As far as I am concerned, the term 'Rashomon set' mentioned in the paper 'stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', is both realistic and can be used to meaningfully capture interpretable models. There are several reasons for this argument.

First of all, it is important to understand the basic meaning of 'Rashomon effect' and 'Rashomon set'. The term 'Rashomon effect' denotes the situation in which there exist many different and approximately-equally accurate to explain a phenomenon. We refer to the paper: define the Rashomon set as the set of reasonably accurate predictive models, for a given data set. If the Rashomon effect is large, it means that there exists a large number of models (the empirical Rashomon set)that perform approximately-equally-well on the training data. When the empirical Rashomon set is large, it is reasonalbe to assume that models with various desirable properties can exist inside it. Due to the finiteness of data, it could admit many close-to-optimal models that predict differently from each other: a large Rashomon set. In reality, it occurs a lot because there are many different machine learning algorithms which use different functional forms perform similarly on the same data set.

Secondly, according to the mathematical foundation provided in the paper 'A study in Rashomon Curves and Volumes: A New Perspective on Generalization and Model Simplicity in Machine Learning' Lesia Semenova et al., 2020, the 'Rashomon ratio' is the cardinality of the Rashomon set divided by the cardinality of all possible models (with varying levels of accuracy). So, Rashomon ratio is defined uniquely for each ML task/dataset pair. When the Rashomon ratio is large, there are several several equally highly accurate ML models to solve the task. Some of these highly accurate models within the Rashomon set might have desirable properties such as interpretability and it would be worthwile to find such models. Thus, Rashomon ratio serves as an indicator of the simplicity of the ML problem. The Rashomon curve is a diagnostic curve connecting the log Rashomon ratio of hierarchy of model classes with increasing complexity as a function of the empirical risk (the error rate bound on the model classes). The Rashomon curve follows a characteristic Γ-shape and the curves connect the Rashomon ratio of increasingly complex model calsses as a function of empirical risk (the observed error of a particular model class). The horizontal part of the Γ-shape corresponds to a decrease in the empirical risk (increase in model accuracy) as we move through the hierarchy of hypothesis spaces (H1 on the top right to H7 on the bottom left). If the ML problem is too complex for a model class considered, only the horizontal part of the Rashomon curve is observed. This is an indication that the model class considered is not complex enough to learn the training data well. On the other hand, if the ML model class considered is too complex for the training data, only the vertical part of the Rashomon curve is observed. As mentioned in the paper above, the Rashomon curve has occurred across all 52 data sets the authors considered.The turning point in the Rashomon curve (‘Rashomon elbow’) is a sweet spot where lower complexity (higher log Rashomon ratio) and higher accuracy (low empirical risk) meet. Thus, among the hierarchy of model classes, those that fall in the vicinity of the Rashomon elbow are likely to have the right level of complexity for achieving the best balance of high accuracy with interpretability.

Finally, some people may question the lack of stability of models chosen by the application of Rashomon set, meaning that small changes in the training data lead to completely different models. However, based on the idea of Rudin, I think the instability in the learning algorithm could be a side-effect of the Rashomon effect mentioned above. Adding regularization to the algorithm will lead to increase of stability, however impose restrictions on the flexibility to chose desirable model of the Rashomon set. As a matter of fact, considering the regularization as a additional input to the Rashomon set. Thus, in the view of increasing interpretability, the instability will be a advantage rather than disadvantage.Given that large Rashomon sets have these interesting properties, it would be worth exploring methods that explicitly try to (re)shape the problem to induce large Rashomon sets. Although we are not aware of any work that has directly done this, there are some existing approaches that may be re-interpreted in this way. For example, one practical technique for producing more robust classifiers is to add noise or smoothing to the training data, e.g., applying a slight blur filter to image data before training. This can be seen as flattening the optimization landscape and potentially increasing the size of the Rashomon set. It is also possible that techniques which inject noise directly into parameter space (Hochreiter and Schmidhuber, 1997) could be interpreted as as having a similar effect.

The fact that injecting noise into the data set and/or optimization potentially leads to larger Rashomon sets is a possible connection to differential privacy and other types of privacy-preserving computation.

In a conclusion, I believe that the Rashomon set mentioned above is realistic and is helpful to capture interpretable models.