Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Comparative investigation of seven supervised regression reference models for chiller fault detection: generalization to unseen days and detection at a matched false alarm rate

Dinh Anh Tuan Tran

trandinhanhtuan@iuh.edu.vn

Industrial University of Ho Chi Minh City

Huong Nguyen Thi Cam

FPT University

Van Tuan Nguyen

Industrial University of Ho Chi Minh City

Thi Khanh Phuong Ho

Industrial University of Ho Chi Minh City

Research Article

Keywords: Chiller fault detection, Supervised regression, Reference model, False alarm rate, Sensor selection

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11370528/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Comparative investigation of seven supervised regression reference models for chiller fault detection: generalization to unseen days and detection at a matched false alarm rate

Dinh Anh Tuan Tran[1*], Huong Nguyen Thi Cam[2], Van Tuan Nguyen[1], Thi Khanh Phuong Ho[1]

[1*]Faculty of Heat and Refrigeration Engineering, Industrial University of Ho Chi Minh City, Ho Chi Minh City, Viet Nam.
[2]FPT University, Ho Chi Minh City, Viet Nam.

*Corresponding author(s). E-mail(s): trandinhanhtuan@iuh.edu.vn;

Abstract

Residual-based chiller fault detection and diagnosis (FDD) relies on a fault-free reference model, but reference models are usually compared on data from the training day and without statistical tests. This study compares seven supervised regression methods implemented in MATLAB (fractional logistic regression, lasso, elastic net, decision tree, random forest, least squares support vector regression and a multilayer perceptron) as reference models for six feature parameters of an ASHRAE RP-1043 centrifugal chiller. Accuracy is evaluated on the training day, on two held-out normal days and by leave-one-day-out prediction with Diebold–Mariano and Friedman–Nemenyi tests; detection is evaluated on 26 fault cases at deployable and matched false alarm rates. Lasso and elastic net gave the lowest unseen-day errors for five of the six feature parameters. With thresholds set from training data, the linear models kept the unseen-day false alarm rate at 1.2–2.7%, whereas the tree-based models reached 6.7–9.2%. At a matched 1% false alarm rate, six models gave similar mean detection rates (33.7–35.7%), and least squares support vector regression was markedly lower. Most of the

residual variance on normal days consisted of offsets persisting over whole operating conditions, which bounds detectability. Condenser fouling became detectable only with 45% of the tubes plugged, because lower levels reduced the condenser UA by only 5–13%, within the fault-free scatter. A sensor ablation showed that the condensing pressure and power measurements carry the refrigerant-side and water-side faults, respectively. The reference model mainly determines how reliably a threshold transfers to new days.

Keywords: Chiller fault detection, Supervised regression, Reference model, False alarm rate, Sensor selection

Nomenclature and abbreviations

Symbol	Description
AUC	Area under the receiver operating characteristic curve
CD	Critical difference of the Nemenyi test
COP	Coefficient of performance (–)
CV(RMSE)	Coefficient of variation of the root mean square error (%)
d	Effect size of a fault on a feature parameter (–)
DM	Diebold–Mariano statistic
DR	Detection rate (%)
EWMA	Exponentially weighted moving average
FAR	False alarm rate (%)
FDD	Fault detection and diagnosis
FP	Feature parameter
kW	Compressor electrical power (kW)
LMTD	Logarithmic mean temperature difference (°C)
LODO	Leave-one-day-out
NMBE	Normalized mean bias error (%)
OOF	Out-of-fold
p	Fraction of plugged condenser tubes (–)
Q	Heat transfer rate (kW)
R^2	Coefficient of determination
RMSE	Root mean square error
T	Temperature (°C)
TCA	Condenser approach temperature (K)
TCI, TCO	Condenser inlet and outlet water temperature (°C)
TEI, TEO	Evaporator inlet and outlet water temperature (°C)

Symbol	Description
T_{sub}	Liquid subcooling (K)
TWE_{set}	Chilled water supply temperature set point (°C)
u_T	Temperature measurement uncertainty (K)
UA	Overall heat transfer conductance (kW/K)
v	Water velocity in the condenser tubes (m/s)
z	Standardized residual (–)
Greek symbols	
α	Mixing parameter of elastic net; significance level
γ	Regularization constant of LSSVR
ε_{sc}	Heat exchanger efficiency of the subcooling section (–)
η	Penalty weight of lasso and elastic net
λ	EWMA weighting factor
μ, σ	Mean and standard deviation of residuals
σ_{seg}	Standard deviation of the segment means of the normal residuals
σ^2	Kernel width of the RBF kernel
ϕ_w	Water-side share of the condenser heat transfer resistance (–)
Subscripts	
cd	Condenser
ev	Evaporator
f	Faulty operation

1 Introduction

Chillers are the largest single energy consumers in the air-conditioning systems of most commercial buildings. Faults in a chiller waste energy, shorten the life of equipment and raise maintenance costs. Automated fault detection and diagnosis (FDD) can find these faults early and has therefore been studied extensively over the last three decades [1, 2].

Chiller FDD methods are commonly divided into model-based and data-driven methods [1, 2]. In the residual-based form of both, a reference model predicts the fault-free value of each feature parameter (FP) from the driving conditions. The residual between the measured and predicted FP is then monitored by a statistical test or a control chart. The accuracy of the reference model sets the width of the control limits

through the residual standard deviation σ. A more accurate model gives narrower limits and a more sensitive FDD strategy [3–5].

For this reason, several reference models have been proposed and compared. Multiple linear regression (MLR) with three inputs was widely used with t-statistic thresholds [6, 7]. Zhao et al. [8] combined support vector regression (SVR) with exponentially weighted moving average (EWMA) control charts. Tran et al. replaced SVR by a radial basis function (RBF) model [3], compared MLR, Kriging and RBF [4], and optimized a least squares SVR (LSSVR) by differential evolution [5]. Yao et al. [9] compared random forest, XGBoost and LightGBM with SVM as reference models on the RP-1043 data and combined LightGBM with a multivariate EWMA chart. In all of these studies, the more flexible model gave the lower error and the better diagnosis rate.

More recent studies have turned to transfer learning and domain adaptation, motivated by the loss of accuracy of data-driven chiller models when they are applied to operating conditions or chillers different from those of the training data [10]. Such methods migrate FDD models between chillers [11] or diagnose faults across working conditions with partial domain adaptation [12] and with ensembles of transferred models with adaptive weights [13]. A recent review identified performance evaluation, benchmarking and transferability as major barriers to the real-building adoption of data-driven FDD [2]. Most of these methods are classifiers that rely on labelled fault data, which are scarce in practice [14], and incipient faults of low severity remain the most difficult to identify [15]. Even a recent physically structured hybrid model of chiller performance lost accuracy outside its validated scope and on a complete unseen day [16]. The reference models of residual-based FDD, which require only fault-free data, have not been re-examined from this perspective.

Four limitations remain in this line of work. First, each study compared only two or three model families, and the regularized linear models were not included. Second,

the reference models were usually trained and tested on data from the same normal test, for example the 'Normal2' data set of ASHRAE RP-1043 [3–5, 17]. Such a test measures interpolation within one day of operation, not the behaviour of the model on a new day, which is the situation of an online FDD system. Third, the differences between models were reported without statistical tests, so it is unclear whether a small gain in R^2 or RMSE is real. Fourth, the detection rates of different reference models were compared at different false alarm rates, because each model sets its own control limits. A model with a wider residual distribution on new days raises both its detection rate and its false alarm rate, and the two effects cannot be separated.

The present study addresses these limitations by a comparative investigation of seven supervised regression methods as chiller reference models. The scope is deliberately restricted to static regression models of steady-state operation, which is the setting of Refs. [3–5, 8]. Sequence models such as recurrent or convolutional networks address transient data and are outside this scope. The contributions are as follows:

1. Seven methods from four families are compared under one protocol: fractional logistic regression, lasso and elastic net (linear); decision tree and random forest (tree-based); LSSVR (kernel); and a multilayer perceptron (neural network).
2. Six feature parameters are modelled, including the subcooling efficiency $\varepsilon_{\mathrm{sc}}$ and the condenser approach temperature TCA used in earlier chiller FDD studies [3, 4, 8].
3. Generalization is evaluated at three levels: a same-day hold-out test, two fully held-out normal days, and leave-one-day-out (LODO) out-of-fold prediction, with Diebold–Mariano and Friedman–Nemenyi tests.
4. Fault detection on 26 fault cases is evaluated in two ways: with thresholds set from training data only, which shows how reliably each model's false alarm rate transfers to new days, and at a matched false alarm rate with ROC analysis, which compares detection power fairly. Both Shewhart and EWMA statistics are used, and cluster bootstrap confidence intervals are reported.

5. The statistical results are interpreted physically: the structure of the fault-free residuals, the measurement resolution of the subcooling, the heat transfer effect of the condenser fouling tests, and a sensor ablation that relates each measurement to the faults it reveals.

The remainder of this paper is organized as follows. Section 2 presents the seven regression reference models. Section 3 describes the methods: the FDD framework, data, feature parameters and evaluation protocol. Section 4 presents and discusses the results for the RP-1043 chiller. Section 5 gives the conclusions, the novelty of the work and a comparison with existing studies.

2 Regression reference models

All reference models are multi-input single-output black-box regressions of a feature parameter y on the driving conditions x. Physical chiller models were not considered, because their accuracy was found similar to empirical models at a higher calibration effort [18]. The input vector is

$$\mathbf{x} = [T_{WE,set},\ T_{EI},\ T_{CI},\ Q_{ev}]^T \tag{1}$$

Each input is standardized by the mean and standard deviation of the training set. The linear models use the quadratic expansion $\phi(\mathrm{x})$, which contains the four inputs, their squares and the six pairwise products (14 terms). This form generalizes Eq. (1) of Ref. [4] and is used so that the linear models are not penalized by the known curvature of chiller behaviour.

The seven methods span four model families of increasing flexibility, from penalized polynomial regression to tree ensembles, kernel and neural-network models. All are available in standard MATLAB toolboxes, the environment of the earlier chiller FDD strategies [3–5], so each can replace the reference model of those strategies directly.

2.1 Fractional logistic regression

Logistic regression is a classification method. It is used here in the fractional response form of Papke and Wooldridge [19]. The target is first scaled to the unit interval with a 10% margin on each side, and its conditional mean is modelled by the logistic function:

$$\widetilde{y} = \frac{y - y_{lo}}{y_{hi} - y_{lo}}, \qquad E\left[\widetilde{y} \,|\, \mathbf{x}\right] = \frac{1}{1 + \exp\left(-\beta^T \varphi\left(\mathbf{x}\right)\right)} \tag{2}$$

The coefficients β are estimated by the Bernoulli quasi-likelihood. The prediction is mapped back to the original scale. The model is bounded, so it cannot extrapolate beyond the scaled range.

2.2 Lasso and elastic net

The lasso [20] and elastic net [21] add a penalty to the least squares fit of the expanded model:

$$\min_{\beta_0, \beta} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta^T \varphi\left(\mathbf{x}_i\right)\right)^2 + \eta \left[\frac{1 - \alpha}{2} \parallel \beta \parallel_2^2 + \alpha \parallel \beta \parallel_1\right] \tag{3}$$

The lasso corresponds to $\alpha = 1$. The penalty η (60 values on a logarithmic path) and, for elastic net, α are selected on the validation set.

2.3 Decision tree and random forest

The regression tree [22] partitions the input space recursively so that the sum of squared errors within the leaves is minimized. Its prediction is the leaf mean. The random forest [23] averages B trees grown on bootstrap samples, with m_{try} inputs drawn at random at each split:

$$\widehat{y}\left(\mathbf{x}\right) = \frac{1}{B} \sum_{b=1}^{B} T_b\left(\mathbf{x}\right) \tag{4}$$

Both models are piecewise constant. They cannot predict values outside the range of the training targets.

2.4 Least squares support vector regression

LSSVR [24] replaces the ε-insensitive loss of SVR by a squared loss. The model parameters follow from the linear Karush–Kuhn–Tucker system

$$\begin{bmatrix} 0 & \mathbf{1}^T \\ \mathbf{1} & \mathbf{K} + \gamma^{-1}\mathbf{I} \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{y} \end{bmatrix}, \qquad K_{ij} = \exp\left(-\frac{\| \mathbf{x}_i - \mathbf{x}_j \|^2}{\sigma^2}\right) \tag{5}$$

The prediction is f(x) = Σ α_i K(x, x_i) + b. Because the cost of solving the system grows with n^3, at most 2000 training samples, taken at uniform intervals, are used. The constants γ and σ^2 are selected by a grid search on the validation set, in the same way as the DE search of Ref. [5]. The grid covers σ^2 from 0.01 to 10 and γ from 1 to 10^5, and the selected values lie inside the grid for five of the six feature parameters ($\sigma^2 = 0.05$–0.2); only $\varepsilon_{\mathrm{sc}}$ selected the upper bound $\sigma^2 = 10$.

2.5 Neural network

The neural network is a multilayer perceptron with two hidden layers of 64 and 32 ReLU units. The target is standardized, an L2 penalty of 10^{-4} is applied, and training stops when the validation loss has not improved for 20 iterations.

2.6 Hyperparameter selection

The hyperparameters of all models are selected on the same validation set (Section 3.3). The search spaces are listed in Table 1.

Table 1: Hyperparameter search spaces of the seven reference models

Model	Hyperparameters and search space
Logistic regression	Quadratic expansion; target margin 10%; no tuning
Lasso	η: 60-point path; $\alpha = 1$
Elastic net	η: 60-point path; $\alpha \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$
Decision tree	Minimum leaf size $\in \{1, 2, 5, 10, 20, 50\}$
Random forest	200 trees; $m_{\text{try}} \in \{2, 4\}$; minimum leaf size $\in \{1, 5\}$
LSSVR	$\gamma \in \{1, 10, 10^2, 10^3, 10^4, 10^5\}$; $\sigma^2 \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10\}$; n ≤ 2000
Neural network	Layers [64, 32]; ReLU; L2 $= 10^{-4}$; early stopping (patience 20)

3 Methods

The comparative framework follows the structure of the residual-based FDD strategies in Refs. [3–5, 8] (Fig. 1). Fault-free data are pre-processed, the FPs are computed, and each reference model is trained to predict the FPs. The residuals of the FPs under normal and faulty operation are then used to assess both accuracy and fault sensitivity.

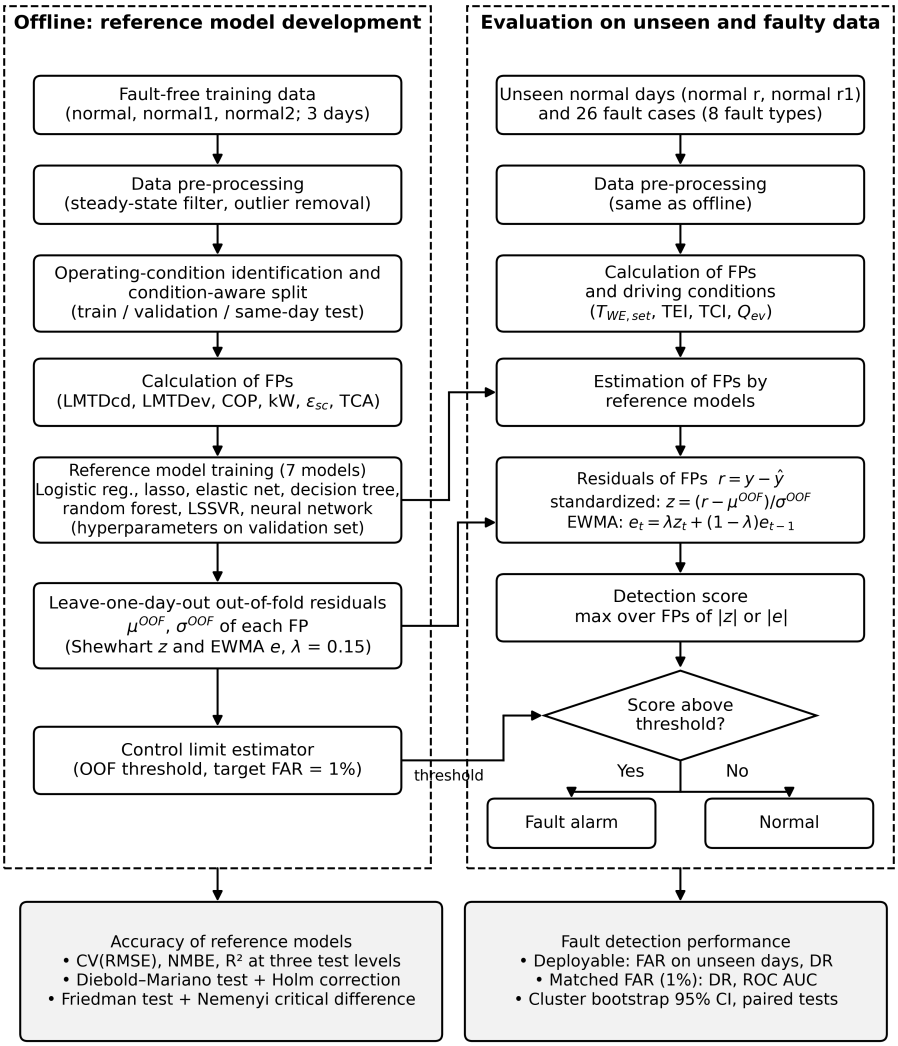


Fig. 1 Structure of the comparative framework. Offline development of the seven reference models on fault-free data, and evaluation of accuracy and fault detection on unseen normal days and fault data.

3.1 Experimental data

The data come from ASHRAE Research Project RP-1043 [17]. The test chiller is a 90-ton (316 kW) centrifugal water-cooled chiller with shell-and-tube evaporator and condenser. Each test covers 27 operating conditions. In contrast to Refs. [3–5], which used the 'Reduced data set' sampled every 2 min, the 'Complete data set' sampled every 10 s is used here.

Three normal tests ('normal', 'normal1' and 'normal2') are used to fit the models. Each test was recorded on a different day. Two further normal tests ('normal r' and 'normal r1') are fully held out and never used for fitting, tuning or threshold estimation. Eight fault types are evaluated at their available severity levels (Table 2). The fault label of each sample is taken from the test file, not from the controller fault code. Condenser fouling levels 1–4 are the cf12, cf20, cf30 and cf45 tests, in which 20, 33, 49 and 74 of the 164 condenser tubes were plugged [17]. The three near-normal tests are not used.

Table 2: Fault cases and number of samples after pre-processing

Fault type	Abbreviation	Severity levels (test)	Samples per level
Condenser fouling	CF	L1 (12%), L2 (20%), L3 (30%), L4 (45%)	2878, 3013, 2889, 2907
Refrigerant leakage	RL	L1–L4 (−10% to −40% charge)	2938, 2915, 2904, 3089
Refrigerant overcharge	RO	L1–L4 (+10% to +40% charge)	2889, 2995, 3025, 2981
Reduced condenser water flow	FWC	L1–L4 (−10% to −40%)	3074, 3138, 2995, 2854
Reduced evaporator water flow	FWE	L1–L4 (−10% to −40%)	3269, 3231, 3169, 3140
Non-condensable gas	NC	L1–L4 (1%, 2%, 3%, 5% nitrogen)	2825, 2689, 2840, 3366
Simultaneous FWE and FWC	FWE+FWC	L2 (−20% each)	3197
Defective pilot valve	DPV	L1	2710
Normal, held-out days	NORMAL_HO	normal r, normal r1	2905, 2783

3.2 Data pre-processing and feature parameters

All temperatures are converted to °C, flow rates to L/s and refrigeration tons to kW. A sample is kept only if the chiller is running (unit status 20), the power and the cooling load exceed 1 kW and 1 ton, the water temperature differences are positive, and both approach temperatures exceed 0.05 K. A steady-state filter then keeps a sample only if the moving standard deviations over 5 min are below 0.15 K for TEO and TCI and below 1.5 kW for the compressor power. For the normal fitting data only, samples with a controller fault code are removed, and outliers are removed by a 10-min moving median with a threshold of three scaled median absolute deviations.

Six FPs are modelled. The condenser and evaporator LMTDs are computed from the condensing temperature T_{cd} = TCO + TCA and the evaporating temperature T_{ev} = TEO − TEA:

$$LMTD_{cd} = \frac{T_{CO} - T_{CI}}{\ln\left[\left(T_{cd} - T_{CI}\right) / \left(T_{cd} - T_{CO}\right)\right]}, \qquad LMTD_{ev} = \frac{T_{EI} - T_{EO}}{\ln\left[\left(T_{EI} - T_{ev}\right) / \left(T_{EO} - T_{ev}\right)\right]} \tag{6}$$

The heat exchanger efficiency of the subcooling section is defined as in Refs. [3, 4]:

$$\varepsilon_{sc} = \frac{T_{sub}}{T_{cd} - T_{CI}} \tag{7}$$

The condenser approach temperature TCA is taken directly from the measurements. LMTDcd, $\varepsilon_{\mathrm{sc}}$ and TCA are the condenser-side FPs used in Refs. [3–5, 8]. LMTDev reflects the evaporator side, and COP and the compressor power (kW) are the measured performance indices.

3.3 Operating conditions and data split

The sample stream of each test is cut into segments at time gaps longer than 2 min. Adjacent segments are merged into one operating condition when their mean TWE_{set},

TCI and TEI differ by less than 0.3, 1.0 and 1.0 K. Within each condition, the last 20% of the samples form the same-day test set. The last 15% of the remaining samples form the validation set, which is used for hyperparameter selection and early stopping. The rest form the training set. This condition-aware split keeps every operating condition in all three sets and avoids the leakage that a random split of autocorrelated samples would cause.

Three test levels are then defined:

1. **Same-day test**: the last 20% of each condition of the three fitting days.
2. **Unseen-day test**: the two held-out normal days, 5688 samples in total.
3. **LODO out-of-fold**: each fitting day is held out in turn, the models are trained and tuned on the other two days, and the held-out day is predicted. The out-of-fold (OOF) residuals are also used to estimate μ and σ for the fault sensitivity analysis.

3.4 Accuracy indices

The accuracy of each model is evaluated by RMSE, MAE, R^2, CV(RMSE) and NMBE. CV(RMSE) and NMBE are recommended by ASHRAE Guideline 14 [25] and are dimensionless, so they can be compared across FPs:

$$CV(RMSE) = \frac{100}{\bar{y}} \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}, \qquad NMBE = \frac{100}{n\,\bar{y}} \sum_{i=1}^{n} (y_i - \widehat{y}_i) \tag{8}$$

where y_i and $\hat{y}_i$ are the measured and predicted values and $\bar{y}$ is the mean of the measured values.

3.5 Statistical tests of accuracy

On each hold-out test set, the Diebold–Mariano (DM) test [26] compares the squared errors of each model with those of the best model:

$$d_t = e_{1,t}^2 - e_{2,t}^2, \qquad DM = \frac{\bar{d}}{\sqrt{\widehat{S}/T}} \tag{9}$$

where $\hat{S}$ is the Newey–West estimate of the long-run variance of d_t [27] with a lag of $\lfloor T^{1/3} \rfloor$, which accounts for the autocorrelation of the residuals. A positive DM means that model 1 has the larger error. The p-values of the six comparisons are adjusted by the Holm procedure [28].

On the LODO OOF predictions, the RMSE of each model is computed for each operating condition with at least 10 samples, which gives N = 80 blocks per FP and 480 blocks in total. The models are ranked within each block, and the Friedman test [29] checks whether the mean ranks differ. Two models differ significantly when their mean ranks differ by more than the Nemenyi critical difference [30]:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \tag{10}$$

with k = 7 models and $q_{0.05} = 2.949$, which gives CD = 1.01 per FP and CD = 0.41 for the pooled blocks.

3.6 Fault detection statistics and their evaluation

For each model and FP k, the residual $r_k = y_k - \hat{y}_k$ is standardized by the mean and standard deviation of the LODO out-of-fold residuals on the fitting days:

$$z_{k,t} = \frac{r_{k,t} - \mu_k^{OOF}}{\sigma_k^{OOF}}, \qquad e_{k,t} = \lambda\, z_{k,t} + (1-\lambda) e_{k,t-1} \tag{11}$$

The second expression is the EWMA statistic of Refs. [3–5, 8, 31] with $\lambda = 0.15$. It is restarted ($e_{k,0} = 0$) at the start of every operating condition and standardized by the standard deviation of its own out-of-fold series. A sample is scored by the largest standardized statistic over the six FPs, |z| for the Shewhart chart and |e| for the EWMA chart, so that a single threshold covers the whole FP set.

Two threshold settings are evaluated. In the deployable setting, the threshold is the 99th percentile of the out-of-fold score on the fitting days, which targets a false alarm rate (FAR) of 1% using training data only. Its realized FAR on the two unseen normal days measures how well the threshold transfers to a new day, and its detection rate (DR) is reported for each fault case. In the matched setting, the threshold is set so that the FAR on the unseen normal days is exactly 1%. This setting is not available in practice, but it removes the effect of FAR on DR and therefore compares the detection power of the models fairly. The receiver operating characteristic (ROC) of the score, with the unseen normal days as negatives and each fault case as positives, is summarized by the area under the curve (AUC).

Uncertainty is quantified by a cluster bootstrap [32] with 1000 replicates, in which whole operating conditions are resampled with replacement within each test. The same resampling indices are used for all models, so that differences in mean DR between models are paired. The p-value of each model against the model with the highest mean DR is adjusted by the Holm procedure [28]. In addition, the fraction of samples with $|z| > 3$ is reported per FP and fault case to identify the most sensitive FP of each fault, as in the earlier rule-based strategies [3, 4].

3.7 Physical interpretation and sensor analysis

Three further analyses relate the statistical results to the physics of the chiller and to its instrumentation. They use the lasso, random forest and neural network models as representatives of the linear, tree-based and neural families.

First, the residuals r on the five normal days (out-of-fold on the three fitting days, final model on the two unseen days) are decomposed into a between-day part, a between-condition part within each day, and a within-condition part:

$$\sum_i (r_i - \bar{r})^2 = \sum_d n_d \left(\bar{r}_d - \bar{r}\right)^2 + \sum_s n_s \left(\bar{r}_s - \bar{r}_{d(s)}\right)^2 + \sum_i \left(r_i - \bar{r}_{s(i)}\right)^2 \tag{12}$$

where d indexes days and s the operating-condition segments of each day. The within-condition part contains sensor noise and short-term fluctuations; the two other parts are systematic offsets that persist over a whole condition or day. The measurement uncertainties are taken from the uncertainty analysis of the test facility [17]: ±0.05°F (0.03 K) for the water-side RTDs (TEI, TEO, TCI, TCO), and ±0.5°F (0.28 K) for TCA and ±0.54°F (0.30 K) for T_{sub}, both of which are computed by the chiller controller from its onboard thermistors and condenser pressure transducer. The uncertainty of $\varepsilon_{\mathrm{sc}}$ follows by first-order propagation.

Second, the shift of each FP under each fault case is expressed as an effect size d $= (\bar{r}_f - \bar{r}_N)/\sigma_{\mathrm{seg}}$, where $\bar{r}_f$ and $\bar{r}_N$ are the mean residuals of the fault case and of the normal days and σ_{seg} is the standard deviation of the segment means of the normal residuals. For condenser fouling, which was produced by plugging 20, 33, 49 and 74 of the 164 condenser tubes (p = 12.2–45.1%) [17], the observed change of UA_{cd} relative to a lasso model of the fault-free UA_{cd} is compared with a simple resistance model. Plugging reduced the condenser water flow only slightly (by about 5% at 45% plugging [17]), so the water velocity in the remaining tubes rises approximately by $1/(1 - \mathrm{p})$. With a water-side coefficient proportional to $v^{0.8}$ [33], the water-side resistance then scales with $(1-p)^{-0.2}$, whereas the refrigerant-side resistance scales with the lost area, $(1-p)^{-1}$:

$$\frac{UA_{cd,f}}{UA_{cd,0}} = \left[\phi_w (1-p)^{-0.2} + (1-\phi_w)\,(1-p)^{-1}\right]^{-1} \tag{13}$$

where ϕ_w is the water-side share of the total resistance, varied from 0.3 to 0.7.

Third, a sensor ablation repeats the matched-FAR evaluation of Section 3.6 with subsets of the FPs: each FP removed in turn, the FPs of one sensor removed together (condensing pressure: LMTDcd, TCA and ε_{sc}; power meter: kW and COP), and each FP alone. The mean DR of each subset is compared with that of the full set by the paired cluster bootstrap with Holm adjustment.

4 Results and discussion

4.1 Accuracy on the same day and on unseen days

The model ranking on the same day did not hold on unseen days (Table 3, Fig. 2). On the same-day test, the flexible models gave the lowest errors: the neural network for LMTDcd (6.37%), COP (0.61%), power (0.70%) and TCA (9.85%), and LSSVR for LMTDev (3.20%) and ε_{sc} (6.17%). On the two held-out days, lasso and elastic net gave the lowest errors for five of the six FPs: 11.19% for LMTDcd, 5.34% for LMTDev, 1.24% for COP, 1.39% for power and 17.41% for TCA.

Table 3: CV(RMSE) (%) of the seven reference models on the same-day and unseen-day tests

Model	LMTDcd	LMTDev	COP	kW	ε_{sc}	TCA
Logistic regression	8.32 / 12.06	4.56 / 5.61	1.10 / 1.29	2.03 / 2.16	6.52 / 7.85	12.40 / 18.51
Lasso	7.92 / **11.19**	4.45 / 5.35	0.99 / **1.24**	1.09 / **1.39**	6.51 / 7.84	12.06 / **17.41**
Elastic net	7.97 / 11.25	4.45 / **5.34**	0.99 / **1.24**	1.09 / 1.40	6.51 / 7.84	12.07 / 17.42
Decision tree	7.43 / 14.26	4.47 / 7.19	0.77 / 3.66	1.02 / 4.25	6.40 / 8.38	11.30 / 21.74
Random forest	6.91 / 11.34	3.41 / 6.60	0.74 / 2.41	0.96 / 3.44	6.29 / 8.23	10.35 / 18.78
LSSVR	6.84 / 15.93	**3.20** / 12.59	1.30 / 6.12	1.38 / 5.27	**6.17** / **7.63**	10.17 / 23.66
Neural network	**6.37** / 12.17	3.48 / 6.23	**0.61** / 1.47	**0.70** / 1.40	6.22 / 7.67	**9.85** / 17.88

Values are given as same-day test / unseen-day test. Bold: lowest value in each column.

The loss of accuracy from the same day to unseen days differed strongly between model families. For COP, the CV(RMSE) of lasso rose by a factor of 1.25 (0.99% to 1.24%), against 3.3 for random forest, 4.7 for LSSVR and 4.8 for the decision tree. The predicted-versus-measured plots show the cause (Figs. 3 and 4; Supplementary Figs. S1–S4 for the other FPs). The tree-based models predict in horizontal bands, because their output is constant within each leaf and limited to the range of the training targets. On a new day, small shifts in the driving conditions move the operating point between bands, whereas the quadratic linear models vary smoothly with the inputs.

LSSVR gave the largest unseen-day errors for five of the six FPs. Its kernel width was selected inside a wide grid (σ^2 = 0.05–0.2 out of 0.01–10), so this result is not an artefact of the search range. Such a narrow kernel fits the training day closely but extrapolates poorly when the operating point of a new day falls between training samples.

ε_{sc} and TCA were the least predictable FPs. Their unseen-day R^2 was only 0.51–0.60 and 0.67–0.82, against 0.84–0.997 for the other FPs. Section 4.7.1 shows that the two have different causes: the ε_{sc} residual is dominated by measurement noise, whereas the TCA residual is dominated by offsets that persist over each operating condition. For ε_{sc}, the seven models differed by less than 0.8 percentage points of CV(RMSE).

All models met the ASHRAE Guideline 14 criteria [25] on unseen days, with CV(RMSE) below 24% and |NMBE| below 5%. These criteria were designed for energy model calibration. They do not separate the models in an FDD context, where the residual standard deviation sets the detection threshold.

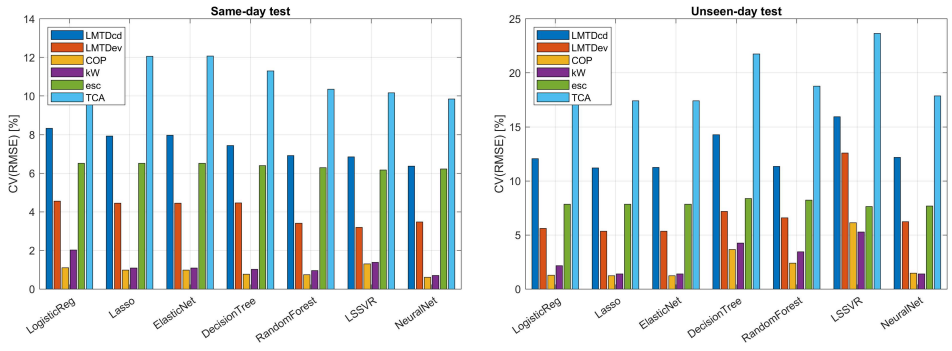


Fig. 2 CV(RMSE) of the seven reference models. (a) Same-day test; (b) unseen-day test.

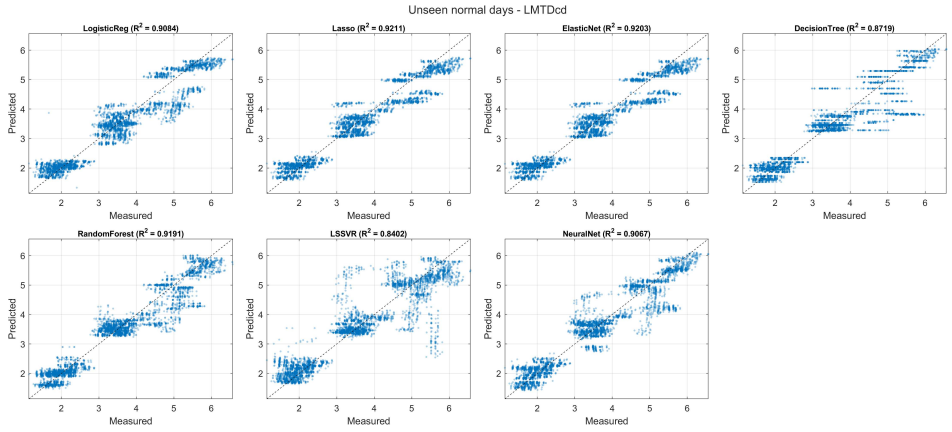


Fig. 3 Predicted versus measured LMTDcd on the unseen normal days.

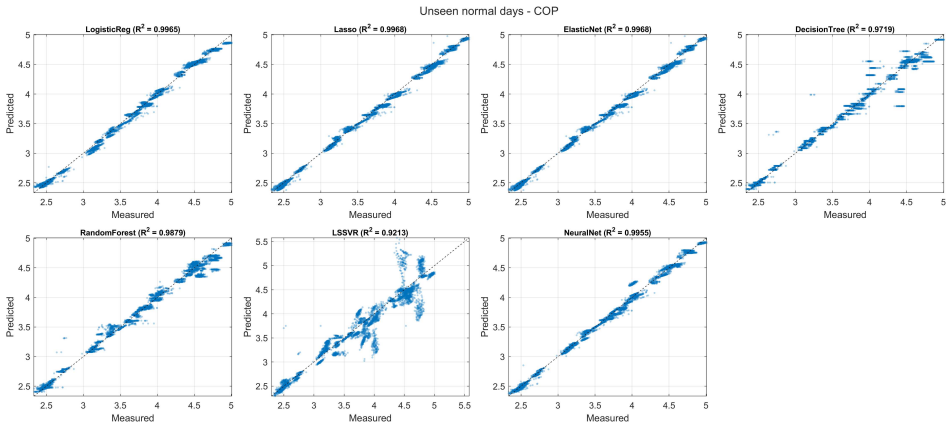


Fig. 4 Predicted versus measured COP on the unseen normal days.

4.2 Significance of the differences on unseen days

The DM test confirmed most of the advantage of the regularized linear models on unseen days (Table 4). For LMTDev, all six models had significantly larger errors than elastic net ($p_{\text{Holm}} < 0.05$). For COP, only logistic regression, lasso and the neural network were not significantly worse. For power, only the neural network was not significantly worse than lasso ($p_{\text{Holm}} = 0.88$). For LMTDcd and TCA, random forest and the neural network were not significantly worse than lasso.

Table 4: RMSE on the unseen-day test and Holm-adjusted p-values of the DM test

Model	LMTDcd (K)	LMTDev (K)	COP (–)	kW (kW)	ε_{sc} (–)	TCA (K)
Logistic regression	0.416 (0.011*)	0.263 (0.016*)	0.0470 (0.316)	1.152 (<0.001*)	0.0528 (0.001*)	0.376 (0.020*)
Lasso	0.386 (—)	0.251 (0.029*)	0.0453 (0.068)	0.742 (—)	0.0528 (0.004*)	0.354 (—)
Elastic net	0.388 (0.012*)	0.251 (—)	0.0453 (—)	0.747 (<0.001*)	0.0528 (0.004*)	0.354 (0.013*)
Decision tree	0.492 (<0.001*)	0.338 (<0.001*)	0.1336 (<0.001*)	2.262 (<0.001*)	0.0564 (<0.001*)	0.442 (0.003*)
Random forest	0.391 (0.724)	0.310 (<0.001*)	0.0877 (<0.001*)	1.831 (<0.001*)	0.0554 (<0.001*)	0.382 (0.088)
LSSVR	0.550 (<0.001*)	0.592 (<0.001*)	0.2233 (<0.001*)	2.806 (<0.001*)	0.0513 (—)	0.481 (0.003*)

Table 4: RMSE on the unseen-day test and Holm-adjusted p-values of the DM test

Model	LMTDcd (K)	LMTDev (K)	COP (–)	kW (kW)	ε_{sc} (–)	TCA (K)
Neural network	0.420 (0.065)	0.292 (<0.001*)	0.0535 (0.075)	0.747 (0.880)	0.0516 (0.309)	0.364 (0.421)

p-values (in parentheses) refer to the DM test against the best model of each FP. Best model: —; * $p_{\text{Holm}} < 0.05$.

Statistical and practical significance must be distinguished. Elastic net was significantly worse than lasso for LMTDcd, power and TCA, although its RMSE was less than 0.7% higher. For ε_{sc}, five models were significantly worse than LSSVR although their RMSE differed by at most 10%. With about 5700 test samples, the DM test detects such small but consistent differences, which should be reported together with their size.

4.3 Out-of-fold accuracy and ranking over operating conditions

The Friedman test rejected equal performance of the seven models for each FP ($p < 10^{-7}$, except ε_{sc} with $p = 0.044$) and for the pooled blocks ($p = 1.7 \times 10^{-40}$, $N = 480$). The neural network had the best pooled mean rank (2.98) and was significantly better than all other models (CD = 0.41), followed by elastic net (3.69), lasso (3.85) and random forest (3.99) (Table 5, Fig. 6).

Table 5: LODO out-of-fold CV(RMSE) (%) and mean rank over the operating conditions

Model	LMTDcd	LMTDev	COP	kW	ε_{sc}	TCA	Pooled rank
Logistic regression	9.65 (5.01)	5.62 (3.39)	1.25 (4.11)	2.37 (5.11)	6.68 (4.12)	14.12 (4.88)	4.44
Lasso	9.36 (4.59)	5.52 (3.54)	1.10 (3.48)	1.21 (3.02)	6.75 (4.08)	14.10 (4.38)	3.85
Elastic net	9.00 (3.84)	5.52 (3.66)	1.10 (3.58)	1.21 (2.94)	6.74 (4.09)	13.65 (4.04)	3.69
Decision tree	8.49 (3.88)	7.13 (5.08)	2.33 (4.85)	2.69 (5.30)	6.89 (3.88)	12.95 (3.59)	4.43

Table 5: LODO out-of-fold CV(RMSE) (%) and mean rank over the operating conditions

Model	LMTDcd	LMTDev	COP	kW	ε_{sc}	TCA	Pooled rank
Random forest	8.20 (3.41)	6.19 (3.89)	1.84 (4.65)	2.35 (4.64)	7.18 (4.30)	11.84 (3.02)	3.99
LSSVR	23.43 (4.40)	9.55 (4.81)	4.01 (4.92)	6.86 (4.85)	26.25 (4.28)	35.29 (4.56)	4.64
Neural network	7.66 (2.88)	5.82 (3.64)	0.99 (2.41)	1.08 (2.14)	6.60 (3.26)	12.16 (3.54)	2.98

Mean ranks in parentheses; lower is better. CD = 1.01 per FP (N = 80) and 0.41 pooled (N = 480).

The ranking depended on the FP. For COP, the neural network (2.41) was significantly better than all other models; for power, only elastic net and lasso lay within the critical difference of the neural network. For LMTDcd, random forest, elastic net and the decision tree lay within the critical difference of the neural network. For TCA, random forest ranked first (3.02). For ε_{sc}, no pair of models differed significantly.

The out-of-fold ranking favours the neural network more than the unseen-day test does, because the three fitting days resemble each other more than they resemble the two held-out days. The two tests therefore measure different degrees of extrapolation, and both are needed.

The mean rank and the pooled CV(RMSE) also measure different properties. A rank counts how often a model wins within a condition, whereas the CV(RMSE) is dominated by the conditions with the largest errors. LSSVR shows the extreme case: its per-condition ranks were moderate (4.3–4.9), but its error reached several times the typical value in a few conditions (Fig. 5), which raised its out-of-fold CV(RMSE) to 23.4% for LMTDcd, 26.3% for ε_{sc} and 35.3% for TCA. For FDD, these worst conditions matter, because they inflate the residual standard deviation that sets the threshold for all conditions (Section 4.4).

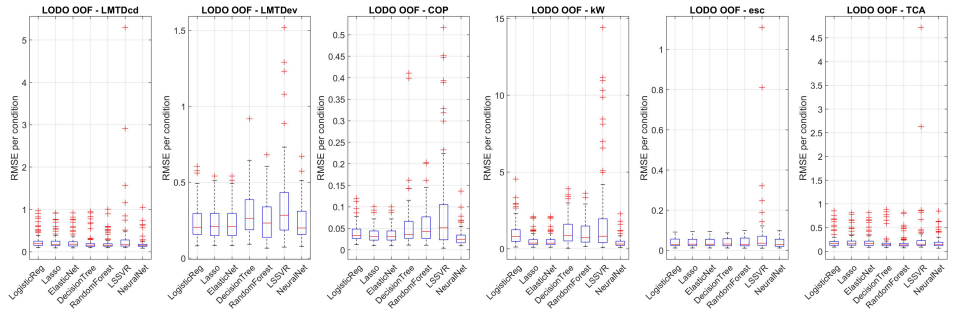


Fig. 5 Per-condition RMSE of the LODO out-of-fold predictions.

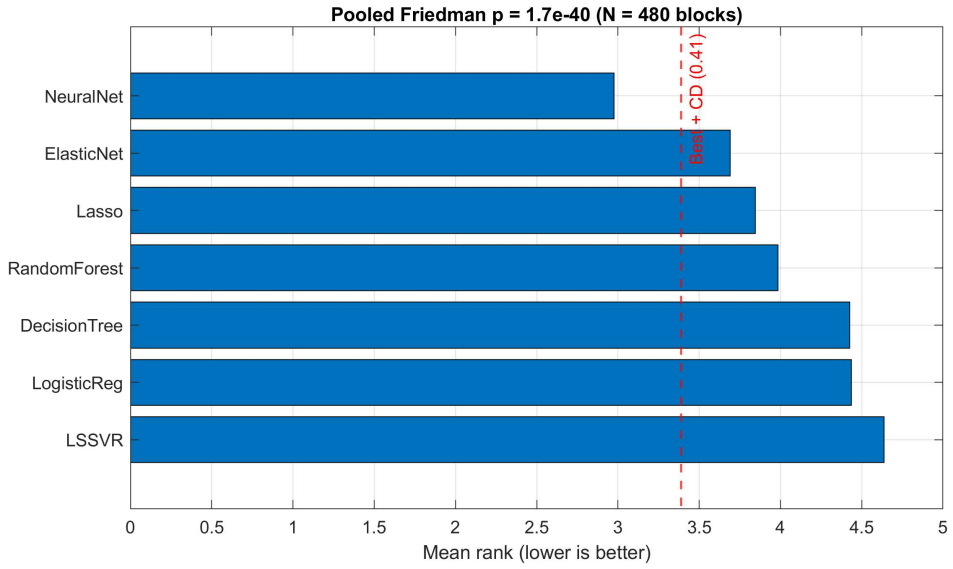


Fig. 6 Pooled mean ranks of the seven models. Dashed line: best mean rank plus the Nemenyi critical difference (CD).

4.4 Fault detection with thresholds set from training data

The main practical difference between the models appeared in the false alarm rate on unseen days (Table 6, Fig. 7). With thresholds set to a 1% FAR on the out-of-fold residuals of the fitting days, the linear models gave realized FARs of 1.2–1.9% (Shewhart) and 2.4–2.7% (EWMA) on the two held-out days. The neural network gave

3.4–3.6%. The decision tree and random forest gave 6.7–9.2%, seven to nine times the target. LSSVR gave almost no false alarms, because its few large out-of-fold errors inflated σ^{OOF} and therefore the threshold.

Table 6: Fault detection summary over 26 fault cases

Model	Deployable FAR, Shewhart (%)	Deployable FAR, EWMA (%)	Deployable DR, EWMA (%)	Matched DR, Shewhart (%)	Matched DR, EWMA (%)	AUC, EWMA	p_{Holm}
Logistic regression	1.9 [0.5, 3.6]	2.7 [0.2, 7.2]	47.4	41.0	35.7 [34.0, 37.6]	0.832	—
Lasso	1.3 [0.5, 2.4]	2.5 [0.1, 6.9]	43.9	36.9	34.3 [32.6, 36.1]	0.804	0.12
Elastic net	1.2 [0.5, 2.3]	2.4 [0.1, 6.8]	43.2	37.0	34.3 [32.7, 36.1]	0.802	0.12
Decision tree	7.6 [3.0, 13.2]	6.7 [2.2, 12.4]	40.3	27.3	33.7 [31.8, 35.9]	0.754	0.12
Random forest	7.8 [3.2, 13.7]	9.2 [3.6, 16.2]	46.4	32.6	34.3 [32.7, 36.1]	0.793	0.12
LSSVR	0.2 [0.0, 0.4]	0.0 [0.0, 0.0]	5.9	10.5	10.6 [9.1, 12.3]	0.688	<0.001
Neural network	3.4 [0.4, 8.5]	3.6 [0.0, 8.6]	39.4	32.7	33.9 [31.9, 35.8]	0.815	0.04

Deployable: threshold from out-of-fold residuals (target FAR 1%); matched: FAR on the unseen normal days fixed at 1%. Brackets: 95% cluster bootstrap CI. p_{Holm}: paired bootstrap test of the matched EWMA DR against logistic regression.

The realized FAR exceeded the 1% target for every model except LSSVR. For a stationary process, a threshold set on out-of-fold residuals should reproduce its target rate on new data. The excess therefore measures the day-to-day shift in the residual distribution that each model cannot absorb. The shift was smallest for the regularized linear models, consistent with their lowest unseen-day errors (Section 4.1), and largest

for the tree-based models. This also explains why studies that set and test thresholds on the same normal day report lower false alarm rates than an online system will experience.

The deployable detection rates must be read together with these FARs. Random forest reached a mean DR of 46.4% with EWMA, close to the 47.4% of logistic regression, but at 9.2% FAR against 2.7%. Part of its detection under faults comes from the same extrapolation error that raises its false alarms. Tree-based ensembles have been recommended as chiller reference models on the basis of their accuracy [9]; the present results show that their thresholds transfer poorly to new days.

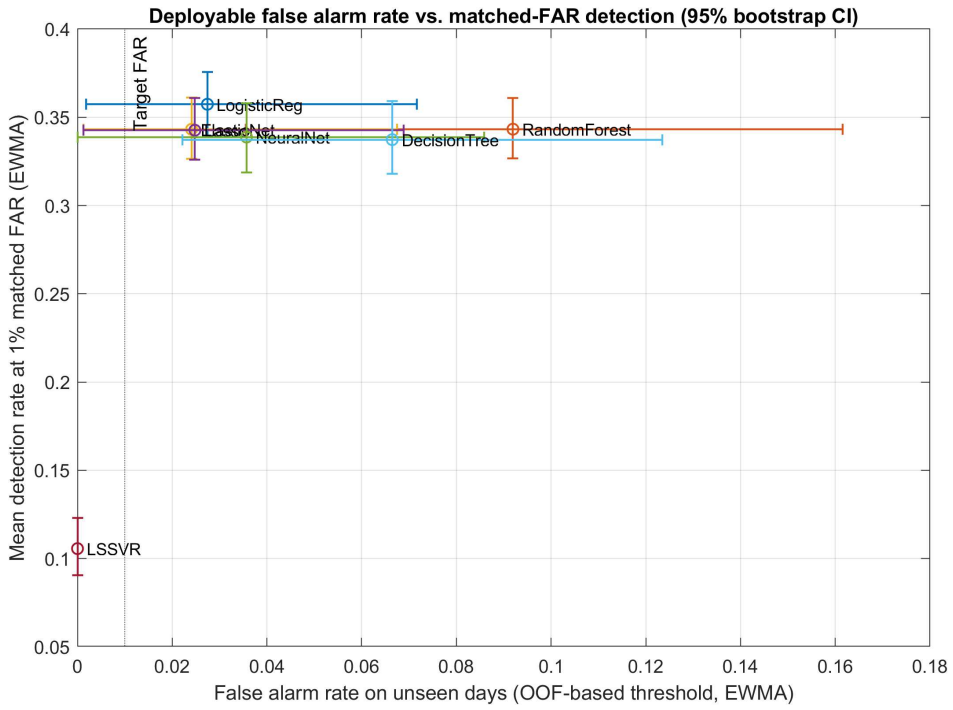


Fig. 7 Deployable false alarm rate versus matched-FAR detection rate. False alarm rate on the unseen days with the out-of-fold threshold (EWMA) versus mean detection rate at a matched 1% false alarm rate; error bars show 95% cluster bootstrap confidence intervals.

4.5 Fault detection at a matched false alarm rate

When the FAR was fixed at 1% on the unseen days, six of the seven models detected faults at similar rates (Table 6). Their mean EWMA detection rates ranged from 33.7% (decision tree) to 35.7% (logistic regression). Lasso, elastic net, the decision tree and random forest did not differ significantly from logistic regression ($p_{\text{Holm}} = 0.12$); the neural network was 1.9 percentage points lower ($p_{\text{Holm}} = 0.04$), a difference smaller than its own run-to-run variation (Section 4.7.3). Only LSSVR was markedly worse (10.6%, $p_{\text{Holm}} < 0.001$). The AUC ranged from 0.754 (decision tree) to 0.832 (logistic regression), with LSSVR at 0.688.

The effect of the EWMA statistic depended on the model. With the Shewhart statistic, the linear models detected more than the tree-based models (36.9–41.0% against 27.3–32.6%). EWMA raised the matched DR of the decision tree from 27.3% to 33.7% but lowered that of lasso from 36.9% to 34.3%. EWMA averages out random residual noise, which benefits the tree-based models with their stepwise predictions. It cannot remove offsets that persist over a whole operating condition, and it accumulates them, which raises the matched threshold. Section 4.7.1 shows that such offsets make up 64–72% of the normal residual variance of five of the six FPs, so for models whose residuals are already smooth, the second effect outweighs the first.

Table 7: Detection rate (%) at a matched 1% false alarm rate per fault type

Fault type	Logistic regression	Lasso	Elastic net	Decision tree	Random forest	LSSVR	Neural network
CF	10.4	14.1	14.3	14.5	10.7	1.3	13.2
RL	38.3	40.1	40.2	35.7	37.2	0.6	37.7
RO	42.4	41.6	41.6	40.3	39.2	6.2	37.9
FWC	28.3	22.7	22.5	18.5	27.5	8.5	21.3
FWE	9.1	0.5	0.5	5.2	7.1	8.8	4.5
NC	96.7	96.6	96.9	96.2	97.2	35.5	96.5
FWE+FWC	14.0	9.2	9.2	11.3	4.4	6.9	19.1
DPV	14.1	19.2	19.2	24.4	12.7	24.4	16.7

EWMA statistic with all six FPs; values are averaged over the severity levels of each fault type.

The detectability depended far more on the fault type and severity than on the model (Table 7, Figs. 8 and 9). Non-condensable gas was detected in 96–97% of the samples by six models. Refrigerant overcharge and leakage reached 36–42%, rising with severity: for leakage, from 8–10% at L1 to 92–97% at L4; for overcharge, from 3–10% at L1 to 82–89% at L4. Reduced condenser water flow reached 19–28%, and the defective pilot valve 13–24%. Reduced evaporator water flow was detected in at most 9% of the samples. Condenser fouling showed the sharpest dependence on severity: at L1–L3 (12–30% of the tubes plugged) it was detected in at most 13% of the samples, whereas at L4 (45% plugged) six models reached 28–43%, with AUC 0.88–0.95.

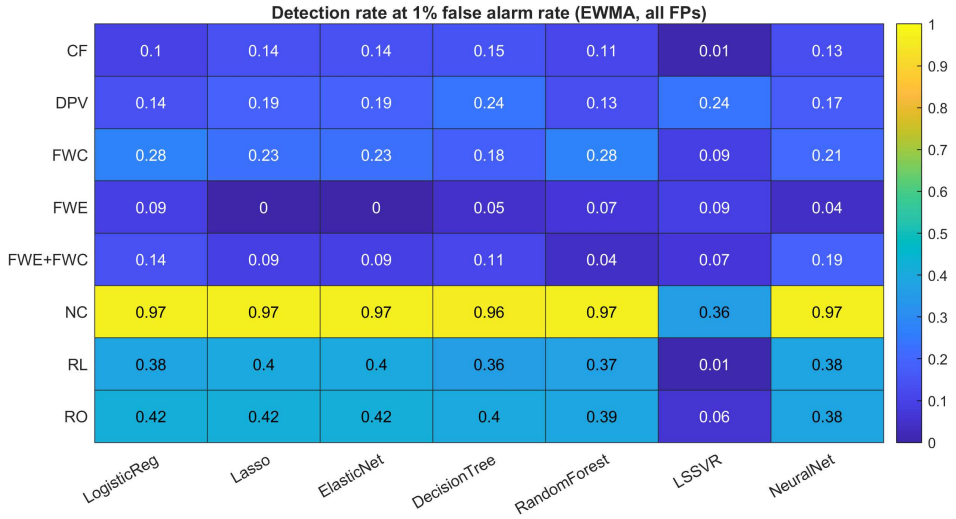


Fig. 8 Detection rate at a matched 1% false alarm rate per fault type and model. EWMA statistic, all FPs.

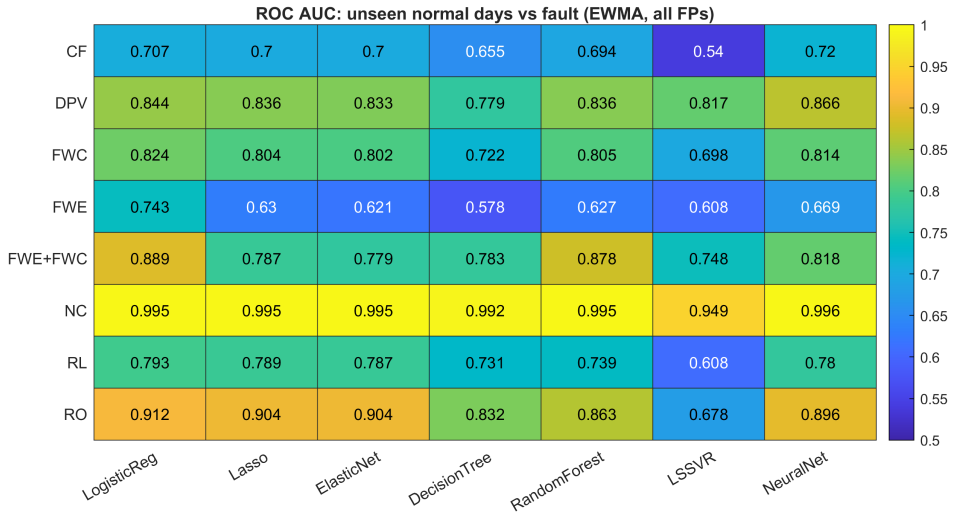


Fig. 9 ROC AUC per fault type and model. Unseen normal days versus fault data; EWMA statistic, all FPs.

4.6 Sensitivity of the feature parameters

The most sensitive FP of each fault agreed with the fault–symptom rules of the earlier strategies [3–5] (Table 8; Supplementary Figs. S5–S10). Non-condensable gas was seen through LMTDcd and TCA in every sample. Refrigerant overcharge and leakage were seen mainly through TCA, LMTDcd and ε_{sc}. At the highest severity, ε_{sc} exceeded its limit in 95–98% of the leakage samples and 91–93% of the overcharge samples for six models. Reduced condenser water flow, the simultaneous fault and the defective pilot valve were seen mainly through COP.

Table 8: Median fraction (%) of samples with $|z| > 3$ per fault type and FP

Fault type	LMTDcd	LMTDev	COP	kW	ε_{sc}	TCA
CF	10	0	13	7	3	12
RL	18	2	5	3	15	17
RO	24	1	13	8	16	30
FWC	11	6	48	7	3	3
FWE	1	2	8	5	2	2
NC	100	3	74	65	35	100
FWE+FWC	9	0	41	28	2	3

Table 8: Median fraction (%) of samples with $|z| > 3$ per fault type and FP

Fault type	LMTDcd	LMTDev	COP	kW	ε_{sc}	TCA
DPV	8	7	33	16	14	5

Median over the seven models; mean over the severity levels of each fault type.

Condenser fouling became visible in the condenser-side FPs only at the highest level. At L4, LMTDcd and TCA exceeded their limits in 51–60% of the samples and ε_{sc} in 27–29% for six models, and COP in 50–51% for the linear models and the neural network; at L1 and L2, no FP exceeded its limit in more than 19% of the samples. This agrees with the original RP-1043 analysis, which found that condenser fouling degrades performance detectably only when about 40% of the heat transfer area is eliminated [17]. The difficulty of CF at low severity was also reported for the RBF-, SVR- and LSSVR-based strategies [3, 5, 8] and, more generally, for incipient chiller faults [15]; its physical cause is analysed in Section 4.7.2.

4.7 Physical interpretation and sensor implications

4.7.1 Structure of the normal residuals

For five of the six FPs, 64–72% of the residual variance on the normal days consisted of offsets that persisted over a whole operating condition or day, and only 28–36% fluctuated within a condition (Table 9, Fig. 10). The between-day part alone was 15–22%. Random forest and the neural network showed the same structure. These offsets are variations of chiller behaviour that the four driving conditions do not describe. They explain three results of Sections 4.4 and 4.5: thresholds set on the fitting days were exceeded more often on new days; EWMA, which averages within a condition, could not remove them; and a fault becomes detectable only when its shift exceeds the segment-level scatter σ_{seg}, not the much smaller within-condition noise.

Table 9: Residual variance decomposition of the lasso models and measurement uncertainty

FP	Residual σ	Between days (%)	Between conditions within a day (%)	Within a condition (%)	σ within a condition	Measurement uncertainty
LMTDcd (K)	0.351	15.6	56.0	28.4	0.187	–
LMTDev (K)	0.257	22.3	43.9	33.8	0.150	–
COP (–)	0.041	15.3	48.4	36.2	0.025	0.117–0.153
kW (kW)	0.682	16.0	49.6	34.5	0.400	1.8
ε_{sc} (–)	0.048	2.7	25.3	72.0	0.041	0.065
TCA (K)	0.315	17.6	53.7	28.8	0.169	0.28

Five normal days. Measurement uncertainty from the facility uncertainty analysis [17].

ε_{sc} behaved differently: 72% of its residual variance lay within conditions and only 2.7% between days. Its inputs are among the least accurate measurements of the facility, which reported uncertainties of 5–15% for the liquid subcooling and 8–25% for the condenser approach temperature [17]. With a median subcooling of 3.5 K (uncertainty 0.30 K) and T_{cd} − TCI of about 5.4 K, the propagated uncertainty of ε_{sc} is about 0.065, or 10% of its value, which is larger than the total residual standard deviation of 0.048. The prediction error of ε_{sc} is therefore already at the limit set by its measurement, and no reference model can reduce it. TCA, with a stated uncertainty of 0.28 K against a residual σ of 0.315 K, is also close to this limit; 71% of its residual variance was nevertheless systematic. Because TCA is computed from the condensing saturation temperature, which the controller infers from the condenser pressure, part of these offsets may stem from the pressure transducer itself: with the chiller off, the inferred saturation temperature matched the water temperature only within ±0.8°F [17]. The water-side quantities, in contrast, are measured by RTDs with an uncertainty of ±0.05°F, and the residual standard deviations of COP and kW were well below their stated uncertainties (Table 9).

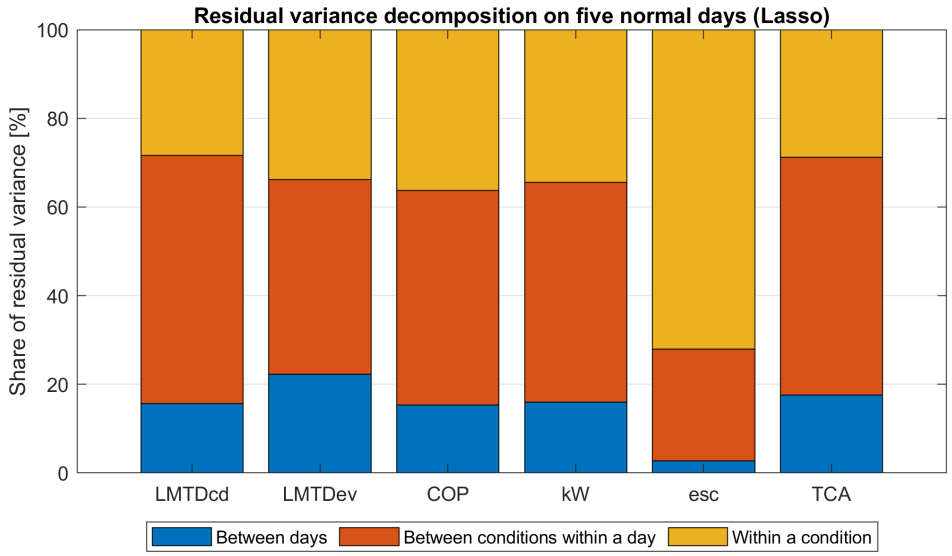


Fig. 10 Residual variance decomposition of the lasso models on the five normal days.

4.7.2 Why condenser fouling is detected only at the highest level

The effect sizes in Fig. 11 quantify how far each fault moves each FP relative to the segment-level scatter of normal operation. Non-condensable gas shifted LMTDcd and TCA by $8.3\sigma_{\mathrm{seg}}$ and $8.8\sigma_{\mathrm{seg}}$ already at L1. Refrigerant overcharge at L4 shifted LMTDcd, $\varepsilon_{\mathrm{sc}}$ and TCA by 8.1–$9.4\sigma_{\mathrm{seg}}$, and refrigerant leakage at L4 lowered $\varepsilon_{\mathrm{sc}}$ by $14.5\sigma_{\mathrm{seg}}$. The signs agree with the fault–symptom rules of Ref. [3]: leakage lowers $\varepsilon_{\mathrm{sc}}$ and LMTDcd, overcharge raises them, and non-condensable gas raises LMTDcd, TCA and $\varepsilon_{\mathrm{sc}}$. Condenser fouling moved the FPs in the expected directions, with higher LMTDcd, TCA and power and lower COP and $\varepsilon_{\mathrm{sc}}$. The shifts were at most $2.6\sigma_{\mathrm{seg}}$ at L1–L3 but 3.5–$4.4\sigma_{\mathrm{seg}}$ at L4. $\varepsilon_{\mathrm{sc}}$ fell because T_{cd} − TCI increased while the subcooling stayed nearly constant, as also observed in the original tests [17].

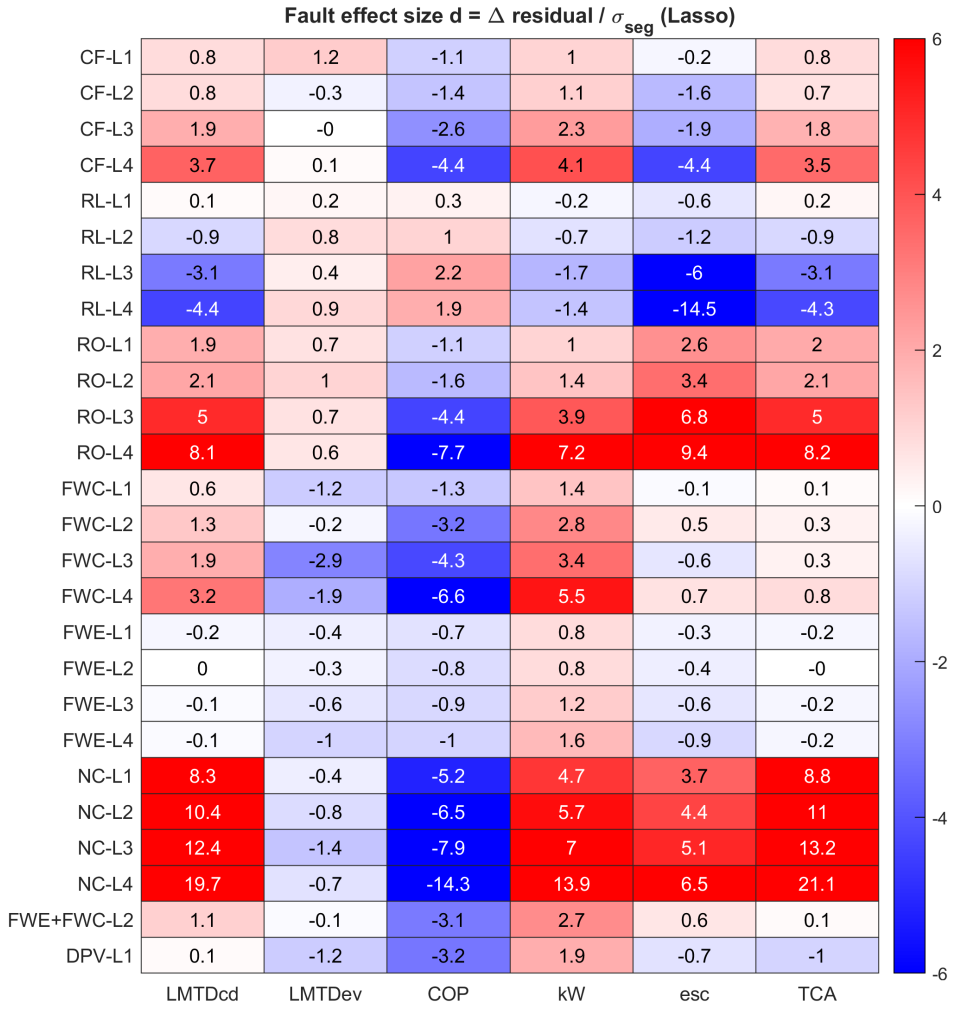


Fig. 11 Effect size of each fault case on each FP. $d = \Delta\text{residual}/\sigma_{\text{seg}}$, lasso models.

The reason lies in the condenser heat transfer itself (Table 10, Fig. 12). Plugging 12.2–45.1% of the tubes reduced UA_{cd} by only 5.4–22.7%, two to four times less than the loss of tube area. The observations lie closest to the resistance model with a water-side share $\phi_w = 0.7$, which is consistent with the higher water velocity in the remaining tubes compensating part of the lost area. At L1–L3, the reductions lie within $\pm 2\sigma_{\text{seg}}$ ($\pm 16.5\%$) of the day-to-day and condition-to-condition scatter of the

fault-free UA_{cd} estimate, and the per-condition values overlap zero at L1 and L2. At these levels, condenser fouling changes the condenser heat transfer less than normal operation varies, and no steady-state residual model with these sensors can separate it reliably from unseen normal days. Only at L4 did the mean reduction (−22.7%, d = −2.8) leave this band, and every operating condition showed a reduction of at least 13%. This coincides with the jump of the detection rate from at most 13% at L3 to 28–43% at L4 (Section 4.5) and with the finding of the original RP-1043 analysis that condenser fouling becomes detectable when about 40% of the heat transfer area is eliminated [17]. The plugged fractions are those recorded in the test log [17]; the slight reduction of the condenser water flow during the tests weakens the velocity effect only marginally.

Table 10: Observed change of UA_{cd} under condenser fouling compared with model predictions

Level	Plugged tubes (%)	Observed ΔUA_{cd} (%)	Range over conditions (%)	d	Area loss only (%)	Model $\phi_w = 0.3$ (%)	Model $\phi_w = 0.5$ (%)	Model $\phi_w = 0.7$ (%)
L1	12.2	−5.5	−14.2 to 4.7	−0.67	−12.2	−9.5	−7.6	−5.7
L2	20.1	−5.4	−15.2 to 9.3	−0.66	−20.1	−16.0	−13.0	−9.7
L3	29.9	−13.0	−22.6 to −1.8	−1.58	−29.9	−24.3	−20.0	−15.2
L4	45.1	−22.7	−32.4 to −13.0	−2.76	−45.1	−38.0	−32.2	−25.1

Predictions from area loss alone and from the resistance model (Section 3.7). Normal segment-level scatter $\sigma_{\mathrm{seg}} = 8.2\%$.

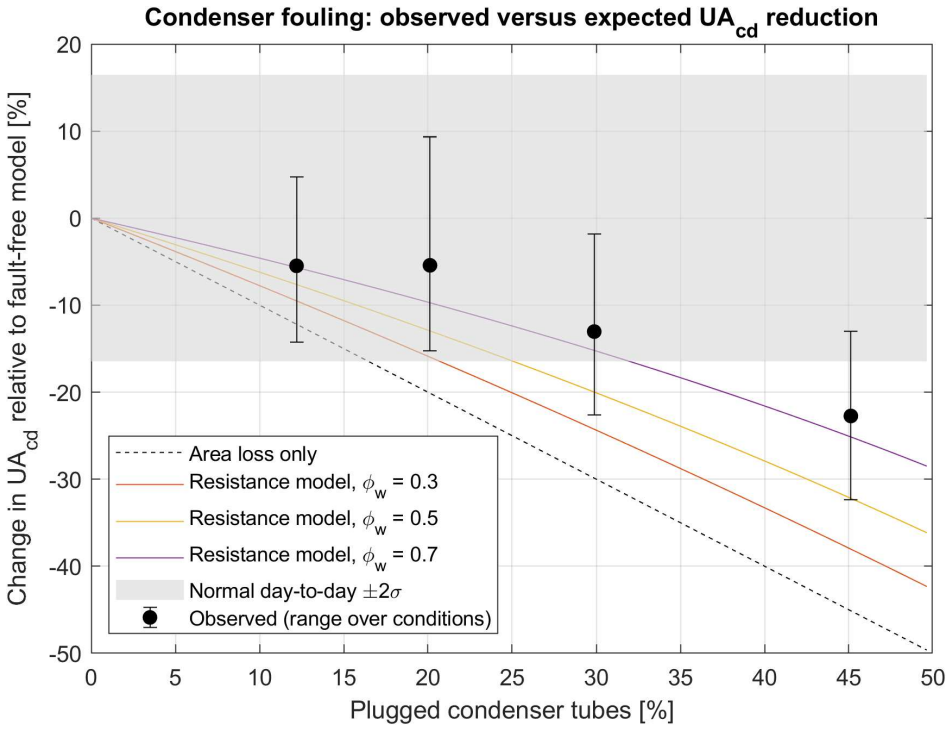


Fig. 12 Observed change of UA_{cd} under condenser fouling. Comparison with area loss, the resistance model and the normal day-to-day scatter ($\pm 2\sigma_{\mathrm{seg}}$) versus the fraction of plugged tubes.

Two practical consequences follow. First, the same velocity increase that masks the thermal effect raises the water-side pressure drop, roughly in proportion to $v^{1.8}$ in turbulent tube flow: by about 25% at 12% plugging, 80–90% at 30% plugging and 170% at 45% plugging, allowing for the small flow reduction. A condenser water pressure-drop measurement, which was not part of the RP-1043 instrumentation [17], would therefore indicate this type of fouling far more directly than any of the thermal FPs. Second, real scale fouling adds a fouling resistance without raising the water velocity, so it lowers UA_{cd} more than tube plugging of the same area fraction. The RP-1043 tests may therefore understate the thermal detectability of real condenser fouling.

4.7.3 Sensor ablation

Removing a single FP changed the mean matched DR by at most 1.8 percentage points for lasso, but the per-fault results show which measurements carry which faults (Table 11, Fig. 13).

Table 11: Mean detection rate (%) at a matched 1% FAR for subsets of the FPs

Configuration	Lasso	Neural network	Random forest
All six FPs	34.3 [32.6, 36.0]	36.0 [34.2, 37.8]	31.0 [29.3, 33.2]
Without LMTDcd	34.3 (0.54)	35.7 (<0.001)	30.7 (0.016)
Without LMTDev	34.3 (1.00)	36.0 (1.00)	31.0 (1.00)
Without COP	32.5 (<0.001)	34.9 (<0.001)	31.6 (0.95)
Without kW	34.1 (<0.001)	35.1 (<0.001)	30.5 (<0.001)
Without ε_{sc}	35.9 (0.48)	31.9 (<0.001)	25.8 (<0.001)
Without TCA	34.1 (<0.001)	39.2 (<0.001)	29.4 (<0.001)
Without condensing pressure (LMTDcd, TCA, ε_{sc})	33.4 (1.00)	34.1 (0.26)	8.9 (<0.001)
Without power meter (kW, COP)	30.7 (<0.001)	32.3 (<0.001)	30.4 (0.95)
Only LMTDcd	30.0 (<0.001)	27.3 (<0.001)	27.2 (<0.001)
Only LMTDev	5.2 (<0.001)	5.3 (<0.001)	8.0 (<0.001)
Only COP	39.8 (<0.001)	41.3 (<0.001)	7.6 (<0.001)
Only kW	26.8 (<0.001)	26.3 (<0.001)	5.6 (<0.001)
Only ε_{sc}	20.1 (<0.001)	21.5 (<0.001)	24.1 (<0.001)
Only TCA	26.9 (<0.001)	25.6 (<0.001)	23.2 (<0.001)

EWMA statistic. Brackets: 95% cluster bootstrap CI for the full set; parentheses: Holm-adjusted p-value of the paired comparison with the full set.

The condensing pressure sensor, from which LMTDcd, TCA and ε_{sc} are derived, carried the refrigerant-side faults. Without it, the lasso DR fell from 40.1% to 15.2% for refrigerant leakage, from 41.6% to 32.7% for overcharge and from 96.6% to 84.6% for non-condensable gas, and the mean DR of random forest collapsed from 31.0% to 8.9%. Within this group, LMTDcd and TCA were redundant, since removing either changed the lasso DR of any fault type by at most 0.7 percentage points, whereas ε_{sc} was indispensable for leakage (40.1% to 23.9% without it). The power meter, from which kW and COP are derived, carried the water-side and control faults and

condenser fouling. Without it, the lasso DR fell from 22.7% to 10.1% for reduced condenser water flow, from 9.2% to 0.1% for the simultaneous fault, from 19.2% to 9.8% for the defective pilot valve and from 14.1% to 10.2% for condenser fouling. LMTDev, which requires the evaporating pressure, contributed nothing: removing it left every DR unchanged, and alone it detected 5–8%.

More FPs did not always mean better detection. COP alone gave a higher mean DR than all six FPs for lasso (39.8% against 34.3%) and for the neural network (41.3% against 36.0%). For condenser fouling, COP alone doubled the lasso DR (29.2% against 14.1%), consistent with the original tests, in which power and the condenser water temperature difference were the quantities most sensitive to this fault [17]. Removing $\varepsilon_{\mathrm{sc}}$ raised the lasso DR for reduced condenser water flow from 22.7% to 34.8% and for condenser fouling from 14.1% to 19.6%, although it lowered the DR for leakage. When the score is the maximum over several FPs and the FAR is fixed, each added FP contributes its own normal-day tail and raises the threshold, so an FP that responds to only a few faults, or a noisy one such as $\varepsilon_{\mathrm{sc}}$, dilutes the detection of the others. FP sets should therefore be matched to the target faults, as also argued by recent feature-selection studies for chiller diagnosis [34], or combined by multivariate statistics that account for their correlation. The interaction with the model is also visible: COP alone failed for random forest (7.6%), whose COP residuals generalized poorly to the unseen days (Table 3).

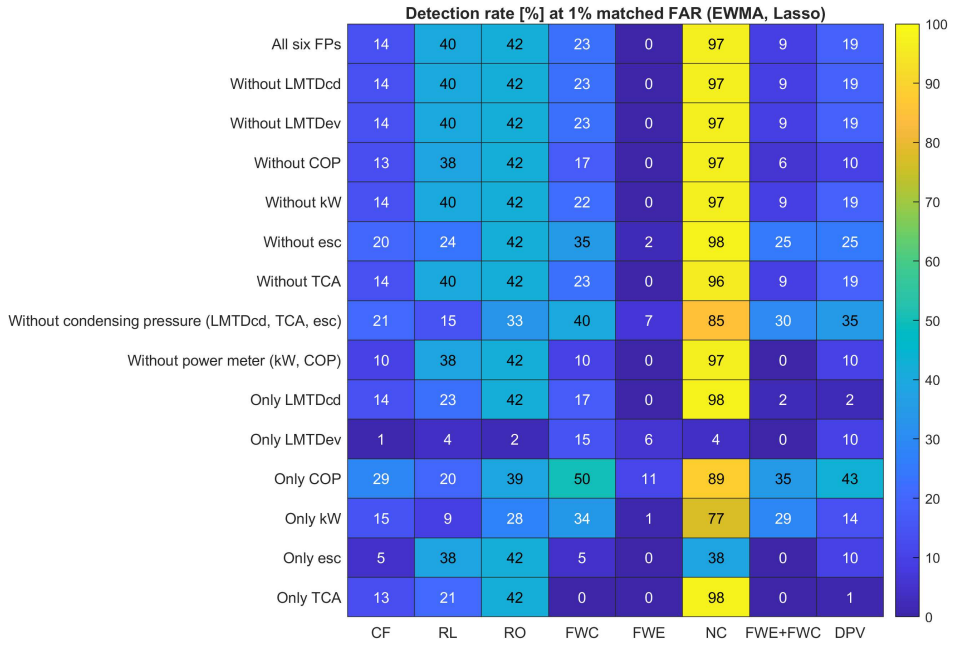


Fig. 13 Detection rate for subsets of the FPs. Matched 1% FAR, EWMA statistic, lasso models; per fault type.

The retraining of random forest and the neural network for this analysis changed their full-set mean DR from 34.3% to 31.0% and from 33.9% to 36.0% (Table 6 versus Table 11), because both depend on random initialization; lasso, which is deterministic, gave 34.3% in both. This run-to-run variation of 2–3 percentage points is larger than the differences between the six models in Table 6, and in this run the neural network even exceeded logistic regression. It supports the conclusion that the six models detect faults equally well at a matched FAR.

For instrumentation, these results suggest: (i) a condensing pressure sensor and a liquid-line temperature sensor for the refrigerant-side faults, measured with calibrated instruments rather than the uncalibrated onboard thermistors and controller-derived saturation temperature, because ε_{sc} and TCA are limited by measurement accuracy; (ii) a compressor power meter for the water-side and control faults; (iii) no need for

the evaporating pressure for the fault types studied here; and (iv) a condenser water pressure-drop measurement for condenser fouling.

4.8 Computational cost

All models are fast enough for online FDD at a 10-s sampling interval. The training time per FP was 0.01–0.08 s for the linear models, 0.1–0.3 s for the decision tree, 0.5–2.7 s for the neural network, about 7–8 s for LSSVR (60-point grid) and 7–16 s for random forest. Prediction for the whole same-day test set took less than 0.2 s for every model.

4.9 Implications and limitations

Four implications follow for the selection of chiller reference models and their instrumentation. First, a model should be selected on its error on unseen days and on the false alarm rate that its training-data threshold produces on unseen days, not on the training-day error. Second, once the false alarm rate is matched, six quite different models detect faults equally well. The reference model therefore mainly determines how reliably a threshold transfers to new days; further gains in detection must come from the measurements, the FP sets and the decision statistics. Third, regularized linear models with quadratic terms offer the best overall balance: the lowest unseen-day errors, the false alarm rate closest to the target and the lowest computational cost. Fourth, detectability is bounded by the condition-level and day-level offsets of the fault-free residuals and by the physical size of the fault: condenser fouling produced by tube plugging changes UA_{cd} less than these offsets until about 45% of the tubes are plugged, and its earlier detection needs a hydraulic rather than a thermal measurement.

The study has limitations. The data come from one laboratory chiller under steady-state conditions. The matched-FAR results use the unseen normal days to set the threshold and are a diagnostic comparison, not a deployable procedure. Even the best

models exceeded the 1% FAR target on unseen days, which indicates that adaptive or periodically recalibrated thresholds are needed in practice. The 480 Friedman blocks share the same operating conditions across the six FPs and are therefore not fully independent. The results of random forest and the neural network vary by a few percentage points between training runs. The physical analysis of condenser fouling relies on a simplified resistance model, and the water-side resistance share was not measured.

5 Conclusions

Seven supervised regression methods were compared as reference models for residual-based chiller FDD on the ASHRAE RP-1043 data. Each model predicted six feature parameters from four driving conditions. Accuracy was evaluated on the training day, on two fully held-out normal days and by leave-one-day-out prediction, and fault detection on 26 fault cases was evaluated with thresholds set from training data and at a matched false alarm rate. The main conclusions are as follows:

1. The accuracy ranking on the training day did not transfer to unseen days. The neural network and LSSVR gave the lowest same-day errors, but lasso and elastic net with quadratic terms gave the lowest unseen-day CV(RMSE) for five of the six FPs, e.g. 1.24% for COP and 1.39% for power.
2. Tree-based models and LSSVR lost the most accuracy on unseen days. For COP, the CV(RMSE) rose by factors of 3.3–4.8, against 1.25 for lasso. LSSVR generalized poorly although its kernel width was selected inside a wide hyperparameter grid.
3. Over 480 operating-condition blocks, the neural network had the significantly best mean rank (2.98; CD = 0.41), followed by elastic net (3.69).
4. With thresholds set from training data for a 1% false alarm rate, the linear models produced 1.2–2.7% false alarms on unseen days, the neural network 3.4–3.6%

and the tree-based models 6.7–9.2%. The reference model mainly determines how reliably a threshold transfers to new days.

5. At a matched 1% false alarm rate, six models gave similar mean detection rates (33.7–35.7% with EWMA; AUC 0.75–0.83), with only the neural network marginally below logistic regression ($p_{\mathrm{Holm}} = 0.04$). LSSVR was markedly lower (10.6%) because rare large errors inflated its residual standard deviation.

6. On the normal days, 64–72% of the residual variance of five FPs consisted of offsets persisting over whole operating conditions or days, which bounds detectability and limits the benefit of EWMA. The $\varepsilon_{\mathrm{sc}}$ residual was instead dominated by the measurement accuracy of the subcooling.

7. Condenser fouling was detected in at most 13% of the samples at levels 1–3 (12–30% of the tubes plugged) but in 28–43% at level 4 (45% plugged; AUC 0.88–0.95). Plugging reduced UA_{cd} by only 5–13% at levels 1–3, within the fault-free scatter, owing to velocity compensation on the water side, and by 23% at level 4, outside it. Non-condensable gas was detected in 96–97% of the samples.

8. The condensing pressure sensor carried the refrigerant-side faults and the power meter the water-side and control faults, while the evaporating pressure added nothing. COP alone detected more than all six FPs for lasso and the neural network, so adding FPs to a maximum-based score can dilute detection.

Novelty and comparison with existing works. Earlier comparisons of chiller reference models [4, 18] mainly ranked the models by their prediction error on normal data, without testing the differences statistically or examining how thresholds transfer to new days. The present study is, to the authors' knowledge, the first to compare reference models for chiller FDD on fully held-out normal days with Diebold–Mariano and Friedman–Nemenyi tests and to compare their fault detection at a matched false alarm rate. This exposed two effects that same-day evaluations hide: the training-day accuracy ranking reverses on unseen days, and the false alarm rate of a training-data

threshold rises up to ninefold above its target for tree-based models. For condenser fouling, earlier rule-based and statistical strategies reported no correct diagnosis at the two lowest severity levels (as summarized in [8]), whereas an EWMA–SVR strategy with a subcooling-section efficiency index reported 7.7–60.7% correctly diagnosed points at levels 1–3 [8]. The present study found at most 13% at levels 1–3 under a matched 1% false alarm rate on unseen days, and the UA_{cd} analysis shows that plugging less than about 45% of the tubes changes the condenser heat transfer less than the normal day-to-day scatter. Low-severity detection rates of condenser fouling obtained with thresholds from the training days should therefore be interpreted with care. Finally, the sensor ablation identifies which measurements carry which faults, a question that previous reference-model comparisons did not address.

Reference models for chiller FDD should therefore be selected on their unseen-day error and on the false alarm rate their training-data thresholds produce on new days, and their FP sets should be chosen for the target faults. Regularized linear models with quadratic terms offer the best balance of accuracy, threshold reliability and cost. Future work will address adaptive thresholds for day-to-day shifts, multivariate statistics for combining FPs, condenser water pressure-drop measurements for fouling, and validation on field data from building chillers.

Acknowledgements. [Acknowledgements, or "Not applicable."]

Declarations

Ethics approval and consent to participate. Not applicable.

Consent for publication. Not applicable.

Availability of data and materials. The dataset supporting the conclusions of this article is the ASHRAE Research Project RP-1043 data set [17], which is available from ASHRAE (American Society of Heating, Refrigerating and Air-Conditioning

Engineers); its use is subject to the terms of ASHRAE. All results that support the findings of this study are included within the article.

Competing interests. The authors declare that they have no competing interests.

Funding. Not applicable.

Authors' contributions. DATT: Conceptualization, Methodology, Formal analysis, Writing – review & editing. HNTC: Software, Formal analysis, Data curation. VTN: Investigation, Writing – original draft. TKPH: Visualization, Resources. All authors read and approved the final manuscript.

References

[1] Zhao Y, Li T, Zhang X, Zhang C (2019) Artificial intelligence-based fault detection and diagnosis methods for building energy systems: advantages, challenges and the future. Renew Sustain Energy Rev 109:85–101

[2] Chen Z, O'Neill Z, Wen J, Pradhan O, Yang T, et al (2023) A review of data-driven fault detection and diagnostics for building HVAC systems. Appl Energy 339:121030

[3] Tran DAT, Chen Y, Chau MQ, Ning B (2015) A robust online fault detection and diagnosis strategy of centrifugal chiller systems for building energy efficiency. Energy Build 108:441–453

[4] Tran DAT, Chen Y, Jiang C (2016) Comparative investigations on reference models for fault detection and diagnosis in centrifugal chiller systems. Energy Build 133:246–256

[5] Tran DAT, Chen Y, Ao HL, Cam HNT (2016) An enhanced chiller FDD strategy based on the combination of the LSSVR-DE model and EWMA control charts.

Int J Refrig 72:81–96

[6] Cui J, Wang S (2005) A model-based online fault detection and diagnosis strategy for centrifugal chiller systems. Int J Therm Sci 44(10):986–999

[7] Xiao F, Zheng C, Wang S (2011) A fault detection and diagnosis strategy with enhanced sensitivity for centrifugal chillers. Appl Therm Eng 31(17–18):3963–3970

[8] Zhao Y, Wang S, Xiao F (2013) A statistical fault detection and diagnosis method for centrifugal chillers based on exponentially-weighted moving average control charts and support vector regression. Appl Therm Eng 51(1–2):560–572

[9] Yao W, Li D, Gao L (2022) Fault detection and diagnosis using tree-based ensemble learning methods and multivariate control charts for centrifugal chillers. J Build Eng 51:104243

[10] Chen Z, Rezgui Y, Zhang R, Zhang X, Zhao W, Li Y (2025) Feature-level interpretability in transfer learning-based chiller fault diagnosis. Build Environ 285:113527

[11] Zhu X, Chen K, Anduv B, Jin X, Du Z (2021) Transfer learning based methodology for migration and application of fault detection and diagnosis between building chillers for improving energy efficiency. Build Environ 200:107957

[12] Liang X, Li P, Chen S, Jin X, Du Z (2022) Partial domain adaption based prediction calibration methodology for fault detection and diagnosis of chillers under variable operational condition scenarios. Build Environ 217:109099

[13] Chen Z, Zhang W, Zhao W, Yang X, Zhang X, Li Y (2024) Cross-condition fault diagnosis of chillers based on an ensemble approach with adaptive weight

allocation. Energy Build 325:115007

[14] Guo Y, Sheng Y, Wang H, Chu P, Xu J (2026) Semi-supervised generative adversarial network with a dynamic discriminator for chiller fault diagnosis under extreme label scarcity. Energy Build 369:117916

[15] Xia Y, Ding Q, Jiang A, Jing N, Zhou W, Wang J (2022) Incipient fault diagnosis for centrifugal chillers using kernel entropy component analysis and voting based extreme learning machine. Korean J Chem Eng 39(3):504–514

[16] Tian X, Liu J, Ma H, Zhang W, Zhang L (2026) A physically structured hybrid model predicts chiller performance under conformal uncertainty and protection events. Sci Rep 16:29727

[17] Comstock MC, Braun JE (1999) Development of analysis tools for the evaluation of fault detection and diagnostics for chillers. ASHRAE Research Project 1043-RP, Report HL 99-20. Ray W. Herrick Laboratories, Purdue University, West Lafayette

[18] Sreedharan P, Haves P (2001) Comparison of chiller models for use in model-based fault detection. In: Proceedings of the International Conference for Enhancing Building Operations (ICEBO), Austin, July 2001

[19] Papke LE, Wooldridge JM (1996) Econometric methods for fractional response variables with an application to 401(k) plan participation rates. J Appl Econom 11(6):619–632

[20] Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol 58(1):267–288

[21] Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol 67(2):301–320

[22] Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth, Belmont

[23] Breiman L (2001) Random forests. Mach Learn 45(1):5–32

[24] Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J (2002) Least squares support vector machines. World Scientific, Singapore

[25] ASHRAE (2014) ASHRAE Guideline 14-2014: Measurement of energy, demand, and water savings. American Society of Heating, Refrigerating and Air-Conditioning Engineers, Atlanta

[26] Diebold FX, Mariano RS (1995) Comparing predictive accuracy. J Bus Econ Stat 13(3):253–263

[27] Newey WK, West KD (1987) A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55(3):703–708

[28] Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6(2):65–70

[29] Friedman M (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J Am Stat Assoc 32(200):675–701

[30] Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7:1–30

[31] Lucas JM, Saccucci MS (1990) Exponentially weighted moving average control schemes: properties and enhancements. Technometrics 32(1):1–12

[32] Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman & Hall, New York

[33] Dittus FW, Boelter LMK (1930) Heat transfer in automobile radiators of the tubular type. Univ Calif Publ Eng 2(13):443–461

[34] Wang Z, Xia P, Guo J, Zhou S, Wang L, Wang Y, Zhang C (2025) Efficient feature selection for enhanced chiller fault diagnosis: a multi-source ranking information-driven ensemble approach. Build Simul 18:141–159