Supplementary Materials of "Quadratic Neuron-empowered Heterogeneous Autoencoder for Unsupervised Anomaly Detection"

Jing-Xiao Liao [1], Bo-Jian Hou[2], Hang-Cheng Dong[1], Hao Zhang[2], Jianwei Ma[3], Jinwei Sun[1], Shiping Zhang[1*], Feng-Lei Fan[2*]

I. Theorems and Proofs

Theorem 1 (Main)**.** *There exist universal constants c_σ, $c_1, c_2, c_3, c_4, C_1, C_2, \epsilon_1, \epsilon_2 > 0$, where c_σ depends on the activation function used in a network, and the following holds. For every dimension d, there exists a measure μ and a function $\tilde{g}_{ip}(\boldsymbol{x}) + \tilde{g}_r(\boldsymbol{x}) : \mathbb{R}^d \rightarrow \mathbb{R}$ with the following properties:*

1. *Every f_1 expressed by a one-hidden-layer conventional network of width at most $c_1 e^{c_2 d}$ has*
$$\mathbb{E}_{\boldsymbol{x}\sim\mu}[f_1 - (\tilde{g}_{ip} + \tilde{g}_r)]^2 \geq \epsilon_2. \tag{1}$$
2. *Every f_2 expressed by a one-hidden-layer quadratic network of width at most $c_3 e^{c_4 d}$ has*
$$\mathbb{E}_{\boldsymbol{x}\sim\mu}[f_2 - (\tilde{g}_{ip} + \tilde{g}_r)]^2 \geq \epsilon_1. \tag{2}$$
3. *There exists a function f_3 expressed by a one-hidden-layer heterogeneous network that can approximate $\tilde{g}_{ip} + \tilde{g}_r$ with $C_1 c_\sigma d^{3.75} + C_2 c_\sigma$ neurons.*

Theorem 2. *There exist universal constants c_σ, $c_1, c_2, C_1, \epsilon_1 > 0$, where c_σ depends on the activation function used in a network, and the following holds. For every dimension d, there exists a measure μ and an inner-product based function $\tilde{g}_{ip}(\boldsymbol{x}) : \mathbb{R}^d \rightarrow \mathbb{R}$ with the following properties:*

1. *Every f_2 that is expressed by a one-hidden-layer quadratic network of width at most $c_1 e^{c_2 d}$ has*
$$\mathbb{E}_{\boldsymbol{x}\sim\mu}[f_2 - \tilde{g}_{ip}]^2 \geq \epsilon_1. \tag{3}$$
2. *There exists a function f_1 that is expressed by a one-hidden-layer conventional network that can approximate $\tilde{g}_r$ with $C_1 c_\sigma$ neurons.*

Theorem 3 (Theorem 1 of [1])**.** *There exist universal constants $c_\sigma, c_3, c_4, C_2, \epsilon_2 > 0$, where c_σ depends on the activation function used in a network, and the following holds. For every dimension d, there exists a measure μ and a radial function $\tilde{g}_r(\boldsymbol{x}) : \mathbb{R}^d \rightarrow \mathbb{R}$ with the following properties:*

1. *Every f_1 that is expressed by a one-hidden-layer conventional network of width at most $c_3 e^{c_4 d}$ has*
$$\mathbb{E}_{\boldsymbol{x}\sim\mu}[f_1 - \tilde{g}_r]^2 \geq \epsilon_2. \tag{4}$$
2. *There exists a function f_2 expressed by a one-hidden-layer quadratic network that can approximate $\tilde{g}_r$ with $C_2 c_\sigma d^{3.75}$ neurons.*

Lemma 1. *Given a unit vector $\boldsymbol{v}$, $\tilde{g}_{ip}(\boldsymbol{x}) = \frac{1}{2\pi}\sqrt{\frac{4R_d}{1-\delta}}\text{sinc}\left(\frac{2R_d}{1-\delta}\boldsymbol{x}^\top\boldsymbol{v}\right)$ satisfies $\int_{2R_d\mathbb{B}_d} \hat{\tilde{g}}_{ip}^2(\boldsymbol{\omega})d\boldsymbol{\omega} = 1-\delta, \delta \in [0,1]$.*

Proof. Let $\{\boldsymbol{\omega} = t\boldsymbol{v} \mid t \in \mathbb{R}\}$ be the support of $\hat{\tilde{g}}_{ip}(\boldsymbol{\omega})$. $\tilde{g}_{ip}(\boldsymbol{x})$ is mathematically formulated as

$$\hat{\tilde{g}}_{ip}(\boldsymbol{\omega}) = \begin{cases} \sqrt{\frac{1-\delta}{4R_d}}, & \boldsymbol{\omega} = t\boldsymbol{v}, t \in [-\frac{2R_d}{1-\delta}, \frac{2R_d}{1-\delta}] \\ 0, & \boldsymbol{\omega} = t\boldsymbol{v}, t \notin [-\frac{2R_d}{1-\delta}, \frac{2R_d}{1-\delta}]. \end{cases} \tag{5}$$

Because $\hat{\tilde{g}}_{ip}$ is a constant over $[-\frac{2R_d}{1-\delta}, \frac{2R_d}{1-\delta}]$, we have $\int_{2R_d\mathbb{B}_d} \hat{\tilde{g}}_{ip}^2(\boldsymbol{\omega})d\boldsymbol{\omega} = \int_{2R_d\mathbb{B}_d} \frac{1-\delta}{4R_d}d\boldsymbol{\omega} = \frac{1-\delta}{4R_d}\int_{[-2R_d,2R_d]} dt = 1-\delta$, which concludes the proof. □

Lemma 2. *Let $g, f : \mathbb{R}^d \rightarrow \mathbb{R}$ be two functions of unit norm. Moreover, $\int_{2R_d\mathbb{B}_d} g^2(\boldsymbol{x})d\boldsymbol{x} \leq 1-\delta$, $\delta \in [0,1]$, g is supported over $\text{Span}\{\boldsymbol{v}\} + R_d\mathbb{B}_d$, and f is a combination of k radial functions, denoted as $f = \sum_{i=1}^k m_i(\|\boldsymbol{x}\|)$, where $m_i(\|\boldsymbol{x}\|)$ is a continuous radial function. Then,*

$$\langle f, g \rangle_{L_2} \leq 1 - \frac{\delta}{2} + k\exp(-cd/2),$$

where $c > 0$ is a constant dependent on f.

Proof. Let $T = \text{Span}\{\boldsymbol{v}\} + R_d\mathbb{B}_d$. For any radius $r > 0$, define $h(r) = \frac{Area(r\mathbb{S}^{d-1}\cap T)}{Area(r\mathbb{S}^{d-1})}$, where $\mathbb{S}^{d-1}$ is the unit Euclidean sphere in $\mathbb{R}^d$, and $Area(\cdot)$ is to compute the area. The geometric meaning of $h(r)$ is the ratio of the area of intersection of T and $r\mathbb{S}^{d-1}$ vs the area of $r\mathbb{S}^{d-1}$. Because T is a hypertube, following the illustration in page 7 of [2], $h(r)$ is exponentially small, *i.e.*, $\exists\ c > 0$, such that

$$h(2R_d) \leq \exp(-cd). \tag{6}$$

As r increases, the area of intersection decreases but the area of the sphere increases. Therefore, $h(r)$ is monotonically decreasing.

*Shiping Zhang (spzhang@hit.edu.cn) and Feng-Lei Fan (hitfanfenglei@gmail.com) are co-corresponding authors.

[1] Jing-Xiao Liao, Hang-Cheng Dong, Jinwei Sun, Shiping Zhang are with School of Instrumentation Science and Engineering, Harbin Institute of Technology, Harbin, China.

[2] Feng-Lei Fan, Bo-Jian Hou, Hao Zhang are with Weill Cornell Medicine, Cornell University, New York City, US.

[3] Jianwei Ma is with School of Mathematics, Harbin Institute of Technology, Harbin, China.

Our goal is to show $\langle f, g\rangle_{L_2}$ can be upper bounded. We decompose the inner product $\langle f, g\rangle_{L_2}$ into the integrals in a hyperball $2R_d\mathbb{B}_d$ and the region out of the hyperball $(2R_d\mathbb{B}_d)^C$. Mathematically,

$$\langle f, g\rangle_{L_2} = \int_{2R_d\mathbb{B}_d} f(\boldsymbol{x})g(\boldsymbol{x})d\boldsymbol{x} + \int_{(2R_d\mathbb{B}_d)^C} f(\boldsymbol{x})g(\boldsymbol{x})d\boldsymbol{x}. \tag{7}$$

Now, we calculate the above two integrals, respectively. For the first integral, we bound it as follows:

$$\int_{2R_d\mathbb{B}_d} f(\boldsymbol{x})\, g(\boldsymbol{x})\, d\boldsymbol{x} \overset{(1)}{\leq} \left\| g\cdot \mathbf{1}_{\{2R_d\mathbb{B}_d\}}\right\|_{L_2} \|f\|_{L_2} \overset{(2)}{\leq} \sqrt{1-\delta} \tag{8}$$

(1) follows from the Cauchy-Schwartz inequality; (2) follows from the fact that $f(\cdot)$ has unit L_2 norm and $\int_{2R_d\mathbb{B}_d} g^2(\boldsymbol{x})d\boldsymbol{x} \leq 1-\delta$.

For the second integral, we have $\int_{(2R_d\mathbb{B}_d)^C} f(\boldsymbol{x})\, g(\boldsymbol{x})\, d\boldsymbol{x} = \int_{(2R_d\mathbb{B}_d)^C \bigcap T} f(\boldsymbol{x})\, g(\boldsymbol{x})\, d\boldsymbol{x}$ because $g(\boldsymbol{x}) = 0$ when $\boldsymbol{x} \notin T$. Then, we have

$$\begin{aligned}
&\int_{(2R_d\mathbb{B}_d)^C \bigcap T} f(\boldsymbol{x})\, g(\boldsymbol{x})\, d\boldsymbol{x}\\
=&\sum_{i=1}^{k}\int_{(2R_d\mathbb{B}_d)^C \bigcap T} m_i(\|\boldsymbol{x}\|)\, g(\boldsymbol{x})\, d\boldsymbol{x}\\
\leq&\sum_{i=1}^{k}\left(\sqrt{\int_{(2R_d\mathbb{B}_d)^C \bigcap T} m_i^2(\|\boldsymbol{x}\|)\, d\boldsymbol{x}}\sqrt{\int_{(2R_d\mathbb{B}_d)^C \bigcap T} g^2(\boldsymbol{x})\, d\boldsymbol{x}}\right)\\
\overset{(1)}{\leq}&\sum_{i=1}^{k}\sqrt{\int_{r\geq 2R_d}\int_{(r\mathbb{S}^{d-1})\bigcap T} m_i^2(r)\, d\mathbb{S}dr}\\
=&\sum_{i=1}^{k}\sqrt{\int_{r\geq 2R_d}\left(\int_{r\mathbb{S}^{d-1}} m_i^2(r)\, d\mathbb{S}\right)\cdot\Big[\frac{\int_{(r\mathbb{S}^{d-1})\cap T} m_i^2(r)\, d\mathbb{S}}{\int_{r\mathbb{S}^{d-1}} m_i^2(r)\, d\mathbb{S}}\Big]dr}\\
\overset{(2)}{=}&\sum_{i=1}^{k}\sqrt{\int_{r\geq 2R_d}\int_{r\mathbb{S}^{d-1}} m_i^2(r)\, d\mathbb{S}\cdot h(r)\, dr}\\
\overset{(3)}{\leq}&\sum_{i=1}^{k}\sqrt{h(2R_d)\int_{r\geq 2R_d}\int_{r\mathbb{S}^{d-1}} m_i^2(r)\, d\mathbb{S}dr}\\
\overset{(4)}{=}&\sum_{i=1}^{k}\sqrt{h(2R_d)} \overset{(5)}{\leq} k\exp(-cd/2)\,.
\end{aligned} \tag{9}$$

In the above, (1) follows from the facts that $g(\cdot)$ has unit L_2 norm; (2) holds because $m_i(r)$ is radial; (3) follows from $h(r)$ is monotonically decreasing; (4) follows from $f(\cdot)$ has a unit norm; (5) follows from Eq. (6).

Combining Eqs. (7), (8) and (9), we have $\langle f, g\rangle_{L_2} \leq \sqrt{1-\delta} + k\exp(-cd/2) \leq 1 - \frac{\delta}{2} + k\exp(-cd/2)$.

□

Proof of ***Theorem*** *2.* The proof of Theorem 2 consists of two parts: (i) the inapproximatability of a quadratic network to $\tilde{g}_{ip}$; (ii) the approximatability of a conventional network to $\tilde{g}_{ip}$.

(i) Define $l = \frac{\widehat{\tilde{g}_{ip}\varphi}}{\|\tilde{g}_{ip}\varphi\|_{L_2}}$. Because for $\tilde{g}_{ip}$, we have $\int_{2R_d\mathbb{B}_d} \tilde{g}_{ip}^2(\boldsymbol{x})d\boldsymbol{x} = 1-\delta$, there must exist a universal constant $c_1 \in [0,1]$ such that $\int_{2R_d\mathbb{B}_d} l^2(\boldsymbol{x})d\boldsymbol{x} \leq 1-c_1$. Define the function $q = \frac{\widehat{f_2\varphi}}{\|f_2\varphi\|_{L_2}}$, where $f_2 = \sum_i^k a_i\sigma(\omega_i\boldsymbol{x}^\top\boldsymbol{x} + b_i)$ is a radial function expressed by a quadratic network. Thus, according to Lemma 2, the functions $l(\cdot)$, $q(\cdot)$ satisfy

$$\langle l(\cdot), q(\cdot)\rangle_{L_2} \leq 1 - \frac{c_1}{2} + k\exp(-c_2 d/2), \tag{10}$$

with $c_2 > 0$ being a universal constant.

For every scalars $\beta_1, \beta_2 > 0$, we have

$$\|\beta_1 l(\cdot) - \beta_2 q(\cdot)\|_{L_2} \geqslant \frac{\beta_2}{2}\|l(\cdot) - q(\cdot)\|_{L_2}\,. \tag{11}$$

The reason why it holds true is as follows.

Without loss of generality, we assume that $\beta_2 = 1$. For two unit vectors u, v in one Hilbert space, it has

$$\begin{aligned}
&\min_{\beta}\|\beta v - u\|^2 = \min_{\beta}\left(\beta^2\|v\|^2 - 2\beta\langle v, u\rangle + \|u\|^2\right)\\
=&\min_{\beta}\left(\beta^2 - 2\beta\langle v, u\rangle + 1\right) = 1 - \langle v, u\rangle^2 = \frac{1}{2}\|v-u\|^2.
\end{aligned}$$

Next, combining Eqs. (10) and (11), and utilizing that q, l have unit L_2 norm, we have

$$\begin{aligned}
&\|f_2 - \tilde{g}_{ip}\|_{L_2(\mu)} = \|f_2\varphi - \tilde{g}_{ip}\varphi\| = \|\widehat{f_2\varphi} - \widehat{\tilde{g}_{ip}\varphi}\|\\
=&\|(\|f_2\varphi\|)\, q(\cdot) - (\|\tilde{g}_{ip}\varphi\|_{L_2})\, l(\cdot)\|\\
\geq&\frac{1}{2}\|\tilde{g}_{ip}\varphi\|\|q(\cdot) - l(\cdot)\| = \frac{1}{2}\|\tilde{g}_{ip}\|_{L_2(\mu)}\|q(\cdot) - l(\cdot)\|_{L_2}\\
\geq&\frac{1}{2}\|\tilde{g}_{ip}\|_{L_2(\mu)}\sqrt{2\left(1 - \langle q, l\rangle_{L_2}\right)}\\
\geq&\|\tilde{g}_{ip}\|_{L_2(\mu)}\sqrt{2\max\left(c_1/2 - k\exp\left(-c_2 d\right), 0\right)},
\end{aligned} \tag{12}$$

where the first inequality is due to Eq. (11). Therefore, as long as $k \leq \frac{c_1}{2}e^{c_2 d}$, $\|f_2 - \tilde{g}_{ip}\|_{L_2(\mu)}$ is larger than a positive constant ϵ_1.

(ii) In contrast, f_1 can approximate $\tilde{g}_{ip}(\boldsymbol{x}) = \frac{1}{2\pi}\sqrt{\frac{4R_d}{1-\delta}}\mathrm{sinc}\left(\frac{2R_d}{1-\delta}\boldsymbol{x}^\top\boldsymbol{v}\right)$ with a polynomial number of neurons. From the proof of Theorem 1 in [3], $\forall H : \mathbb{R} \to \mathbb{R}$ which is constant outside a bounded interval $[r_1, r_2]$, there exist scalars $a, \{\alpha_i, \beta_i, \gamma_i\}_{i=1}^k$, where $k \leq c_\sigma\frac{(r_2-r_1)L}{\epsilon}$ such that the function $h(t) = a + \sum_{i=1}^k \alpha_i\cdot\sigma(\beta_i t - \gamma_i)$ satisfies that $\sup_{t\in\mathbb{R}} |H(t) - h(t)| \leq \epsilon$.

According to this theorem, we define an auxiliary function $\tilde{g}_{ip}^{(t)}(\boldsymbol{x})$ by truncating $\tilde{g}_{ip}(\boldsymbol{x})$ when $\boldsymbol{x}^\top\boldsymbol{v} \notin [-\frac{1-\delta}{R_d\epsilon}, \frac{1-\delta}{R_d\epsilon}]$ to be zero, then $|\tilde{g}_{ip}^{(t)}(\boldsymbol{x}) - \tilde{g}_{ip}(\boldsymbol{x})| \leq |\tilde{g}_{ip}(\boldsymbol{x})| \leq \frac{1}{2\pi}\sqrt{\frac{4R_d}{1-\delta}}/(\frac{2R_d}{1-\delta}\cdot\boldsymbol{x}^\top\boldsymbol{v}) \leq \epsilon/2$ when $\boldsymbol{x}^\top\boldsymbol{v} \notin [-\frac{\sqrt{1-\delta}}{\pi\sqrt{R_d}\epsilon}, \frac{\sqrt{1-\delta}}{\pi\sqrt{R_d}\epsilon}]$. Applying the aforementioned theorem to $\tilde{g}_{ip}^{(t)}(\boldsymbol{x})$, we have $f_1(\boldsymbol{x}) = a + \sum_{i=1}^k \alpha_i\cdot\sigma(\beta_i\boldsymbol{x}^\top\boldsymbol{v} - \gamma_i)$ that can approximate $\tilde{g}_{ip}^{(t)}$ in terms of $\sup_{\boldsymbol{x}\in\mathbb{R}^d} |\tilde{g}_{ip}^{(t)}(\boldsymbol{x}) - f_1(\boldsymbol{x})| \leq \epsilon/2$. Here, $r_1 = -\frac{\sqrt{1-\delta}}{\pi\sqrt{R_d}\epsilon}$ and $r_2 = \frac{\sqrt{1-\delta}}{\pi\sqrt{R_d}\epsilon}$, the number of neurons needed is no more than $c_\sigma\frac{2\sqrt{1-\delta}L}{\pi\sqrt{R_d}\epsilon} \leq c_\sigma\frac{2\times\sqrt{5.264}\sqrt{1-\delta}L}{\pi\epsilon} = C_1 c_\sigma$. Because the geometric meaning of $1/R_d$ is the volume of d-hyperball, which is upper bounded by $(1/R_d)_{d=5} \approx 5.264$[1], the number of needed neurons is irrelevant to d. Furthermore, f_1 fulfills that

$$\begin{aligned}
&\sup_{\boldsymbol{x}\in\mathbb{R}^d}|\tilde{g}_{ip}(\boldsymbol{x}) - f_1(\boldsymbol{x})|\\
\leq&\sup_{\boldsymbol{x}\in\mathbb{R}^d}|\tilde{g}_{ip}(\boldsymbol{x}) - \tilde{g}_{ip}^{(t)}(\boldsymbol{x})| + \sup_{\boldsymbol{x}\in\mathbb{R}^d}|\tilde{g}_{ip}^{(t)}(\boldsymbol{x}) - f_1(\boldsymbol{x})| \leq \epsilon.
\end{aligned} \tag{13}$$

[1] https://en.wikipedia.org/wiki/Volume_of_an_n-ball

□

*Proof of **Theorem** 1.* Theorems 3 and 2 suggest that a conventional network f_1 can express $\tilde{g}_{ip}$ with a polynomial number of neurons, and a quadratic network f_2 can express $\tilde{g}_r$ with a polynomial number of neurons. Let a heterogeneous network be $f_3 = f_1 + f_2$, f_3 can straightforwardly express $\tilde{g}_{ip} + \tilde{g}_r$ with a polynomial number of neurons.

Next, we need to show how to deduce from $\mathbb{E}_{\boldsymbol{x}\sim\mu}[f_2 - \tilde{g}_{ip}]^2 \geq \epsilon_1$ to $\mathbb{E}_{\boldsymbol{x}\sim\mu}[f_2 - (\tilde{g}_{ip} + \tilde{g}_r)]^2 \geq \epsilon_1$. Here, we slightly abuse ϵ_1 for succinctness. Both formulas mean that there exists a gap between functions. We can rewrite $\tilde{g}_r$ as

$$\tilde{g}_r(\|\boldsymbol{x}\|) = \tilde{g}_r \circ s(\|\boldsymbol{x}\|^2), \tag{14}$$

where $s(\cdot)$ is $\sqrt{\cdot}$. Thus, we can express $\tilde{g}_r$ by a quadratic neuron using a special activation function $\sigma' = \tilde{g}_r \circ s$. It holds that $\sigma'(x) \leq C'(1 + |x|^{\alpha'})$, since both $\tilde{g}_r$ and s are bounded. As a result, regarding $f_2 - \tilde{g}_r$ as a quadratic network, we can get $\mathbb{E}_{\boldsymbol{x}\sim\mu}[f_2 - (\tilde{g}_{ip} + \tilde{g}_r)]^2 = \mathbb{E}_{\boldsymbol{x}\sim\mu}[(f_2 - \tilde{g}_r) - \tilde{g}_{ip}]^2 \geq \epsilon_1$.

Similarly, $\tilde{g}_{ip}$ can be re-expressed by a conventional neuron with a bounded activation. We can also get $\mathbb{E}_{\boldsymbol{x}\sim\mu}[f_1 - (\tilde{g}_{ip} + \tilde{g}_r)]^2 \geq \epsilon_2$ from $\mathbb{E}_{\boldsymbol{x}\sim\mu}[f_1 - \tilde{g}_r]^2 \geq \epsilon_2$.

□.

II. Training strategies.

The training process of a quadratic network suffers the risk of collapse due to the high nonlinearity. For example, the output function of a quadratic network with L layers will be a polynomial of 2^L degrees, which is too high to keep the network stable during training. To fix this issue, Fan *et al.* [4] proposed the so-called ReLinear algorithm, where the parameters in a quadratic neuron are initialized as $\boldsymbol{w}^g = 0, \boldsymbol{w}^b = 0, \boldsymbol{c} = 0$ and $\boldsymbol{b}^g = 1$, while $\boldsymbol{w}^r$ and b^r follow the random initialization. Consequently, during the initialization stage, a quadratic neuron degenerates to a conventional neuron. Then, during the training stage, different learning rates are cast for $(\boldsymbol{w}^r, b^r)$ and $(\boldsymbol{w}^g, \boldsymbol{w}^b, c, b^g)$: a normal learning rate for the former and a relatively smaller learning rate for the latter. By doing so, the nonlinearity of quadratic terms are constrained, and the learned quadratic network can avoid the training instability. The schematic illustration of the ReLinear algorithm is shown in Figure 1.

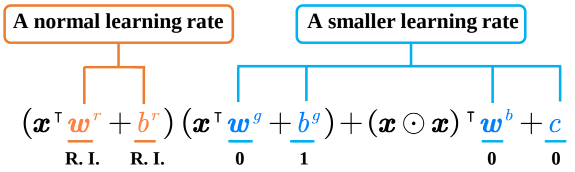

A normal learning rate
A smaller learning rate
$(\boldsymbol{x}^\top \boldsymbol{w}^r + b^r)(\boldsymbol{x}^\top \boldsymbol{w}^g + b^g) + (\boldsymbol{x} \odot \boldsymbol{x})^\top \boldsymbol{w}^b + c$
R. I.
R. I.
0
1
0
0

Fig. 1: ReLinear: a training strategy of quadratic networks (R. I. means random initialization).

III. Bearing Fault Anomaly Detection Experiments

We conduct experiments on bearing fault detection, which is a high-dimensional problem. Bearings are widely used in rotating machinery, but they are prone to damage. Anomaly detection determines whether the bearing is working under normal conditions, which is the key to bearing health management. Different from tabular data, bearing fault detection is based on long vibration signals. We use two datasets of run-to-failure bearing data to validate our method: one from the public[5] and the other provided by a nuclear power plant in Hainan Province, China.

Dataset descriptions. 1) XJTU-SY dataset is collected by the Institute of Design Science and Basic Component at Xi'an Jiaotong University (XJTU) and the Changxing Sumyoung Technology Co., Ltd. (SY) [5]. Two accelerometers of type PCB 352C33 (sampling frequency set to 25.6 kHz) are mounted at 90° on the housing of the tested bearings, which measure the vibration of the bearing. The accelerometers record 1.28s (32768 points) per minute until the bearing fails completely. We select the Bearing2_5 data which runs at 2250 rpm (37.5 Hz) and 11 kN load. The bearing's lifetime is 5 h 39 min and ends with an outer race fault.

2) The seawater booster pump (SBP) data are gathered from Hainan Nuclear Power Co., Ltd. The SBP transfers cooling water to steam turbines and serves as a crucial component of the auxiliary cooling water system in nuclear power plants. Given the intricate nature of nuclear plants, mechanical systems require periodic monitoring to ensure their reliability. Figure 2 illustrates that acceleration sensors are placed at four points on both the pump and the motor when performing measurement. The vibration signal for the seawater booster pumps is measured every two months, with a sampling rate of 16.8 kHz and a recording period of 0.4 seconds, which results in 6,721 data points. The faulty and healthy signals are shown in Figure 3. Data in this study are collected from 2015 to 2023. During this period, a bearing fault at the outer race is observed.

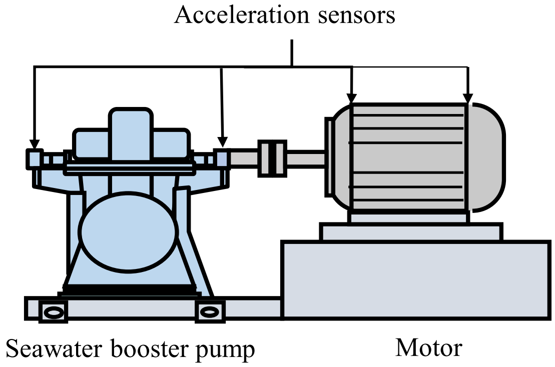


Fig. 2: The measurement structure of seawater booster pump.

We resample all signals to 1024 points per sample and use Fast Fourier Transform to split a single side of it into 512 points. The summary of the two datasets is in Table I. Outliers in the XJTU dataset are identified based on the method described in [6]. However, it should be noted that unlike simulation tests, the availability of faulty data from real industrial scenes is very limited. As a result, only 6 samples were available for outlier detection in the SBP dataset, which

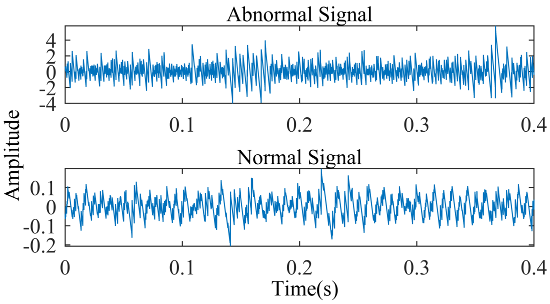


Fig. 3: Abnormal and normal signals of the seawater booster pump.

TABLE I: The summary of two bearing datasets. a×b represents the number of samples and sample dimensions.

Datasets	XJTU	SBP
#Raw Sample	334×32768	57×6721
#Abnormal Raw Sample	222×32768	1×6721
#Resample	8544×512	342×512
#Abnormal Resample	4671×512	6×512
Outlier Ratio	54.66%	1.75%

makes this task pretty challenging.

Table II presents the classification results, indicating that HAE-based methods outperform other competitors in both datasets. It should be noted that DAGMM is not included as a comparison method because its training fails. The SUOD model performs well in the SBP but achieves the limited performance in XJTU dataset, while the GAAL is the opposite. The RCA model demonstrates the best precision in the XJTU dataset and the best recall in the SBP dataset. However, it fails in some other metrics. Autoencoder-based models demonstrated better results in both datasets, implying that autoencoder-based models are relatively suitable for handling higher dimensional data. In particular, HAE models are consistently better than AE and QAE, which shows that the combination of different neuron types is a wise strategy.

IV. Ablation Study

Neuron types in HAEs. Earlier, we compared the performance between HAEs and purely conventional or quadratic autoencoders. However, the proposed HAEs use different architectures from AE and QAE to better synergize the power of conventional and quadratic neurons. Consequently, it is not sure if the superior performance of HAEs is due to the synergy of different neurons or due to the architecture. To resolve this ambiguity, we replace neurons in HAEs to prototype homogeneous autoencoder, as Table III shows. Then, we conduct experiments on 15 anomaly detection datasets to compare the difference in performance between HAEs and homogeneous autoencoders using the same architecture.

TABLE II: Classification results of all baseline methods over two bearing fault datasets. Bold-faced numbers are better results.

Datasets	Methods	AUC	Pre	Recall	F1
XJTU	SUOD	0.874	0.714	0.668	0.686
	SO-GAAL	0.958	0.896	0.724	0.777
	DeepSVDD	0.828	0.633	0.652	0.641
	RCA	0.946	**1.000**	0.528	0.691
	AE	0.942	0.755	0.712	0.731
	QAE	0.944	0.790	0.715	0.744
	HAE-X	0.957	0.956	**0.740**	**0.803**
	HAE-Y	0.958	0.957	0.737	0.800
	HAE-I	**0.961**	0.959	0.730	0.795
SBP	SUOD	**1.000**	0.875	0.985	0.921
	SO-GAAL	0.831	0.723	0.894	0.781
	DeepSVDD	**1.000**	0.700	0.967	0.769
	RCA	**1.000**	0.018	**1.000**	0.035
	AE	**1.000**	0.850	0.956	0.911
	QAE	**1.000**	0.873	0.963	0.904
	HAE-X	**1.000**	0.876	0.997	0.933
	HAE-Y	**1.000**	**0.883**	0.995	**0.935**
	HAE-I	**1.000**	0.881	0.990	0.932

The results are presented in Tables IV and V. First, both AUC and precision metrics show that the performance of HAEs is superior to their homogeneous counterparts. Specifically, HAE-X achieves the highest AUC scores on 11 datasets, while HAE-Y performs best on 8 datasets and HAE-I outperforms its counterparts on 10 datasets. All quadratic neuron-based autoencoders are better than conventional ones, which underscores the powerful feature representation capabilities of quadratic neurons. At last, we observe that HAE-I with quadratic neurons in the first layer outperforms those with conventional neurons (HAE-Ic), suggesting that quadratic neurons are more adept at extracting hidden features.

TABLE III: The structure of autoencoders with different neurons. Q represents quadratic neurons and C represents conventional neurons.

Structures	Encoder	Decoder
HAE-X	Q+C	Q+C
QAE-X	Q+Q	Q+Q
AE-X	C+C	C+C
HAE-Y	Q+C	Q
QAE-Y	Q+Q	Q
AE-Y	C+C	C
HAE-Yc	Q+C	C
HAE-I	Q→C	C→Q
HAE-Ic	C→Q	Q→C

The form of quadratic neurons. Quadratic neurons have different variants, such as only keeping the power terms and replacing the interaction term with a linear term. What we use in a quadratic neuron is the standard form of the quadratic function. But is it suitable for anomaly detection? Here, we investigate if the standard form of quadratic neurons fits. We have chosen several datasets for comparison, with the results for AUC and precision presented in Tables VI and VII, respectively. We highlight that models using standard quadratic neurons display superior performance in the majority

TABLE IV: Comparison of AUCs for all structures. Bold-faced numbers are the best in an identical autoencoder structure with different neurons.

Datasets	HAE-X	QAE-X	AE-X	HAE-Y	QAE-Y	AE-Y	HAE-Yc	HAE-Iq	HAE-Ic
arrhythmia	**0.832**	0.819	0.815	0.816	**0.826**	0.813	0.816	**0.817**	0.814
glass	**0.605**	0.600	0.551	**0.608**	0.585	0.554	0.559	**0.591**	0.581
musk	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
optdigits	**0.772**	0.604	0.476	**0.787**	0.656	0.475	0.487	**0.668**	0.510
pendigits	**0.962**	0.945	0.931	0.943	**0.947**	0.934	0.935	**0.967**	0.937
pima	**0.739**	0.690	0.577	**0.695**	0.606	0.582	0.581	**0.701**	0.581
verterbral	**0.574**	0.481	0.571	**0.573**	0.561	0.570	0.571	**0.573**	0.572
wbc	**0.926**	0.909	0.857	**0.876**	0.868	0.857	0.856	**0.915**	0.855
ALOI	**0.556**	0.553	0.546	**0.547**	0.554	0.546	**0.547**	**0.553**	0.547
Ionosphere	**0.929**	0.920	0.906	0.910	**0.919**	0.905	0.907	**0.910**	0.908
KDDCUP99	0.989	**0.994**	0.989	0.989	0.989	0.989	0.989	0.989	0.989
Shuttle	0.473	0.385	**0.488**	**0.500**	0.453	0.491	0.491	0.495	0.495
Waveform	**0.692**	0.681	0.643	0.651	**0.688**	0.655	0.647	**0.680**	0.657
WDBC	**0.985**	0.981	0.978	0.978	**0.980**	0.979	**0.980**	0.980	**0.981**
WPBC	0.438	0.440	**0.445**	**0.443**	0.439	0.442	0.442	0.439	**0.442**
Avg.	**0.765**	0.733	0.718	**0.754**	0.738	0.719	0.721	**0.752**	0.725

TABLE V: Comparison of precision for all structures. Bold-faced numbers are the best in an identical autoencoder structure with different neurons.

Datasets	HAE-X	QAE-X	AE-X	HAE-Y	QAE-Y	AE-Y	HAE_Yc	HAE_Iq	HAE_Ic
arrhythmia	**0.758**	0.742	0.729	**0.740**	0.735	0.724	0.736	**0.745**	0.734
glass	**0.582**	0.576	0.556	0.589	**0.636**	0.574	0.574	**0.682**	0.572
musk	0.775	**0.776**	0.770	0.773	**0.917**	0.771	0.773	**0.773**	0.770
optdigits	**0.514**	0.443	0.431	**0.515**	0.433	0.432	0.431	**0.481**	0.431
pendigits	**0.757**	0.717	0.712	0.726	**0.818**	0.711	0.715	**0.753**	0.715
pima	**0.650**	0.626	0.557	**0.629**	0.573	0.555	0.555	**0.639**	0.553
verterbral	0.495	0.504	**0.582**	0.574	0.572	0.586	**0.589**	**0.593**	0.589
wbc	**0.829**	0.723	0.699	**0.704**	0.695	0.700	0.703	**0.712**	0.700
ALOI	**0.518**	**0.518**	0.515	**0.529**	0.517	0.516	0.516	**0.528**	0.516
Ionosphere	**0.832**	0.827	0.812	**0.816**	0.815	0.806	0.806	**0.816**	0.807
KDDCUP99	**0.870**	0.713	0.724	**0.872**	0.727	0.729	**0.872**	**0.871**	0.726
Shuttle	0.513	0.509	**0.518**	0.512	0.511	**0.519**	0.469	0.512	**0.522**
Waveform	**0.557**	0.546	0.537	0.548	**0.558**	0.541	0.544	**0.547**	0.537
WDBC	**0.859**	0.747	0.747	0.748	0.749	0.746	**0.855**	0.747	0.747
WPBC	0.493	0.487	**0.494**	**0.495**	0.490	0.484	0.487	0.486	**0.488**
Avg.	**0.667**	0.630	0.626	**0.651**	0.650	0.626	0.642	**0.659**	0.627

TABLE VI: Comparison of AUC for different quadratic functions on some datasets. **Bold-faced** numbers are the best compared in every structure.

Methods	Quadratic Function	Arrhythmia	Glass	Optdigits	ALOI	Shuffle	Waveform	XJTU	Avg.
HAE-X	$(\boldsymbol{x}^\top \mathbf{w}^r + b^r)(\boldsymbol{x}^\top \mathbf{w}^g + b^g)$	0.817	0.591	0.627	0.552	0.471	0.682	0.942	0.669
	$\boldsymbol{x}^\top \mathbf{w}^r + (\boldsymbol{x} \odot \boldsymbol{x})^\top \mathbf{w}^b + c$	0.830	0.580	0.676	0.554	0.471	0.687	0.945	0.678
	Standard	**0.832**	**0.605**	**0.772**	**0.556**	**0.473**	**0.692**	**0.957**	**0.698**
HAE-Y	$(\boldsymbol{x}^\top \mathbf{w}^r + b^r)(\boldsymbol{x}^\top \mathbf{w}^g + b^g)$	0.820	0.578	0.659	0.553	0.453	0.689	0.942	0.671
	$\boldsymbol{x}^\top \mathbf{w}^r + (\boldsymbol{x} \odot \boldsymbol{x})^\top \mathbf{w}^b + c$	**0.829**	0.571	0.661	**0.554**	0.452	**0.695**	0.943	0.672
	Standard	0.816	**0.608**	**0.787**	0.547	**0.500**	0.651	**0.958**	**0.695**
HAE-I	$(\boldsymbol{x}^\top \mathbf{w}^r + b^r)(\boldsymbol{x}^\top \mathbf{w}^g + b^g)$	**0.818**	0.567	0.627	0.551	0.450	0.671	0.944	0.661
	$\boldsymbol{x}^\top \mathbf{w}^r + (\boldsymbol{x} \odot \boldsymbol{x})^\top \mathbf{w}^b + c$	0.815	0.568	0.628	**0.553**	0.450	0.670	**0.966**	0.664
	Standard	0.817	**0.591**	**0.668**	**0.553**	**0.495**	**0.680**	0.961	**0.681**

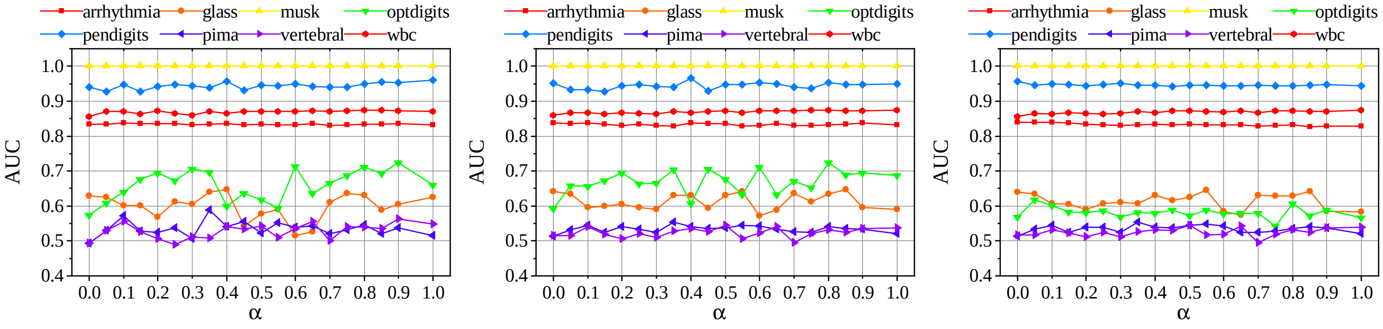
of cases. This outcome strongly suggests that the standard quadratic neuron, with its inclusion of both power and inner-product terms, possesses the best representation capabilities. Additionally, it is worth noting that models utilizing the power term generally perform better than those omitting it. This observation highlights the important role of the power term within the quadratic neuron.

V. Analysis Experiments

Hyperparameter sensitivity. As mentioned earlier, let γ_r and $\gamma_{g,b}$ be the learning rates of $(\boldsymbol{w}^r, b^r)$ and $(\boldsymbol{w}^g, \boldsymbol{w}^b, c, b^g)$, respectively. We assume that $\gamma_{g,b} = \alpha \cdot \gamma_r$, where α controls how slow $(\boldsymbol{w}^g, \boldsymbol{w}^b, c, b^g)$ are updated relative to $(\boldsymbol{w}^r, b^r)$. Since α is an important parameter, it is necessary to find how α affects the model. We train heterogeneous autoencoders by varying α from 0 to 1 with a step of 0.05. Figure 4 shows the results. On most datasets, three HAEs have a consistent performance, which indicates we don't need to fine-tune α to get a better result. By comparing three HAEs, we find out that HAE-X and HAE-Y show unstable in optidigits and glass, especially, when $\alpha \in (0.4, 0.8)$. Such that, we recommend setting α from 0.0 to 0.4 to have a consistent performance.

TABLE VII: Comparison of precision for different quadratic functions on some datasets. **Bold-faced** numbers are the best compared in every structure.

Methods	Quadratic Function	Arrhythmia	Glass	Optdigits	ALOI	Shuffle	Waveform	XJTU	Avg.
HAE-X	$(\boldsymbol{x}^\top \mathbf{w}^r + b^r)(\boldsymbol{x}^\top \mathbf{w}^g + b^g)$	0.716	**0.634**	0.457	0.516	0.517	0.548	0.747	0.591
	$\boldsymbol{x}^\top \mathbf{w}^r + (\boldsymbol{x} \odot \boldsymbol{x})^\top \mathbf{w}^b + c$	0.742	0.568	0.481	0.515	**0.523**	0.554	0.818	0.600
	Origin	**0.758**	0.582	**0.514**	**0.581**	0.513	**0.557**	**0.956**	**0.637**
HAE-Y	$(\boldsymbol{x}^\top \mathbf{w}^r + b^r)(\boldsymbol{x}^\top \mathbf{w}^g + b^g)$	**0.753**	0.572	0.460	0.516	**0.522**	0.550	0.746	0.588
	$\boldsymbol{x}^\top \mathbf{w}^r + (\boldsymbol{x} \odot \boldsymbol{x})^\top \mathbf{w}^b + c$	0.744	0.572	0.475	0.515	0.520	**0.557**	0.852	0.605
	Origin	0.740	**0.589**	**0.515**	**0.529**	0.512	0.548	**0.957**	**0.627**
HAE-I	$(\boldsymbol{x}^\top \mathbf{w}^r + b^r)(\boldsymbol{x}^\top \mathbf{w}^g + b^g)$	0.741	0.573	0.455	0.514	**0.551**	0.539	0.728	0.586
	$\boldsymbol{x}^\top \mathbf{w}^r + (\boldsymbol{x} \odot \boldsymbol{x})^\top \mathbf{w}^b + c$	0.742	0.573	0.453	0.514	0.511	0.541	0.923	0.608
	Origin	**0.745**	**0.682**	**0.481**	**0.528**	0.512	**0.547**	**0.959**	**0.636**



(a) HAE-X

(b) HAE-Y

(c) HAE-I

Fig. 4: Sensitivity of α. AUC denotes the area under the receiver operating characteristic curve.

Network depth. As shown in Table VIII, we investigate the performance of HAEs in response to different depths on the optidigits dataset. Note that increasing network layers will not tend to a better performance when the scale of the neural network is large enough. It may lead to overfitting. Therefore, all autoencoders are based on V1 ([*dim(x)*/2-*dim(x)*/4]) structure in our follows experiments.

TABLE VIII: Comparison of AUC for different network structure on optidigits. **Bold** numbers are the best result compared in every column. For hidden neurons, V1 represents [*dim(x)*/2-*dim(x)*/4], V2 is [*dim(x)*/2-*dim(x)*/3-*dim(x)*/4], V3 is [*dim(x)*/2-*dim(x)*/3-*dim(x)*/4-*dim(x)*/4] and V4 is [*dim(x)*/2-*dim(x)*/3-*dim(x)*/3-*dim(x)*/4-*dim(x)*/4].

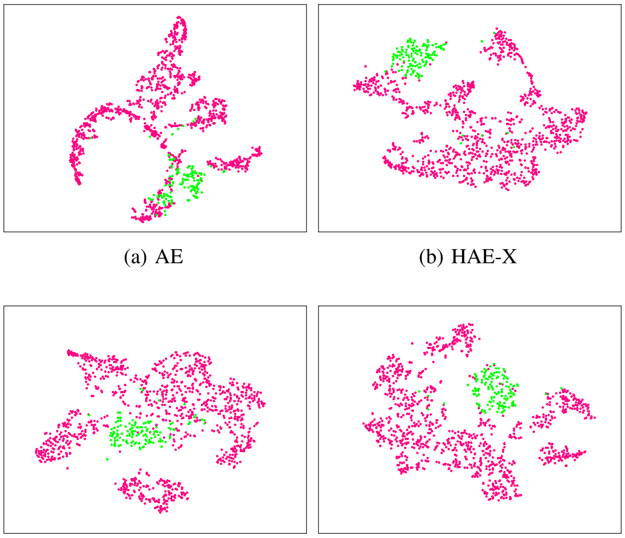
	AE	QAE	HAE-X	HAE-Y	HAE-I
V1	**0.652±0.017**	**0.599±0.023**	**0.680±0.054**	**0.681±0.056**	0.617±0.031
V2	0.613±0.010	0.528±0.045	0.669±0.043	0.653±0.027	**0.636±0.023**
V3	0.582±0.028	0.447±0.054	0.671±0.036	0.666±0.068	0.625±0.026
V4	0.572±0.008	0.316±0.052	0.647±0.029	0.653±0.040	0.613±0.041

Network width. How to find the best width arrangement for autoencoders is still an open question. If the width is large, the compressing effect is compromised, which causes the encoder to fail to learn the most representative features. If the width is small, the learned features are insufficient to completely represent the original data, which makes the detection based on these features inaccurate. We conduct an experiment to evaluate the validity of different width arrangements. The results are presented in Table IX. While our recommended structure exhibits superior performance in the majority of cases, we also note that the results are quite similar. We believe that, in the context of anomaly detection where large-scale datasets are often unavailable, the number of neurons employed does not significantly impact results.

TABLE IX: Comparison of HAEs using different width arrangements in several datasets. Bold-faced numbers are better.

Dataset		Arrhythmia		Shuffle		XJTU	
Method	Structure	AUC	PRE	AUC	PRE	AUC	PRE
HAE-X	d-d/2-d/4	**0.832**	**0.758**	**0.473**	0.513	**0.957**	**0.956**
	d-d/4-d/8	0.820	0.744	0.435	0.511	0.946	0.925
	d-d-d/2	0.828	0.741	0.452	**0.520**	0.945	0.913
HAE-Y	d-d/2-d/4	0.816	0.740	**0.500**	**0.512**	**0.958**	**0.957**
	d-d/4-d/8	0.816	0.732	0.449	0.511	0.948	0.935
	d-d-d/2	**0.827**	**0.741**	0.421	0.509	0.945	0.908
HAE-I	d-d/2-d/4	**0.817**	**0.745**	**0.495**	**0.512**	**0.961**	**0.959**
	d-d/4-d/8	**0.817**	0.738	0.454	0.511	0.962	0.937
	d-d-d/2	0.814	0.737	0.445	0.508	0.958	0.947

Visualization. Because HAE variants show a better performance than a conventional one, we perform a learned embedding visualization to understand what these methods have learned. Here, with the optidigits dataset, we conduct t-SNE [7] to visualize the latent space learned by AE and HAEs into 2D and 3D space. As shown in Figure 5 and Figure 6, compared to the AE result, abnormal data show distinguishable and concentration clusters by HAEs, suggesting that a heterogeneous autoencoder better HAE better expresses the divergences between the different categories of samples.



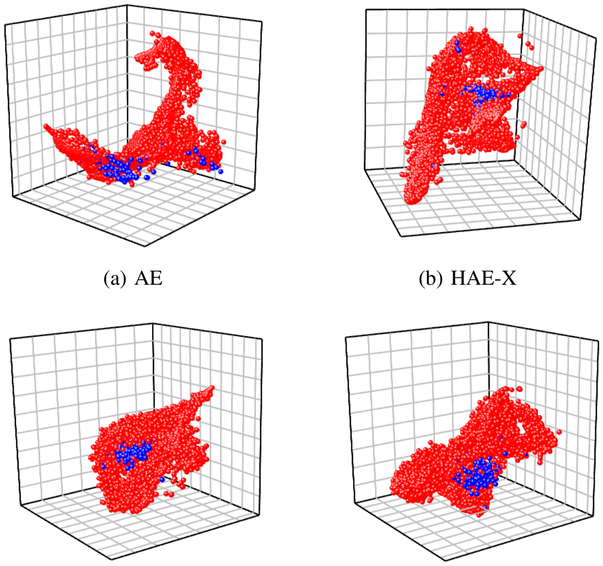
(a) AE

(b) HAE-X

(c) HAE-Y

(d) HAE-I

Fig. 5: The t-SNE results of representation learned by different methods on optidigits. The red points represent the learned embedding from abnormal data, and green points are learned latent variables from normal data.



(a) AE

(b) HAE-X

(c) HAE-Y

(d) HAE-I

Fig. 6: The t-SNE results of four autoencoders on optidigits. The red and blue points are learned latent variables from abnormal and normal data, respectively. It is observed that the outliers are more concentrated in HAEs than AE.

References

[1] F. Fan, J. Xiong, and G. Wang, “Universal approximation with quadratic deep networks,” *Neural Networks*, vol. 124, pp. 383–392, 2020.

[2] R. Eldan and O. Shamir, “The power of depth for feedforward neural networks,” in *Conference on learning theory*. PMLR, 2016, pp. 907–940.

[3] C. Debao, “Degree of approximation by superpositions of a sigmoidal function,” *Approximation Theory and its Applications*, vol. 9, no. 3, pp. 17–28, 1993.

[4] F.-L. Fan, M. Li, F. Wang, R. Lai, and G. Wang, “Expressivity and trainability of quadratic networks,” *arXiv preprint*, 2021.

[5] B. Wang, Y. Lei, N. Li, and N. Li, “A hybrid prognostics approach for estimating remaining useful life of rolling element bearings,” *IEEE Transactions on Reliability*, vol. 69, no. 1, pp. 401–412, 2018.

[6] X. Huang, G. Wen, S. Dong, H. Zhou, Z. Lei, Z. Zhang, and X. Chen, “Memory residual regression autoencoder for bearing fault detection,” *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2021.

[7] G. E. Hinton, “Visualizing high-dimensional data using t-sne,” *Vigiliae Christianae*, vol. 9, pp. 2579–2605, 2008.