We developed a four-stage pipeline for cervical cell image classification using a Vision Transformer foundation model (EVA-02). First, we fine-tuned EVA-02, pre-trained on ImageNet, to classify Pap-smear cell images into Healthy, Unhealthy, and Rubbish classes, merging “Both cells” into Unhealthy for stability. Second, we extracted representations from the model’s Class and Image tokens and averaged them to obtain 768-dimensional feature vectors per image. Third, feature importance was estimated through Logistic Regression, Random Forest, and Gradient Boosting to remove redundant features and mitigate overfitting. The selected features were then used to train a deeper Artificial Neural Network (three hidden layers: 1024–512–256 neurons) with optional class-balanced loss weighting. Finally, we applied Kernel SHAP to interpret model decisions and identify morphological or staining features correlated with predictions. The proposed framework achieved an F1-score of 0.852 on the public leaderboard—surpassing the EVA-02 baseline (0.849)—while providing interpretable insights into relevant cellular features.