First, we determined the necessary number of synthetic units to balance the proportion of units in both classes (efficient vs. inefficient units). There is no exact proportion that guarantees an ideal balance in the dataset. Weiss and Provost (2003) suggest testing performance with different percentages of minority class examples to identify the optimal class distribution, or an optimal range of class distribution, for the training set. They conclude that, depending on the metric selected, the best minority-class distribution can vary. Moreover, they observed that for highly imbalanced datasets, a balanced class distribution (50%-50%) was often not the optimal choice when minimizing the error rate, and the best range varied across datasets.

In our context, the efficient (minority) class represents units located along the best-practice frontier and is dispersed across the input-output space. This dispersion increases the complexity of the dataset since the model may struggle to learn when these units are very far apart. In contrast to the efficient class, the inefficient (majority) class units tend to be more densely packed, forming clusters with no units from the other class within this concept. For this reason, if there are not enough efficient units for the model to consider, the majority class tends to exert greater influence in that region of the space, making it harder for the model to discriminate between classes and ultimately affecting its performance. Consequently, in scenarios with a very low percentage of minority-class examples, model performance may suffer due to this issue. Following their findings, in our approach, we evaluate the performance of minority-class distributions at 20%, 25%, 30%, 35% and 40% generating synthetic units for each scenario.