Appendix S1

SUPPLEMENTAL METHODS

Model Training

1-1) Dataset

Our dataset was highly imbalanced, and it could be trained biased toward the majority class. To prevent this, we resampled the data using the synthetic sampling techniques synthesizing minority oversampling technology combined with SMOTETomek method, which combines over- and undersampling. SMOTE first selects a random example from the minority class, calculates the Euclidean distance between the example and the k-nearest neighbors of the example, and then creates synthesized data by multiplying a random number between 0 and 1. A Tomek link removes a majority example from a pair of data that belong to the majority and minority classes and are the nearest neighbors to each other. As a characteristic of the above two sampling methods, the SMOTETomek links method makes the class boundary clear and facilitate classification because they remove some of the majority class examples that have invaded the expanded minority class space by oversampling (10). The result of visualization after reducing the resampled data with a k_neighbors parameter of six to two dimensions is shown in eFigure 1.

1-2) Model training

We chose the random forest algorithm because of the following characteristics: 1) it is an ensemble model that reduces variance while combining weak learners with low bias but high variance, 2) weak learners learn independently in parallel, and 3) there is less risk of overfitting than boosting-based models (11). The dataset was divided into training and testing sets at a ratio of 7:3, and hyperparameters were optimized by applying a grid search algorithm and three-fold cross validation. The optimal parameters were n_estimator=200, criterion='gini', and max_depth=5, and the default values were used for the remaining parameters. The relevant source code for developing and validating the random forest ensemble and data sampling has been published in a public repository (https://github.com/dbssk6904/Morphea-Over-sampling-ML.git).

We calculated a confusion matrix to evaluate the performance of the test set, as shown in eFigure 2. The SHapley Additive exPlanations (SHAP) method was then applied to a random forest algorithm to interpret the output of the model. SHAP is based on the concept of the Shapley value, which comes from coalitional game theory and is the average marginal contribution of one variable from all possible combinations of variables. We used three SHAP algorithms and summary plots to identify the importance of each variable.