations on our ability to rigorously evaluate model performance.

We have focused little on optimising hyperparameter and architecture choices, using deep CNNs for both encoder and decoder that we have deliberately kept small relative to the deep generative learning literature. However, we expect the gains from increasing the network size or complexity, for example, to be marginal in this application domain. Indeed, increasing the learning capacity of the decoder may mean that the latent variables are ignored and do not encode information about the data. One hyperparameter we have investigated is the latent dimension. This is particularly interesting in view of unsupervised representation learning, provided we are able to learn disentangled interpretable features encoded in the latent space.

We chose the ABP signal as a test case for several reasons. Firstly, it is particularly artefact-prone due to the effects of movement and flushing on the fluid filled catheter system typically employed. Secondly, ABP is of universal physiological importance: especially the extremes of ABP (hypo- and hypertension), which are particularly influenced by artefacts. Finally, ABP morphology tends to have good signal to noise properties. The performance of our model with other types of physiological signal is an area of future investigation and optimisation. We have trained and tested our system on available data from a patient in sinus rhythm which is strongly periodic over the sample lengths considered. The performance of this framework on data from patients with more irregular (e.g. patients with arrhythmias, such as atrial fibrillation or multiple ectopics) has not been determined and will be the subject of future work. It is important to establish whether a generative model trained on the waveforms from one patient is able to generalise to other patients and maintain similar performance. If retraining is required, it is possible to initialise from a previous model and so not necessary to train from scratch. Such transfer learning between patients either requires identical recording frequencies, which is common with standardised monitoring software, or an interpolation scheme such that the length of samples corresponding to a fixed time interval is constant.

This work offers a promising alternative to the identification of signal artefacts in physiological waveforms from intensive care multimodality monitoring. This study proposes DeepClean, a self-supervised deep generative learning approach to identify and reject these artefacts. This method demonstrated high sensitivity and specificity for identification of samples containing an artefact from an ICU ABP waveform. High-frequency waveforms are central to clinical prognosis of the patient state, and clinical parameters associated with outcome are derived from features of the signals. Analysis that may inform care and treatment is susceptible to biases that arise from unidentified signal artefacts, and subsequent potential misdiagnosis of clinical events could result in aggressive yet unnecessary clinical intervention. Further, by removing artefacts within the signal, an improvement in the prognostic ability of a real clinical event will reduce the high false positive rate of alarms in ICU monitoring systems, improving the conditions in which clinical staff can provide the best possible care for

patients. Real-time identification of artefacts and signal irregularities is absolutely critical and DeepClean suggests that deep generative learning can have an important role in this task.

Data and code availability

The data used in this work were recorded as part of routine clinical care and not available for open access. However, the authors will consider requests for collaborative research projects. The code used in this study is available at https://github.com/tedinburgh/deepclean.

Acknowledgements

Jim Stone for helpful discussion. TE is funded by Engineering and Physical Sciences Research Council (EPSRC) National Productivity Investment Fund (NPIF), reference 2089662, and Cantab Capital Institute for Mathematics of Information.

References

[1] J.-L. Vincent, "The coming era of precision medicine for intensive care," *Crit. Care*, vol. 21, no. Suppl 3, p. 314, Dec. 2017. (DOI:10.1186/s13054-017-1910-z)

[2] M. J. H. Aries *et al.*, "Continuous determination of optimal cerebral perfusion pressure in traumatic brain injury*," *Crit. Care Med.*, vol. 40, no. 8, pp. 2456–2463, Aug. 2012. (DOI:10.1097/CCM.0b013e3182514eb6)

[3] S. N. Karmali, A. Sciusco, S. M. May, and G. L. Ackland, "Heart rate variability in critical care medicine: a systematic review," *Intensive Care Med Exp*, vol. 5, no. 1, p. 33, Dec. 2017. (DOI:10.1186/s40635-017-0146-1)

[4] S. M. Bishop, S. I. Yarham, V. U. Navapurkar, D. K. Menon, and A. Ercole, "Multifractal analysis of hemodynamic behavior: intraoperative instability and its pharmacological manipulation," *Anesthesiology*, vol. 117, no. 4, pp. 810–821, Oct. 2012. (DOI:10.1097/ALN.0b013e31826a4aa2)

[5] L. Gao, P. Smielewski, M. Czosnyka, and A. Ercole, "Early Asymmetric Cardio-Cerebral Causality and Outcome after Severe Traumatic Brain Injury," *J. Neurotrauma*, vol. 34, no. 19, pp. 2743–2752, Oct. 2017. (DOI:10.1089/neu.2016.4787)

[6] E. Beqiri *et al.*, "Feasibility of individualised severe traumatic brain injury management using an automated assessment of optimal cerebral perfusion pressure: the COGiTATE phase II study protocol," *BMJ Open*, vol. 9, no. 9, p. e030727, Sep. 2019. (DOI: 10.1136/bmjopen-2019-030727)

[7] A. M. Sullivan, H. Xia, J. C. Mc Bride, and X. Zhao, "Reconstruction of Missing Physiological Signals Using Artificial Neural Networks," *Comput. Cardiol.*, vol. 37, pp. 317–320, 2010.

[8] F. Scalzo and X. Hu, "Semi-supervised detection of intracranial pressure alarms using waveform dynamics," *Physiol. Meas.*, vol. 34, no. 4, pp. 465–478, Apr. 2013. (DOI:10.1088/0967-3334/34/4/465)

[9] M. C. Chambrin, "Alarms in the intensive care unit: how can the number of false alarms be reduced?," *Crit. Care*, vol. 5, no. 4, pp. 184–188, Aug. 2001. (DOI:10.1186/cc1021)

[10] J. X. Sun, A. T. Reisner, and R. G. Mark, "A signal abnormality index for arterial blood pressure waveforms," in *2006 Computers in Cardiology*, 2006, pp. 13–16. (DOI:10.1097/CCM.0b013e3181930174)

[11] M. Megjhani *et al.*, "An active learning framework for enhancing identification of non-artifactual intracranial pressure waveforms," *Physiol. Meas.*, vol. 40, no. 1, p. 15002, Jan. 2019. (DOI:10.1088/1361-6579/aaf979)

[12] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 15:1--15:58, Jul. 2009. (DOI:10.1145/1541880.1541882)

[13] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv*, Dec. 2013. https://arxiv.org/abs/1711.00464.

[14] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," *Spec. Lect. IE*, vol. 2, pp. 1–18, 2015.

[15] J. V Stone, *Artificial Intelligence Engines: A Tutorial Introduction to the Mathematics of Deep Learning*. Sebtel Press,

2019.

[16] D. P. Kingma and M. Welling, “An Introduction to Variational Autoencoders,” *arXiv*, Jun. 2019. https://arxiv.org/abs/1906.02691.

[17] C. Doersch, “Tutorial on Variational Autoencoders,” *arXiv*, Jun. 2016. Preprint at https://arxiv.org/abs/1606.05908.

[18] T. Edinburgh, P. Smielewski, M. Czosnyka, S. J. Eglen, and A. Ercole, “DeepClean - self-supervised artefact rejection for intensive care waveform data using generative deep learning,” *arXiv*, Aug. 2019. https://arxiv.org/abs/1908.03129.

[19] F. Chollet and Others, “Keras.” 2015.

[20] M. Abadi *et al.*, “TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems,” *arXiv*, Mar. 2016. Preprint at https://arxiv.org/abs/1312.5663.

[21] S.-B. Lee *et al.*, “Artifact removal from neurophysiological signals: impact on intracranial and arterial pressure monitoring in traumatic brain injury,” *J. Neurosurg.*, pp. 1–9, May 2019. (DOI:10.3171/2019.2.JNS182260)

[22] I. Higgins *et al.*, “beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework,” *ICLR*, vol. 2, no. 5, p. 6, 2017.

[23] D. J. Rezende and F. Viola, “Taming VAEs,” *arXiv*, Oct. 2018. Preprint at https://arxiv.org/abs/1810.00597.

[24] A. Ercole, “Attenuation in invasive blood pressure measurement systems,” *Br. J. Anaesth.*, vol. 96, no. 5, pp. 560–562, May 2006. (DOI:10.1093/bja/ael070)

[25] M. D. Hoffman and M. J. Johnson, “ELBO surgery: yet another way to carve up the variational evidence lower bound,” in *{NIPS} Workshop in Advances in Approximate Bayesian Inference*, 2016.

Figures

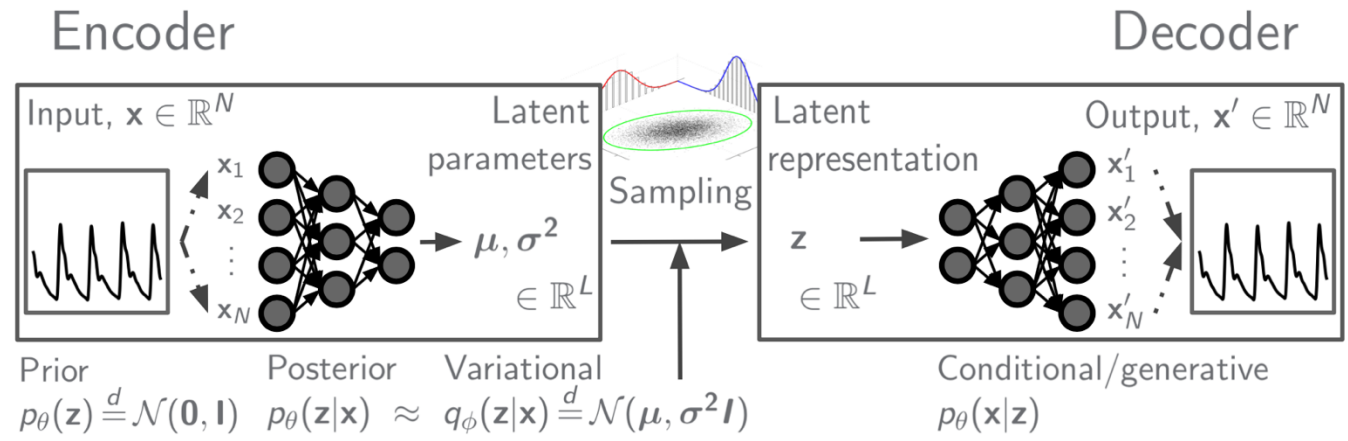

Figure 1: A variational autoencoder learns a generative model, $p_\theta(x|z)$, and a variational approximation, $q_\phi(z|x)$, to the posterior distribution for the latent variables, $p_\theta(x|z)$, and we can generate new data by sampling from the latent and generative distributions in order. The weights of the encoder are the variational parameters, ϕ, are the weights of the decoder are the generative parameters, θ. The output of the encoder acts as a mean and standard deviation to the Gaussian variational distribution. Figure also in [18].

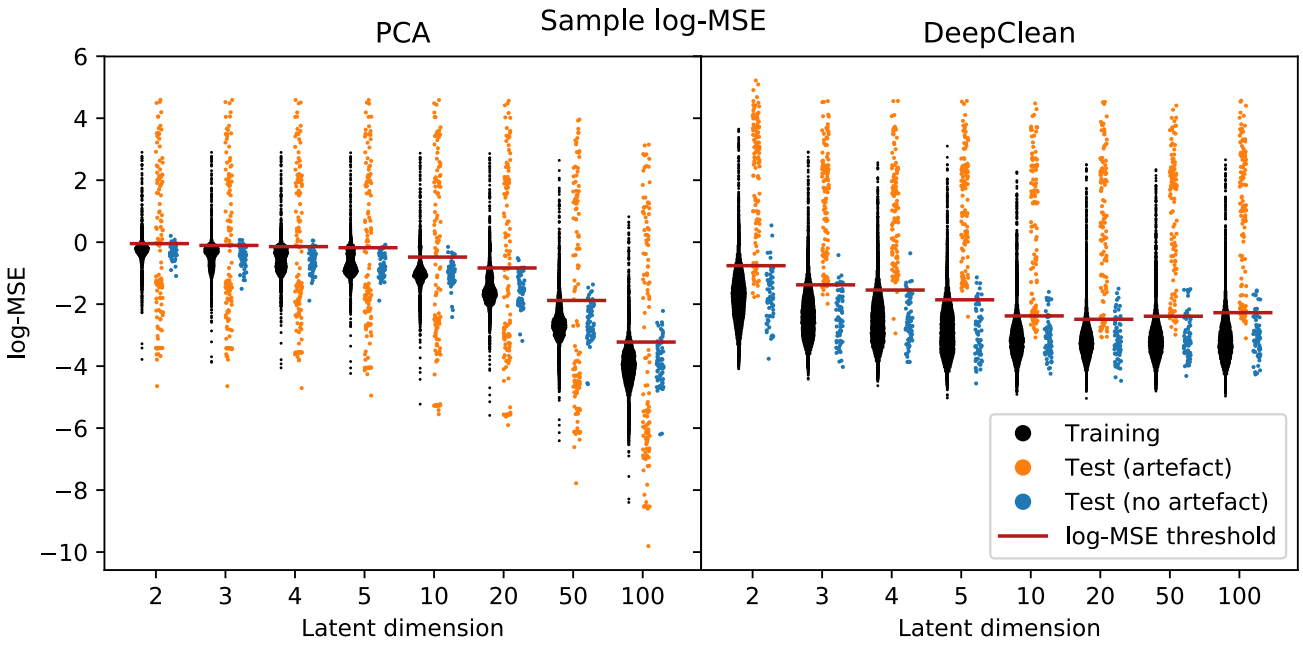

Figure 2: Log-MSE values for both PCA and DeepClean, with increasing latent dimension. The distribution of training set log-MSE is shown with the training set 90th percentile threshold. The test samples are split into two, shown separately: those annotated as containing an artefact and not. The `ground truth' artefacts that lie below the threshold and non-artefacts above the threshold were samples that were incorrectly identified, and were false negatives and false positives respectively. Therefore, the proportion of test artefacts points above the threshold is the sensitivity and the proportion of test non-artefacts below the threshold is the specificity. Figure reproduced under creative commons licence from [18].

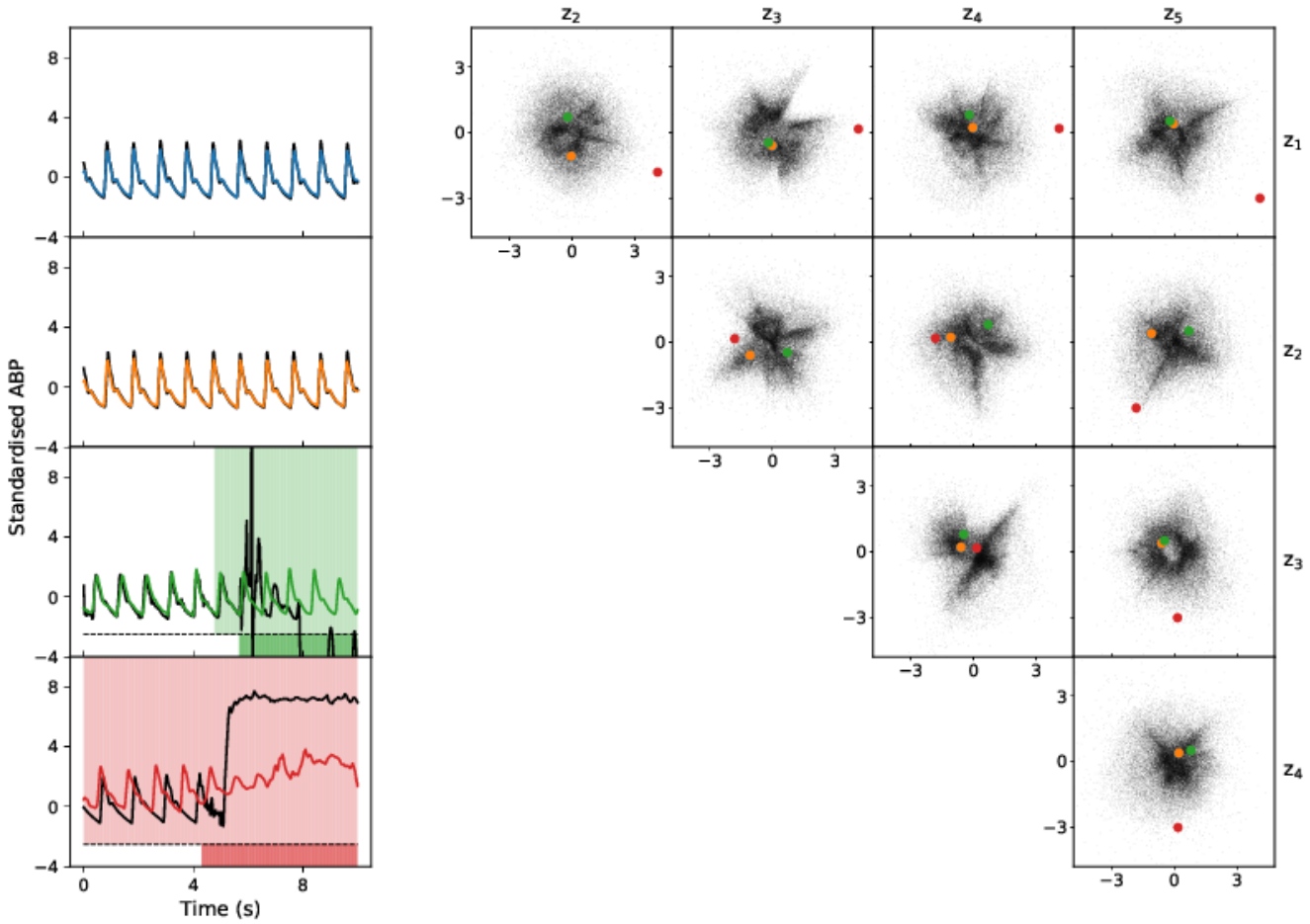

Figure 3: Illustration of key principles involving the latent distribution, via examples. Each example test input is given in black (left) with the corresponding latent representation within the latent distribution (right) and reconstruction (left) in (matching) colour. The shaded region above the dotted line (left) shows where DeepClean identifies an artefact (within the artefactual sample, using a sliding 1-second window MSE calculation as in [18]) and the shaded region below the dotted line shows the artefactual region as defined by the manual annotation. The latent distribution is the average encoding distribution, and is approximately equal to the distribution of the prior, a multivariate unit Gaussian [25]. It shows the latent representations of the training samples, where each subplot shows the marginal distribution of a pair of latent variables as a scatter plot, after marginalising over all other latent variables. First, almost identical inputs (the first two panels on the left; blue and orange) have almost identical latent representations (in fact, the blue marker corresponding to the first panel is barely visible as the orange marker has been overlaid on top) and similarly almost identical reconstructions. Second, inputs containing an artefact but with high probability mass in the latent distribution (the third panel; green) have `valid' reconstructions, synthetic waveforms that are of the underlying process described by the generative model, and we can therefore impute the section marked as artefact using this reconstruction. Conversely, an input with very low probability mass in the learnt latent distribution (fourth panel; red) does not have a valid reconstruction and we must resort to using an alternative latent representation if we want to use the generative model for imputation in this case.

Author contributions and conflicts of interest

TE, SE and AE contributed to development of code and analysis. PS, MC and MC contributed to the acquisition of data. TE prepared the manuscript, which was edited by SE and AE and approved by all authors. Correspondence requests should be made to TE and AE. Competing interests: The authors declare no competing interests.

	Accuracy		Sensitivity		Specificity		ROC AUC	
Latent dim	PCA	VAE	PCA	VAE	PCA	VAE	PCA	VAE
2	0.61	0.88	0.454	0.854	0.9	0.929	0.471	0.967
3	0.615	0.925	0.462	0.977	0.9	0.886	0.478	0.984
4	0.62	0.945	0.462	0.977	0.914	0.886	0.482	0.987
5	0.62	0.945	0.462	0.992	0.914	0.857	0.487	0.994
10	0.62	0.86	0.470	0.869	0.9	0.843	0.496	0.953
20	0.63	0.855	0.470	0.877	0.929	0.814	0.528	0.96
50	0.605	0.905	0.454	0.931	0.886	0.857	0.474	0.976
100	0.58	0.895	0.446	0.908	0.828	0.871	0.478	0.969
Mean	0.613	0.901	0.460	0.919	0.896	0.868	0.487	0.973

Table 1: Assessment of the classification of samples as containing an artefact or not, for both PCA and DeepClean (VAE). Both methods had comparable specificity, i.e. correctly identifying non-artefact data as such, with PCA performing slightly better. However, DeepClean alone can identify artefactual data as such, with high sensitivity.