is highly dependent on the input of the policy/value function. Learning is slow and convergence within a limited time is difficult if a massive amount of information with no typical characteristics is used. We proposed a clustering reinforcement learning model named parallel advantage actor-critic with K-means (PAAC-K) in our work. This model can extract filter characteristics and effectively reduce the action space. The PAAC-K model achieves an end-to-end closed-loop training process of designing, simulating, and optimizing. The model is an assembled reinforcement learning model used to design irregular filters automatically, which was proven with four application examples.

2 Related work

As computing power has rapidly increased, artificial intelligence has expanded at an unprecedented rate in recent years. Many researchers have used this method to solve problems in relevant fields [10]. A neural network-based structure parameter estimation approach was used to optimize the structures of BPFs presented by Dai [11]. The designed bandpass lumped circuits were accurately divided into four basic blocks, eliminating the tedious work of manual optimization. Yahya [12] designed a high-performance microstrip duplexer using an artificial neural network (ANN) for improved bandstop characteristics. Salehi [13], using an ANN and an adaptive neuro-fuzzy inference system (ANFIS), designed a subsystem consisting of a microstrip bandpass filter and microstrip low-noise amplifier (LNA). Luo [7] proposed a convolutional neural network with a transfer function for microstrip filter electromagnetic (EM) modelling. Na [14] proposed an algorithm to further reduce the amount of training data through an adaptive sampling process, and demonstrated it with a microwave modelling example. Pinchuk [15] designed and trained a convolutional neural network (CNN). CNN-based direction-of-origin (DoO) filters are superior to existing source direction filters in a range of metrics. The multilayer perceptron feedforward back propagation neural network (MLPFFBP-ANN) presented by Singh [16] was utilized to construct a neural network model and judge the bandwidth of a microstrip antenna. Through several experiments, Sharma [17] used a Gaussian process regression (GPR) model and proved that the fabricated antenna performance is nearest to that designed and predicted by GPR. Both the literature mentioned above and the traditional training methods in neural networks require manual involvement in data extraction and are based on a specific structural improvement. The results by the method depend on the training data, which are extracted manually before the experiment. As there are not sufficient data to support the network model, the engineer's requirements are not met. When the structure of the filter is changed, the data has to be extracted manually again, which is tedious. This limitation makes it difficult to generate a paradigm model.

In addition to the machine learning and neural network methods mentioned above, reinforcement learning (RL), where an agent learns in an interactive environment, is used by researchers. Reinforcement learning is a paradigm and a machine learning approach designed to learn strategies to maximize rewards or achieve specific task targets in agent interactions with the environment. Unlike supervised learning, reinforcement learning does not require preliminary data but requires interaction between the environment and the actors. Sophisticated reinforcement learning algorithms have superior intelligence that can be used to solve complex problems. For example, there are different applications in computer vision [18–21], machine translation [22–24], natural language processing [25–28], data mining [29–34], drug design [35–37], board games [38–40], etc. In the field of microwave integrated circuits (MWICs), reinforcement learning algorithms are one way to automate design. Liu proposed a relationship-based MWIC design network. This design process is transformed from the manual extraction of parameters to the direct interaction of an agent with its environment. The approaches proposed in the literature further reduce manual participation but also have shortcomings.

1. When designing circuits of different frequencies, the agent will repeatedly explore the central frequencies of the circuits sufficiently.
2. If the training queue is not empty and the model is updated, learning will no longer be policy-based (experience at this point comes from the old policy) and causes asynchronous training instability [41, 42]. Our work improved the above shortcomings.

3 Proposed model

3.1 K-means

K-means is one of the most commonly used unsupervised learning algorithms to resolve multiple clustering problems. $X = |\{x_1, \ldots, x_n|\}$ a dataset in a d-dimensional Euclidean space $\mathbb{R}^d$. $z = [z_{ik}]_{n\times c}$, where z_{ik} is a binary variable (i.e., $z_{ik} \in 0, 1$). $A = |\{a_1, \ldots, a_c|\}$ is the c cluster center. If x_i belongs to the k-th$(k = 1, \cdots, c)$ cluster, the k-means objective function is $J(z, A) = \sum_{i=1}^{n} \sum_{k=1}^{c} z_{ik} \|x_i - a_k\|^2$. The objective function $J(z, A)$ iterated for minimiz using