IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 June 2022, accepted 12 July 2022, date of publication 18 July 2022, date of current version 29 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3192134

RESEARCH ARTICLE

Do Autoencoders Need a Bottleneck for Anomaly Detection?

BANG XIANG YONG AND ALEXANDRA BRINTRUP
Department of Engineering, University of Cambridge, Cambridge CB3 0FS, U.K.

Corresponding author: Bang Xiang Yong (bxy20@cam.ac.uk)

This work was supported in part by the Research England for the Pitch-In Project (Promoting the Internet of Things via Collaborations Between Higher Education Institutes (HEIs) and Industry), and in part by the European Metrology Programme for Innovation and Research (EMPIR) for the Metrology for the Factory of the Future (MET4FOF) Project.

ABSTRACT A common belief in designing deep autoencoders (AEs), a type of unsupervised neural network, is that a bottleneck is required to prevent learning the identity function. Learning the identity function renders the AEs useless for anomaly detection. In this work, we challenge this limiting belief and investigate the value of non-bottlenecked AEs. The bottleneck can be removed in two ways: (1) overparameterising the latent layer, and (2) introducing skip connections. However, limited works have reported on the use of one of the ways. For the first time, we carry out extensive experiments covering various combinations of bottleneck removal schemes and datasets using variants of Bayesian AEs. In addition, we propose the infinitely-wide AEs as an extreme example of non-bottlenecked AEs. Their improvement over the baseline implies learning the identity function is not trivial as previously assumed. Moreover, we find that non-bottlenecked architectures (highest AUROC=0.905) can outperform their bottlenecked counterparts (highest AUROC=0.714) on a recent benchmark of CIFAR (inliers) vs SVHN (anomalies), among other tasks, shedding light on the potential of developing non-bottlenecked AEs for improving anomaly detection.

INDEX TERMS Anomaly detection, autoencoders, bottleneck, unsupervised neural network.

I. INTRODUCTION

Numerous works have demonstrated the successful use of autoencoders (AEs), a type of unsupervised neural network (NN), for anomaly detection [1]. AEs are optimised to reconstruct a set of training data with minimal error. When given anomalous data which have high dissimilarity from the training data, the AEs reconstruct them with high error. Therefore, the reconstruction error is a measure of data anomalousness; by placing a threshold, we can effectively classify data points as inliers or anomalies.

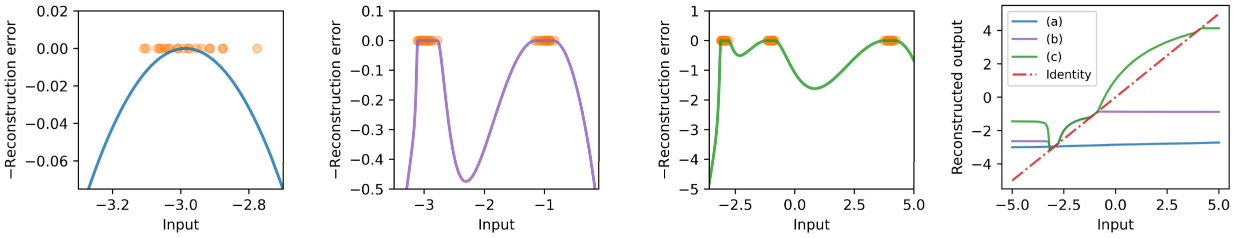
Extant works claim that AEs will trivially learn the identity function when no constraints are placed [2], [3]. If this were to occur, AEs will perfectly reconstruct any inputs (regardless whether it is anomalous or not), and hence the reconstruction loss will be low for all inputs, leading to unreliable anomaly detection. To prevent this, it is common to impose a bottleneck in the architecture, resulting in an undercomplete architecture: the output of the encoder has much lower dimensions than the input. However, most works describe the need for a bottleneck analogically and report only the empirical performance of bottlenecked AEs, without comparing them against non-bottlenecked AEs [2], [4]–[7].

Recently, a study by Nalisnick *et al.* [8] has shed light on a surprising failure of AEs on a seemingly trivial benchmark: distinguishing common images of vehicles or animals in CIFAR dataset (as inliers) from house numbers in the SVHN dataset (as anomalies), which differences are obvious to humans. Since then, several works have followed up by proposing ad-hoc fixes such as likelihood ratio [9], Watanabe-Akaike Information Criterion (WAIC) [10], and density of states estimation (DOSE) [11]. The failure of AEs on the benchmark task is surprising and profoundly questions our understanding of AEs, especially since they have been applied to many industrial applications [1].

A. WHY SHOULD WE CARE ABOUT NON-BOTTLENECKED AES?

By limiting to bottlenecked architectures, we miss the potential of achieving better performance with non-bottlenecked

The associate editor coordinating the review of this manuscript and approving it for publication was Jolanta Mizera-Pietraszko.



(a) (b) (c) (d)

FIGURE 1. (a-c) Negative reconstruction errors (i.e. log-likelihood) from BAEs with five layers of infinitely many parameters on 1D toy datasets, resembling reasonable density estimation. Orange dots represent the training data points. (d) The reconstructed outputs (last panel) clearly differ from the identity function. All layers use GELU [20] as activation functions, except the last, which uses the sigmoid function; the min-max scaler [21] is used for pre-processing.

AEs. Therefore, in this work, we study the use of non-bottlenecked AEs for anomaly detection. We investigate combinations of ways for removing the bottleneck, including (1) expanding the latent dimensions, also known as an overcomplete architecture, and (2) introducing skip connections. Furthermore, we propose the infinitely-wide AEs as an extreme example; this work is the first to provide empirical results on their effectiveness as anomaly detectors. Extensive experiments demonstrate the empirical success of non-bottlenecked AEs in detecting anomalies over the baseline and the bottlenecked AEs, indicating the non-bottlenecked AEs have failed to learn the identity function, contrary to conventional belief.

Several works have considered the use of overcomplete architectures or skip connection with additional modifications such as rough neurons proposed by Khodayar *et al.* [12], dictionary learning [13], dropout and denoising mechanisms [14] to prevent learning the identity function by encouraging sparsity. Sparsity refers to having neuron activations or weights with zeros which is equivalent to turning off unused neurons, and is often implemented in a form of L_1 regularisation that prevents the AE from overfitting to the training data [3]. While introducing sparsity can improve the robustness of the AE, a work by Baur *et al.* [15] has briefly reported an observation that random weight initialisation alone in the classic deterministic AE is sufficient to prevent learning the identity function. It is worth noting that a recent theoretical study [16] has shown that overparameterised supervised NNs do not tend to overfit and can generalise well with a conventional L_2 regularisation, in agreement with our findings on unsupervised AEs.

We suggest that rethinking about AEs is needed. In this effort, we adopt the probabilistic formulation of Bayesian autoencoders (BAEs), viewing them as regularised density estimators that benefit from having higher expressivity allowed by non-bottlenecked architectures (see Fig. 1 for an example). The Bayesian framework also provides a sound foundation for theoretical analysis of these architectures in future work, while drawing clear connection to the classic deterministic AE.

In particular, our contributions toward understanding the use of AEs for anomaly detection are as follows:

1) **Development of non-bottlenecked AEs.** Inspired by our observation that AEs do not learn the identity function on low-dimensional datasets despite being overparameterised, this study investigates deeper into the use of non-bottlenecked architectures. Combinations of overcomplete structure and skip connections are applied to various layer types including fully-connected dense layers, convolutional 1D and 2D layers, and variants of AEs including variational autoencoder (VAE) [17] and BAEs [18]. An ablation study reveals that removing the bottleneck of the deterministic AE demonstrates significant improvements on various datasets including the recent benchmark of CIFAR vs SVHN, which bottlenecked AEs have shown to fail [8].
2) **Development of infinitely-wide BAE.** For the first time, we present a study on the applications of infinitely-wide BAEs for anomaly detection. The Bayesian framework permits us to view AEs as probability density estimators instead of conventional reconstruction models, and draws connection to the well-studied Gaussian Process (GP) at the infinite-width limit. In addition to achieving higher modelling capability from the infinitely-wide layers, these BAEs demonstrate improved robustness due to Bayesian model averaging [19].
3) **Applications of non-bottlenecked AEs to industrial sensor datasets.** A wide range of non-bottlenecked AEs is applied to real sensor data for condition monitoring of a hydraulic system and quality inspection of a radial forging process, demonstrating that improvements on benchmark datasets can be generalised and are impactful to industrial applications. The industrial testbeds also shed light on the additional computational time incurred by the non-bottlenecked AEs.

This paper is organised as follows: Section III formulates AEs from a Bayesian perspective and describes ways to remove the bottleneck. Our experimental setup is described in Section IV followed by results and discussion in Section V. We relate to previous works in Section II and state our limitations in Section VI. We close with a summary and future directions in Section VII.

II. RELATED WORK

Several works have investigated the use of skip connections in AEs for tasks such as image denoising [22], [23] and audio separation [24]. Our work differs from current works on overcomplete and skip-AEs for anomaly detection [12]–[15], [25]: we have investigated a wider range of non-bottlenecked AEs, in which overcomplete or skip-AEs are only one type, and experimented with more datasets.

Snoek *et al.* [26] has proposed the autoencoder with an infinitely-wide decoder while keeping its encoder finite, and demonstrates its effectiveness for supervised classification and learning latent representations. Nguyen *et al.* [27] has theoretically studied infinitely-wide and shallow (2 layers) AEs, providing insights into their behaviours. To the best of our knowledge, we are the first to propose a deep BAE (up to 13 layers) with all layers being infinitely-wide, and provide empirical results on anomaly detection.

Radhakrishnan *et al.* [28] and Zhang *et al.* [29] have observed that overcomplete AEs exhibit *memorisation*, a phenomenon where the AEs reconstruct the closest training examples instead of the inputs. We suggest that a possible link exists between memorisation and the success of detecting anomalies using overcomplete AEs: when given an anomalous input, the AEs reconstruct the closest training example of inliers. This leads to a more discriminating, larger reconstruction error with the anomalous input than if the input were to be an inlier.

III. METHODS

This section introduces the AEs from a generalised Bayesian perspective, followed by methods for removing the bottleneck.

A. BAYESIAN AUTOENCODERS

Suppose we have a set of data $X^{train} = \{x_1, x_2, x_3, \dots x_N\}$, $x_i \in I\!R^D$. An AE is an NN parameterised by θ, and consists of two parts: an encoder for mapping input data x to a latent embedding, $z = f_{\text{encoder}}(x)$, and a decoder f_{decoder} for mapping the latent embedding to a reconstructed signal of the input $\hat{x}$ (i.e. $\hat{x} = f_\theta(x) = f_{\text{decoder}}(f_{\text{encoder}}(x))$) [30].

Bayes' rule can be applied to the parameters of the AE to create a BAE,

$$p(\theta|X^{train}) = \frac{p(X^{train}|\theta)\,p(\theta)}{p(X^{train})}, \tag{1}$$

where $p(X^{train}|\theta)$ is the likelihood and $p(\theta)$ is the prior distribution of the AE parameters. The negative log-likelihood (NLL) for an isotropic Gaussian distribution with variance=1 is

$$-\log p(x|\theta) = \frac{1}{D}\sum_{i=1}^{D}(x_i - \hat{x}_i)^2 \tag{2}$$

Note that the NLL is proportional to the mean-squared error (MSE) function, also known as the reconstruction loss in classic AE parlance.

For the prior, an isotropic Gaussian prior distribution is employed, effectively leading to L_2 regularisation of weights. When a Laplace distribution is used instead, this leads to L_1 regularisation used in sparse AEs [3], [13]; regularisation penalises the large-valued weights and encourage smaller valued weights to prevent overfitting. L_1 regularisation typically has higher robustness in selecting features due to its ability to encourage sparsity [31].

Since Equation 1 is analytically intractable for a deep NN, various approximate methods have been developed such as Stochastic Gradient Markov Chain Monte Carlo (SGHMC) [32], Monte Carlo Dropout (MCD) [33], Bayes by Backprop (BBB) [34], and anchored ensembling [35] to sample from the posterior distribution (see Fig. 2 for an overview).

In anchored ensembling [35], posteriors are approximated by Bayesian inference under the family of methods called randomised maximum a posteriori (MAP) sampling, where model parameters are regularised by values drawn from a so-called *anchored prior distribution*. Assume our ensemble consists of M independent AEs θ_m where $m \in \{1, 2, \dots, M\}$. The *anchored weights* θ_m^{anc} are the randomly initialised NN parameters for which the Kaiming [36] or Xavier initialisation schemes [37] are used as default in deep learning libraries such as Pytorch [38] and Tensorflow [39]; the values of θ_m^{anc} remain fixed throughout the training procedure. For each member of the ensemble, the loss to be optimised is

$$\mathcal{L}(\theta_m, X) = \underbrace{-\log p(X|\theta_m)}_{\text{likelihood loss}} + \underbrace{\lambda||\theta_m - \theta_m^{\text{anc}}||^2}_{\text{prior loss}} \tag{3}$$

where λ is a hyperparameter for scaling the regulariser term arising from the prior, also known as weight decay. In contrast, a classic deterministic AE is a single maximum likelihood estimate (MLE) or MAP estimate (i.e. $M = 1$) when regularisation is introduced; the training loss for such AE is

$$\mathcal{L}(\theta^{\text{MAP}}, X) = \underbrace{-\log p(X|\theta^{\text{MAP}})}_{\text{likelihood loss}} + \underbrace{\lambda||\theta^{\text{MAP}}||^2}_{\text{prior loss}} \tag{4}$$

In short, the training phase of BAE entails using one of the sampling methods to obtain a set of approximate posterior samples of AE parameters $\{\hat{\theta}_m\}_{m=1}^M$ (i.e. an ensemble of AEs) essential for the prediction phase (Algorithm 1). On the other hand, the deterministic AE is a single MAP estimate of the posterior distribution and does not benefit from Bayesian model averaging [19]. The variational autoencoder (VAE) [17] and BAE are AEs formulated differently within a probabilistic framework: in the VAE, only the latent embedding is stochastic while the f_{encoder} and f_{decoder} are deterministic and the model is trained using variational inference; on the other hand, the BAE, as an unsupervised Bayesian neural network (BNN), has distributions over all parameters of f_{encoder} and f_{decoder}.

Then, during the prediction phase, we use the posterior samples to compute M estimates of the NLL. The negative log predictive density of a new data point $\mathbf{x}^*$ can be approximated

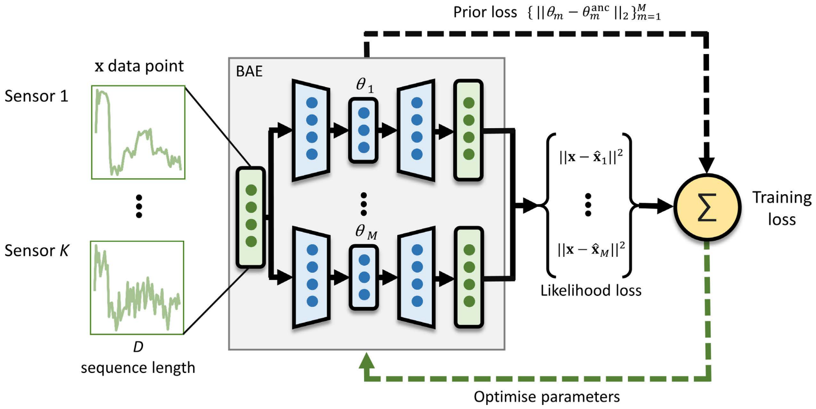

Prior loss $\{||\theta_m - \theta_m^{anc}||_2\}_{m=1}^M$
x data point
Sensor 1
Sensor K
D
sequence length
BAE
θ_1
θ_M
$||\mathbf{x} - \hat{\mathbf{x}}_1||^2$
$||\mathbf{x} - \hat{\mathbf{x}}_M||^2$
Likelihood loss
Σ
Training loss
Optimise parameters

FIGURE 2. Overview of training a BAE with anchored ensembling; the training loss comprises the likelihood and prior losses arising from M samples of BAE parameters. The example is shown for a sensor data input (K sensors × D measurements).

as the mean of the posterior NLL estimates,

$$-\log p(x^*|X^{train}) = \mathbb{E}_\theta[-\log p(x^*|\theta)\, p(\theta|X^{train})]$$
$$\approx -\frac{1}{M}\sum_{m=1}^{M} \log p(x^*|\hat{\theta}_m) \qquad (5)$$

For convenience, we denote $-\log p(x^*|X^{train})$ as $\mathbb{E}_\theta$ [NLL]. The Bayesian formulation allows us to view AEs as regularised probability density estimators: they model the training data distribution, assigning lower density scores to data which have higher dissimilarity from the training data.

Algorithm 1 BAE: Training and Predictions

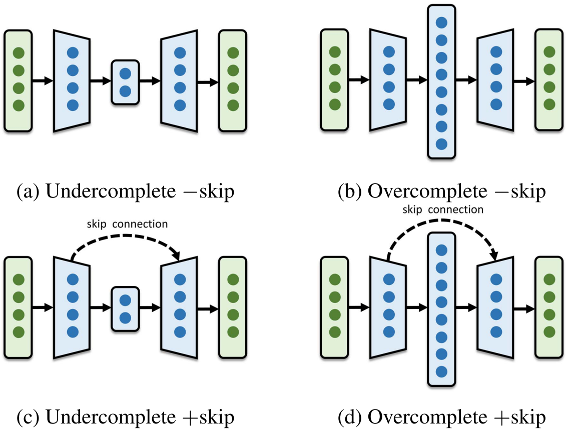
```
for number of training epochs do                        ▷ Train BAE
    Minimise loss on $X^{train}$ (Eq. 3)
end for
$NLL_M \leftarrow \{\}$                                   ▷ Prediction
for each ensemble member, $m$=1 to $M$ do
    $\hat{\mathbf{x}}_m \leftarrow f_{\hat{\theta}_m}(\mathbf{x}^*)$          ▷ Reconstruct signal
    $NLL_M$ append $||\mathbf{x}^* - \hat{\mathbf{x}}_m||^2$          ▷ Store NLL
end for
$\mathbb{E}_\theta$ [NLL] $\leftarrow$ mean($NLL_M$)             ▷ Anomaly score
return $\mathbb{E}_\theta$ [NLL]
```

B. HOW TO REMOVE THE BOTTLENECK?

The identity function is successfully learnt when $f_\theta(x) = x$ holds true for all $\mathbf{x}$ and therefore the reconstruction loss or NLL is always 0, rendering it useless for distinguishing anomalies from inliers. In an effort to mitigate this, a bottleneck is implemented at the latent layer (encoder's final layer) by having the latent dimensions smaller than the input dimensions, $\dim(z) < \dim(x)$, and there is no way for any output of the intermediate layers to bypass the bottleneck layer. It is straightforward to eliminate the bottleneck by doing the opposite: (1) simply expand the size of the latent dimensions to $\dim(z) \geq \dim(x)$, also known as an overcomplete architecture, and/or (2) introduce long-range skip connections from the encoder to the decoder akin to a U-Net architecture [40], thereby allowing each layer's data flow to bypass the bottleneck; for clarity, see Table 1 and Fig. 3.

TABLE 1. Categorising architectures into with or without a bottleneck depends on the latent dimensions and the presence of skip connections.

Architecture type	Latent dimensions	Skip connections
Bottlenecked		
A	Undercomplete	✗
Non-bottlenecked		
B	Overcomplete	✗
C	Undercomplete	✓
D	Overcomplete	✓



(a) Undercomplete −skip
(b) Overcomplete −skip
(c) Undercomplete +skip
(d) Overcomplete +skip

FIGURE 3. Overview of architectures of (a) bottlenecked and (b-d) non-bottlenecked AEs. Blue blocks represent NN layers; green blocks depict input or reconstructed vectors. The bottleneck is removed by expanding the latent layer (center layer) to be wider than the input dimension and/or introducing skip connections.

1) WHY SKIP CONNECTIONS?

Skip connections allow a better flow of information in NNs with many layers, leading to a smoother loss landscape [41] and easier optimisation, without additional computational complexity [42]. Recent works [14], [15], [25] have reported that AEs with skip connections outperform those without on image anomaly detection. In preventing the skip-AEs from learning the identity function, Collin and

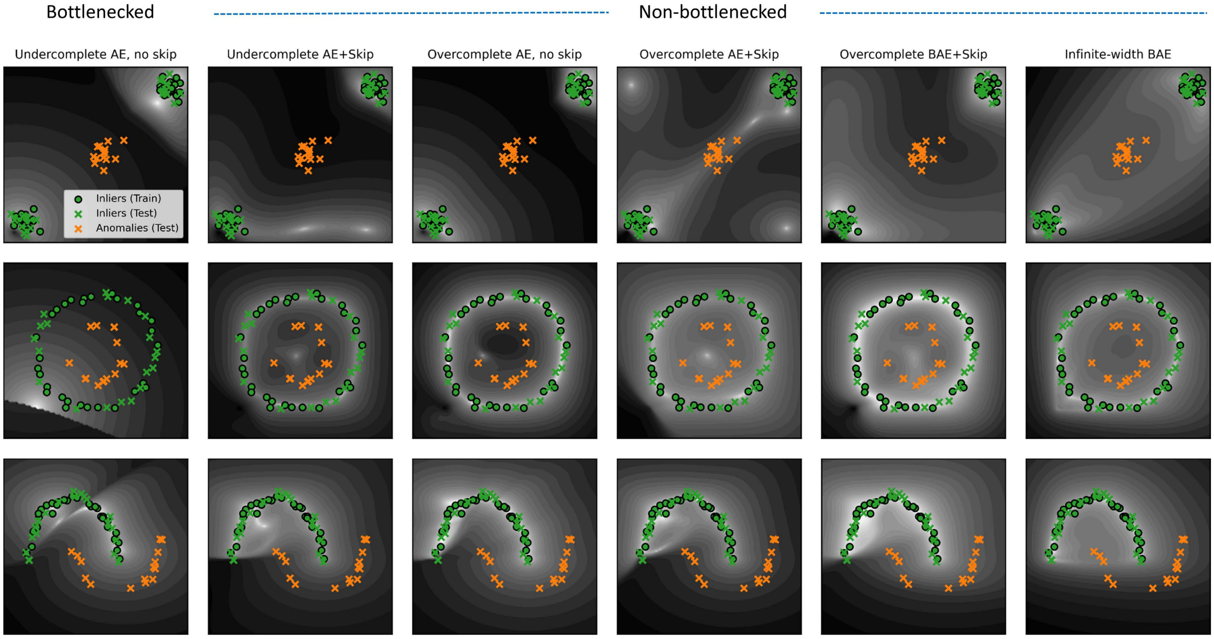


FIGURE 4. $\mathbb{E}_\theta$ **[NLL] using deterministic AE and BAEs with bottlenecked and non-bottlenecked architectures on a toy dataset. Brighter region has lower** $\mathbb{E}_\theta$ **[NLL] values and log-scale of contour is used to increase visibility. The darker contours away from the training points show that the deterministic AE and BAE do not learn the identity function despite being overparameterised and having skip connections. The encoder architecture has fully-connected layers with nodes of 2-50-50-50-dim(z) where dim(z)=1 for undercomplete and dim(z)=100 for overcomplete architectures. We use SELU activation [43] for every layer and sigmoid activation for the final layer. The BAE-**∞ **has a similar number of layers, each with infinite parameters.**

Vleeschouwer [14] and Baur *et al.* [15] have implemented a denoising scheme and a dropout mechanism, respectively. Notably, Baur *et al.* [15] have reported that random weight initialisation alone is sufficient to prevent learning the identity function, rendering the dropout redundant.

2) INFINITELY-WIDE BAE

In the infinite-width limit, a fully-connected BNN is equivalent to a neural network Gaussian process (NNGP) [44]. The results have been extended to modern architectures such as convolutional NNs, recurrent NNs, and transformers [45]–[47] in recent years. We propose extending the NNGP to the AE to create an infinitely-wide BAE (BAE-∞), which opposes the conventional bottleneck design. Viewing the BAE as a density estimator motivates this; it is not unconventional for density estimators to have infinite parameters as they benefit from higher expressivity to model an arbitrary distribution well [48], [49].

There are two primary advantages of the NNGP: having a closed-form solution and modelling a BNN with infinitely many parameters. The first facilitates a theoretical understanding by linking to the well-studied GP model, and the second *potentially* improves performance since deep NNs succeed over traditional ML models via increasing model parameters [1]. Nonetheless, empirically, infinite NNs do not always outperform finite NNs; reasons for their underperformance remain an active research topic [50], [51]. Another drawback is their computational complexity of $\mathcal{O}(N^3)$, where N is the number of training examples, reducing scalability to large datasets.

Surprisingly, when we examine the behaviours of AEs on 2D toy data sets (Fig. 4), we find that the identity function is not learnt despite using various types of non-bottlenecked AEs. Consequently, this observation on low-dimensional data implies it is more unlikely to learn the identity function on high-dimensional data due to higher degrees of freedom.

We suggest several reasons hindering AEs from the identity mapping: high degree of non-linearity in the AE and regularisation induced by mini-batching, the deep learning optimiser (e.g. Adam [52]) and the prior over parameters. Since these are usually implicit in training the AE, no additional, explicit efforts are necessary (e.g. denoising or dropout mechanisms).

IV. EXPERIMENTAL SETUP

This section describes the datasets, preprocessing steps, and models used in our experiments. All datasets and code used in this work are publicly available[1]; the dimensions of training and test sets after preprocessing are tabulated in Table 2.

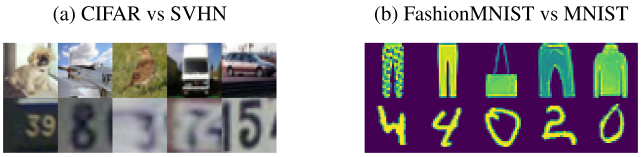
A. BENCHMARK DATASETS

Recent studies on anomaly detection with deep learning models have often benchmarked on pairs of high-dimensional image datasets; one dataset is labelled as inliers and the other as anomalies. These datasets have larger number of features than the traditional datasets in the ODDS collection, posing higher difficulty in achieving performant models. Two pairs

[1] Download links Image datasets: pytorch.org/vision/stable/datasets ODDS: odds.cs.stonybrook.edu ZeMA: doi.org/10.5281/zenodo.1323611 STRATH: doi.org/10.5281/zenodo.3405265 Code: github.com/bangxiang yong/bottleneck_ae

TABLE 2. Number of examples in train and test sets, and number of features for each image dataset pair, and for each task in ODDS, ZeMA and STRATH datasets.

Datasets & tasks	Train (inliers)	Test (inliers)	Test (anomalies)	Features
Images				
CIFAR vs SVHN	50000	10000	26032	3072
F.MNIST vs MNIST	60000	10000	10000	784
ODDS				
Cardio	1324	331	176	21
Ionosphere	180	45	126	33
Lympho	113	29	6	18
Optdigits	4052	1014	150	64
Pendigits	5371	1343	156	16
Pima	400	100	268	8
Thyroid	2943	736	93	6
Vowels	1124	282	50	12
ZeMA				
Cooler	518	223	1464	120
Valve	787	338	1080	60
Pump	854	367	984	120
Accumulator	419	180	1606	120
STRATH				
Radial forge	51	23	7	113


(a) CIFAR vs SVHN

(b) FashionMNIST vs MNIST

FIGURE 5. Image samples from benchmarks of (a) CIFAR vs SVHN and (b) FashionMNIST vs MNIST. Top row is the inlier distribution and bottom row is the anomalous distribution.

of image datasets are of importance due to the surprising finding of Nalisnick *et al.* [8] that deep learning models such as AEs have failed to perform well, despite their visual differences (Fig. 5) which are obvious to the human eye: (1) FashionMNIST [53] vs MNIST [54], and (2) CIFAR [55] vs SVHN [56].

The MNIST [54] dataset consists of 28 × 28 grayscale images of handwritten digits ranging from 0 to 9. The images in FashionMNIST [53] are of similar format with a difference that the images depict ten classes of fashion products: T-shirt, trouser, pullover, dress, coat, sandals, shirt, sneaker, bag, and ankle boots. The CIFAR [55] dataset comprises 32×32 colour images with 10 classes: automobile, bird, cat, deer, dog, frog, horse, ship, truck. The SVHN [56] images share similar format with those of CIFAR (32 × 32 and colored) and depicts actual house numbers captured from Google Street View. In comparison to FashionMNIST vs MNIST, the task of CIFAR vs SVHN is a more difficult problem since these datasets contain natural scene images which are less regular and cover more colour channels (red, blue, green) than grayscale images.

The ODDS [57] collection consists of smaller datasets with fewer features than FashionMNIST, MNIST, CIFAR, and SVHN. These datasets have been used extensively in extant studies to benchmark anomaly detection methods [58]–[60]. The Cardio [61] dataset consists of fetal heart rate and uterine contraction measurements classified by expert obstetricians as healthy or malignant. Similarly, Lympho [62], Thyroid [63], and Pima [64] datasets consists of medical measurements relevant to lymph nodes, thyroid, and diabetes diseases, respectively. For these datasets, the healthy condition forms the inliers, while the pathologic class is treated as anomalies. Optdigits [65] and Pendigits [66] consist of handwritten digits (0-9) with the class "0" treated as anomaly, while the others are categorised as inliers; unlike MNIST, however, these datasets are of much smaller dimensions. The Ionosphere [67] dataset comprises radar data collected from 16 high-frequency antennas; inliers are good signals that indicate structure in the ionosphere, while the anomalies are bad signals that pass through the ionosphere. The Vowels [68] dataset consists of time series features of utterances from 9 Japanese speakers; the anomaly class consists of features from speaker 1, while inliers comprise speakers 6,7, and 8; the other speakers have been discarded in the provided dataset.

1) DATA PREPROCESSING

For FashionMNIST vs MNIST and CIFAR vs SVHN datasets, we use the default split of train-test sets available in Pytorch [38] which have pixel values scaled to [0,1]. For ODDS, we split the inliers into train-test sets with a ratio of 70:30 and with random shuffling. Subsequently, min-max scaling [21] is fitted to the training set and transforms each feature in train-test sets to range in [0,1].

B. INDUSTRIAL DATASETS

To demonstrate the applicability of non-bottlenecked AEs to real datasets, sensor data gathered from two industrial testbeds are used: (1) condition monitoring of a hydraulic test rig [69] and (2) quality inspection of a radial forging process [70]. The datasets gathered from these testbeds have the dimensions of ($N \times K \times D$), where N is the number of cycles (i.e. data examples), D is the sequence of measurements in each cycle, and K is the number of sensors.

1) CONDITION MONITORING OF A HYDRAULIC SYSTEM

Hydraulic systems transmit power using pressurised liquids such as oils [71]; these systems are often found in heavy machinery such as elevators, cranes, aircraft, ships, construction vehicles, and are essential in many industries including petroleum, railway, construction, and manufacturing. In practice, hydraulic systems can suffer from malfunctions and degradation of vital components, reducing their efficiency in transmitting power and increasing risk of occupational hazard.

Due to their relevance to industrial applications, condition monitoring of hydraulic systems has gained increasing importance to anticipate machine failure, reduce machine downtime and maintenance costs [72]. Adopting a data-driven approach to condition monitoring is promising due to the advances in deep AEs. In addition to boosting predictive accuracy, deep AEs reduce the need of feature engineering, in contrast to conventional modelling which involve detailed physical and mathematical analysis of a complex system [1], [73].

To this end, we investigate the application of unsupervised AEs on a hydraulic test rig [72], [74] developed in the Center for Mechatronics and Automation Technology

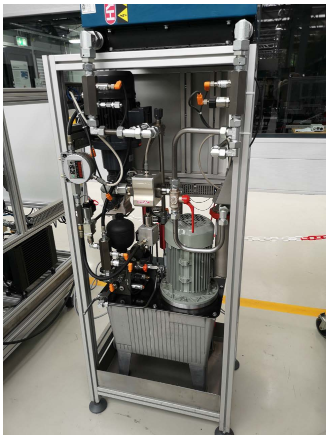


FIGURE 6. Hydraulic system test rig for condition monitoring at ZeMA, Germany.

TABLE 3. Conditions of components in the ZeMA hydraulic system and descriptions.

Component	State	Description
Cooler (efficiency)	100%	Maximum efficiency
	20%	Reduced efficiency
	3%	Near failure
Valve (efficiency)	100%	Optimal switching
	90%	Small lag
	80%	Severe lag
	73%	Near failure
Internal pump (leakage)	0	No leakage
	1	Weak leakage
	2	Severe leakage
Accumulator (pressure/bar)	130	Optimum pressure
	115	Slightly reduced pressure
	100	Severely reduced pressure
	90	Near failure

GmbH (ZeMA) in Saarland, Germany (Fig. 6). The test rig permits safe and non-destructive changes to the states of vital components to emulate faults and degradation. There are 4 detection tasks in the ZeMA dataset; each task entails using sensor measurements as inputs to detect anomalies in key components, namely, (i) the cooler, (ii) the valve, (iii) the pump, and (iv) the accumulator. The functions of these components are described as follows: the cooler prevents the liquid from overheating and maintain its viscocity; the pressure relief valve limits maximum pressure by providing an alternative flow path; the pump converts mechanical power into fluid power and controls the flow rate from a tank; the accumulator acts as an energy storage that can collect and feed the fluid into the system when needed. The fault severity levels are configurable by setting the duty cycle of the cooler, controlling the set-point current of the valve, switching bypass orifices to the pump, and switching flow to accumulators with different pre-charge pressures. A total of 2000 fixed working cycles have been recorded under a combination of varied component states; each cycle lasts for 60 seconds. The optimal operating state of each component is regarded as inlier, while the degraded states are labelled as anomalies. The states of each component are tabulated in Table 3.

TABLE 4. Description of sensors in the ZeMA hydraulic system.

Sensor ID	Quantity	Unit	Sampling frequency (Hz)
TS1-4	Temperature	°C	1
VS1	Vibration	mm/s	1
SE	Efficiency factor	%	1
CE	Virtual cooling efficiency	%	1
CP	Virtual cooling power	kW	1
FS1-2	Volume flow	l/min	10
PS1-6	Pressure	bar	100
EPS1	Motor power	W	100

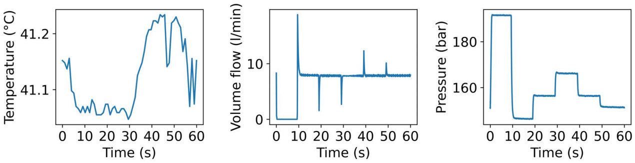


FIGURE 7. Sequence of sensor measurements during a working cycle of the ZeMA hydraulic system.

The system has 17 sensors installed to monitor the pressure, flow, temperature, power, vibration, volume flow, and efficiency (Table 4), each with a different sampling frequency from 1Hz to 100Hz. Samples of the process data during a working cycle are depicted in Fig. 7. The process data were buffered using programmable logic controller (PLC) before transferring to a main computer via EtherCAT. The Pearson correlation coefficients (PCC) between sensors are computed and sensor pairs with greater than 0.95 PCC are considered highly redundant; a sensor in each of these pairs is randomly dropped from subsequent analysis.

2) QUALITY INSPECTION OF RADIAL FORGING PROCESS

Quality inspection is an important task in manufacturing and refers to the classification of a manufactured product into within-tolerance or out-of-tolerance based on a measured quality of the product [75]. Ideally, manufacturers would want to inspect the quality of every product and subsequently reject products which are of unacceptable quality. However, the costs are overly prohibitive which means that only a small number of arbitrarily chosen products in a batch are inspected. For instance, it may consume much time to measure the quality of each forged product with a Coordinate Measuring Machine (CMM) and hence it is impractical to inspect all products. Furthermore, some procedures may be intrusive or destructive to the product. One route to improving the quality inspection is by leveraging the sensor data gathered during the forging process, feeding them as inputs to an anomaly detector which classifies a forged part as either within-tolerance or out-of-tolerance. Essentially, the anomaly detector would serve as a virtual sensor, enabling quality inspection of every forged part in a batch of production.

Radial forging, also known as swaging, reduces the diameter of a metal workpiece by applying radial forces towards its center [76]. Situated at the Advanced Forming Research

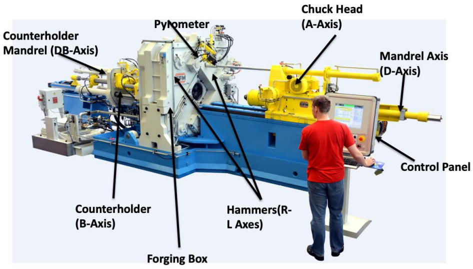


FIGURE 8. Radial forging machine (GFM SKK10/R) at the AFRC of University of Strathclyde, UK.

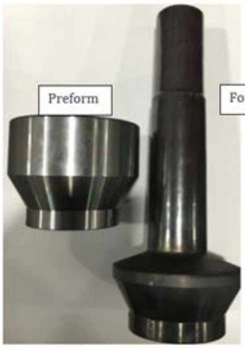


FIGURE 9. Preform and forged part with a Factory Acceptance Trial (FAT) geometry, which is used to test and demonstrate radial forging capability.

Centre (AFRC) of the University of Strathclyde (STRATH) in Glasgow, UK, a GFM SKK10/R radial forging machine is used as a testbed for data-driven quality inspection (Fig. 8). Equipped with four hammers, it is capable of providing a maximum forging force of 1,500kN and hammer speed of 1,200 strokes/min. In practice, the radial forging machine is used to forge shafts and rods from metals such as steel, titanium, and inconel for industrial sectors such as medical, railway, and aerospace manufacturing.

In a test case at the AFRC, the radial forging machine is used to forge 81 parts into a Factory Acceptance Trial (FAT) geometry (Fig. 9) during which process data from 99 sensors are recorded; 8 sensors are found to be relevant to the forging phase after elicitation from domain experts and analysis from a previous work [77], tabulated in (Table 5). The process of forging each part lasts for 202 seconds with a sampling interval of 10ms (100Hz), and can be segmented into heating, transfer, and forging phases (Fig. 10). During the heating phase, the workpiece is heated to a temperature of 800°C using induction coils. Next, it is transferred to the chuck head using a robotic arm. The forging process involves moving the heated part into the forging box, and the part is hammered in oscillations as the chuck head grips and rotates it.

Each forged part is sent to a CMM to measure its geometrical dimensions as quality indicators. We focus our analysis on the diameter of the forged part (labelled as *38 diameter@200* in the provided dataset), an important quality for radial forging process emphasised by the domain experts at the AFRC. To flag out-of-tolerance parts as anomalies, the Tukey's fences method [78] is applied on the absolute error between the measured and nominal dimensions; the remaining within-tolerance parts are labelled as inliers.

TABLE 5. Description of sensors relevant to the forging phase in the AFRC radial forging facility. All sensors have a sampling frequency of 100Hz.

Sensor ID	Unit	Description
A-ACTpos	mm	Position of chuck head from the forging box (A-axis)
DB-ACTpos	mm	Position of mandrel head from the forging box (DB-axis)
Power	kW	Forging drive power
L-ACTpos	mm	Position of the pair of hammers (L-Axis)
Feedback-R	%	Servo feedback of hammers (R-axis)
Force	kN	Hammer force
Feedback-SPA	%	Servo feedback of chuck jaws (A-axis)
W1-Durchfluss	l	Volume of mandrel coolant

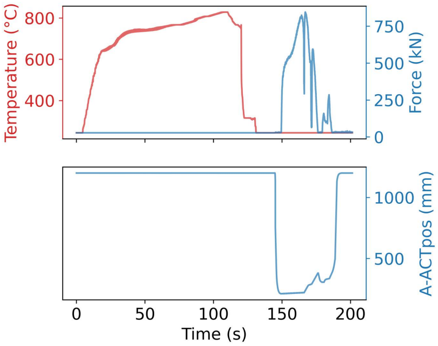


FIGURE 10. Sensor measurements of a single cycle of the radial forging process comprising heating (red) and forging phases (blue).

Data preprocessing of ZeMA and STRATH Datasets: The inliers are split into train-test sets of 70:30 ratio with random shuffling, and all anomalous examples are included in the test set. Min-max scaling is applied by obtaining the min-max values from the train set for each sensor independently, instead of each feature, to retain the shape of signal. Following Jiang *et al.* [79], train-test bias is prevented by fitting the scaler to the train set only instead of the entire dataset. Measurements of the forging phase are segmented from each cycle in the STRATH dataset, while the full sequence is used in the ZeMA dataset. In order to reduce the data size to fit into memory, the sequence of process measurements is downsampled via decimation with a low-pass Chebyshev type I filter of eighth-order which mitigates the distortion caused by aliasing [80], implemented as default in the *scipy* package [81]; the resulting sequence lengths of each cycle are reduced to $D = 60$ and $D = 112$ for ZeMA and STRATH datasets, respectively.

To optimise and reduce the number of sensors in the datasets, a sensor selection scheme is employed by evaluating a bottlenecked deterministic AE on each sensor independently, and the sensors are ranked by their AUROC scores (Fig. 11); the optimal combinations of sensors are selected by maximising the AUROC, tabulated in Table 6.

C. MODELS

Multiple variants of AEs are trained: deterministic AE, VAE, BAE-MCD, BAE-BBB, and BAE-Ensemble. The posterior samples for VAE, BAE-MCD, and BAE-BBB are set to $M = 50$, while $M = 10$ samples are drawn for the BAE-Ensemble,

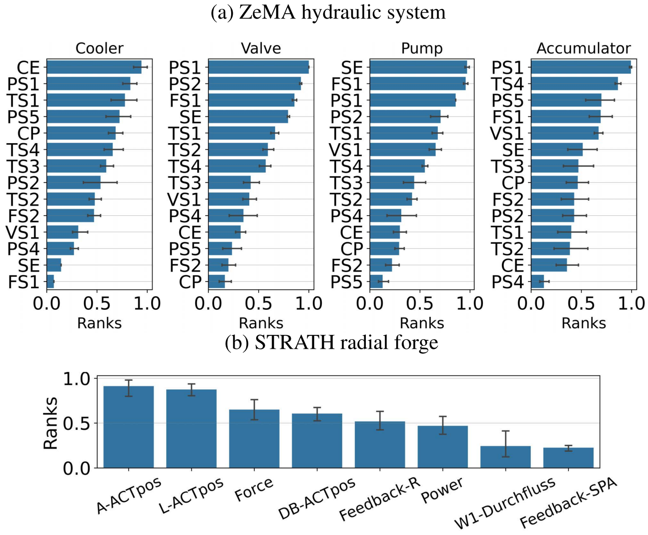


FIGURE 11. Normalised ranks with 95% confidence interval for sensor selection in (a) ZeMA and (b) STRATH datasets.

TABLE 6. Optimal sensor selection based on sensor ranks for ZeMA and STRATH datasets.

Dataset	Component / system	Selected sensors
ZeMA	Cooler	CE, PS5
	Valve	PS1
	Pump	SE, FS1
	Accumulator	PS1, TS4
STRATH	Radial forge	A-ACTpos

TABLE 7. Encoder architecture of finite-width AEs with two encoding layers $L_{encoder} = 2$. The decoder is a reflection of the encoder, in which the Conv1D and Conv2D layers are replaced by Conv1D- and Conv2D-Transpose layers.

(a) CIFAR and FashionMNIST

Layer	Output channels/nodes	Kernel	Strides
Conv2D	10	2 x 2	2 x 2
Conv2D	32	2 x 2	1 x 1
Reshape	-	-	-
Dense	100	-	-
Dense	Latent dimensions	-	-

(b) ODDS

Layer	Output channels/nodes	Kernel	Strides
Dense	Input dimensions ×4	-	-
Dense	Input dimensions ×4	-	-
Dense	Latent dimensions	-	-

(c) ZeMA and STRATH

Layer	Output channels/nodes	Kernel	Strides
Conv1D	10	8	2
Conv1D	20	2	2
Reshape	-	-	-
Dense	1000	-	-
Dense	Latent dimensions	-	-

which are marginally higher than the recommended minimum samples for these methods [33], [35]. The BAE-∞ is implemented as NNGPs with infinitely-wide dense layers using the Neural Tangent Kernel library [82] and with weight standard deviation optimised across levels of $\{0.5, 0.75, \ldots, 1.5\}$. Isotropic Gaussian prior is used for all models.

The model architectures are tabulated in Table 7; the number of convolutional filters and nodes in dense layers are set appropriately with sufficient capacity to ensure the training loss converges. Considering the range of preprocessed data is [0,1], the Sigmoid activation function [83] is applied at the final decoder layer, while all intermediate layers apply Leaky Rectified Linear Unit (LeakyReLU) [84]. In addition, the bias term is turned off for all layers following the recommendation by Ruff *et al.* [2] to prevent the output collapsing to a constant mean function and to improve training convergence. The option of layer normalisation [85] is applied only when it increases the performance.

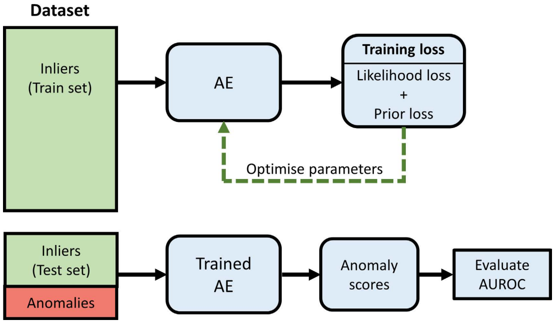


FIGURE 12. General evaluation setup of AE for anomaly detection.

A factorial experiment is conducted to investigate the effect of removing the bottleneck on architectures of various depths; the number of convolutional layers (image and sensor datasets) or dense layers (ODDS dataset) in the encoder are varied $L_{encoder} \in \{2, 4, 6\}$, where each additional layer has the same output channels, kernels, and strides as the one before; skip connections are turned on or off $\{+, -\}$; the latent dimensions are varied over factors of $\{\times\frac{1}{10}, \times\frac{1}{2}, \times 1, \times 2, \times 10\}$ with regards to the input dimensions. The experiment is repeated using variants of finite AEs. The BAE-∞ does not have skip connections and finite latent dimensions; we vary the $L_{encoder}$ at similar levels, resulting in architectures with a total of 5, 9, and 13 infinitely-wide layers (including the latent layer and decoder), respectively.

D. TRAINING AND EVALUATION SETUP

The general experimental setup is depicted in Fig. 12. The setup of anomaly detection involves training the AE on a set of data labelled as inliers. Then, during prediction phase, the AE is required to output anomaly scores to distinguish inliers and anomalies in the test set, and the performance is evaluated with AUROC.

Since each experiment run initialises the AE weights with different values, and the train-test split of inliers is subject to random shuffling, the experimental uncertainty is accounted for by repeating the experiments: 10 runs on ODDS, ZeMA, and STRATH datasets, and 5 runs on CIFAR and FashionMNIST datasets. From the repeated experiments, the mean and statistical spread of performance scores are reported.

A learning rate finder [86] is employed to search for the optimal learning rate (Fig. 13), which is capped at 0.001 for stability. The models are trained using the Adam optimiser [52] with early-stopping for a maximum of 20,

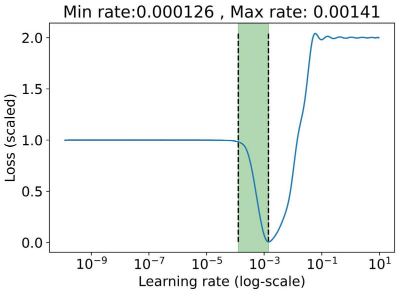


FIGURE 13. Example of a learning rate finder diagram [86]; the learning rate is then set as the optimal maximum learning rate for improved convergence.

20, 300, 200, and 200 epochs for FashionMNIST, CIFAR, ODDS, ZeMA, and STRATH datasets, respectively. The weight decay controls the strength of the prior and is set as $\lambda = 1 \times 10^{-10}$ which is low enough to prevent the training loss from stalling.

To evaluate the diagnostic performance of anomaly detectors, the area under the receiver-operating characteristic curve (AUROC) [87] is often used by extant studies [2], [73]. The AUROC yields intuitive interpretations suggested by Hosmer *et al.* [88]: a score of 0.5 suggests no discriminatory ability which corresponds to randomly tossing a coin, 0.7 to 0.8 is acceptable, and outstanding performance is attained when the AUROC exceeds 0.9.

Sensitivity Analysis: Following the study by Kalapanidas *et al.* [89] and Atla *et al.* [90], the noise sensitivity curve is employed to evaluate the drop in detection performance after injecting additive noise at different levels. From the curve, the average reduction in AUROC scores is used as a summary statistic of noise sensitivity,

$$\mathbb{E}_{\sigma^+}[\Delta\text{AUROC}] = \frac{1}{N}\sum_{n=1}^{N}(\text{AUROC}_{\sigma_n^+} - \text{AUROC}_0) \quad (6)$$

where $\text{AUROC}_{\sigma_n^+}$ denotes the AUROC score obtained at the σ_n^+ level of injected noise, and N is the number of noise levels. A lower $\mathbb{E}_{\sigma^+}[\Delta\text{AUROC}]$ value indicates a greater drop in accuracy and hence higher sensitivity to noise. Thereafter, a robustness metric can be defined by taking the exponent,

$$\text{Robustness} = \exp(\mathbb{E}_{\sigma^+}[\Delta\text{AUROC}]) \quad (7)$$

such that a higher value indicates higher robustness and a value of 1 indicates a perfect score (i.e. zero reduction in accuracy due to additive noise). Additive noise levels similar to extant studies [89], [90] are applied; the preprocessed test inputs (inliers and anomalies) which range in [0,1] are injected with random samples drawn from Gaussian $\mathcal{N}(0, \sigma^{+\text{Gauss}})$ or Uniform distributions $\mathcal{U}(-\sigma^{+\text{Uni}}, \sigma^{+\text{Uni}})$ where $\sigma^{+\text{Gauss}} \in \{0.1, 0.2, \ldots, 0.5\}$ and $\sigma^{+\text{Uni}} \in \{0.1, 0.2, \ldots, 0.5\}$. The sensitivity analysis is repeated for 5 times on CIFAR and FashionMNIST, and 10 times on remaining datasets to account for experimental uncertainty due to random sampling.

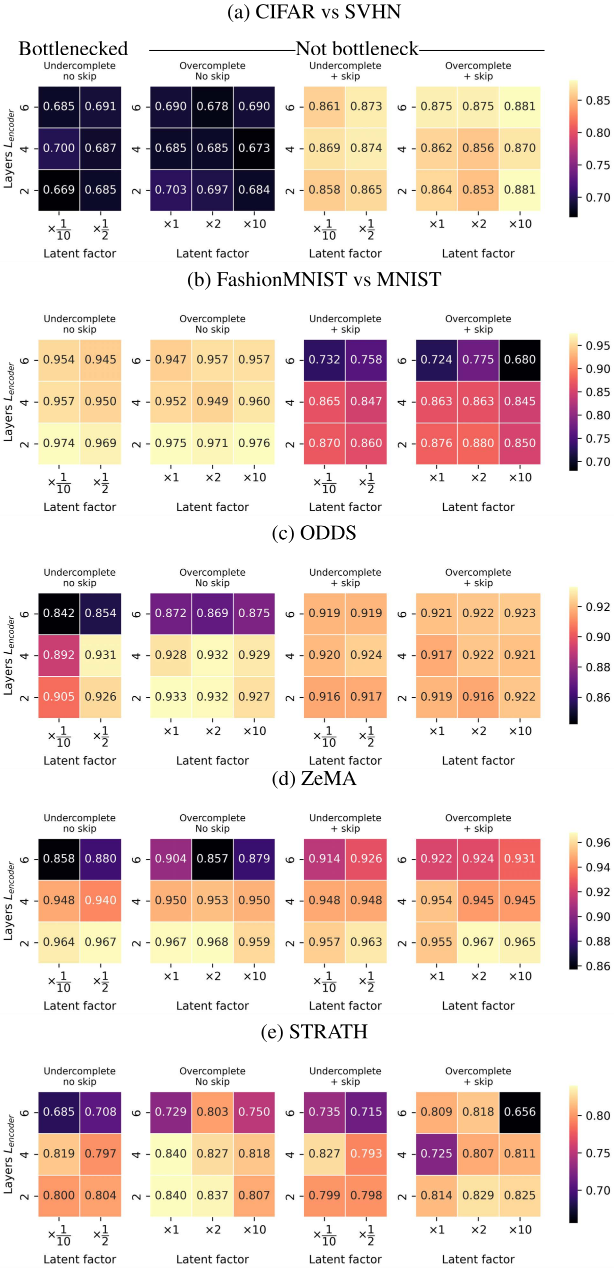


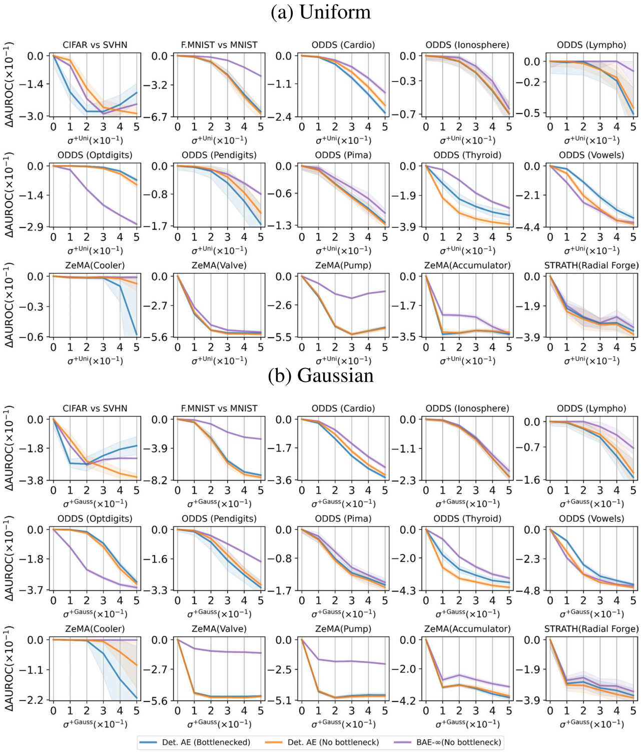
FIGURE 14. Hyperparameter selection of deterministic AE: mean AUROC scores of bottlenecked and non-bottlenecked architectures with varied number of encoder layers. Scores for ODDS and ZeMA are averaged over the tasks in those datasets for space compactness. Bright yellow boxes indicate the best performances.

V. RESULTS AND DISCUSSION

In this section, we discuss and analyse the results of experiments described in Section IV.

A. HYPERPARAMETER SELECTION

For each class of bottlenecked and non-bottlenecked architectures, the best combination of number of layers and latent factor is selected. When there are ties in AUROC performance, an architecture with fewer layers or smaller latent factor is preferred to reduce the computational cost. In Fig. 14, it is observed that the best performing architectures are


(a) Uniform

(b) Gaussian

FIGURE 15. (a) Uniform and (b) Gaussian noise sensitivity curves of bottlenecked and non-bottlenecked deterministic AEs and the BAE-∞ on various anomaly detection tasks and datasets. A lower and steeper curve indicates a higher sensitivity to noise.

often found to be non-bottlenecked. Furthermore, the non-bottlenecked architectures of various depths perform well and are capable of achieving $\geq$ 0.9 despite using a classic deterministic AE without sparsity or Bayesian sampling, implying the identity function has not been learnt. This holds true even on the real sensor data for condition monitoring of hydraulic system (ZeMA) and quality inspection of radial forging process (STRATH), demonstrating their potential value to industrial applications.

For CIFAR vs SVHN, applying skip connections have a strong effect as we find the performance is improved from AUROC$\geq$ 0.6 to AUROC$\geq$ 0.8 for all layer depth and latent factors. Therefore, this stands as a solution to the reported poor performance of AEs by Nalisnick *et al.* [8]. However, the effect is reversed on FashionMNIST vs MNIST, harming the accuracy instead. Hence, in alignment with the study by Kim *et al.* [25], this could indicate that skip connections have strong effect on image data types. The effect of applying skip connections is not as obvious on sensor data (ZeMA and STRATH) or tabular data in ODDS.

Another observation is that having too many layers can harm the performance on most datasets, and fewer layers are sufficient to perform well. On the contrary, we find increasing layers are helpful for CIFAR vs SVHN. One possible reason is the higher complexity of the CIFAR dataset which compose of 10 classes of natural images with different views, and hence adding the number of layers help in better extracting hierarchical features. Importantly, this improvement is realised only when skip connections are applied; without skip connections, vanishing gradients [91] are more likely to occur in deep architectures, preventing the gradient flow and hampering the training process.

B. SENSITIVITY ANALYSIS

The noise sensitivity curves [89], [90] depict the decrease in AUROC scores as the level of injected noise increases in Fig. 15. From these sensitivity curves, the average drop is summarised as a robustness measure, depicted in Fig. 16 for the deterministic AE and BAE. Although we find that removing the bottleneck of the classic deterministic AE often improves anomaly detection accuracy, the sensitivity analysis reveals that it is possible that this increases the sensitivity towards noise as observed on Thyroid, Vowels, and Optdigits tasks. Therefore, when higher robustness is required, various modifications can be applied such as including sparsity with deep temporal dictionary learning (DTDL) proposed by Khodayar *et al.* [13], rough neurons [12], or the Bayesian sampling methods used extensively in our study.

Several works have shed light on the relationship between accuracy and robustness [92]; whether it is strictly a trade-off remains debatable as attempts to achieve both criteria are an active area of research [93], [94]. The BAE-∞ offers higher modelling capability with its infinite parameters to boost accuracy; this flexibility is balanced with Bayesian inference that accounts for model uncertainty [19]. The BAE offers greater robustness than the deterministic AE on most datasets including FashionMNIST vs MNIST, ODDS, ZeMA and STRATH (Fig. 16), evaluated using significance tests: the null hypothesis of the Friedman test (p-value=3.97×10^{-5}) is rejected at a 95% confidence level, indicating a significant difference in robustness between the deterministic AE and BAE; post-hoc Nemenyi tests are then applied which indicate a significantly higher robustness of the infinitely-wide BAE compared to the bottlenecked and non-bottlenecked deterministic AEs with p-values of 0.001. However, on a few datasets, we do find the BAE-∞ to be lower in robustness, suggesting a need to improve the robustness of non-bottlenecked AEs without lowering the accuracy.

It is observed in Fig. 17 that increasing the strength of the prior regularisation through the weight decay increases the robustness at the expense of accuracy; this trade-off pattern exists whether the L_2 regularisation or the L_1 regularisation is used. Another observation is that the L_1 regularisation has generally higher robustness, which is due to the sparsity induced that discards irrelevant features [3]. The findings hold for both additive Gaussian and Uniform noise.

An important note is that non-bottlenecked AEs can perform well without sparsity and denoising mechanisms in spite of claims [13], [14] that these are strictly necessary for non-bottlenecked AEs. In Fig. 17, despite turning off the prior regularisation ($\lambda = 0$), the deterministic AE's accuracy remains high and has not decreased to below 0.5 AUROC, indicating the identity function has not been learnt. This is aligned with the observation of Baur *et al.* [15] that random

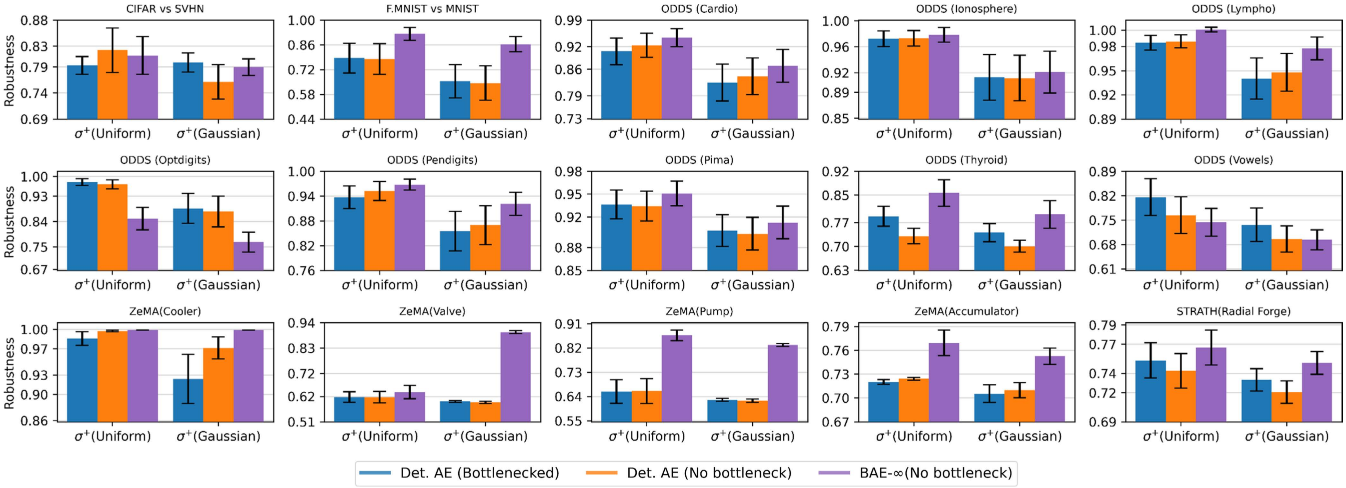
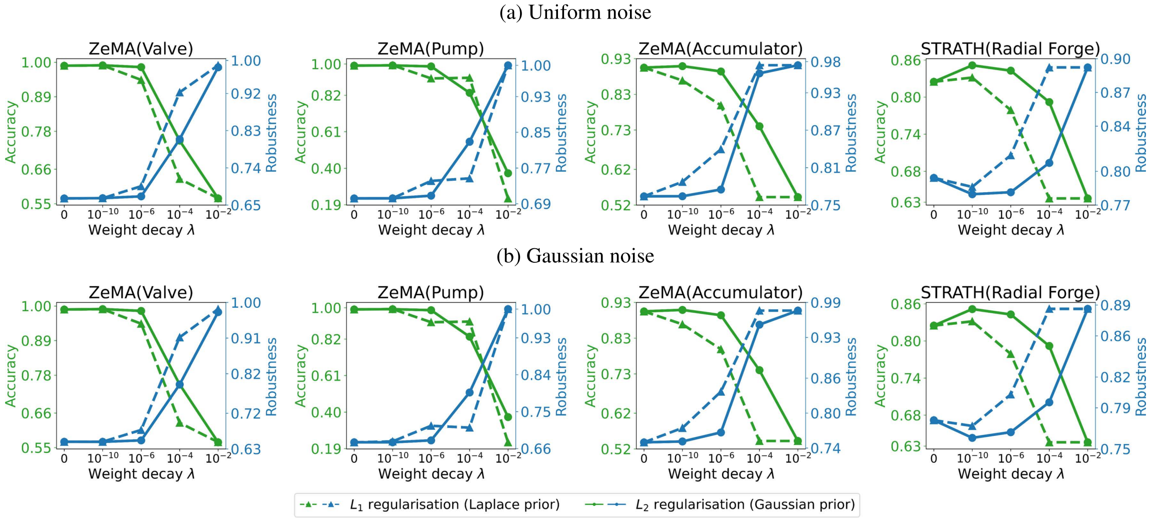


FIGURE 16. Robustness ($\exp(\mathbb{E}_{\sigma^+}[\Delta\text{AUROC}])$) of deterministic AE and BAE towards Uniform and Gaussian noise on various anomaly detection tasks and datasets.


(a) Uniform noise

(b) Gaussian noise

FIGURE 17. Trade-off between accuracy (AUROC) and robustness ($\exp(\mathbb{E}_{\sigma^+}[\Delta\text{AUROC}])$) towards (a) Uniform and (b) Gaussian noise as weight decay λ is varied for a non-bottlenecked deterministic AE with L_1 (dashed) or L_2 regularisation (solid).

weight initialisation is sufficient in preventing the identity function. In short, we posit that additional regularisation is beneficial when higher robustness is required, but not mandatory for a non-bottlenecked AE to detect anomalies accurately.

C. PERFORMANCE EVALUATION

In Table 8, the best non-bottlenecked models (type B, C and D) beat the baseline with mean AUROC $\geq$ 0.8 on all datasets. This observation indicates the identity function has not been learnt by any variant of AE despite being overparameterised and having skip connections. In addition, the best mean AUROC scores on most datasets have been achieved by the non-bottlenecked models.

Focusing on CIFAR vs SVHN, our results provide new insights into previous works which reported poor performances [8], [10]. Notably, the best non-bottlenecked model (BAE-Ensemble, type C, AUROC=0.905) and the BAE-∞ (AUROC=0.756) outperform the best bottlenecked model (BAE-Ensemble, type A, AUROC=0.714). These results imply the poor performance could be fixed if previous works were to consider non-bottlenecked architectures.

In an ablation study (Fig. 18), we note that switching to an overcomplete architecture yields consistent improvements over the baseline of bottlenecked deterministic AE. On the other hand, adding skip connections can either greatly increase or decrease the AUROC score. This indicates the unreliability of having skip connections which performance gain is highly data dependent. However, modellers should not leave this option since it has potential for strong gains as demonstrated on CIFAR vs SVHN which could not be improved much by having only an overcomplete latent layer. Switching from a deterministic AE to a BAE improves performance as the best performing BAEs achieve the highest AUROC scores on all datasets. The performance gain is attributed to Bayesian model averaging [95], which accounts for uncertainty in model parameters. The best BAEs also

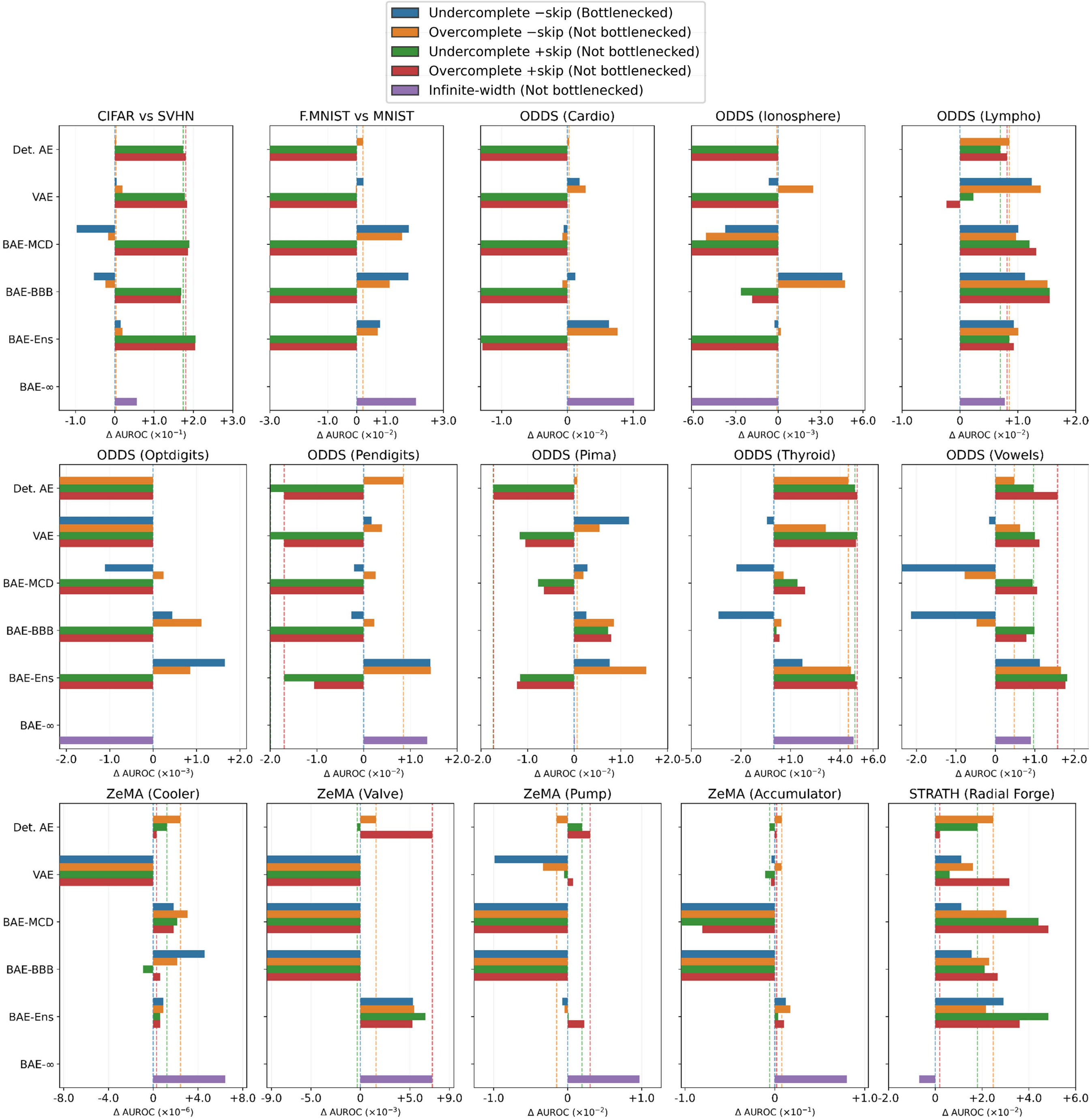


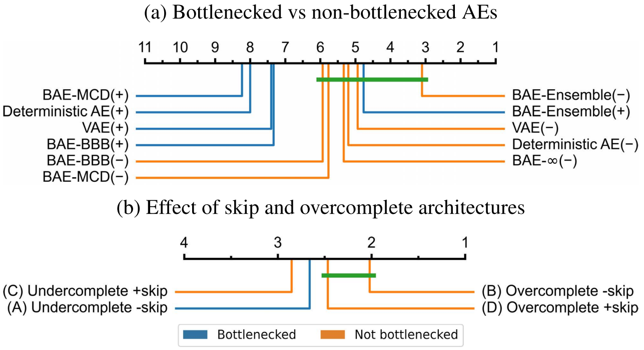
FIGURE 18. **Ablation study: difference in AUROC scores of non-bottlenecked AEs compared to the conventional bottlenecked deterministic AE as a baseline (Δ AUROC=0). For clearer comparisons, vertical dotted lines are drawn for the deterministic AEs and the negative axis is truncated when it is too long. The best non-bottlenecked AEs can outperform the baseline bottlenecked AEs, although not always, demonstrating their effectiveness for anomaly detection.**

outperform the VAEs, evidencing the advantage of addressing the uncertainty over parameters of the entire model instead of considering only the latent layer. However, from Fig. 18, it can be seen that the BAE-BBB, BAE-MCD, and VAE perform worse than the deterministic AE on several datasets (CIFAR vs SVHN, Optdigits, Pendigits, Vowels, and ZeMA), while the BAE-Ensemble consistently perform better than the deterministic AE (both bottlenecked and non-bottlenecked), in alignment with the finding of extant studies [35], [96] that ensembling yields better posterior sampling quality.

Although the BAE-∞ does not score the highest AUROC for all datasets, on a positive note, there are specific tasks on which the BAE-∞ outperforms other models with the highest AUROC. It outperforms all AEs on the ZeMA tasks, demonstrating its effectiveness for condition monitoring. Furthermore, these achievements in accuracy are accompanied with higher robustness than the deterministic AE as indicated before in Section V-B. However, the performance gain is not demonstrated on some tasks in ODDS and STRATH datasets.

TABLE 8. Mean ± standard error AUROC scores for deterministic AEs, VAEs and BAEs with bottlenecked and non-bottlenecked architectures. Model architecture with highest AUROC is bolded for each dataset.

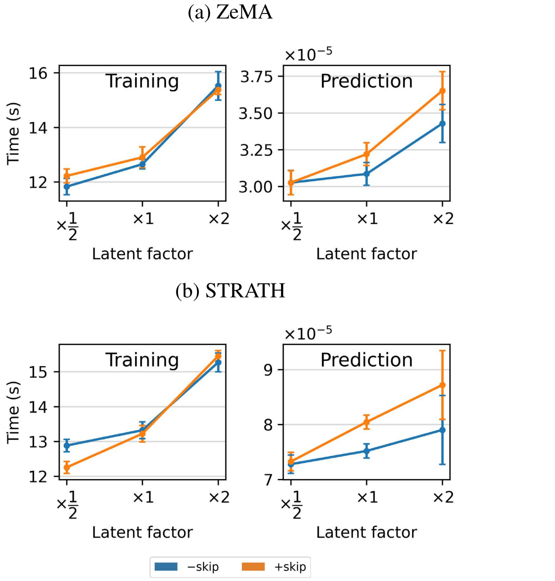
	—Bottlenecked—	—Not bottlenecked—		
Model	A Undercomplete No skip	B Overcomplete No skip	C Undercomplete + skip	D Overcomplete + skip
CIFAR vs SVHN, five runs				
Deterministic AE	0.700±0.024	0.703±0.017	0.874±0.005	0.881±0.009
VAE	0.704±0.015	0.719±0.024	0.878±0.005	0.884±0.005
BAE-MCD	0.603±0.025	0.683±0.029	0.890±0.005	0.886±0.006
BAE-BBB	0.647±0.020	0.676±0.008	0.869±0.006	0.868±0.004
BAE-Ensemble	0.714±0.018	0.720±0.019	**0.905±0.002**	0.904±0.003
BAE-∞	-	0.756±0.002	-	-
FashionMNIST vs MNIST, five runs				
Deterministic AE	0.974±0.002	0.976±0.003	0.870±0.012	0.880±0.008
VAE	0.976±0.002	0.974±0.002	0.821±0.013	0.842±0.018
BAE-MCD	0.992±0.001	0.990±0.001	0.818±0.025	0.822±0.031
BAE-BBB	0.992±0.001	0.986±0.001	0.919±0.004	0.899±0.016
BAE-Ensemble	0.982±0.001	0.981±0.001	0.929±0.005	0.938±0.006
BAE-∞	**-**	**0.995±0.000**	-	-
ODDS, eight tasks, ten runs				
Deterministic AE	0.939±0.035	0.948±0.035	0.927±0.036	0.929±0.036
VAE	0.942±0.033	0.947±0.035	0.925±0.035	0.927±0.035
BAE-MCD	0.933±0.034	0.940±0.034	0.926±0.034	0.928±0.034
BAE-BBB	0.935±0.035	0.943±0.034	0.930±0.032	0.932±0.032
BAE-Ensemble	0.948±0.034	**0.953±0.034**	0.934±0.035	0.936±0.036
BAE-∞	-	0.940±0.035	-	-
ZeMA-Hyd, four tasks, ten runs				
Deterministic AE	0.968±0.023	0.970±0.021	0.967±0.024	0.971±0.023
VAE	0.957±0.022	0.964±0.020	0.959±0.024	0.960±0.023
BAE-MCD	0.885±0.048	0.872±0.047	0.879±0.053	0.887±0.047
BAE-BBB	0.840±0.053	0.839±0.055	0.858±0.053	0.850±0.055
BAE-Ensemble	0.972±0.020	0.974±0.019	0.971±0.022	0.973±0.020
BAE-∞	**-**	**0.992±0.004**	-	-
STRATH, ten runs				
Deterministic AE	0.827±0.025	0.852±0.021	0.845±0.026	0.829±0.016
VAE	0.838±0.026	0.843±0.026	0.833±0.028	0.858±0.023
BAE-MCD	0.838±0.027	0.857±0.032	0.871±0.027	**0.875±0.020**
BAE-BBB	0.842±0.031	0.850±0.022	0.848±0.021	0.853±0.025
BAE-Ensemble	0.856±0.021	0.848±0.038	**0.875±0.021**	0.863±0.030
BAE-∞	-	0.820±0.016	-	-


(a) Bottlenecked vs non-bottlenecked AEs

(b) Effect of skip and overcomplete architectures

FIGURE 19. Critical difference diagrams for comparing (a) bottlenecked and non-bottlenecked AEs, and (b) effects of adding skip connections and overcomplete latent layer on multiple datasets. The models on the right have higher ranks than those on the left. The highest ranked models or treatments which do not show significant differences using post-hoc Nemenyi tests at a 95% confidence level are grouped with a thick green line. In (a), a plus (+) symbol indicates a bottlenecked architecture is used, while (-) indicates otherwise.

The best bottlenecked and non-bottlenecked AEs are compared on each dataset with critical difference diagrams [97], which are commonly used to visualise comparisons of machine learning classifiers. The null hypothesis of the non-parametric Friedman test [98] is rejected with a p-value of 0.00018, indicating a statistical difference exists between one of the AE variants at a 95% confidence level. Then, post-hoc Nemenyi tests [97], [99] are applied, and the highest ranked models which do not show significant differences at a 95% confidence level are grouped with a thick green line in Fig. 19a. It is found that non-bottlenecked AEs have generally higher ranks than their bottlenecked counterparts, indicating the positive effect of removing the bottleneck; the non-bottlenecked BAE-Ensemble holds the highest rank and significantly improves over the bottlenecked deterministic AE.


(a) ZeMA

(b) STRATH

FIGURE 20. Time taken for training 200 epochs and prediction (per example) with a single deterministic AE on sensor datasets: (a) ZeMA and (b) STRATH using a consumer-grade i9-9900K CPU with a Nvidia GeForce 2080 GPU.

In Fig. 19b, similar tests are applied to the architecture classes (type A, B, C or D) which treatments are conditioned on every AE and dataset. It is found that removing the bottleneck with an overcomplete architecture (type B) improves significantly from a bottlenecked architecture (type A), while adding skip connections (type C and D) yields less consistent improvement.

One practical drawback of non-bottlenecked AEs is the increase in computational cost as we increase the latent dimensions. The number of weights in the latent layer connecting the encoder and decoder increase, and hence more time are required for training and prediction, shown in Fig. 20. On a positive note, the increase in computational time is not linear, for instance, increasing the latent factor from $\times\frac{1}{2}$ to $\times 1$ does not double the training time as the computations are parallelised by default in conventional deep learning libraries. While adding skip connections do not affect the training time, the prediction time is visibly increased. In addition, the prediction time using the non-bottlenecked AEs remains much under the working cycles of 60s and 202s in the ZeMA and STRATH testbeds, respectively, demonstrating the predictions can be made in near real-time for every working cycle. Although converting the deterministic AE to a BAE-Ensemble yields 10 times more computational time, the prediction time still remains within the working cycles for these industrial use cases.

VI. LIMITATIONS

Our study has focused on unsupervised anomaly detection and implies nothing about other use cases (e.g. clustering and dimensionality reduction), for which a bottleneck is necessary. Our experiments have covered various data types, however, there may exist datasets where learning the identity function is trivial for the AE. While we lack theoretical proof that non-bottlenecked AEs never learn the identity function, the contrary is true; there is no proof, to the best of our knowledge, that they always learn the identity function.

VII. CONCLUSION

With visualisations on low-dimensional toy data and extensive experiments covering high-dimensional datasets for anomaly detection, we find that non-bottlenecked AEs (including the BAE-∞) can perform reasonably well over the baseline. This result holds even on a conventional deterministic AE which does not have sparsity or Bayesian sampling. The major implications of our work are (1) learning the identity function is not as trivial as previously assumed and (2) modellers should not restrict to only bottlenecked architectures since non-bottlenecked architectures can perform better. The applications on real sensor datasets demonstrate the effectiveness of non-bottlenecked AEs on industrial applications of condition monitoring and quality inspection.

In light of the potential of non-bottlenecked AEs, future work should develop more variants. The tractable solutions of BAE-∞ can facilitate further theoretical work on understanding and proving the conditions for not learning the identity function. Possible directions include understanding the connection between BAEs as predictive density models and kernel density estimation [100], [101]. Future works should also develop methods to improve the robustness of non-bottlenecked AEs without lowering the accuracy, and promising directions are incorporating the rough neurons [12] and informative prior through dictionary learning [13].

ACKNOWLEDGMENT

This work was supported by Research Englands Connecting Capability Fund award (CCF18-7157) for the Promoting the Internet of Things via Collaboration between Higher Education Institutes (HEIs) and Industry (Pitch-In) project and by the European Metrology Programme for Innovation and Research (EMPIR) for the Metrology for the Factory of the Future (MET4FOF) project (17IND12). The authors express gratitude to Tanja Dorst, Andreas Schutze, Christos Tachtatzis, Gordon Gourlay, and Ivan Andonovic for providing pictures and details of the ZeMA and STRATH datasets.

REFERENCES

[1] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection: A review," *ACM Comput. Surveys*, vol. 54, no. 2, Mar. 2021, doi: 10.1145/3439950.

[2] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep one-class classification," in *Proc. Mach. Learn. Res. (PMLR)*, Stockholm, Sweden, vol. 80, Jul. 2018, pp. 4393–4402. [Online]. Available: https://proceedings.mlr.press/v80/ruff18a.html

[3] A. Ng, "Sparse autoencoder," *CS294A Lect. Notes*, vol. 72, pp. 1–19, Jan. 2011.

[4] Z. Chen, C. K. Yeo, B. S. Lee, and C. T. Lau, "Autoencoder-based network anomaly detection," in *Proc. Wireless Telecommun. Symp. (WTS)*, Apr. 2018, pp. 1–5.

[5] J. K. Chow, Z. Su, J. Wu, P. S. Tan, X. Mao, and Y. H. Wang, "Anomaly detection of defects on concrete structures with the convolutional autoencoder," *Adv. Eng. Informat.*, vol. 45, Aug. 2020, Art. no. 101105. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1474034620300744

[6] S. Kim, W. Jo, and T. Shon, "APAD: Autoencoder-based payload anomaly detection for industrial IoE," *Appl. Soft Comput.*, vol. 88, Mar. 2020, Art. no. 106017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494619307999

[7] A. Mujeeb, W. Dai, M. Erdt, and A. Sourin, "One class based feature learning approach for defect detection using deep autoencoders," *Adv. Eng. Informat.*, vol. 42, Oct. 2019, Art. no. 100933. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1474034619301259

[8] E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan, "Do deep generative models know what they don't know?" in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–87.

[9] J. Ren, P. J. Liu, E. Fertig, J. Snoek, R. Poplin, M. Depristo, J. Dillon, and B. Lakshminarayanan, "Likelihood ratios for out-of-distribution detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 14680–14691.

[10] H. Choi, E. Jang, and A. A. Alemi, "WAIC, but why? Generative ensembles for robust anomaly detection," 2018, *arXiv:1810.01392*.

[11] W. Morningstar, C. Ham, A. Gallagher, B. Lakshminarayanan, A. Alemi, and J. Dillon, "Density of states estimation for out of distribution detection," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 3232–3240.

[12] M. Khodayar, O. Kaynak, and M. E. Khodayar, "Rough deep neural architecture for short-term wind speed forecasting," *IEEE Trans. Ind. Informat.*, vol. 13, no. 6, pp. 2770–2779, Dec. 2017.

[13] M. Khodayar, J. Wang, and Z. Wang, "Energy disaggregation via deep temporal dictionary learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1696–1709, May 2020.

[14] A.-S. Collin and C. D. Vleeschouwer, "Improved anomaly detection by training an autoencoder with skip connections on images corrupted with stain-shaped noise," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 7915–7922.

[15] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab, "Bayesian skip-autoencoders for unsupervised hyperintense anomaly detection in high resolution brain mri," in *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2020, pp. 1905–1909.

[16] J. W. Rocks and P. Mehta, "Memorizing without overfitting: Bias, variance, and interpolation in overparameterized models," *Phys. Rev. Res.*, vol. 4, no. 1, Mar. 2022, Art. no. 013201.

[17] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[18] B. X. Yong, Y. Fathy, and A. Brintrup, "Bayesian autoencoders for drift detection in industrial environments," in *Proc. IEEE Int. Workshop Metrol. Ind. (IoT)*, Jun. 2020, pp. 627–631.

[19] M. Hinne, Q. F. Gronau, D. van den Bergh, and E.-J. Wagenmakers, "A conceptual introduction to Bayesian model averaging," *Adv. Methods Practices Psychol. Sci.*, vol. 3, no. 2, pp. 200–215, Jun. 2020.

[20] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[22] G. Zhao, J. Liu, J. Jiang, H. Guan, and J.-R. Wen, "Skip-connected deep convolutional autoencoder for restoration of document images," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 2935–2940.

[23] X. Mao, C. Shen, and Y.-B. Yang, "Image restoration using very deep convolutional encoder–decoder networks with symmetric skip connections," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2016, pp. 2810–2818. [Online]. Available: https://proceedings.neurips.cc/paper/2016/file/0ed9422357395a0d4879191c66f4faa2-Paper.pdf

[24] J.-Y. Liu and Y.-H. Yang, "Denoising auto-encoder with recurrent skip connections and residual regression for music source separation," in *Proc. 17th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2018, pp. 773–778.

[25] J. Kim, J. Ko, H. Choi, and H. Kim, "Printed circuit board defect detection using deep learning via a skip-connected convolutional autoencoder," *Sensors*, vol. 21, no. 15, p. 4968, Jul. 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/15/4968

[26] J. Snoek, R. Adams, and H. Larochelle, "On nonparametric guidance for learning autoencoder representations," in *Proc. Artif. Intell. Statist.*, 2012, pp. 1073–1080.

[27] T. V. Nguyen, R. K. W. Wong, and C. Hegde, "Benefits of jointly training autoencoders: An improved neural tangent kernel analysis," *IEEE Trans. Inf. Theory*, vol. 67, no. 7, pp. 4669–4692, Jul. 2021.

[28] A. Radhakrishnan, K. Yang, M. Belkin, and C. Uhler, "Memorization in overparameterized autoencoders," 2018, *arXiv:1810.10333*.

[29] C. Zhang, S. Bengio, M. Hardt, M. C. Mozer, and Y. Singer, "Identity crisis: Memorization and generalization under extreme overparameterization," 2019, *arXiv:1902.04698*.

[30] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[31] A. Y. Ng, "Feature selection, L_1 vs. L_2 regularization, and rotational invariance," in *Proc. 21st Int. Conf. Mach. Learn. (ICML)*, 2004, p. 78.

[32] T. Chen, E. Fox, and C. Guestrin, "Stochastic gradient Hamiltonian Monte Carlo," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1683–1691.

[33] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.

[34] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1613–1622.

[35] T. Pearce, F. Leibfried, and A. Brintrup, "Uncertainty in neural networks: Approximately Bayesian ensembling," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 234–244.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[37] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

[38] A. Paszke, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

[39] M. Abadi, "Tensorflow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Operating Syst. Design Implement. (OSDI)*, 2016, pp. 265–283.

[40] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham, Switzerland: Springer, 2015, pp. 234–241.

[41] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, "Visualizing the loss landscape of neural nets," 2017, *arXiv:1712.09913*.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[43] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 972–981.

[44] R. Neal, "Bayesian learning for neural networks," Ph.D. dissertation, Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 1995.

[45] R. Novak, L. Xiao, J. Lee, Y. Bahri, G. Yang, J. Hron, D. A. Abolafia, J. Pennington, and J. Sohl-Dickstein, "Bayesian deep convolutional networks with many channels are Gaussian processes," 2018, *arXiv:1810.05148*.

[46] G. Yang, "Wide feedforward or recurrent neural networks of any architecture are Gaussian processes," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2019, pp. 9951–9960.

[47] J. Hron, Y. Bahri, J. Sohl-Dickstein, and R. Novak, "Infinite attention: NNGP and NTK for deep attention networks," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 4376–4386.

[48] T. Chen, J. Morris, and E. Martin, "Probability density estimation via an infinite Gaussian mixture model: Application to statistical process monitoring," *J. Roy. Statist. Soc. C, Appl. Statist.*, vol. 55, no. 5, pp. 699–715, 2006.

[49] B. Sriperumbudur, K. Fukumizu, A. Gretton, A. Hyvärinen, and R. Kumar, "Density estimation in infinite dimensional exponential families," *J. Mach. Learn. Res.*, vol. 18, pp. 1–59, Jul. 2017.

[50] L. Aitchison, "Why bigger is not always better: On finite and infinite neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 156–164.

[51] J. Lee, S. Schoenholz, J. Pennington, B. Adlam, L. Xiao, R. Novak, and J. Sohl-Dickstein, "Finite versus infinite neural networks: An empirical study," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 15156–15172.

[52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[53] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," 2017, *arXiv:1708.07747*.

[54] L. Deng, "The MNIST database of handwritten digit images for machine learning research," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 141–142, Nov. 2012.

[55] A. Krizhevsky, "Learning multiple layers of features from tiny images," Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, Tech. Rep., 2009.

[56] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *Proc. NIPS Workshop Deep Learn. Unsupervised Feature Learn.*, 2011. [Online]. Available: https://research.google/pubs/pub37648

[57] S. Rayana. (2016). *ODDS Library*. [Online]. Available: http://odds.cs.stonybrook.edu

[58] S. Sathe and C. Aggarwal, "LODES: Local density meets spectral outlier detection," in *Proc. SIAM Int. Conf. Data Mining*, Jun. 2016, pp. 171–179.

[59] C. C. Aggarwal and S. Sathe, "Theoretical foundations and algorithms for outlier ensembles," *ACM SIGKDD Explor. Newslett.*, vol. 17, no. 1, pp. 24–47, Jun. 2015.

[60] H.-P. Kriegel, P. Kroger, E. Schubert, and A. Zimek, "Interpreting and unifying outlier scores," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2011, pp. 13–24.

[61] D. Ayres-de-Campos, J. Bernardes, A. Garrido, J. Marques-de-Sá, and L. Pereira-Leite, "Sisporto 2.0: A program for automated analysis of cardiotocograms," *J. Maternal-Fetal Med.*, vol. 9, no. 5, pp. 311–318, 2000.

[62] R. S. Michalski, I. Mozetic, J. Hong, and N. Lavrac, "The multi-purpose incremental learning system AQ15 and its testing application to three medical domains," in *Proc. AAAI*, 1986, pp. 1–41.

[63] J. R. Quinlan, P. J. Compton, K. Horn, and L. Lazarus, "Inductive knowledge acquisition: A case study," in *Proc. 2nd Austral. Conf. Appl. expert Syst.*, 1987, pp. 137–156.

[64] J. W. Smith, J. E. Everhart, W. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in *Proc. Annu. Symp. Comput. Appl. Med. Care*, 1988, p. 261.

[65] L. Xu, A. Krzyzak, and C. Y. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," *IEEE Trans. Syst., Man, Cybern.*, vol. 22, no. 3, pp. 418–435, May 1992.

[66] F. Alimoglu and E. Alpaydin, "Methods of combining multiple classifiers based on different representations for pen-based handwritten digit recognition," in *Proc. 50th Turkish Artif. Intell. Artif. Neural Netw. Symp. (TAINN)*, 1996. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.47.6383

[67] V. G. Sigillito, S. P. Wing, L. V. Hutton, and K. B. Baker, "Classification of radar returns from the ionosphere using neural networks," *Johns Hopkins APL Tech. Dig.*, vol. 10, pp. 262–266, Jul. 1989.

[68] M. Kudo, J. Toyama, and M. Shimbo, "Multidimensional curve classification using passing-through regions," *Pattern Recognit. Lett.*, vol. 20, nos. 11–13, pp. 1103–1111, Nov. 1999.

[69] T. Schneider, S. Klein, and M. Bastuck, "Condition monitoring of hydraulic systems data set at ZeMA," Apr. 2018, doi: 10.5281/zenodo.1323611.

[70] C. Tachtatzis, G. Gourlay, I. Andonovic, and O. Panni, "Sensor data set radial forging at AFRC testbed v2," Sep. 2019, doi: 10.5281/zenodo.3405265.

[71] R. Doddannavar, A. Barnard, and J. Ganesh, *Practical Hydraulic Systems: Operation Troubleshooting for Engineers Technicians*. Amsterdam, The Netherlands: Elsevier, 2005.

[72] N. Helwig, E. Pignanelli, and A. Schütze, "Condition monitoring of a complex hydraulic system using multivariate statistics," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (I2MTC)*, May 2015, pp. 210–215.

[73] L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, and K.-R. Müller, ''A unifying review of deep and shallow anomaly detection,'' *Proc. IEEE*, vol. 109, no. 5, pp. 756–795, May 2021.

[74] T. Dorst, ''Sensor data set of 3 electromechanical cylinder at ZeMA testbed (2kHz),'' May 2019.

[75] W. Karwowski, ''Quality inspection task in modern manufacturing,'' in *International Encyclopedia of Ergonomics and Human Factors-3 Volume Set*. Boca Raton, FL, USA: CRC Press, 2006, pp. 2308–2311.

[76] E. Rauschnabel and V. Schmidt, ''Modern applications of radial forging and swaging in the automotive industry,'' *J. Mater. Process. Technol.*, vol. 35, nos. 3–4, pp. 371–383, Oct. 1992. [Online]. Available: https://www.sciencedirect.com/science/article/pii/092401369290328P

[77] Y. Luo and P. Harris, ''Uncertainty in data analysis for STRATH testbed,'' in *Proc. IEEE Int. Workshop Metrol. Ind. (IoT)*, Jun. 2020, pp. 95–100.

[78] J. W. Tukey *et al.*, *Exploratory Data Analysis*, vol. 2, Reading, MA, USA, 1977.

[79] G. Jiang, P. Xie, H. He, and J. Yan, ''Wind turbine fault detection using a denoising autoencoder with temporal information,'' *IEEE/ASME Trans. Mechatronics*, vol. 23, no. 1, pp. 89–100, Feb. 2018.

[80] F. J. Harris, *Multirate Signal Processing for Communication Systems*. Copenhagen, Denmark: River Publishers, 2021.

[81] P. Virtanen *et al.*, ''SciPy 1.0: Fundamental algorithms for scientific computing in Python,'' *Nature Methods*, vol. 17, pp. 261–272, Feb. 2020.

[82] R. Novak, L. Xiao, J. Hron, J. Lee, A. A. Alemi, J. Sohl-Dickstein, and S. S. Schoenholz, ''Neural tangents: Fast and easy infinite neural networks in Python,'' in *Proc. Int. Conf. Learn. Represent.*, 2020. [Online]. Available: https://research.google/pubs/pub48824/

[83] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, ''Learning representations by back-propagating errors,'' *Nature*, vol. 323, pp. 533–536, Oct. 1986.

[84] A. L. Maas, A. Y. Hannun, and A. Y. Ng, ''Rectifier nonlinearities improve neural network acoustic models,'' in *Proc. ICML Workshop Deep Learn. Audio, Speech Lang. Process.*, 2013. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.693.1422

[85] J. L. Ba, J. R. Kiros, and G. E. Hinton, ''Layer normalization,'' 2016, *arXiv:1607.06450*.

[86] L. N. Smith, ''Cyclical learning rates for training neural networks,'' in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 464–472.

[87] F. Melo, *Area under ROC Curve*. New York, NY, USA: Springer, 2013, pp. 38–39.

[88] D. W. Hosmer, Jr., S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, vol. 398. Hoboken, NJ, USA: Wiley, 2013.

[89] E. Kalapanidas, N. Avouris, M. Craciun, and D. Neagu, ''Machine learning algorithms: A study on noise sensitivity,'' in *Proc. 1st Balcan Conf. Informat.*, 2003, pp. 356–365.

[90] A. Atla, R. Tada, V. Sheng, and N. Singireddy, ''Sensitivity of different machine learning algorithms to noise,'' *J. Comput. Sci. Colleges*, vol. 26, no. 5, pp. 96–103, 2011.

[91] S. Basodi, C. Ji, H. Zhang, and Y. Pan, ''Gradient amplification: An efficient way to train deep neural networks,'' *Big Data Mining Anal.*, vol. 3, no. 3, pp. 196–207, Sep. 2020.

[92] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, ''Robustness may be at odds with accuracy,'' 2018, *arXiv:1805.12152*.

[93] Y.-Y. Yang, C. Rashtchian, H. Zhang, R. R. Salakhutdinov, and K. Chaudhuri, ''A closer look at accuracy vs. robustness,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 8588–8601.

[94] A. Rozsa, M. Gunther, and T. E. Boult, ''Are accuracy and robustness correlated,'' in *Proc. 15th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2016, pp. 227–232.

[95] M. Hinne, Q. F. Gronau, D. van den Bergh, and E.-J. Wagenmakers, ''A conceptual introduction to Bayesian model averaging,'' *Adv. Methods Practices Psychol. Sci.*, vol. 3, no. 2, pp. 200–215, Jun. 2020, doi: 10.1177/2515245919898657.

[96] J. Yao, W. Pan, S. Ghosh, and F. Doshi-Velez, ''Quality of uncertainty quantification for Bayesian neural network inference,'' 2019, *arXiv:1906.09686*.

[97] J. Demšar, ''Statistical comparisons of classifiers over multiple data sets,'' *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Dec. 2006.

[98] M. Friedman, ''The use of ranks to avoid the assumption of normality implicit in the analysis of variance,'' *J. Amer. Statist. Assoc.*, vol. 32, no. 200, pp. 675–701, Dec. 1937.

[99] P. B. Nemenyi, ''Distribution-free multiple comparisons,'' Ph.D. dissertation, Dept. Statist., Princeton Univ., Princeton, NJ, USA, 1963.

[100] M. Rosenblatt, ''Remarks on some nonparametric estimates of a density function,'' *Ann. Math. Statist.*, vol. 27, no. 3, pp. 832–837, 1956, doi: 10.1214/aoms/1177728190.

[101] E. Parzen, ''On estimation of a probability density function and mode,'' *Ann. Math. Statist.*, vol. 33, no. 3, pp. 1065–1076, Sep. 1962.

BANG XIANG YONG received the B.Eng. degree in electronics engineering from Universiti Teknologi Malaysia. He is currently pursuing the Ph.D. degree in engineering with the University of Cambridge. His research interests include Bayesian deep learning, uncertainty quantification, and anomaly detection.

ALEXANDRA BRINTRUP received the Ph.D. degree from Cranfield University, Cranfield, U.K. She is currently a Lecturer in digital manufacturing with the University of Cambridge, Cambridge, U.K. She develops intelligent systems to help organizations navigate through complexity. Her main work in this area includes system development for digitized product lifecycle management. She uses artificial intelligence paradigms, particularly for data analytics and automated decision making. She has a postdoctoral and a fellowship appointments with the University of Cambridge and the University of Oxford. She teaches operations management and decision engineering. Her research interests include modeling, analysis, and control of dynamical and functional properties of emergent manufacturing networks.