# MLF-RFDC:Multi-type Load Forecasting Model Algorithm based on Random Forest and Density Clustering

This is the supplementary file for the paper entitled *MLF-RFDC: Multi-type Load Forecasting Model Algorithm based on Random Forest and Density Clustering* in *IEEE Transactions on Industrial Informatics*. Additional sections, figures, and tables are put into this file and cited by the paper.

## S1. SUPPLEMENTARY FIGURE

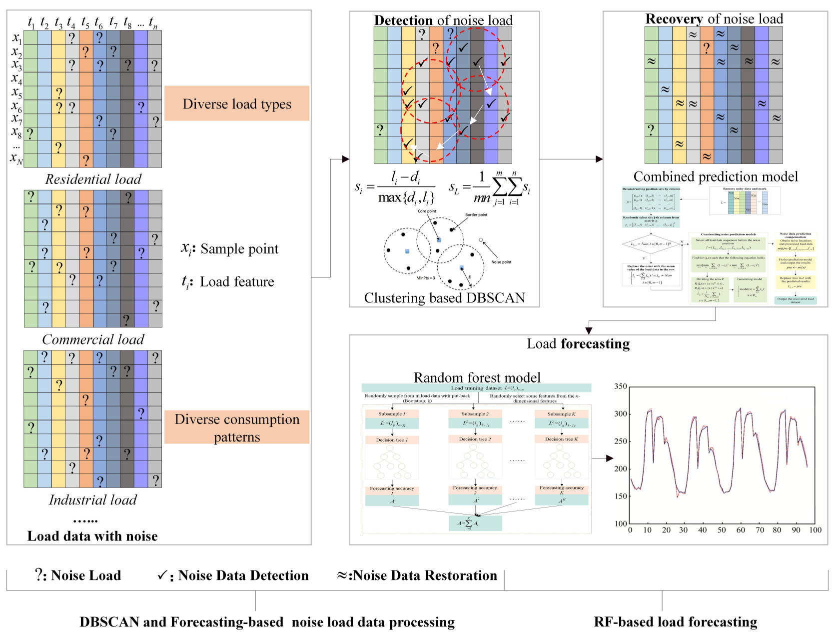


Fig. 1. Logic idea of MLF-RFDC.

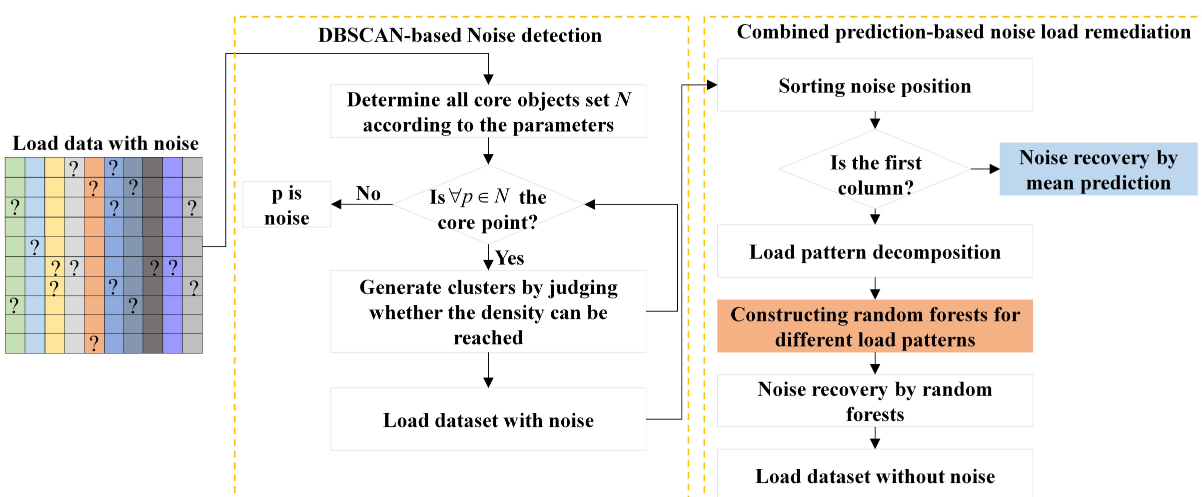


Fig. 2. The logical flow of noise load detection and repair.

## S2. SUPPLEMENTARY PROOF

**Proof.** According to Definition 5 of the original paper, we have the following equation holds

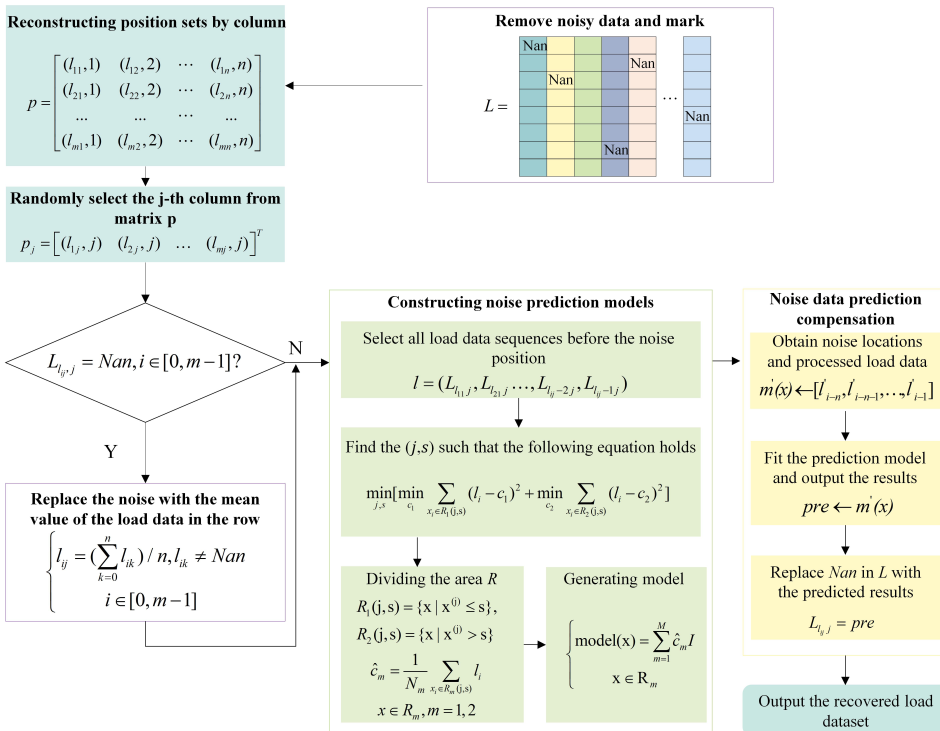


Fig. 3. The flow of NLS-CP.

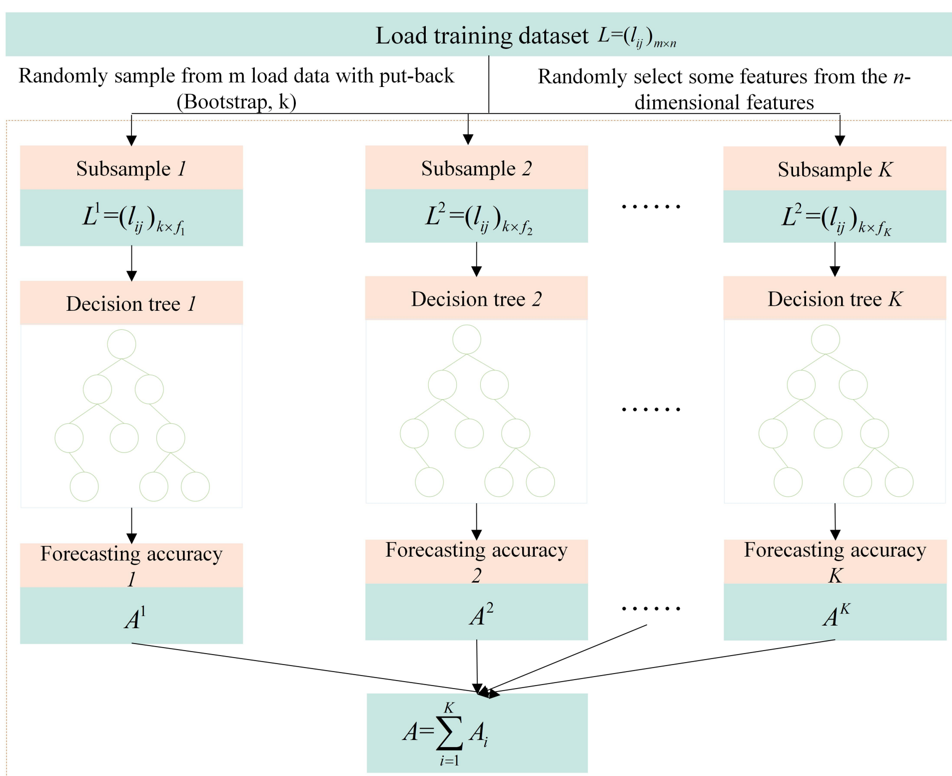


Fig. 4. The schematic diagram of the LFR-MF.

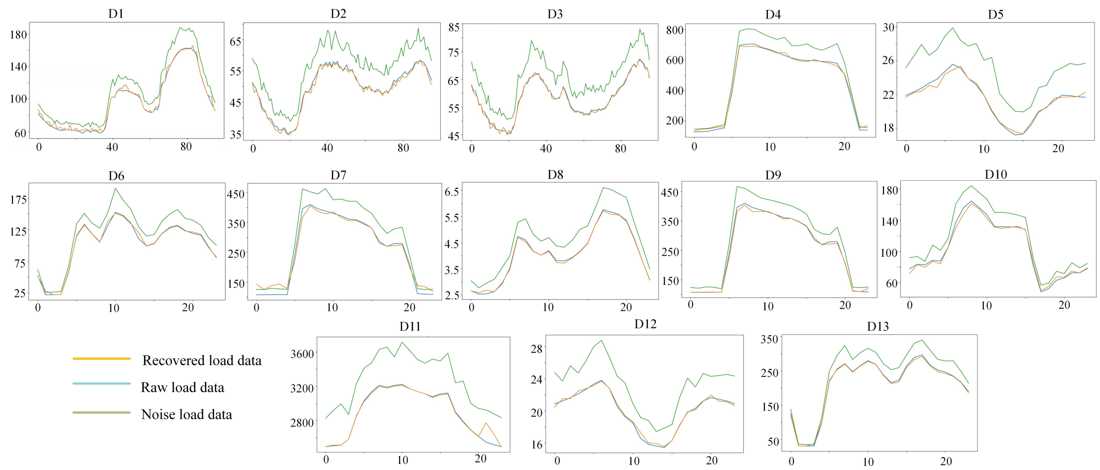


Fig. 5. Comparison of curves between original load data, restored load data and noisy load data.

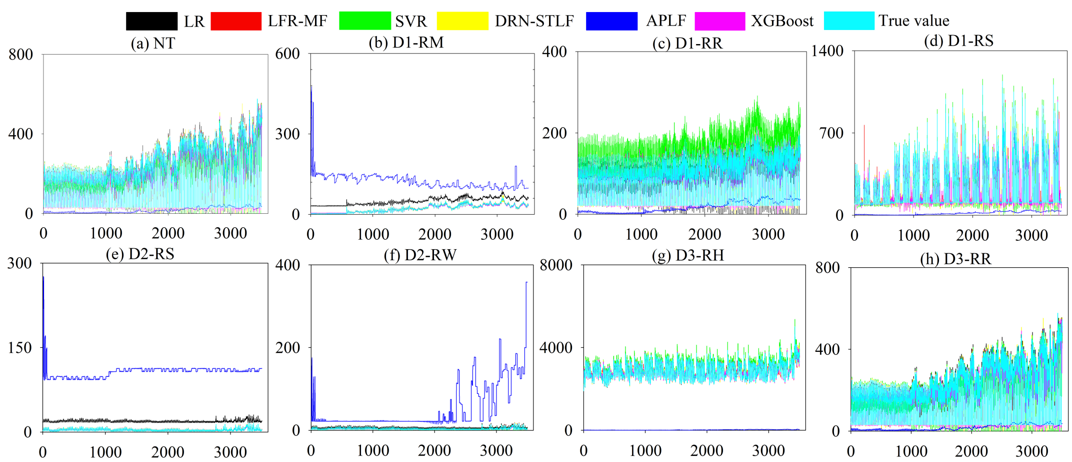


Fig. 6. Comparison of forecasting and true curves between different load forecasting models.

$$\begin{aligned} m_r(X,Y) &= a_{v_k}(T_k(X)=Y) - \max_{j\neq Y} a_{v_k}(T_k(X)=j) \\ &= P_\Theta(T(X,\Theta)=Y) - \max_{j\neq Y} P_\Theta(T(X,\Theta)=j). \end{aligned} \tag{1}$$

Due to $GE = P_{X,Y}(m_r(X,Y) < 0) = P_{X,Y}(m_r(X,Y) - E_{X,Y}(m_r(X,Y)) < -E_{X,Y}(m_r(X,Y)))$, according to Chebyshev's inequality, we have the following inequality holds

$$GE \leq \frac{Var(m_r(X,Y))}{(E_{X,Y}(m_r(X,Y)))^2} = \frac{Var(m_r(X,Y))}{s^2}. \tag{2}$$

Where $s = E_{X,Y}(m_r(X,Y))$. Let $\hat{j}(X,Y) = \arg\max_{j\neq Y} P_\Theta(T(X,\Theta)=j)$, then Eq. 1 can be changed to

$$\begin{aligned} m_r(X,Y) &= P_\Theta(T(X,\Theta)=Y) - P_\Theta(T(X,\Theta)=\hat{j}(X,Y)) \\ &= E_\Theta(I(T(X,\Theta)=Y) - I(T(X,\Theta)=\hat{j}(X,Y))). \end{aligned} \tag{3}$$

For a vector $\Theta$, we have $E(r_{m_r}(\Theta,X,Y)) = m_r(X,Y)$. Since $\Theta$ and $\Theta^{'}$ are two independent and identically distributed random vectors, for any function $f$, the following equation holds

$$(E(f(\Theta)))^2 = E_\Theta(f(\Theta)) \cdot E_{\Theta^{'}}(f(\Theta^{'})) = E_{\Theta,\Theta^{'}}(f(\Theta)\cdot f(\Theta^{'})). \tag{4}$$

Thus, $(m_r(X,Y))^2 = (E(r_{m_r}(\Theta,X,Y)))^2 = E_{\Theta,\Theta'}(r_{m_r}(\Theta,X,Y)\cdot r_{m_r}(\Theta',X,Y))$, then we have

$$\begin{array}{c} Var(m_r(X,Y)) = cov_{X,Y}(E(r_{m_r}(\Theta,X,Y)), \\ E(r_{m_r}(\Theta,X,Y))){=}E_{\Theta,\Theta'}(cov_{X,Y}(r_{m_r}(\Theta,X,Y), \\ r_{m_r}(\Theta,X,Y))) = E_{\Theta,\Theta'}(\rho(\Theta,\Theta')\cdot s_d(\Theta)\cdot s_d(\Theta')) \\ = E_{\Theta,\Theta'}(\rho(\Theta,\Theta'))\cdot E_{\Theta,\Theta'}(s_d(\Theta)\cdot s_d(\Theta')) \end{array}. \tag{5}$$

Where $s_d(\Theta)$ is the standard deviation of $r_{m_r}(\Theta,X,Y)$. Then, Eq. 6 holds

$$\begin{array}{c} Var(m_r(X,Y)) = \bar{\rho}\cdot(E_\Theta(s_d(\Theta)))^2 \\ \le \bar{\rho}(E_\Theta(s_d(\Theta)^2)) \\ = \bar{\rho}(E_\Theta(Var(\Theta))) \end{array}. \tag{6}$$

Where

$$\begin{array}{c} Var(\Theta) = Var(r_{m_r}(\Theta,X,Y)) \\ = E_{X,Y}(r_{m_r}(\Theta,X,Y))^2 - (E_{X,Y}(r_{m_r}(\Theta,X,Y)))^2 \\ = E_{X,Y}(r_{m_r}(\Theta,X,Y))^2 - s^2 \end{array}. \tag{7}$$

And since $r_{m_r}(\Theta,X,Y) = I(T(X,\Theta)=Y) - I(T(X,\Theta)=\hat{j}(X,Y)) \le 1,$ . From Eq. 7, we have $Var(\Theta)\le 1-s^2$. Then, the following inequality holds

$$E_\Theta(Var(\Theta)) \le 1 - s^2. \tag{8}$$

We bring Eq. 8 into Eq. 6, and obtain

$$Var(m_r(X,Y)) \le \bar{\rho}(1-s^2). \tag{9}$$

We then bring Eq. 9 into Eq. 2 to obtain the following inequality holds

$$GE \le \frac{Var(m_r(X,Y))}{s^2} \le \frac{\bar{\rho}(1-s^2)}{s^2}. \tag{10}$$

□

## S3. SUPPLEMENTARY TABLE

TABLE I
EXPERIMENTAL DATASETS

| Abbreviations | Name | # Instance | # Features |
|---|---|---|---|
| D1 | Nantong-Resident | 159 | 96 |
| D2 | Nantong-Commerce | 430 | 96 |
| D3 | Nantong-Mixed | 159 | 96 |
| D4 | District1-Market | 365 | 24 |
| D5 | District1-Resident | 365 | 24 |
| D6 | District1-Restaurant | 365 | 24 |
| D7 | District1-School | 365 | 24 |
| D8 | District2-Resident | 365 | 24 |
| D9 | District2-School | 365 | 24 |
| D10 | District2-Warehouse | 365 | 24 |
| D11 | District3-Hospital | 365 | 24 |
| D12 | District3-Resident | 365 | 24 |
| D13 | District3-Restaurant | 365 | 24 |

## S4. SUPPLEMENTARY DEFINITION

**Definition 1.** *$\forall p \in L$, the neighborhood with $p$ as the center and $\varepsilon$ as the radius is called the $\varepsilon$-neighborhood of $p$, denoted as $N\_\varepsilon(p) = \{q \in L | d(p,q) \le \varepsilon\}$.*

**Definition 2.** *$\forall p \in L$, $p_i, i \in [1,k]$ is in the $\varepsilon$-neighborhood of the load data $p$. If $k > Minpts$, then we call that the load data $p$ is a core point.*

**Definition 3.** *Let the load data* p *be the core point. For any $q$ in $L$, $q_i, i \in [1,k]$ is in the $\varepsilon$-neighborhood of the load data $q$. If $k < Minpts$ and $p_i \in N\_\varepsilon(p)$, then $q$ is said to be a boundary point.*

**Definition 4.** *$\forall p, q \in L$, if $p \in N\_\varepsilon(q)$ and $|N\_\varepsilon(q)| \ge Minpts$, then $q$ is said to be a direct density up to $p$, noted as $DDR(p,q)$.*

TABLE II
EXPERIMENTAL DATASETS WITH NOISE LOAD DATA

| Abbreviations | Name | # Instance | # Features | # Noise |
|---|---|---|---|---|
| D1 | Nantong-Resident | 159 | 96 | 768 |
| D2 | Nantong-Commerce | 430 | 96 | 2112 |
| D3 | Nantong-Mixed | 159 | 96 | 768 |
| D4 | District1-Market | 365 | 24 | 456 |
| D5 | District1-Resident | 365 | 24 | 456 |
| D6 | District1-Restaurant | 365 | 24 | 456 |
| D7 | District1-School | 365 | 24 | 456 |
| D8 | District2-Resident | 365 | 24 | 456 |
| D9 | District2-School | 365 | 24 | 456 |
| D10 | District2-Warehouse | 365 | 24 | 456 |
| D11 | District3-Hospital | 365 | 24 | 456 |
| D12 | District3-Resident | 365 | 24 | 456 |
| D13 | District3-Restaurant | 365 | 24 | 456 |

**Definition 5.** *$\forall p, q \in L$, $\exists p_1, p_2, ..., p_k$, if $p_1 = p$ and $p_n = q$, where $p_i$ can be directly densely accessible to $p_{i+1}$. Then $q$ is said to be densely accessible to $p$, denoted $DR(p, q)$.*

**Definition 6.** *$\forall p, q \in L$, $\exists r \in L$, If $r$ can be densely connected up to $p$ and $q$ at the same time, then $p$ and $q$ are said to be density connected, denoted as $DC(p, q)$.*

**Definition 7.** *$\forall p, q \in L$, let $C$ be load dataset, if $p$, $q$, $C$ satisfy the following properties: (1) if $p \in C$ and $DC(p, q)$, $q \in C$; (2)$DC(p, q)$. Then $C \subset L$, we call that $C$ is a cluster of $L$.*

**Definition 8.** *Let $N$ be the set of load data, and for a given $\varepsilon$ and $Minpts$, there exist $k$ clusters $C_1, C_2, ..., C_k$, and if $N = \{p \in L | p \notin C_i\}$, then is called the set of noisy load.*

## S5. SUPPLEMENTARY ALGORITHM

**Algorithm 1:** NLD-DBSPA

**Input:** $L = (L_1, L_2, ..., L_m)$,$\varepsilon$,$Minpts$
**Output:** Noise load dataset $N$

$L^{'} \leftarrow ReconCol(L)$;//Reconstructing the load dataset by column
Marking all load points in $L^{'}$ with an unprocessed status;
$(s_L, N_c) \leftarrow Init(L, \varepsilon, Minpts)$;
**while** *($|s_L - 1| > \sigma_1$ or $|1 - s_L| > \sigma_2$) and $|N_c - 1| > \sigma_3$* **do**
    **for** *each $p$ in $L$* **do**
        For object $p$, check the $\varepsilon$-neighborhood $N\_\varepsilon(p)$;
        **if** *the number of objects contained in $N\_\varepsilon(p)$ is less than $MinPts$* **then**
            $p$ is marked as a boundary point or noise point
        **else**
            $p$ is marked as a core point, and a new cluster $C$ is created, and all points in the neighborhood of $p$ are added to $C$;
            **for** *all $q \notin N\_\varepsilon(p)$* **do**
                **if** *$N\_\varepsilon(p)$ contains $MinPts$ objects* **then**
                    we add to $C$ the objects in $N\_\varepsilon(p)$ that are not classified;
**for** *each $p$ in $L$* **do**
    **if** *$p \in L$ and $p \notin C_i$* **then**
        $N \leftarrow N.insert(p)$;
**Return** $N$;

## S6. SUPPLEMENTARY EXPERIMENT

There are $N$ experimental datasets and $k$ comparison algorithms, and let denote the average ordinal value of the $i$-th algorithm. In the experiment, we set the variable $\Gamma_F$ to obey the $F$ distribution with $k - 1$ and $(k-1)(N - 1)$ degrees of freedom. For

**Algorithm 2:** NLS-CP

**Input:** $L = (L_1, L_2, ..., L_m)$, Noise Locations Set $P_N = [p_1, p_2, \cdots, p_N]$
**Output:** Recovered load dataset $R_L$

```
P_N ← ReconCol(P_N);//Reconstructing position sets P_N by column
Mark all load points in L'.columns as unprocessed;
for P_N[i] in P_N do
    L'[P_N[i]] = Nan;
for cols in L'.columns do
    i = P_N[i][0], j = P_N[i][1];
    if ∃L'_{i,j} = Nan, i ∈ [0, n−1] then
        L'_{ij} = (Σ_{k=0}^{m} L'_{ik})/m, p_{ik} ≠ Nan;
    else
        l = (L'_{i−n,j}, L'_{i−n−1,j}, ..., L'_{i−2,j}, L'_{i−1,j});
        l' = {(x_1,y_1), (x_2,y_2), ..., (x_n,y_n)} ← l;
    Find the (j,s) such that the following equation holds min_{j,s}[min_{c_1} Σ_{x_i∈R_1(j,s)} (l'_i − c_1)^2 + min_{c_2} Σ_{x_i∈R_2(j,s)} (l'_i − c_2)^2];
    (R_1(j,s), R_2(j,s)) ← Divid(R);
    ĉ_m ← (1/N_m) Σ_{x_i∈R_m(j,s)} l'_i;
    m(x) ← model(x) = Σ_{m=1}^{M} ĉ_m I;
    m'(x) ← [l'_{i−n}, l'_{i−n−1}, ..., l'_{i−1}];
    pre ← m'(x);
    L'_{ij} ← pre;
    R_L ← L';
Return R_L;
```

**Algorithm 3:** MLF-CC

**Input:** Load dataset: $L^I_{m\times n}$, $L^C_{k\times n}$
**Output:** The fused load dataset $L$

```
for i ← 1 to m do
    for j ← 1 to k do
        L^I_i ← (l^I_{i1}, l^I_{i2}, ..., l^I_{in});
        L^C_i ← (l^C_{i1}, l^C_{i2}, ..., l^C_{in});
        r_{L^I_i L^C_i} ← (n Σ_{j=1}^{n} l^I_{ij} l^C_{ij} − Σ_{j=1}^{n} l^I_{ij} Σ_{j=1}^{n} l^C_{ij}) / ( sqrt(n Σ_{j=1}^{n} (l^I_{ij})^2 − (Σ_{j=1}^{n} l^I_{ij})^2) · sqrt(n Σ_{j=1}^{n} (l^C_{ij})^2 − (Σ_{j=1}^{n} l^C_{ij})^2) );
        if r_{L^I_i L^C_i} ≈ 0 then
            L ← L.insert(L^I_i, L^C_i);
Return L;
```

the datasets after multi-type data fusion, this paper conducts Friedman test and Nemenyi post-hoc test on the prediction results of LFR-MF, LR, XGBoost, SVR, DRN_STLF and APLF.

The null hypotheses ($H_0$) and opposite hypotheses ($H_1$) of the Friedman test on the experimental data are denoted as follows.

1) $H_0$: There is no significant difference in the performance of multiple algorithms on multiple datasets;
2) $H_1$: There are obvious differences in the performance of multiple algorithms on multiple datasets.

If the difference of average ordinal value between any two algorithms exceeds the critical value range, the following inequality holds.

$$(r_i - r_j) > q_\alpha \sqrt{\frac{k(k+1)}{6N}} \tag{11}$$

Where $q_\alpha$ is the test critical value. Then we reject the null hypothesis $H_0$.

If the test result rejects $H_0$, it means that the performance of the $k$ algorithms is significantly different, and then the performance of the $k$ algorithms is further distinguished based on the Nemenyi post-hoc test.

For the datasets after multi-type data fusion, the Friedman test results are shown in Table III.

TABLE III
FRIEDMAN TEST RESULTS

| Index | Test statistic value | p-value |
|---|---|---|
| RMSE | 23.728 | 0.00127 |
| MAE | 27.485 | 0.00027 |
| MAPE | 14.626 | 0.0411 |
| $R^2$ | 16.157 | 0.02372 |

The experimental results in Table III demonstrate that p-values corresponding to RMSE, MAE, MAPE and $R^2$ are less than 0.05, respectively. Then we reject the null hypothesis $H_0$. In other words, the performance of each load forecasting algorithm is significantly different, and Nemenyi post-hoc test is required to further distinguish each algorithm. Fig. 7 shows the test results of six load forecasting algorithms based on Nemenyi post-hoc test for different evaluation metrics.

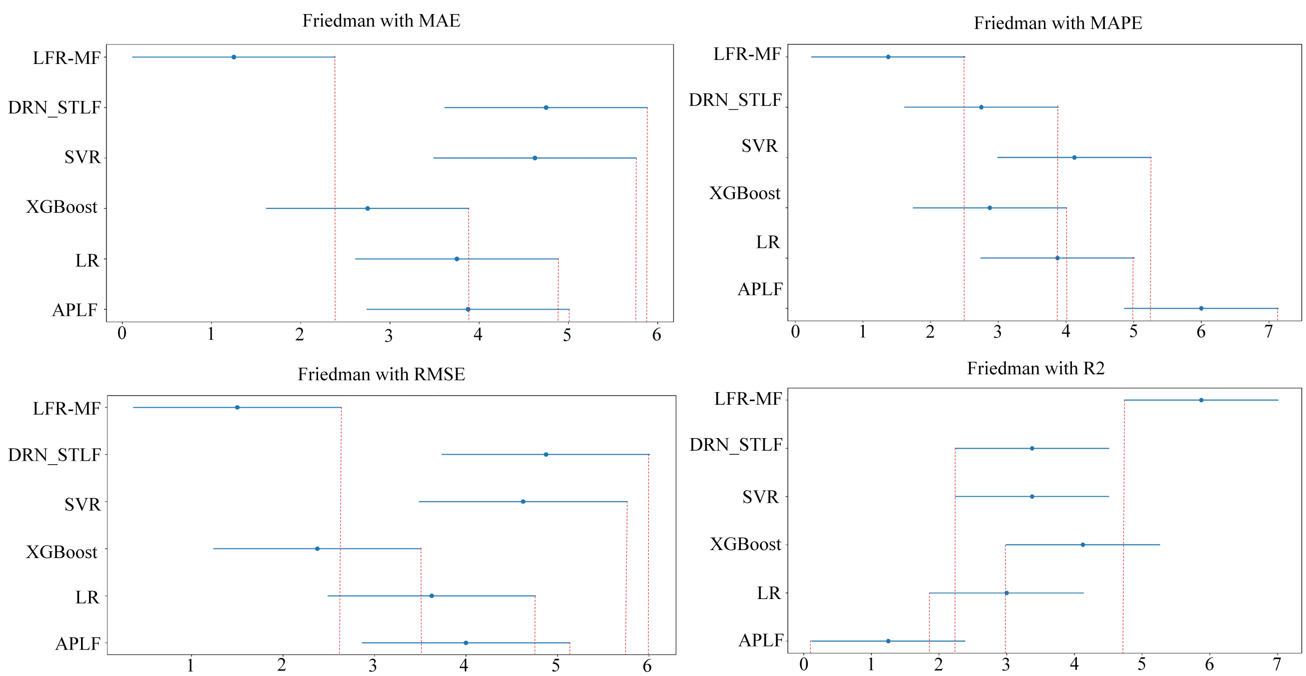


Fig. 7. Comparison of test results of six load forecasting algorithms based on Nemenyi post-hoc test.

As can be seen in Fig. 7, our proposed LFR-MF algorithm has no overlapping regions with DRN_STLF, SVR and APLF in terms of MAE, RMSE and , which indicates that the LFR-MF is significantly different from the above three models. And compared with the LR, the LFR-MF has only a few overlapping regions with it on the RMSE. Although the results in Fig. 7 show that there is no significant difference in the performance of the LFR-MF compared with the XGBoost model. However, judging from the forecasting results in Fig. 9 and Fig. 10 of the original paper, it still outperforms XGBoost.

## S7. SUPPLEMENTARY EVALUATION CRITERIA

### 1) Metrics for noisy data detection

To effectively evaluate the performance of various noise data detection algorithms, we set $T_n$ to denote the number of normal data identified as normal data, $F_n$ to denote the number of noisy data identified as normal data, $T_a$ to denote the number of noisy data identified as noisy data, and $F_a$ to denote the number of normal data identified as noisy data.

$P_r = \frac{T_n+T_a}{T_n+T_a+F_n+F_a}$ is called the precision of the noise data detection algorithm. The higher the $P_r$, the better the performance of the noise data detection algorithm.

$F_r = \frac{F_a}{T_n+F_a}$ is called the false detection rate of the noisy data detection algorithm. The smaller the $F_r$, the better the performance of the noisy data detection algorithm.

$M_r=\frac{F_n}{T_a+F_n}$ is called the missing detection rate of the noisy data detection algorithm. The smaller the , the better the performance of the noisy data detection algorithm.

**2) Metrics for noisy data restoration**

Let $\bar{e}_{(d_r,d_t)} = \sum_{i=1}^{m} |\frac{|(d_t-d_r)|}{d_t} - 1|/m$ denote the average data recovery rate of the noisy data restoration algorithm, where $d_r$ is the recovered value of the load data at time $t$, $d_t$ is the actual load value at time $t$, and $m$ is the number of noisy data in each experiment.

**3) Metrics for load forecasting**

And in this paper, we use RMSE, MAE, MAPE, and $R^2$ as shown in the following equations to evaluate the performance of the load forecasting model, respectively.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2} \tag{12}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\widehat{y}_i - y_i| \tag{13}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\widehat{y}_i - y_i}{y_i}\right| \tag{14}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{N}\left(y_i - \frac{1}{N}\sum_{i=1}^{N}y_i\right)^2} \tag{15}$$

Where $N$ denotes the size of the load dataset, $y_i$ denotes the true value of the $i$-th load data, and $\widehat{y}_i$ denotes the predicted value of the $i$-th load data.