Journal of Electronic Testing (2024) 40:557–572
https://doi.org/10.1007/s10836-024-06134-6

RESEARCH

# High-Dimensional Feature Fault Diagnosis Method Based on HEFS-LGBM

**Gen Li[1,2] · Wenhai Li[1] · Tianzhu Wen[1] · Weichao Sun[1] · Xi Tang[1]**

Received: 13 April 2024 / Accepted: 9 August 2024 / Published online: 5 September 2024


## Abstract

The challenge caused by redundant feature interference in high-dimensional fault feature data of analog circuits, will undermines the efficacy of conventional analog circuit fault diagnosis techniques, Thus, a novel approach termed Heterogeneous Ensemble Feature Selection (HEFS) is proposed in this paper. This approach is synergistically integrated with the Light Gradient Boosting Machine (LGBM) for pattern recognition, facilitating the prioritization and selection of significant high-dimensional features in analog circuit test data before classification. The methodology commences with the deployment of a heterogeneous ensemble learning strategy for the discernment of crucial high-dimensional features based on their significance. This is followed by the application of the LGBM technique for the pattern recognition classification of the earmarked features. Furthermore, the Tree-structured Parzen Estimator (TPE) optimization method, and five-fold cross-validation, are used for hyperparameter optimization to improve the model's performance. Diagnostic evaluations are conducted on both University of California Irvine (UCI) datasets and analog circuits to underscore the superior diagnostic precision of the proposed HEFS-LGBM method compared with the existing techniques.

**Keywords** High-dimensional features · Feature selection · Ensemble learning · Heterogeneous ensemble · Analog circuits · Fault diagnosis

## 1 Introduction

Analog circuits are pivotal in processing analog signals and are widely used in technological domains, including communication, navigation, control systems, and power supply units. Despite constituting an average of 20% of all circuit architectures, analog circuits are disproportionately affected by faults, accounting for over 80% of total circuit malfunctions [1]. Due to the inherent non-linearity and tolerance variations of components in analog circuits, fault diagnosis techniques for these circuits remain underdeveloped. Consequently, this technology does not yet meet the practical requirements of engineering applications [2]. In existing research on analog circuit fault diagnosis, Huang et al. [3] developed a method that can precisely identify and locate local defects in analog circuits. Pavlidis et al. [4] proposed a fault diagnosis method for analog circuits based on Built-In Self-Test (BIST), while Melis et al. [5] introduced a photonic emission tracking technique for diagnosing faults in analog circuits used in automotive applications. Each of these studies approaches analog circuit fault diagnosis from different perspectives, yet there is relatively little research addressing the diagnosis of analog circuit faults in the context of high-dimensional features. High-dimensional features are crucial because they can provide a more comprehensive representation of the circuit's state, capturing subtle variations that may be indicative of faults. However, the challenge lies in effectively managing these high-dimensional features to avoid issues such as redundancy and irrelevance, which can degrade diagnostic performance. Our proposed methods are specifically designed to address these challenges by employing advanced feature selection techniques to identify the most relevant features and enhance the accuracy and efficiency of fault diagnosis.

Responsible Editor: H-G. Stratigopoulos.

✉ Gen Li
504123921@qq.com

[1] Aviation Combat Service Academy, Naval Aviation University, Yantai 264001, People's Republic of China

[2] Repair Brigade, 77120 Unit of the Chinese People's Liberation Army, Chengdu 611930, People's Republic of China

In the realm of AI-driven analog circuit fault diagnosis, the process is frequently conceptualized as a pattern recognition challenge. Two pivotal components in diagnosing faults within analog circuits characterized by high-dimensional features are feature selection and fault mode classification. Feature selection, the extraction of a comprehensive set of high-dimensional features from analog circuit states, typically provides more valuable information. However, the process may also introduce redundant and irrelevant features. Redundant features can lead the model to disproportionately focus on these aspects, overshadowing other more distinctive features. Conversely, irrelevant features, which lack a meaningful relationship with the target variable, can detract from the model's accuracy. Employing feature selection techniques to identify and eliminate these superfluous or irrelevant features can significantly mitigate data noise, enhance the model's predictive accuracy, and bolster its generalization capabilities [6]. Moreover, feature selection aids in reducing feature dimensionality, curbing the risk of overfitting, diminishing the computational demands during classification, and elucidating the significance of various circuit characteristics. Various feature selection methodologies have been introduced, including SelectKBest [7], Recursive Feature Elimination (RFE) [8], Mutual Information [9], Random Forest [10], and so on. In the domain of pattern classification, prevalent approaches include Support Vector Machine (SVM) [11], Logistic Regression [12], eXtreme Gradient Boosting (XGBoost) [13], and so on. Recently, ensemble learning-based feature selection has emerged as a pivotal technique, particularly in the medical and biological sciences [14]. Additionally, the Light Gradient Boosting Machine (LGBM), an ensemble learning classifier, has garnered considerable attention for its exceptional efficacy [15].

This paper focuses on the challenge of redundant and irrelevant features in high-dimensional feature scenarios within analog circuits, which compromise fault diagnosis and lead to suboptimal diagnostic outcomes. A novel fault diagnosis methodology that leverages Heterogeneous Ensemble Feature Selection (HEFS) and Light Gradient Boosting Machine (LGBM) is introduced. Firstly, the deployment of HEFS is conducted to identify and select the most critical features from the data samples. Followingly, LGBM is employed for the classification of faults. Finally, the Tree-structured Parzen Estimator (TPE) is used for comprehensive hyperparameter optimization of both HEFS and LGBM, to enhance diagnostic precision and reduce labor costs. The innovation and progress of this method are manifested through:

(1) The adoption of a heterogeneous ensemble method and feature importance ranks strategy for feature selection in high-dimensional samples, effectively eliminating redundant and irrelevant features;
(2) The integration of HEFS and LGBM within a unified hyperparameter tunes framework through the TPE optimization method and stratified cross-validation, markedly improving the model's overall efficacy;
(3) The validation of the proposed method is conducted by five public UCI datasets and an analog circuit, demonstrating superior diagnostic accuracy in comparison to other feature selection and classification approaches.

The structure of the paper is methodically organized as follows: it begins with the introduction of the HEFS ensemble feature selection method, followed by the establishment of the HEFS-LGBM high-dimensional fault diagnosis model. Subsequently, the implementation framework and evaluation metrics for analog circuit fault diagnosis are delineated. The paper concludes with diagnostic experiments conducted on UCI datasets and an analog circuit, accompanied by a comprehensive discussion.

## 2 HEFS-LGBM Fault Diagnosis Model

### 2.1 HEFS Ensemble Feature Selection

The strength of ensemble learning methodologies resides in their capacity to harness effective features from an array of base classifiers, thereby enhancing the probability of discerning the underlying patterns within data. Predominantly, existing ensemble feature selection approaches are homogeneous, utilizing identical base classifiers [16, 17]. This paper advocates for the adoption of a diverse ensemble of heterogeneous base classifiers to infuse greater diversity into ensemble learning, with the objective of fostering more robust and universally applicable performance. The core concept of the HEFS (Heterogeneous Ensemble Feature Selection) approach involves initially training a variety of heterogeneous base classifiers, subsequently deriving a list of feature importance rankings from each classifier, and ultimately consolidating the most significant features through a specific aggregation methodology. Drawing upon the inclusion of 22 distinct advanced classification algorithms from the literature [18], this study undertakes 10 experiments on each algorithm using an analog circuit dataset. To ensure optimal adaptability to datasets of varying characteristics, the selection of base classifiers spanned a broad spectrum of algorithmic categories, ranging from simple to complex and linear to nonlinear. The nine algorithms demonstrating the highest mean accuracy were chosen as base classifiers, encompassing Bagging [19], Extra Trees [20], GBDT (Gradient Boosting Decision Tree) [21], Random Forest [10], Logistic Regression [12], Passive Aggressive [22], Ridge [23], SGD (Stochastic Gradient Descent) [21], and SVM (Support Vector Machine) [11].

The base classifiers selected for the heterogeneous ensemble demonstrate marked diversity, employing methodologies for feature importance extraction that can be categorized into two distinct groups: tree model ensemble algorithms (such as Bagging, Extra Trees, GBDT, and Random Forest) and linear model algorithms (such as Logistic Regression, Passive Aggressive, Ridge, SGD, and SVM). For tree model ensemble algorithms, feature importance is ascertained by tallying and ranking the frequency at which features are utilized in tree splits, from the most frequent to the least, thereby deriving the importance scores and rankings. Conversely, for linear model algorithms, the relative importance of each feature is indicated by the coefficient values attributed to each feature, organized from the highest to the lowest absolute value, thus facilitating the determination of importance scores and rankings.

Given the disparate methodologies for determining feature importance, the two approaches yield heterogeneous values that are not directly comparable or aggregable. To address this challenge, this paper introduces an innovative method that recalibrates the importance scores assigned by the base classifier algorithms, leveraging both the ranking of feature importance and the frequency of occurrence within the importance ranking lists. Subsequently, it aggregates these recalibrated feature importance scores derived from the base classifiers. This strategy not only facilitates the effective integration of the base classifiers' insights but also streamlines the computational process.

Figure 1 depicts the workflow of the HEFS (Heterogeneous Ensemble Feature Selection) process. Let's consider the scenario where the training dataset undergoes $N$ splits, there are $K$ base classifiers, $P$ represents the feature selection proportion factor, and the dataset initially contains $F$ features. To mitigate the effects of training errors, the dataset is subjected to N-fold stratified splitting. In this setup, during each iteration, one fold is designated as the validation set while the remaining N-1 folds are utilized for training. Consequently, this arrangement enables each base classifier to participate in $N$ iterations of stratified cross-validation experiments.

Firstly, we define the function TopFeatures($l$, $m$) as the process of selecting the top $m$ features based on the highest average importance scores from the sorted list $l$. In this context, $m$ represents the number of features determined by the ceiling operation $\lceil \cdot \rceil$.

Thus, for each base classifier $k$ and each fold $n$ of the validation dataset, $l_{kn}$ denotes the feature importance ranking list generated by that base classifier on that specific validation set.

$$T = TopFeatures\left(l_{kn}, \lceil F \cdot P \rceil\right) \tag{1}$$

$T$ signifies the selection of the top $\lceil F \cdot P \rceil$ features with the highest importance scores from the $l_{kn}$ list, where $F$ represents the total number of features, and $P$ is the selection proportion factor. In instances where a base classifier assigns the $\lceil F \cdot P \rceil$-th feature an importance score identical to others, the features are handled using a method that maintains their sequence as in the original list.

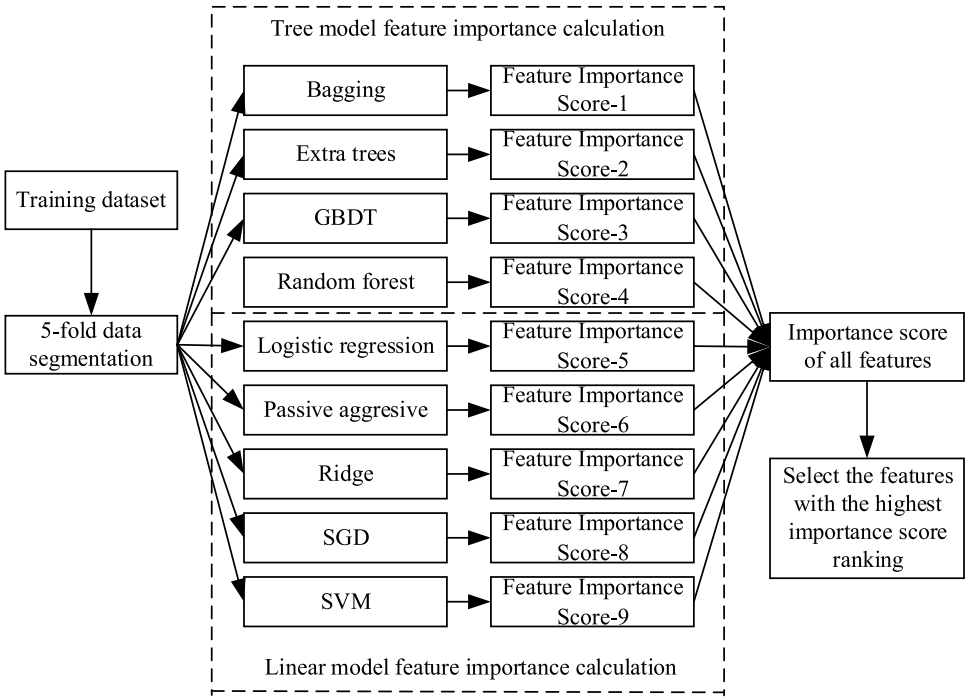



**Fig. 1** The basic process of HEFS feature selection

Let $w_{fnk}$ represent the weight assigned to feature $f$, where the value of $w_{fnk}$ is defined as follows:

$$w_{fnk} = \begin{cases} 1, \ if \ f \in T \\ 0, \ if \ f \notin T \end{cases} \tag{2}$$

Let $N_t$ denote the total importance score of feature $f$, where the value of $N_t$ is calculated as:

$$Nt = \sum_{n=1}^{N} \sum_{k=1}^{K} w_{fnk} \tag{3}$$

Let $N_c$ represent the theoretical maximum value of the total importance score for feature $f$, where the value of $N_c$ is given by:

$$N_c = KN \tag{4}$$

In this study, with the number of base classifiers $K=9$ and the number of splits in the training dataset $N=5$, the value of $N_c$ is calculated as $K \times N = 9 \times 5 = 45$.

The average importance score for each feature $f$ is calculated as follows:

$$s_f = N_t / N_c \tag{5}$$

Ultimately, utilizing the computed average importance scores, all features are arranged in descending order. The top $\lceil F \cdot P \rceil$ features, boasting the highest $s_f$ scores, are then selected. In scenarios where the average importance score of the $\lceil F \cdot P \rceil$-th feature matches that of others, the features are managed through a methodology that retains their sequence from the original list, mirroring the strategy employed earlier.

The pseudo-code for implementing feature selection via the HEFS (Heterogeneous Ensemble Feature Selection) algorithm is outlined as Algorithm 1.

**Algorithm 1** Heterogeneous Ensemble Feature Selection (HEFS).

| | |
|---|---|
| Input: | $N$—Number of folds for splitting the training dataset |
| | $K$—Number of base classifiers |
| | $P$—Proportion for feature dimension selection |
| | $F$—Initial feature dimension |
| Output: | Selection of the top $\lceil F \cdot P \rceil$ features with the highest importance scores |
| 1 | **For** each n-th fold of the dataset, where $n$=1, 2, …, $N$: |
| 2 | **For** each base classifier $k$=1, 2, …, $K$: |
| | Train base classifier $k_n$ using all features from the N-1 folds of data excluding the n-th fold; |
| | Test base classifier $k_n$ on the n-th fold dataset; |
| | Obtain the feature importance ranking list $l_{kn}$ from $k_n$; |
| | Assign weights $w_{fnk}$ to each of the $F$ features. |
| 3 | **For** each feature $f$=1, 2, …, $F$: |
| | **If** $f$ is among the top $\lceil F \cdot P \rceil$ features in the $l_{kn}$ list: |
| | $w_{fnk}$=1; |
| | **Else** |
| | $w_{fnk}$=0; |
| 4 | Calculate $Nc=N \cdot K$; |
| 5 | **For** each feature $f$=1, 2, …, $F$: |
| | Calculate $N_t=\sum_{n=1}^{N} \sum_{k=1}^{K} w_{fnk}$; |
| | Calculate $s_f= N_t/N_c$; |
| 6 | Select the top $\lceil F \cdot P \rceil$ features with the highest $s_f$ scores |

### 2.2 LGBM Fault Mode Classification

The Light Gradient Boosting Machine (LGBM), introduced by Ke Guolin and colleagues at Microsoft Research Asia in 2017 [24], is a distributed decision ensemble learning method renowned for its exceptional classification capabilities [25]. Following the elimination of redundant features via the HEFS ensemble feature selection model, the refined dataset, now streamlined through feature selection, is fed into the LGBM model for training. This simplification of data alleviates the training load on the LGBM model, and the exclusion of superfluous features is anticipated to enhance classification performance.

LGBM is built upon the foundation of the Gradient Boosting Decision Tree (GBDT). In the iterative process of GBDT, consider the training dataset to be $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}, x_i \in X, y_i \in Y$, where $X$ represents the input features of the samples, and $Y$ denotes the corresponding labels. The loss function is represented by $L(y, F(x))$, where $F(x)$ is the robust model ultimately derived, consisting of all models linearly combined with respective weights. The formula for $F(x)$ is as follows:

$$F(x) = \sum_{m=1}^{M} \beta_m f(x;\gamma_m) \tag{6}$$

Herein, $f(x;\gamma_m)$ represents the weak model, $\beta_m$ is the weight coefficient, and $\gamma_m$ denotes the parameters of the weak model.

During the iteration process, if the model at the m-1st iteration is denoted by $f_{m-1}(x)$, and the loss function by $L(y_{m-1}, f_{m-1}(x))$, then the objective of the m-th iteration is to identify a weak learner, $h_m(x)$, that minimizes the loss function $L(y_m, f_m(x)) = L(y_m, f_{m-1}(x) + h_m(x))$, in the current iteration. This implies that in the m-th iteration, the aim is to discover a decision tree model, $h_m(x)$, that reduces the sample loss to the greatest extent possible. Typically, the negative gradient of the loss function is employed to approximate the residual at the current value of the function $f(x) = f_{m-1}(x)$, and the formula is expressed as:

$$r_{mi} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]\Big|_{f(x)=f_{m-1}(x)} \tag{7}$$

Based on the data $(x_i, r_{mi})$, the m-th decision tree can be fitted, with its corresponding leaf node region denoted by $R_{mj}$, where $j = 1,2,...,J$ represents the number of leaf nodes. For each sample within a leaf node, the output value $c_{mj}$ that minimizes the loss function can be determined through linear search, which is articulated as:

$$c_{mj} = arg\ \min_{c} \sum_{xi \in R_{mj}} L(y_i, f_{m-1}(x_i) + c) \tag{8}$$

where $c$ is a constant value that minimizes the loss function.

At this juncture, the decision tree model for the m-th iteration is described as:

$$h_m(x) = \sum_{j=1}^{J} c_{mj} I(x_i \in R_{mj}) \tag{9}$$

where $I(x_i \in R_{mj})$ is an indicator function that equals 1 if $x_i$ falls within the region $R_{mj}$, and 0 otherwise.

The model has thus been updated to become:

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J} c_{mj} I(x_i \in R_{mj}) \tag{10}$$

By aggregating the initial decision tree with the decision trees from each iteration, the final outcome is derived as:

$$F(x) = f_0(x) + \sum_{m=1}^{M} \sum_{j=1}^{J} c_{mj} I(x_i \in R_{mj}) \tag{11}$$

Expanding on GBDT, LGBM integrates innovative techniques like Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), which capitalize on data insights to boost training efficiency significantly. Throughout the decision tree fitting phase, the Light Gradient Boosting Machine transforms continuous floating-point feature values into $q$ integers, creating a histogram of width $q$. This approach facilitates the pinpointing of the most advantageous splitting point. During node splitting, LGBM employs a leaf-wise growth strategy, opting to split the leaf that offers the maximum split gain. This method not only shortens the search duration but also improves the model's accuracy.

### 2.3 TPE Hyperparameter Optimization

In 2011, James Bergstra and colleagues [26] from Harvard University introduced the Tree-structured Parzen Estimator (TPE), employing a tree structure to delineate the interrelations among hyperparameters. This method excels in addressing multi-dimensional optimization challenges and, crucially, is capable of identifying satisfactory hyperparameters within a relatively limited number of iterations. Consequently, TPE emerges as an optimal strategy for hyperparameter tuning within the HEFS-LGBM framework. This technique is versatile, accommodating various objective functions and enhancing the caliber of hyperparameter selection.

The TPE algorithm delineates $p(x|y)$ through the utilization of two density functions:

$$p(x|y) = \begin{cases} l(x) & if \ y < y^* \\ g(x) & if \ y \geq y^* \end{cases} \tag{12}$$

In this framework, $x$ signifies the observation point, namely, the hyperparameter vector of the model undergoing optimization; $y$ denotes the observation value, that is, the outcome of the objective function (either the loss function or evaluation function) for the given parameters $x$; $y^*$ represents a threshold, corresponding to a specific percentile within the TPE algorithm, employed to segregate $l(x)$ and $g(x)$, with its value spanning the interval (0,1). The TPE algorithm ascertains $y^*$ based on the accumulated observation points, subsequently categorizing these points into two density functions:$l(x)$, the density function constituted by the observation value $\{x_{(i)}\}$ with loss function $f(x_{(i)}) < y^*$, and $g(x)$, the density function composed of the residual observation values. By leveraging the principles of Bayesian optimization, TPE aims to minimize the scope of ineffective search areas.

The TPE algorithm adopts the Expected Improvement (EI) strategy, generating new observation points by maximizing the EI value. Utilizing Bayesian techniques to optimize the acquisition function EI, the TPE algorithm aligns $p(y)p(x|y)$ with $p(x, y)$. Consequently, the acquisition function EI is defined as follows:

$$EI_{y^*}(x) = \int_{-\infty}^{y^*} (y^* - y)p(y|x)dy = \int_{-\infty}^{y^*} (y^* - y)\frac{p(x|y)p(y)}{p(x)}dy \tag{13}$$

Setting $\gamma = p(y < y^*)$ to simplify the equation above, we construct $p(x) = \int p(x|y)p(y)dy = \gamma l(x) + (1-\gamma)g(x)$ for the denominator.

For the numerator, we can derive the following:

$$\begin{aligned} &\int_{-\infty}^{y^*} (y^* - y)p(x|y)p(y)dy = l(x) \\ &\int_{-\infty}^{y^*} (y^* - y)p(y)dy = \gamma y^* l(x) - l(x)\int_{-\infty}^{y^*} p(y)dy \end{aligned} \tag{14}$$

Ultimately, EI can be simplified to the following expression:

$$EI_{y^*}(x) = \frac{\gamma y^* l(x) - l(x)\int_{-\infty}^{y^*} p(y)dy}{\gamma l(x) + (1-\gamma)g(x)} \propto \left(\gamma + \frac{g(x)}{l(x)}(1-\gamma)\right)^{-1} \tag{15}$$

From the equation, it becomes clear that to maximize EI, it is preferable to have a high probability of $l(x)$ under the hyperparameter $x$ while maintaining a low probability for $g(x)$. In each iteration, the algorithm identifies and returns the candidate hyperparameter $x^*$ that exhibits the highest EI value, denoted as $x^* = argmaxEI_{y^*}$. This implies that the maximum EI value is achieved when the hyperparameter $x$ attains the highest probability $l(x)$ and the lowest probability $g(x)$. The TPE optimization algorithm leverages the ratio of $l(x)$ to $g(x)$ to generate hyperparameter samples. Within the context of hyperparameter optimization for the HEFS-LGBM model, this process entails utilizing the newly proposed hyperparameter $x$ to fine-tune the parameters of the HEFS-LGBM diagnostic model. After training, the observation $y$ is obtained, and this new sample observation is compared with the existing observations to update the probability model. Consequently, this iterative process facilitates the selection of the optimal hyperparameter $x$ that correlates with the most favorable outcome.

K-fold cross-validation partitions the dataset into $K$ equally sized subsets, with one subset reserved for evaluating the model's performance and the remaining K-1 subsets utilized for training. Each subset sequentially serves as the validation set, culminating in $K$ evaluations. This study adopts the StratifiedKFold technique for stratified sampling cross-validation, which ensures the sample proportion within each fold mirrors that of the original dataset [27]. This method maximizes data usage and mitigates the risk of overfitting. In this research, $K$ is established at 5, indicating the implementation of five-fold cross-validation. The dataset is randomly segmented into 5 equal sections, where 4 folds are allocated for training and onefold for validation during each cycle. Following 5 rounds of validation, the mean of these 5 validation outcomes is computed to represent the cross-validation result.

In this study, the accuracy derived from five-fold cross-validation serves as the objective function. The TPE optimization method is utilized to identify the set of parameters that maximize accuracy within the five-fold cross-validation framework, deeming these as the optimal hyperparameters. The TPE algorithm concurrently searches for the optimal hyperparameters for both HEFS and LGBM. The determined HEFS parameters are subsequently applied to the test set for feature selection, while the hyperparameters for LGBM are integrated into the LGBM model for classification and diagnosis. The structural framework of the HEFS-LGBM model for multi-class fault diagnosis is depicted in Fig. 2.

The entire diagnostic process, from feature selection to hyperparameter optimization and model training, is fully automated. This automation is achieved through the integration of the HEFS, LGBM, and TPE algorithms within a unified framework. The automation process includes the following steps:

(1) Automated Feature Selection: The HEFS algorithm automatically selects the most relevant features from the

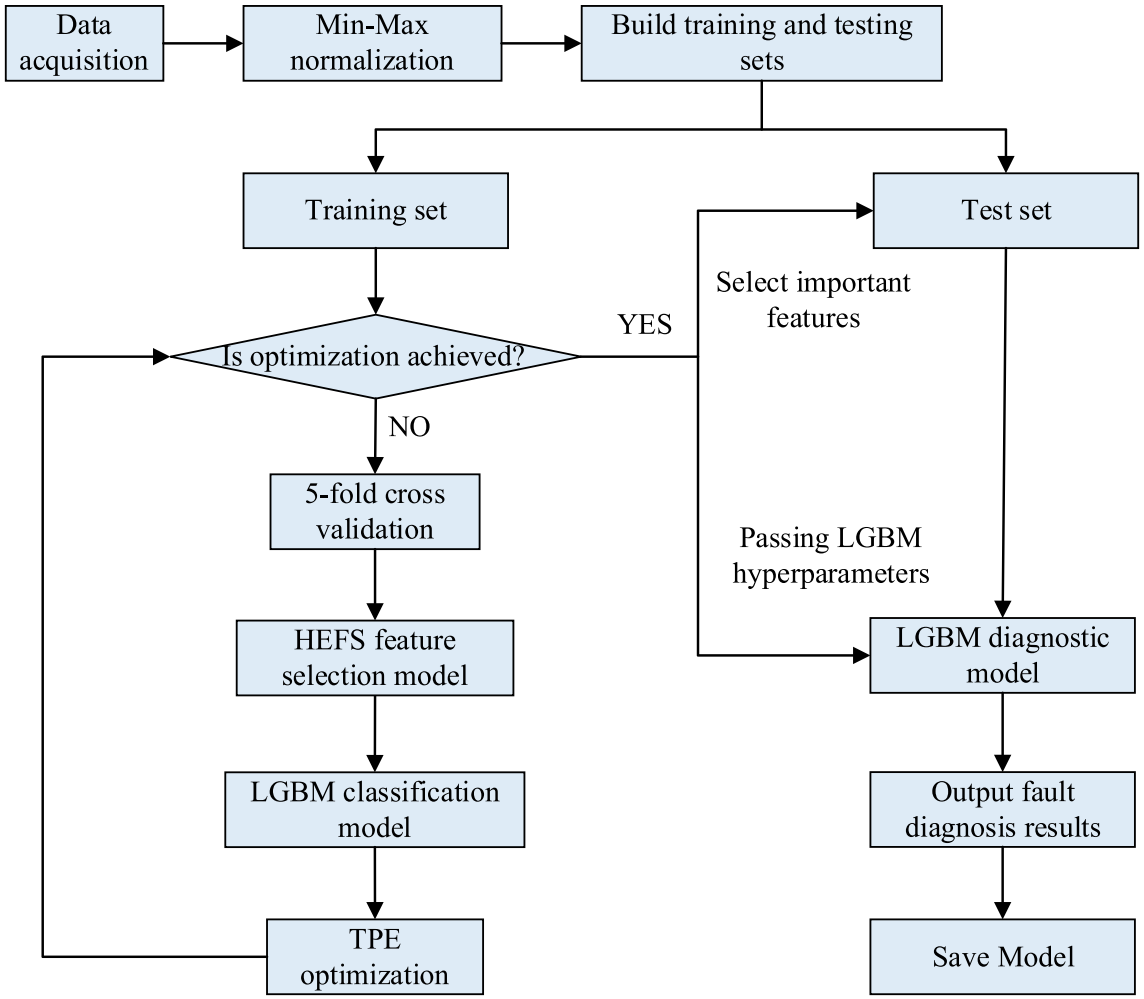



**Fig. 2** HEFS-LGBM model framework

high-dimensional dataset, eliminating redundant and irrelevant features without manual intervention.
(2) Automated Hyperparameter Optimization: The TPE algorithm automatically tunes the hyperparameters of both the HEFS and LGBM models, ensuring optimal performance.
(3) Automated Model Training and Evaluation: The LGBM model is trained and evaluated using the selected features and optimized hyperparameters, with the entire process being automated to ensure consistency and efficiency.

This automation significantly enhances the efficiency and reliability of the fault diagnosis system, reducing the need for manual intervention and ensuring consistent and reproducible results. The automated process is particularly beneficial for handling large datasets and complex feature spaces, making the HEFS-LGBM method highly suitable for practical applications in analog circuit fault diagnosis.

## 3 Framework for Analog Circuit Fault Diagnosis and Evaluation Metrics

### 3.1 Implementation Framework for Analog Circuit Fault Diagnosis

Analog circuit faults are typically classified into two main types: hard faults and soft faults. Hard faults involve alterations to the circuit topology that inhibit its fundamental operations, such as open or short circuits. Soft faults arise when the parameter values of circuit components deviate beyond their specified tolerance ranges, resulting in diminished circuit performance. Given that the fault feature selection method and fault mode classifier training in the proposed model necessitates a substantial number of component fault samples that accurately represent actual fault conditions, and considering that naturally occurring fault samples in analog circuits are insufficient in quantity, this paper suggests employing Electronic Design Automation

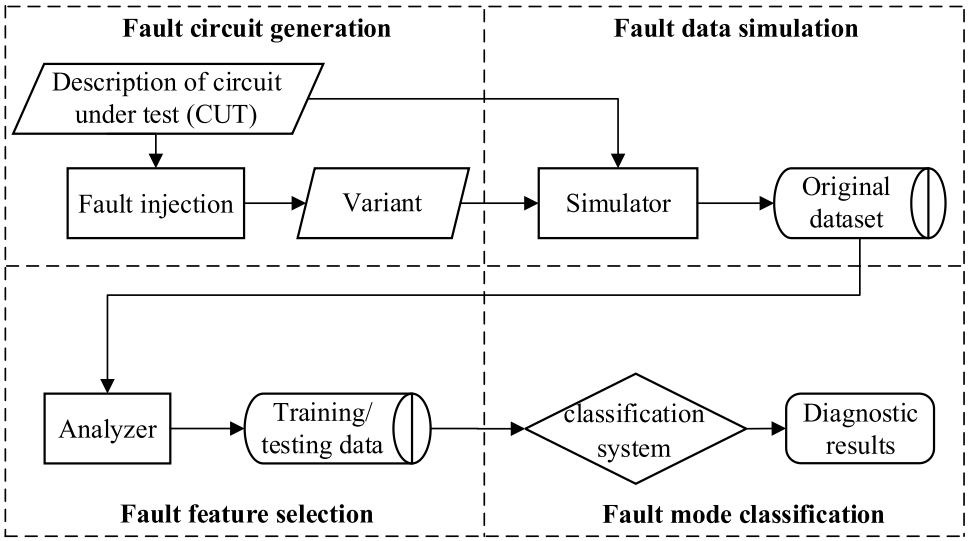



**Fig. 3** Framework for implementing analog circuit fault diagnosis

**Table 1** Forms of implementation for mutation operators

| Operator | Name | Implementation Form |
|---|---|---|
| PCH + /PCH- | Positive/Negative Parameter Change | Alters a component's parameter to exceed or fall below its tolerance range |
| ROP | Resistor Open | Connects a high-value resistor across a component's terminals |
| LRB | Local Resistor Bridging | Connects a low-value resistor across the terminals of the same component |
| GRB | Global Resistor Bridging | Connects a low-value resistor between nodes of different components |
| NSP | Node Splitting | Divides a node into two and connects a high-value resistor at the split node |

(EDA) technology for fault injection in analog circuits. Fault injection serves as a method to acquire fault samples for electronic devices. Tang Xiaofeng and colleagues [28] have introduced a mutation generation technique for the automatic injection of analog circuit component fault samples that emulate real faults and have developed the Electronic Fault Analyzer (EFA) software. This study utilizes the EFA software created by them as the fault sample generation tool.

The framework for analog circuit fault diagnosis implementation is depicted in Fig. 3.

#### 3.1.1 (1) Fault Circuit Generation

Initially, the basic information of the Circuit Under Test (CUT) is inputted, followed by the injection of mutation operators into the CUT to generate the corresponding fault circuits. The implementation details of these mutation operators are outlined in Table 1. Within this context:

PCH signifies a soft fault mutation operator, where the component parameter values are either too high (PCH + ) or too low (PCH-), adhering to uniform distributions U($\Theta + 2\varepsilon$, $2\Theta$) and U($0.1\Theta$, $\Theta$-$2\varepsilon$) respectively, with $\Theta$ and $\varepsilon$ denoting the component's nominal value and tolerance value. ROP, LRB, GRB, and NSP are indicative of hard fault mutation operators. The mutation operation by ROP and NSP involves the connection of resistors with resistance values distributed uniformly according to U(100 kΩ, 100 MΩ). Conversely, the mutation operation by LRB and GRB involves the connection of resistors with resistance values distributed uniformly according to U(10 Ω, 1 kΩ).

**Table 2** Confusion matrix for binary classification outcomes

| Actual Condition | Classification Result | |
|---|---|---|
| | Normal | Abnormal |
| Normal | TP (True Positive) | FN (False Negative) |
| Abnormal | FP (False Positive) | TN (True Negative) |

#### 3.1.2 (2) Fault Data Simulation

Upon the creation of mutants for different fault circuit types, the circuit schematics for both the fault-free state and each mutant variant are submitted to a simulation platform equipped with a Pspice core. During the simulations, a sweep signal is employed as the stimulus, spanning the full operational frequency range of the circuit under examination. To facilitate feature selection, K frequency points with equal interval and discrete distribution in the swept signal

are extracted as the feature selection range. The parameters for each component are configured to conform to a Gaussian distribution, centered around the nominal value $\Theta$ with a standard deviation of $\sigma = \varepsilon\Theta/3$, and a set relative tolerance $\varepsilon$ of 10%. Subsequently, Monte Carlo simulations are conducted across the spectrum of circuit states, thus acquiring characteristic data for the circuit in both its faultless state and under various fault conditions.

#### 3.1.3 (3) Fault Feature Selection

Use the heterogeneous ensemble feature selection method proposed in Sect. 2.1 to perform feature selection on the circuit response data corresponding to the K different frequency points, remove redundant features, and generate low-dimensional training samples.

#### 3.1.4 (4) Fault Mode Classification

The LGBM diagnostic model, as detailed in Sect. 2.2, is employed for the classification of fault modes in the analog circuit.

### 3.2 Metrics for Evaluating Fault Diagnosis Performance

Table 2 presents the confusion matrix applicable to binary classification scenarios.

For a multi-class classification problem with $p$ classes, the definitions are slightly modified. Accuracy is defined as:

$$Accuracy = \frac{\sum_{i=1}^{p}(TP_i + TN_i)}{\sum_{i=1}^{p}(TP_i + FP_i + TN_i + FN_i)} \tag{16}$$

Recall is defined as:

$$Recall = \frac{1}{p}\sum_{i=1}^{p}\frac{TP_i}{TP_i + FN_i} \tag{17}$$

The F1 score is defined as:

$$F1 = \frac{1}{p}\sum_{i=1}^{p}\frac{2 \cdot TP_i}{2 \cdot TP_i + FP_i + FN_i} \tag{18}$$

Both Recall and F1 score employ macro averaging, which considers each class with equal importance. Utilizing these three evaluation metrics in tandem provides a thorough assessment of the diagnostic performance of the analog circuit.

## 4 Experiment and Discussion

The HEFS-LGBM approach introduced in this study was initially evaluated for its efficacy using the University of California Irvine (UCI) dataset, followed by diagnostic tests on analog circuits. It was benchmarked against two traditional feature selection and classification methods, SelectKBest with Support Vector Machine (SKB-SVM) and Recursive Feature Elimination with Logistic Regression (RFE-LR), as well as two homogeneous ensemble feature selection and classification methods, RandomForest with XGBoost (RF-XGB) and XGBoost with CatBoost (XGB-CAB), and the LGBM classification algorithm utilizing all features. These methods are denoted as SKB-SVM, RFE-LR, RF-XGB, XGB-CAB, and LGBM, respectively. In the experiments, all methods employed the TPE optimization algorithm to determine the optimal hyperparameters, setting the feature selection proportion optimization range between [0.05, 0.9]. The SKB-SVM and RFE-LR methods were executed using the scikit-learn library in Python, RF-XGB was implemented with scikit-learn and xgboost libraries, and XGB-CAB was realized using xgboost and catboost libraries. The TPE optimization algorithm was implemented via the Optuna hyperparameter tuning framework, introduced by Takuji Akiba et al. [29] from the Japanese company PFN in 2019. Furthermore, the base classifiers in HEFS were implemented using the scikit-learn library, with ensemble decision tree methods such as Bagging, Extra Trees, GBDT, and Random Forest all configured to use 10 trees, and the SGD and SVM methods employing a linear kernel.

**Table 3** UCI datasets overview

| Dataset Name | Sample Count | Class Count | Feature Count | Class Distribution |
|---|---|---|---|---|
| Arrhythmia | 420 | 2 | 278 | 237,183 |
| Colon | 62 | 2 | 1908 | 22,40 |
| Multiple-features | 1994 | 10 | 649 | 200,199,200,200,199,200,199,198,199,200 |
| Semeion | 1593 | 10 | 256 | 161,158,162,159,159,161,159,161,158,155 |
| Sonar | 208 | 2 | 60 | 111,97 |

**Table 4** Experimental results of UCI datasets

| Dataset | Algorithm | Accuracy | Recall | F1 Score |
|---|---|---|---|---|
| Arrhythmia | SKB-SVM | 0.797 6 | 0.783 0 | 0.786 2 |
| | RFE-LR | 0.785 7 | 0.763 5 | 0.769 9 |
| | RF-XGB | 0.809 5 | 0.802 4 | 0.802 4 |
| | XGB-CAB | 0.773 8 | 0.758 2 | 0.761 6 |
| | LGBM | 0.773 8 | 0.758 2 | 0.761 6 |
| | HEFS-LGBM | **0.821 4** | **0.807 6** | **0.811 8** |
| Colon | SKB-SVM | **0.923 1** | **0.944 4** | **0.915 0** |
| | RFE-LR | 0.846 2 | 0.819 4 | 0.819 4 |
| | RF-XGB | 0.884 7 | 0.8819 | 0.867 2 |
| | XGB-CAB | **0.923 1** | 0.875 0 | 0.902 3 |
| | LGBM | 0.769 2 | 0.694 4 | 0.706 8 |
| | HEFS-LGBM | 0.903 9 | 0.878 5 | 0.884 8 |
| Multiple-features | SKB-SVM | 0.987 5 | 0.986 9 | 0.987 4 |
| | RFE-LR | 0.987 5 | 0.986 7 | 0.987 4 |
| | RF-XGB | 0.982 5 | 0.981 9 | 0.981 8 |
| | XGB-CAB | 0.985 0 | 0.984 7 | 0.985 2 |
| | LGBM | 0.985 0 | 0.983 9 | 0.984 8 |
| | HEFS-LGBM | **0.990 0** | **0.989 2** | **0.989 6** |
| Semeion | SKB-SVM | 0.943 6 | 0.942 3 | 0.943 1 |
| | RFE-LR | 0.921 6 | 0.922 9 | 0.922 3 |
| | RF-XGB | 0.893 4 | 0.894 7 | 0.892 9 |
| | XGB-CAB | 0.940 4 | 0.941 4 | 0.939 9 |
| | LGBM | 0.937 3 | 0.937 4 | 0.936 2 |
| | HEFS-LGBM | **0.946 7** | **0.946 9** | **0.945 4** |
| Sonar | SKB-SVM | 0.690 5 | 0.708 3 | 0.690 3 |
| | RFE-LR | 0.714 3 | 0.729 2 | 0.714 3 |
| | RF-XGB | 0.785 7 | 0.805 6 | 0.785 6 |
| | MI-XGB | 0.857 1 | 0.875 0 | 0.857 1 |
| | LGBM | **0.904 8** | **0.916 7** | **0.904 5** |
| | HEFS-LGBM | 0.881 0 | 0.888 9 | 0.880 3 |

The experimental setup was equipped with an Intel Core i9-13900HX processor, featuring a base frequency of 2.20 GHz and a maximum turbo frequency of 5.60 GHz, complemented by 32 GB of RAM. On the software front, the system ran on the Windows 11 operating system and utilized the PyCharm 2023.1 integrated development environment, with Python 3.8.5 as the programming language.

### 4.1 Experiments on UCI Datasets

The UCI dataset is a widely recognized public dataset commonly employed for benchmarking algorithm performance in machine learning. To evaluate the efficacy of the proposed algorithm, five UCI datasets characterized by relatively high feature dimensions—namely, Arrhythmia, Colon, Multiple-features, Semeion, and Sonar—were chosen for experimental analysis. Details pertaining to these five datasets are summarized in Table 3.

Randomly allocating 80% of the data for training set samples and the remaining 20% for test set samples, and to mitigate the influence of algorithmic randomness, each method was subjected to 10 trials on the 5 UCI datasets. The average of these 10 experimental outcomes was recorded as the definitive result. The results garnered from these experiments are compiled in Table 4, with the optimal values for each metric emphasized in bold within the table.

As observed in Table 4, the HEFS-LGBM method outperformed in terms of accuracy, recall, and F1 score on three datasets: Arrhythmia, Multiple-features, and Semeion. It secured the third position in average metrics on the Colon dataset and was surpassed only by LGBM on the Sonar dataset. This performance underscores the efficacy of integrating the HEFS ensemble feature selection method, adept at pinpointing significant features, with the LGBM ensemble learning method, recognized for its robust pattern

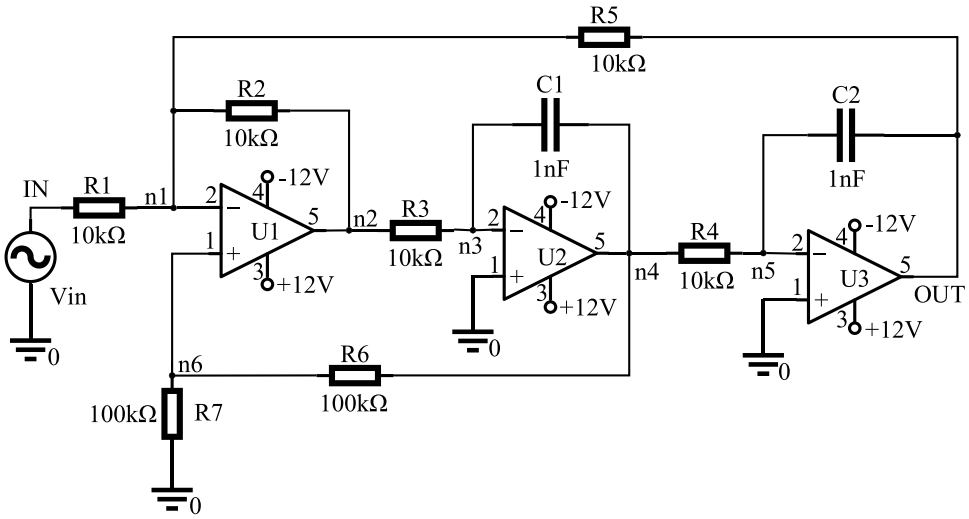



**Fig. 4** Schematic diagram of the Biquad low-pass filter circuit

**Table 5** Fault states of the Biquad low-pass filter circuit

| No | ID | State Description |
|---|---|---|
| 1 | F0 | Fault-free |
| 2 | F1 | PCH (C1↑) |
| 3 | F4 | PCH (R2↓) |
| 4 | F7 | ROP (C1, +) |
| 5 | F14 | LRB (C2, +, -) |
| 6 | F26 | GRB (n6, OUT) |
| 7 | F28 | NSP (n1, [2,4]–[1.3]) |

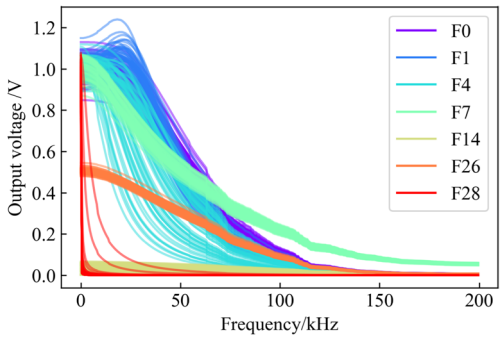



**Fig. 5** Feature relationship graph for each class of samples

recognition capabilities, in enhancing classification outcomes for datasets characterized by high dimensionality or feature redundancy. The proposed method's commendable results across multiple datasets underscore the versatility of the HEFS-LGBM approach. Notably, on the Sonar dataset, all feature selection classification methods fell short compared to the LGBM method without feature selection. This implies that feature selection does not universally enhance accuracy for high-dimensional data. Typically, datasets with a significant amount of redundant and irrelevant features benefit from feature selection. Conversely, if the dataset comprises solely relevant and beneficial features, feature selection might inadvertently diminish classification accuracy. Hence, the presence of redundant or irrelevant features should be a key consideration when performing feature selection.

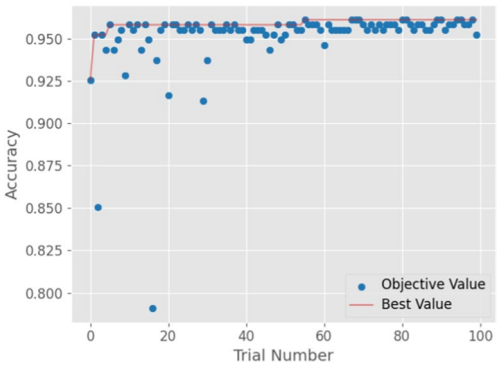



**Fig. 6** Variation of the TPE fitness function with the number of iterations

## 4.2 Biquad Low-pass Filter Circuit Fault Diagnosis Experiment

Given that the proposed fault diagnosis method is based on machine learning techniques, it possesses a high degree of universality and adaptability. This method is applicable across various types of circuits, not limited by the specific characteristics or configurations of the circuits involved. The core strength of this approach lies in its ability to extract critical information from circuit responses and to identify abnormal patterns effectively.Therefore, this section conducts experiments using the widely utilized Biquad low-pass filter circuit, which has a relatively large scale and includes both soft and hard faults. The configuration of the Biquad low-pass filter circuit is depicted in Fig. 4. In the figure, U1 to U3 denote integrated operational amplifiers; IN is the input port; n1 to n6 are the interconnection nodes; and the numbers 1 to 5 indicate the pins of the operational amplifiers.

The Biquad low-pass filter circuit was subjected to a range of fault injections, encompassing both soft and hard fault types. For soft faults, the mutation operators PCH↑ (Positive Change) and PCH↓ (Negative Change) were used to introduce faults with component parameters distributed according to uniform distributions U($\Theta+2\varepsilon$, $2\Theta$) and

**Table 6** Hyperparameters optimized by TPE algorithm

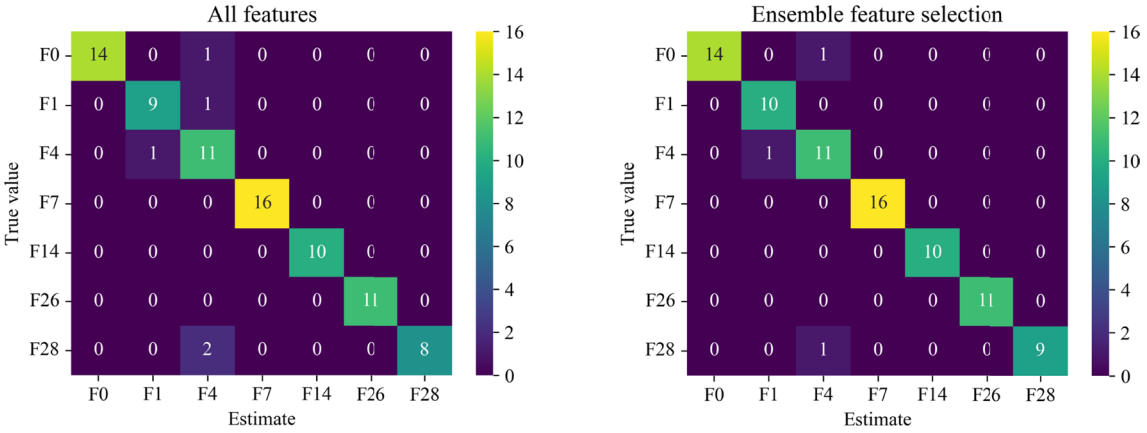
| Parameter Category | Model Parameter | Description of Parameter | Range of Parameter | Type of Parameter Value | Optimal Value of Hyperparameter |
|---|---|---|---|---|---|
| HEFS Parameters | featureselection_ratio | Ratio for Feature Selection | (0.05,0.9) | Float | 0.08 |
| LGBM Parameters | lambda_l1 | L1 Regularization Penalty Coefficient | (0,10) | Float | 0.106 3 |
| | lambda_l2 | L2 Regularization Penalty Coefficient | (0,10) | Float | 0.215 6 |
| | num_leaves | Number of Leaf Nodes | (3,128) | Integer | 17 |
| | learning_rate | Learning Rate | (0,1) | Float | 0.124 9 |

U(0.1$\Theta$, $\Theta$-2$\varepsilon$), respectively. In the case of hard faults, the mutation operators ROP (Resistor Open) and NSP (Node Splitting) were employed to inject faults with parameters uniformly distributed within U(100 kΩ, 100 MΩ), whereas LRB (Local Resistor Bridging) and GRB (Global Resistor Bridging) introduced faults with parameters following U(10 Ω, 1 kΩ). The specific soft and hard faults introduced are itemized in Table 5.

The diverse fault mutants are fed into the fault simulator, with each category producing 60 distinct mutants. A sinusoidal sweep signal, comprising 350 uniformly distributed frequency points spanning the range [1 Hz, 200 kHz], serves as the stimulus. This process yields an initial dataset of 420 samples, each characterized by 350-dimensional feature vectors. Figure 5 illustrates the correlation between frequency and output voltage for each class of samples.

Randomly allocating 80% of the data to the training sample set and reserving the remaining 20% for the test sample set, the TPE optimization algorithm—guided by the Optuna hyperparameter tuning framework—and stratified fivefold cross-validation are jointly employed to fine-tune the parameters of the HEFS ensemble feature selection algorithm and the LGBM pattern recognition method. HEFS has one parameter subject to tuning, whereas LGBM has four parameters that necessitate optimization. The optimal hyperparameters determined through this optimization process are detailed in Table 6.

Figure 6 shows the change in the fitness function (accuracy in this case) with the number of iterations when using the TPE optimization algorithm for hyperparameter optimization. Each blue dot in the figure represents the result of a trial, i.e., the accuracy obtained by evaluating the hyperparameter combination in that trial. The red line represents the change in the global best value (Best Value) with the number of trials, showing the best accuracy found after each trial. In the initial stage of the trials (the first few trials), the red line rises rapidly. This indicates that in the initial stage, the algorithm quickly found better hyperparameter combinations than before, significantly improving the model's accuracy. In the middle stage, the distribution of blue dots gradually becomes concentrated, indicating that the algorithm is gradually converging and finding some relatively stable hyperparameter combinations. The red line gradually levels off, indicating that the best accuracy found has not significantly improved. In the later stage, the red line remains basically unchanged,




(**a**) Result of all features

(**b**) Result of ensemble features selection

**Fig. 7** Confusion matrix comparison of diagnostic results

**Table 7** Comparative experimental results for the Biquad low-pass filter circuit

| Algorithm | Accuracy (Mean ± SD) | Recall (Mean ± SD) | F1 (Mean ± SD) | Training time/s | Test time/s |
|---|---|---|---|---|---|
| SKB-SVM | 0.946 5 ± 0.055 1 | 0.947 0 ± 0.051 1 | 0.947 4 ± 0.054 5 | 0.988 1 | 0.000 6 |
| RFE-LR | 0.949 5 ± **0.012 4** | 0.948 5 ± **0.011 0** | 0.948 9 ± **0.011 8** | 171.123 3 | 0.000 1 |
| RF-XGB | 0.940 5 ± 0.022 7 | 0.931 0 ± 0.023 4 | 0.935 1 ± 0.024 2 | 139.800 3 | 0.002 0 |
| XGB-CAB | 0.952 4 ± 0.023 8 | 0.952 1 ± 0.022 6 | 0.953 0 ± 0.024 0 | 166.541 3 | 0.004 5 |
| LGBM | 0.940 5 ± 0.015 1 | 0.935 7 ± 0.014 7 | 0.938 5 ± 0.014 6 | 58.847 2 | 0.001 1 |
| HEFS-LGBM | **0.964 3** ± 0.020 1 | **0.964 3** ± 0.017 0 | **0.963 6** ± 0.017 6 | 232.143 7 | 0.001 0 |

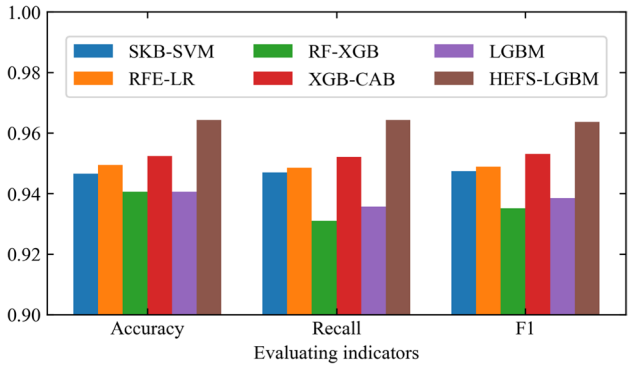



**Fig. 8** Average performance metrics of the Biquad low-pass filter circuit experiment

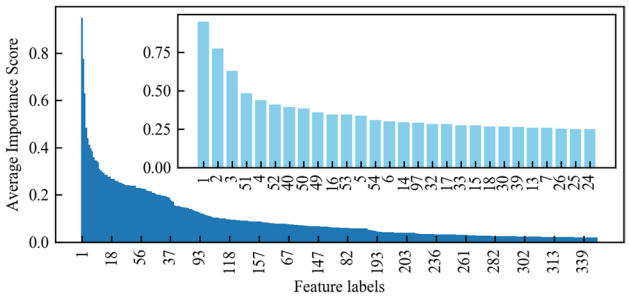



**Fig. 9** Average importance scores of each feature label

indicating that the algorithm has almost converged and found a relatively stable best hyperparameter combination. The fluctuation of the blue dots is small, indicating that the results of most trials are close to the global best value. Overall, the TPE algorithm quickly found good hyperparameter combinations in the initial stage and gradually converged in the middle and later stages, finding stable best hyperparameter combinations.

The optimization yielded a feature selection ratio of 0.08 for HEFS, indicating that the top $\lceil 350*0.08 \rceil = 28$ features ranked by importance should be retained. Figure 7 compares the confusion matrices of test results obtained using all features against those using the top 28 features identified by the ensemble feature selection. It is evident that the prediction error count when utilizing all features stands at 5, whereas it drops to 3 when applying the ensemble feature selection. This reduction in prediction errors underscores the efficacy of the HEFS-LGBM method. Moreover, there was a substantial decrease in the number of features post-ensemble feature selection. By archiving the most critical selected features offline, future diagnostics can be expedited using these features, thereby obviating the need for re-extracting a vast array of features and significantly cutting down on labor costs.

Additional comparative experiments were carried out between the HEFS-LGBM method and alternative approaches, including SKB-SVM, RFE-LR, RF-XGB, MI-XGB, and LGBM. Each method underwent 10 iterations to ascertain the mean values and standard deviations. The experimental findings are compiled in Table 7, with the top-performing values accentuated in bold. A graphical depiction of these average values is provided in Fig. 9.

From the experimental results presented in Table 7 and Fig. 8, the HEFS-LGBM method achieved the best performance in the Biquad low-pass filter circuit fault diagnosis experiment, with average accuracy, recall, and F1 scores of 96.43%, 96.43%, and 96.36%, respectively, outperforming other models. Although the RFE-LR model showed the best results in terms of standard deviation, the standard deviation of the HEFS-LGBM model was not significantly different. In fault diagnosis, accuracy is typically prioritized over standard deviation, making the HEFS-LGBM model more advantageous. The experiment demonstrates that the HEFS-LGBM model can effectively meet the requirements for fault diagnosis in analog circuits. In terms of runtime, the HEFS-LGBM model has the longest training duration among all listed algorithms, clocking in at 232.1437 s. This extended training time is attributed to the model's implementation of a more complex feature selection strategy, which demands substantial computational resources and time. Despite the lengthy training process, the HEFS-LGBM model excels in testing speed, taking only 0.0010 s to make predictions on new data. This indicates that once the model is trained, it can quickly respond to new inputs. While the longer training time of the HEFS-LGBM model may be seen as a limitation, it is important to consider the context in which the model will be used. In many practical applications, the training phase is performed offline, and the primary concern is the model's performance during the testing phase. The HEFS-LGBM model's exceptional accuracy and rapid prediction capabilities justify the investment in training time. Moreover, the improved diagnostic performance can lead to more reliable and efficient fault detection, which is critical in high-stakes environments such as analog circuit applications.

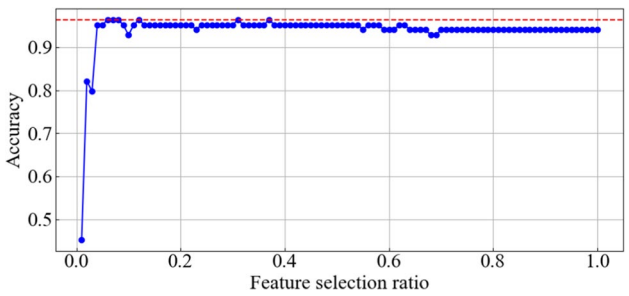



**Fig. 10** Variation of Accuracy with Feature Selection Ratio

To investigate the significance of features within the test samples, the average importance scores for each feature were charted, as depicted in Fig. 9. The horizontal axis, labeled 1–350, corresponds to the discretized feature labels from the Biquad test data, while the vertical axis displays the average importance scores for each feature label, as determined by the HEFS method. Figure 9 primarily illustrates the importance scores for 350 global discretized feature labels of the Biquad analog circuit, as calculated using the HEFS approach. Given the figure's limited display space, feature labels are marked at 20-unit intervals. The upper right section of Fig. 9 highlights the average feature importance scores for the 28 feature labels selected by HEFS, indicating that, among all 350 features, the lower-ranked feature labels exhibit significantly low average importance scores. In contrast, the 28 selected feature labels all have average importance scores exceeding 0.25. This suggests that the lower-ranked feature labels, potentially serving as redundant features, could disrupt sample classification, underscoring the value of implementing ensemble feature selection.

To further assess the impact of the feature selection ratio on the HEFS method, with all other parameters held constant post-optimization, the feature selection ratio was incrementally adjusted from 0.01 to 1, with increments of 0.01. A graph illustrating the variation in accuracy relative to the feature selection ratio was plotted, as shown in Fig. 10. The curve predominantly exhibits a pattern of rapid increase, followed by a gradual decline and eventual stabilization. This observation reinforces the notion that applying ensemble feature selection to samples characterized by high dimensionality and redundant features significantly enhances classification accuracy. Furthermore, the feature selection ratio of 0.08, as optimized by the TPE algorithm, falls within the high-performance range for ensemble feature selection, highlighting the TPE algorithm's remarkable efficiency in optimization.

## 5 Conclusion and Future Work

This study tackles the challenge of low diagnostic accuracy in analog circuits characterized by high-dimensional features by introducing a fault diagnosis model that leverages heterogeneous ensemble feature selection combined with Light Gradient Boosting Machine (LGBM). To validate the efficacy of the proposed model, experiments were conducted using both UCI datasets and analog circuits. The findings from these experiments indicate that:

(1) The method proposed in this study is capable of effectively diagnosing faults in analog circuits, offering a viable approach for fault diagnosis in systems with high-dimensional features and complex fault patterns.

(2) The strategic integration of ensemble feature selection with pattern recognition techniques plays a crucial role in minimizing the number of features extracted from analog circuits, thereby reducing labor costs.

While the results are promising, several areas for future research and improvement have been identified:

(1) Accuracy and Impact of Fault Model Deviations. The accuracy of the fault model presented in Table 1 is critical for the reliability of our fault diagnosis method. Small deviations in the fault model parameters may impact the overall diagnostic flow, potentially introducing noise and reducing precision. Discrepancies between simulated and actual faults could affect the generalizability of our results. Future research should focus on improving the accuracy of fault models to closely mimic real-world conditions, investigating the robustness of our method to parameter deviations. We plan to verify this through simulation experiments in the next step. This will ensure that our HEFS-LGBM method remains effective and reliable under various fault conditions.
(2) Application in More Complex Analog Circuits. Our current study primarily focuses on a relatively simple Biquad LP filter and assumes that the operational amplifiers (OPAMPs) are ideal. This choice was made to clearly demonstrate the fundamental capabilities and advantages of the HEFS-LGBM method in a controlled environment. Applying our technique to more complex analog circuits, which may include multiple stages, non-ideal components, and interactions between various elements, would provide a more challenging and informative testbed for our method. In such scenarios, the heterogeneity and robustness of the HEFS approach could be rigorously evaluated, especially in cases where faults occur within more integral components like the OPAMPs themselves.
(3) Consideration of Process, Voltage, and Temperature (PVT) Effects. In the current scope of our study, we primarily focused on the development and validation of the HEFS-LGBM method under controlled conditions, and thus, PVT variations were not explicitly modeled in our simulations. Addressing PVT effects would undoubtedly enhance the robustness and applicability of our method. In future work, we plan to incorporate PVT variations into our simulation environment to test and adapt our method under these more realistic and challenging conditions. This will involve adjusting our feature selection and classification algorithms to account for the variability and uncertainty introduced by PVT effects, potentially through techniques such as robust feature selection, enhanced data augmentation, or the development of PVT-aware models.

**Acknowledgements** We thank the authors of Scikit-learn, LightGBM and Optuna open-source code. Our code is based on their open-source code for improvement.

**Funding** This research was funded by the Mount Taishan Scholar Construction Project in Shandong Province, China, grant number: tstp20221146.

**Data Availability** The UCI dataset used in this study are openly available in the database at https://archive.ics.uci.edu/datasets (accessed on 1 April 2024).

## Declarations

**Competing Interests** The authors declare no competing interest.

## References

1. Zhang CL, He YG, Yuan LF, Xiang S (2018) Analog circuit incipient fault diagnosis method using DBN based features extraction. IEEE Access 6(5):23053–23064. https://doi.org/10.1109/ACCESS.2018.2823765
2. Wang SD, Liu ZB, Jia Z, Li ZH (2023) Incipient fault diagnosis of analog circuit with ensemble HKELM based on fused multi-channel and multi-scale features. Eng Appl Artif Intell 117:105633. https://doi.org/10.1016/j.engappai.2022.105633
3. Huang K, Stratigopoulos HG, Mir S, Hora C, Xing YZ, Kruseman B (2012) Diagnosis of local spot defects in analog circuits. IEEE Trans Instrum Meas 61(10):2701–2712. https://doi.org/10.1109/TIM.2012.2196390
4. Pavlidis A, Faehn E, Louërat MM, Stratigopoulos HG (2021) BIST-assisted analog fault diagnosis. Proc. of 2021 IEEE European Test Symposium (ETS). IEEE 2021:1–6. https://doi.org/10.1109/ETS50041.2021.9465386
5. Melis T, Simeu E, Auvray E, Saury L (2023) Light Emission Tracking and Measurements for Analog Circuits Fault Diagnosis in Automotive Applications. J Electron Test 39(2):171–187. https://doi.org/10.1007/s10836-023-06059-6
6. Liang H, Zhu YM, Zhang DY, Chang L, Lu YM, Zhao XF, Guo Y (2021) Analog circuit fault diagnosis based on support vector machine classifier and fuzzy feature selection. Electronics 10(12):1496. https://doi.org/10.3390/electronics10121496
7. Naidu SV, Mullapudi C, Patil HY (2021) Early Diabetes Detection Using Combination Polynomial Features and SelectKBest Classifier. SPAST Abstracts 1(01).
8. Senan EM, Al-Adhaileh MH, Alsaade FW, Aldhyani THH, Alqarni AA, Alsharif N, Uddin MI, Alahmadi AH, Jadhav ME, Alzahrani MY (2021) Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques. J Healthc Eng 2021:1004767. https://doi.org/10.1155/2021/1004767
9. Kraskov A, Stögbauer H, Grassberger P (2004) Estimating mutual information. Physical Rev E 69(6):066138. https://doi.org/10.1103/PhysRevE.69.066138
10. Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324
11. Hearst MA, Dumais ST, Osman E, Platt J, Scholkopf B (1998) Support vector machines. IEEE Intell Syst Appl 13(4):18–28. https://doi.org/10.1109/5254.708428
12. Cox DR (1958) The regression analysis of binary sequences. J Roy Stat Soc: Ser B (Methodol) 20(2):215–232. https://doi.org/10.1111/j.2517-6161.1958.tb00292.x
13. Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. Proc. of the 22nd acm sigkdd international conference on knowledge discovery and data mining 2016: 785–794. https://doi.org/10.1145/2939672.2939785
14. Mienye ID, Sun Y (2022) A survey of ensemble learning: Concepts, algorithms, applications, and prospects. IEEE Access 10:99129–99149. https://doi.org/10.1109/ACCESS.2022.3207287
15. Lao ZP, He DQ, Wei ZX, Shang H, Jin ZZ, Miao J, Ren CC (2023) Intelligent fault diagnosis for rail transit switch machine based on adaptive feature selection and improved LightGBM. Eng Fail Anal 148:107219. https://doi.org/10.1016/j.engfailanal.2023.107219
16. Seijo-Pardo B, Porto-Diaz I, Bolon-Canedo VA, Alonso-Betanzos A (2017) Ensemble feature selection: Homogeneous and heterogeneous approaches. Knowl-Based Syst 118:124–139. https://doi.org/10.1016/j.knosys.2016.11.017
17. Saeys Y, Abeel T, Van PY (2008) Robust feature selection using ensemble feature selection techniques. Proc. of Joint European Conference on Machine Learning and Knowledge Discovery in Databases, vol. 5212.Springer LINK; 2008. p. 313–25. https://doi.org/10.1007/978-3-540-87481-2_21
18. Rincon AL, Tonda A, Elati M, Schwander O, Piwowarski B, Gallinari P (2018) Evolutionary optimization of convolutional neural networks for cancer miRNA biomarkers classification. Appl Soft Comput 65:91–100. https://doi.org/10.1016/j.asoc.2017.12.036
19. Breiman L (1999) Pasting small votes for classification in large databases and on-line. Mach Learn 36(1–2):85–103. https://doi.org/10.1023/A:1007563306331
20. Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. Mach Learn 63(1):3–42. https://doi.org/10.1007/s10994-006-6226-1
21. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232
22. Crammer K, Dekel O, Keshet J, Shalev-Shwartz S, Singer Y (2006) Online passive-aggressive algorithms. J Mach Learn Res 7:551–585
23. Tikhonov AN (1943) On the stability of inverse problems. Dokl akad nauk sssr 39:195–198
24. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY (2017) Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems 30. https://dl.acm.org/doi/https://doi.org/10.5555/3294996.3295074
25. Tang M, Zhao Q, Ding SX, Wu HW, Li LL, Long W, Huang B (2020) An improved LightGBM algorithm for online fault detection of wind turbine gearboxes. Energies 13(4):807. https://doi.org/10.3390/en13040807
26. Bergstra J, Bardenet R, Bengio Y, Kegl B (2011) Algorithms for hyper-parameter optimization. Advances in neural information processing systems 24. https://dl.acm.org/doi/https://doi.org/10.5555/2986459.2986743
27. Prusty S, Patnaik S, Dash SK (2022) SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. Front Nanotechnol 4:972421. https://doi.org/10.3389/fnano.2022.972421
28. Tang XF, Xu AQ, Li RF, Zhu M, Dai JL (2018) Simulation-based diagnostic model for automatic testability analysis of analog circuits. IEEE Trans Comput Aided Des Integr Circuits Syst 37(7):1483–1493. https://doi.org/10.1109/TCAD.2017.2762647
29. Akiba T, Sano S, Yanase T, Ohta T, Koyama M (2019) Optuna: A Next-generation Hyperparameter Optimization Framework. Proc. of the 25th ACM SIGKDD International Conference on

Knowledge Discovery & Data Mining (KDD '19) 2623–2631. https://doi.org/10.1145/3292500.3330701
30. Zhang T (2004) Solving large scale linear prediction problems using stochastic gradient descent algorithms. Proc. of the Twenty-first International Conference on Machine Learning. New York: ACM, 2004. p.116. https://doi.org/10.1145/1015330.1015332



**Gen Li** received his Masters degree in Control Engineering from the Naval Aviation University in 2021. He is currently pursuing a Ph.D. in Military Equipment at the Naval Aviation University. His main research interests include intelligent fault diagnosis and prediction.

**Wenhai Li** received his Ph.D. in Information and Communication Engineering from the Naval Aviation University in 2011. He is now a professor at the Naval Aviation University. His main research interests include intelligent fault diagnosis and testability methods.

**Tianzhu Wen** received his Ph.D. in Military Equipment from the Naval Aviation University in 2015. He is now a lecturer at the Naval Aviation University. His main research interests include fault prediction and health management.

**Weichao Sun** received his Ph.D. in Information and Communication Engineering from the Naval Aviation University in 2015. He is now a lecturer at the Naval Aviation University. His main research interests include the design of intelligent fault diagnosis platforms.

**Xi Tang** received his Masters degree in Instrument Science and Technology from the Naval Aviation University in 2017. He is currently pursuing a Ph.D. in Military Equipment at the Naval Aviation University. His main research interests include intelligent fault diagnosis and testing.