Results:

Model	Balanced Accuracy
Ridge	0.2380
Logistic	0.2875
NeuralNet	0.4448
RandomForest	0.6557

The target balanced accuracy is 0.77 on nPrint, and the RandomForest model achieved amazing results for this 13-class classification task.

Analysis:

The balanced accuracies reveal clear differences across model types. Linear models (Ridge: 0.2380, Logistic: 0.2875) struggle significantly, performing only marginally better than random chance ($1/13 \approx 0.077$) and indicating that OS fingerprints do not exhibit a linear relationship to the IAT/STATS feature space. The Neural Network (0.4448) demonstrates modest improvement by capturing non-linear relationships, but still fails to achieve strong classification performance. Random Forest dominates with 0.6557 balanced accuracy, nearly 50% better than the neural network.

Even though neural networks can model nonlinear patterns, they usually need lots of data and many rich features to work well. In this case, the model only sees a set of simple timing and summary statistics from each flow, so a neural net doesn't have much to learn from and tends to overfit on noise instead. A Random Forest, on the other hand, is good at looking for clear, rule-like splits that distinguish one operating system from another. OS behavior often shows up as a mix of "if the timing is this and packet size is that, it's likely this OS," which the random forest model captures easily.

Future Work:

Future improvements could include expanding the feature set beyond IAT and STATS to incorporate packet-level headers and payload patterns, similar to nPrint's approach. Collecting larger, more diverse datasets across different network conditions and OS versions would help models generalize better. Additionally, testing ensemble methods that combine Random Forest with other classifiers could potentially push accuracy higher while maintaining interpretability.