Dermatology Practical & Conceptual

Supplemental File 1. Random Forest and Shapley Plot Methodology:

To facilitate public evaluation of our algorithm, we have provided a detailed methodology of the machine-learning analysis. AK patients were dichotomized by SCC status, with SCC diagnosis defined as within three years of AK diagnosis. For AK patients with SCC, demographic and laboratory data were collected after AK diagnosis and before SCC diagnosis. Missing values were imputed using iterative imputation with K-neighbours regressor (15 neighbors). A class imbalance was seen with 9,915 AK patients without SCC and 202 AK patients with SCC. Classification was performed using Easy Ensemble classifier, which inherently mitigates potential class imbalance bias by constructing multiple balanced random subsets from the majority group during training for the Random Forest estimators. Classification using the Easy Ensemble classifier utilized 25 Random Forest estimators of 80% training and 20% testing data with 5-fold cross validation after Grid-Search (scikit-learn) optimized hyperparameters. The following parameters were utilized for each classifier: max_features = 'sqrt', n_estimators = 92, min_samples_split = 10, max_depth = 5, max_lead_nodes = 3. Permutation-based feature importance was averaged across all five test sets.

In order to evaluate the overall purity of the classification process and effectiveness of feature splits, we employed a Gini impurity analysis of the Random Forest Classifier nodes. Gini impurity values range from 0 (low) to 0.5 (high) and are used as a criterion to assess node purity and feature splits effectiveness in the Random Forest classifier. Low Gini impurity values across features supports the strength of the model, which indicates that decisions were made on clear, distinctive patterns in the data and is more influential in the decision-making process of the model. Therefore, the alignment of low Gini impurity values with high permutation values in our results confirms that the most influential features contribute to strong and clear classifications, further validating the model's efficacy.

Shapley plots were performed for feature importance visualization. All trends across all Shapley plots were the same, and one plot was included. All analyses were conducted in Python (v 3.9). Although the specific dataset utilized in this study is not publicly available due to privacy constraints from the All of Us database, we recommend researchers use similar, comparable datasets to evaluate the model's generalizability.

Table S1. Demographic Data on AK Patients with and without SCC.

Variable		AK without SCC (N = 9,915)	AK with SCC (N = 202)
Ethnicity	*Not Hispanic or Latino*	9173 (92.52%)	189 (93.56%)
	Hispanic or Latino	352 (3.55%)	1 (0.49%)
	None Of These	69 (0.69%)	4 (1.98%)
Gender	*Female*	4739 (47.80%)	82 (40.59%)
	Male	4974 (50.17%)	115 (56.93%)
	Other	22 (0.22%)	0 (0%)
Sex	*Female*	4743 (47.84%)	80 (39.60%)
	Male	4976 (50.19%)	117 (57.92%)
Race	*White*	9027 (91.04%)	189 (93.56%)
	Black or African American	101 (1.02%)	0 (0%)
	Asian	22 (0.22%)	0 (0%)
	Middle Eastern or North African	18 (0.18%)	0 (0%)
	Native Hawaiian or Other Pacific Islander	2 (0.02%)	0 (0%)

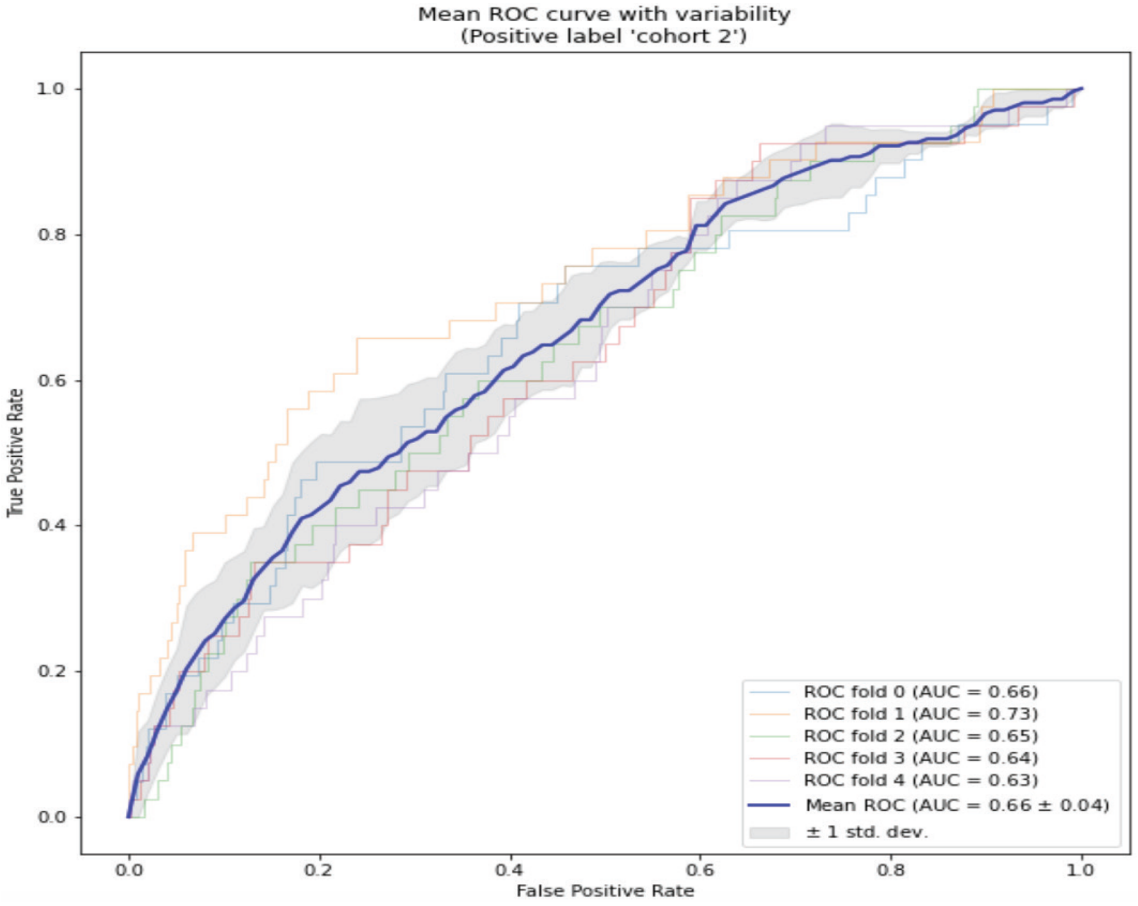

Figure S1. ROC/AUC curve for the Random Forest machine-learning model. AUC scores were evaluated on a 5-fold cross-validation test with the final AUC being the mean of all five cross-validation tests. Blue line shows the mean ROC over the 5-fold cross-validation test. Note: Positive label 'cohort 2' refers to AK patients with SCC.

Figure S2. Gini-impurity based feature importances in the Random Forest model. Gini-impurity values were determined for each primary feature in the Random Forest model and were consistent within the 5-fold cross validation. All impurity values were low (less than 0.1), which is indicative of strong and clear classifications for each feature.