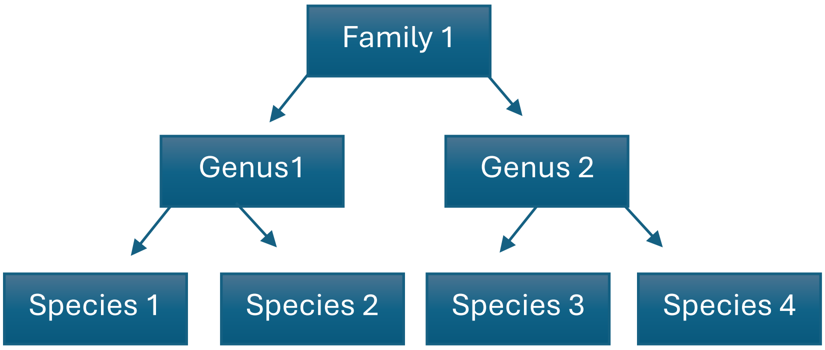
Wood-species Identification: Stakeholder Report

The Australian Customs Service is responsible for validating the authenticity of wood species declared in import and export documentation.

This process is critical for both biosecurity compliance and trade integrity, false declarations may conceal protected, invasive, or high-value timber species. The data derived from microscopic wood images comprises 292,830 samples across:

- 58 unique families, 7 with less than 100, 14 with more than 10,000 samples.
- 191 unique genera, 26 with less than 100, 4 with more than 10,000 samples.
- 925 unique species, 454 with less than or equal to 100 samples, and 7 with counts higher than 5,000.

Each record contains three class columns (family, genus, species). The problem is indicative of a highly imbalanced hierarchal multi-classification problem where:



The goal is to identify species with calibrated rarity awareness, maintaining low false positives for common species whilst maximising detection of rare or protected wood types. The objective varies with context, is this historic anomaly detection, or live classification, for the methodology we pursued the latter.

Each feature file was merged to form a master set, where there were:

- Zero Constant columns.
- Zero Column duplicates.
- ~130,000 duplicated rows.

For visualised distributions see **[1]**.

Duplicated rows were retained, as deduplication risked erasing under-represented specie patterns. Features were scaled appropriately.

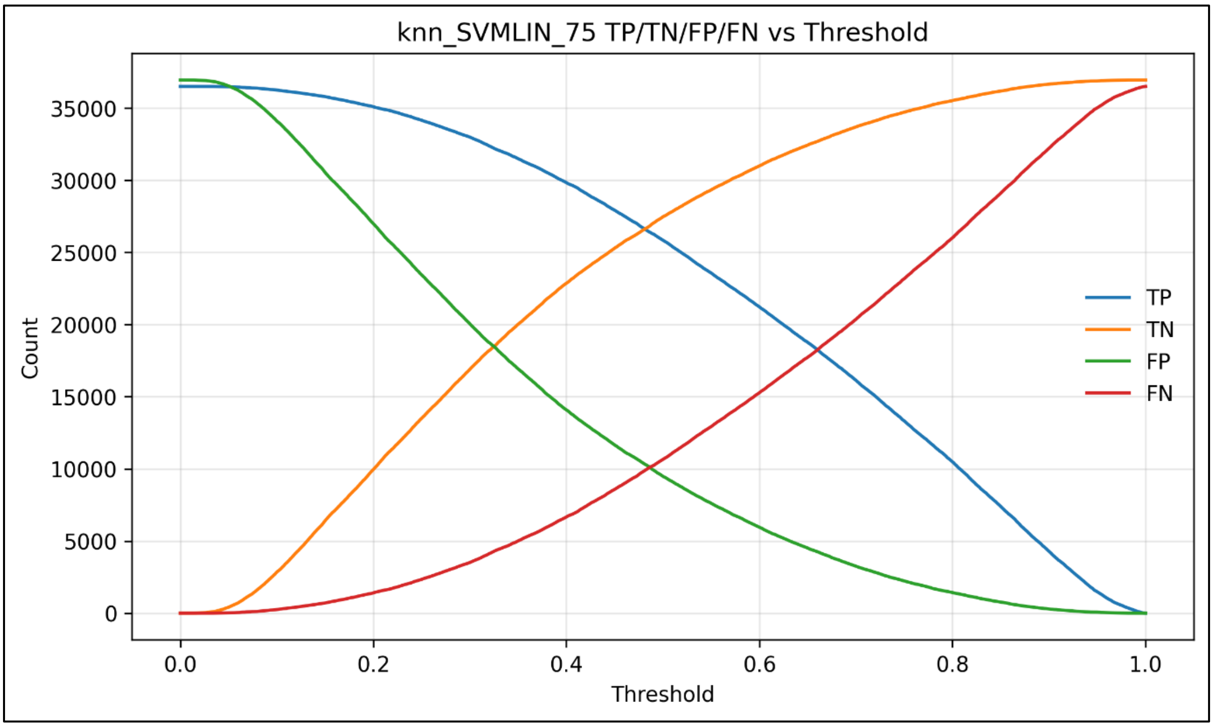
Model Architecture:

The classifier operates in three sequential stages; to exploit hierarchal structure each component is designed to mirror the taxonomic approach of wood classification, handling class imbalance.

Stage 1 – Gating: Classifying Rarity through SVM thresholding:

Stage 1 ask’s the question, “is this species rare?” to separate common vs rare wood samples to guide later stages. Rarity estimated via K-means, Gaussian Mixture (GMM), and K-nearest neighbours’ density over log-scaled species counts, generating rarity probability. Further Support Vector Machine methods (Linear, Radial-Basis Functions) were implemented for classification, Youden’s J threshold optimisation balances selection of best method through distance-threshold trade-off between True-Positive Rates and False-Positive Rates over selected thresholds.

It was found that the SVM with a linear kernel with KNN, yielded the highest true classifications (optimal threshold: 0.500865720134036).



Further in **[4]**. The figure above identifies the threshold that maximises the distance between True Positives (correct classification of rare) and False Positives, and vice versa with True Negatives (correct classification of common) and False Negatives (misclassification of rare species). At optimal threshold AUC was found to be 0.801.

Stage 2 Family and Genus Ensembles:

After establishing if the specie is potentially rare, stage two sorts out to determine *what it is*. First its Family, then its Genus. This provides an interpretable mid-tier before the model attempt the much harder species-level identification (1/925).

Two ensembles utilising Random Forest variations. Then a logistic regression layer wraps the results to learn how to weight each of the ensembles probabilies. The rarity probability from Stage 1 is added as an auxiliary feature, allowing the wrapper to adjust for imbalance between rare and common cases.

Additionally, a family -> genus constraint ensures consistency with hierarchal nature. If a predicted genus is not valid for a chosen family, its probability is zeroed, and the remaining probabilities renormalised, see **[6]** for metrics. We saw features from all files were amongst the most informative features validating dataset merge proved fruitful.

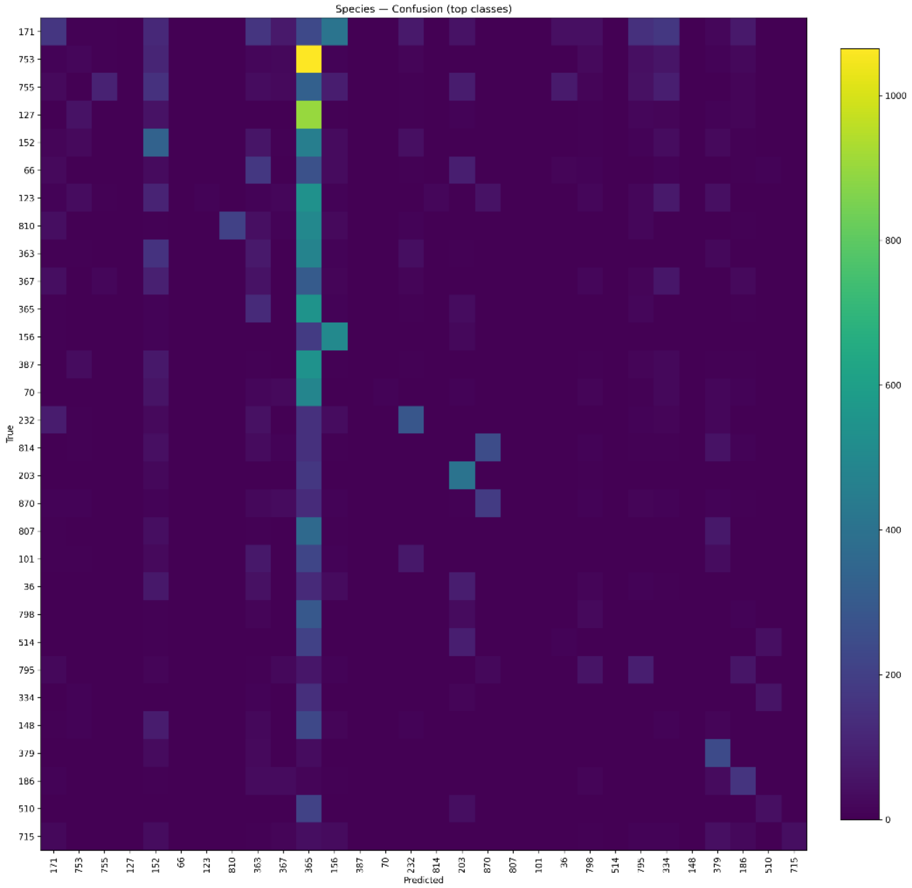
Overall the initial family classification performed better then the sub-genus classification, that said, predicted genus with > 0.70 probability saw 100% accuracy.

Stage 3 Boosted Species Classifier:

Stage 3 consolidates everything measured thus far, the rarity signal, family and genus probabilities to predict the final species label. This stage mimics an expert who first identifies the general group, then upon investigation identifies the precise species.

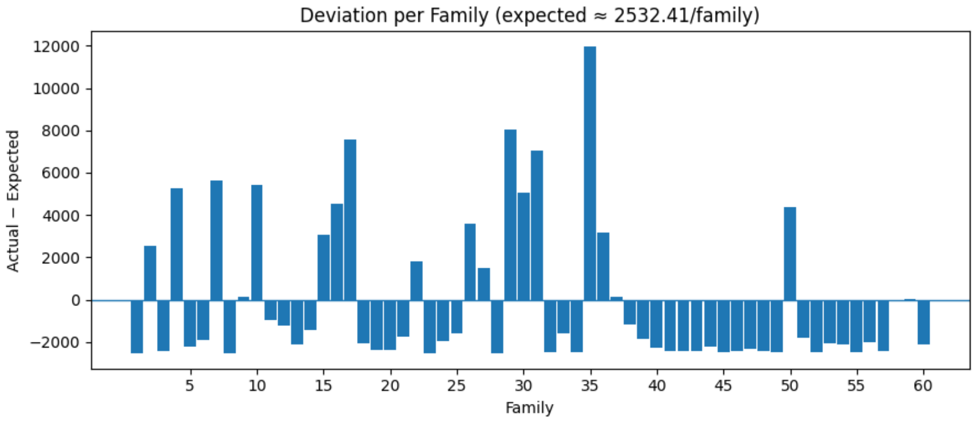
Features include a rarity score, 58 family-level probability + 191 genus-level probability.

The classifier saw an over-representation of misclassification when classifying rare species, which is advantageous for the project's objective. However, could prove problematic for wholesale timber export/importers, but minimises biosecurity risk.



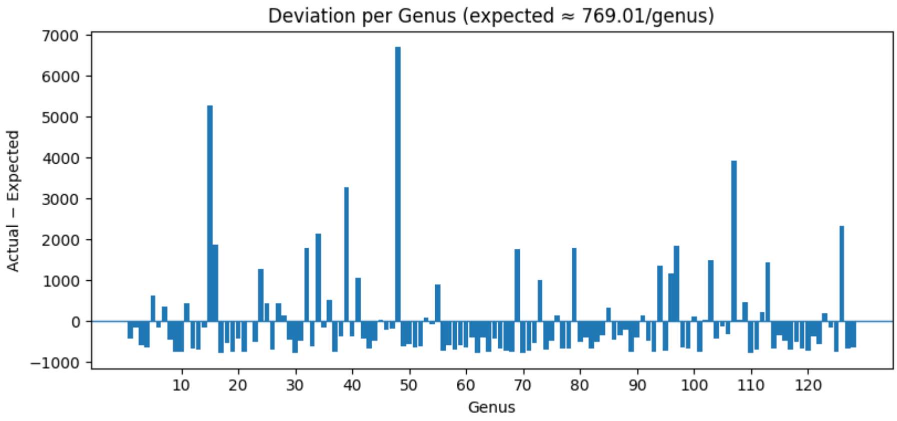
Appendix:

[1] Family: Observed vs Expected



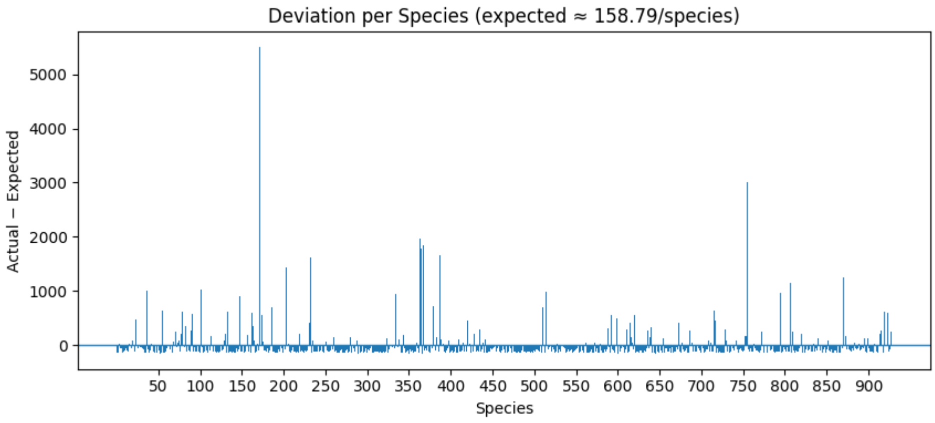
Missing family: 1, 28

[2] Genus: Observed vs Expected



Missing genus: 30, 62, 70, 110

[3] Species: Observed vs Expected

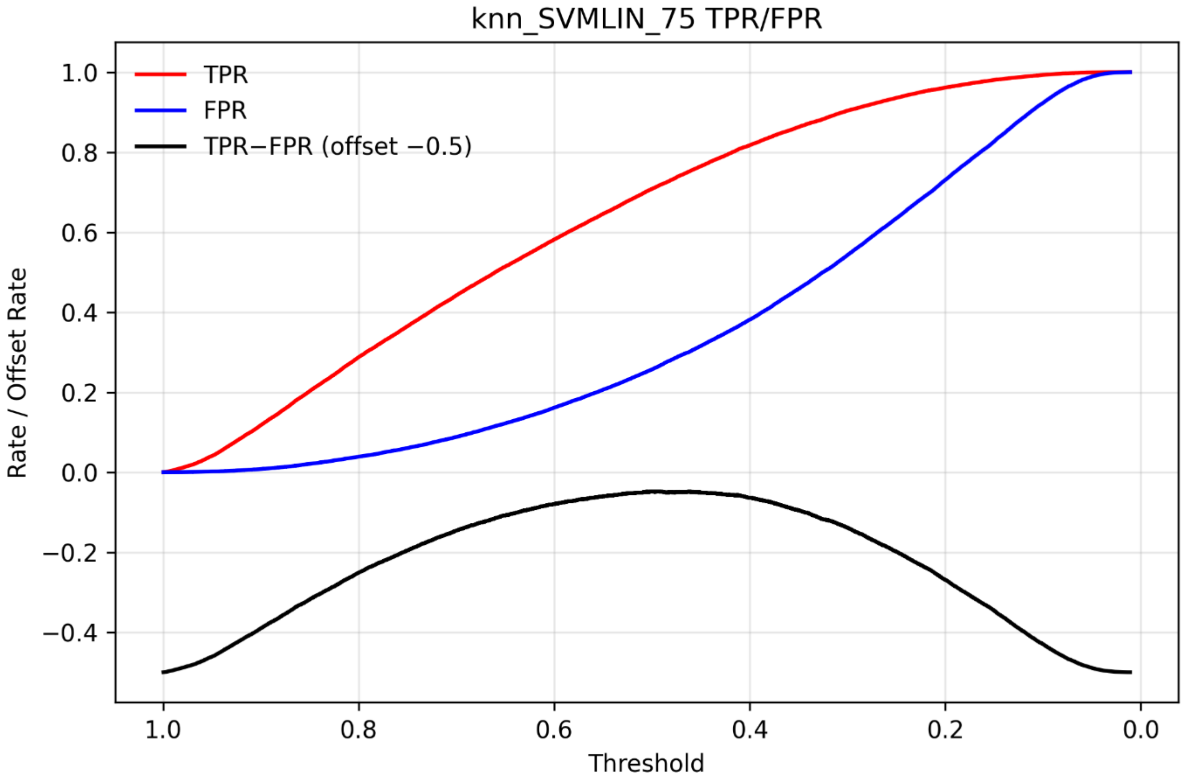


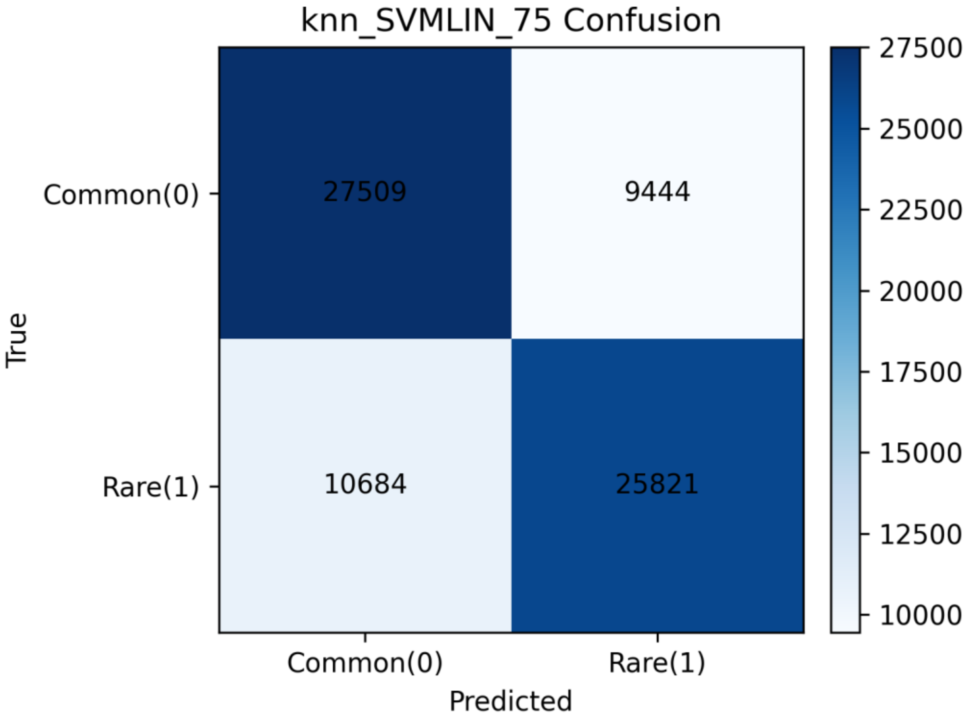
Missing species: 114, 229

Where expected:

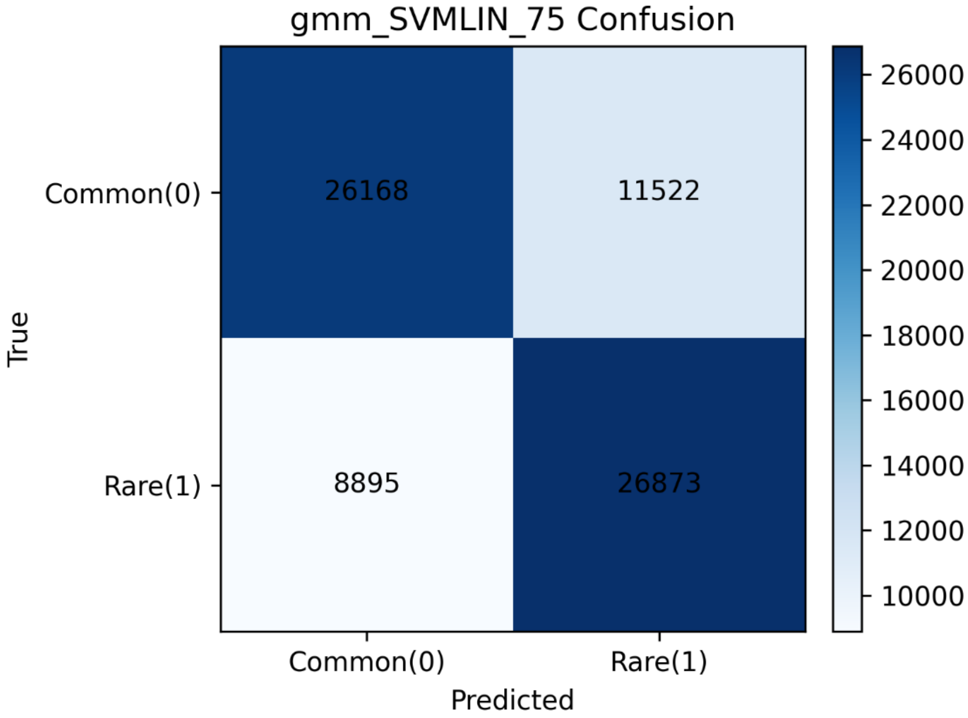
$$\frac{n_{samples_{class}}}{\sum samples}$$

[4] KNN Linear Metrics

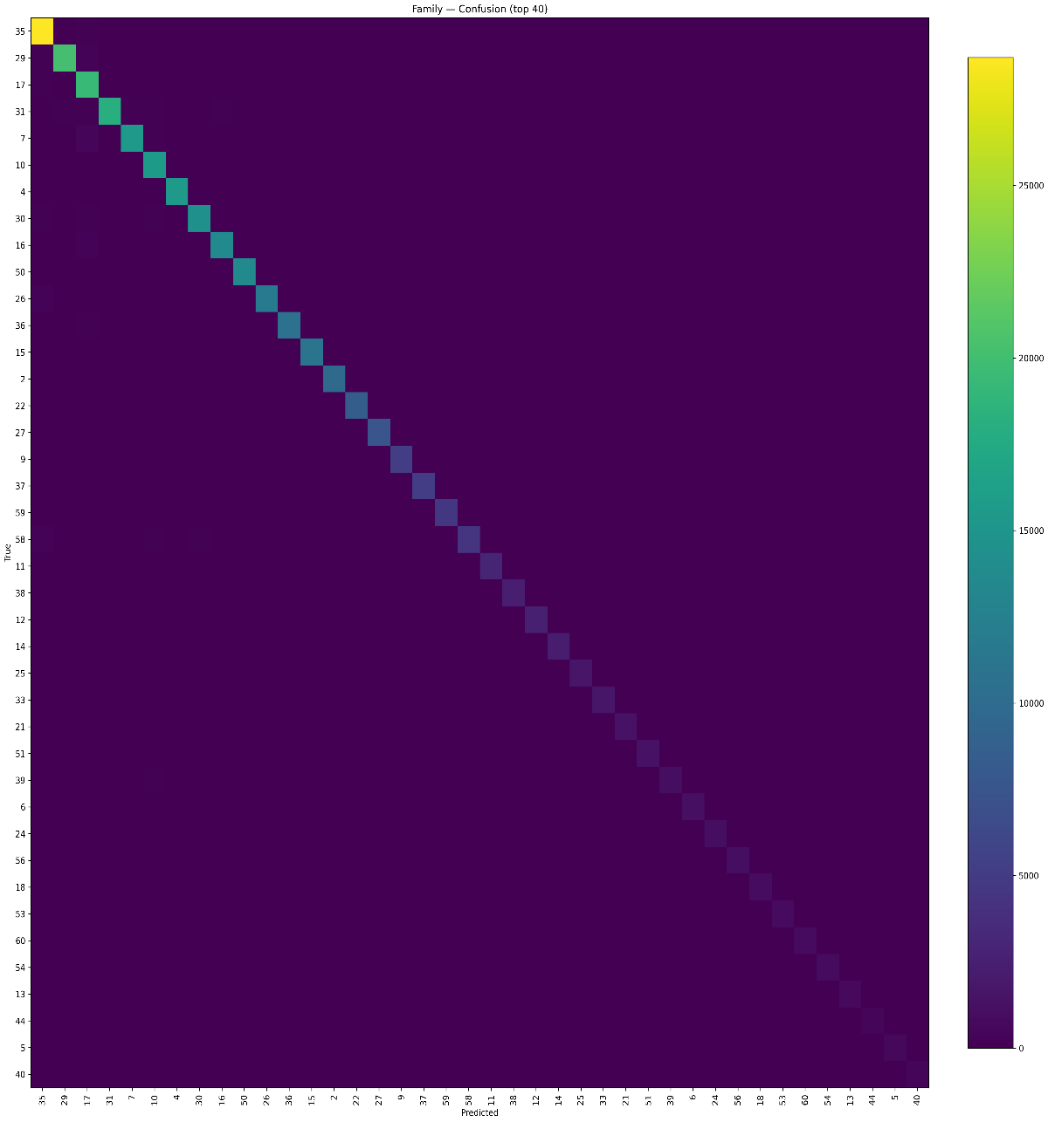


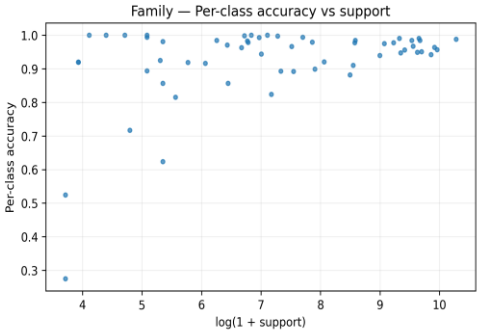


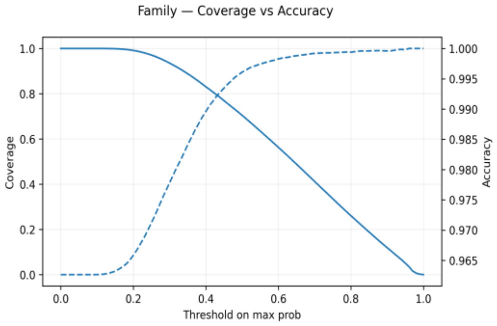
[5] Notable Stage 1 Metrics

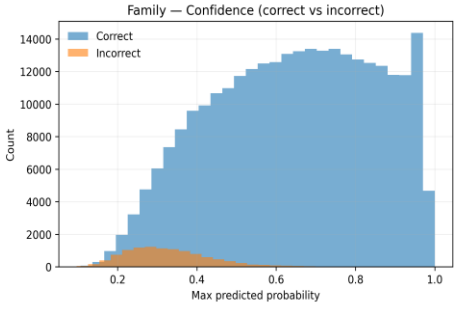


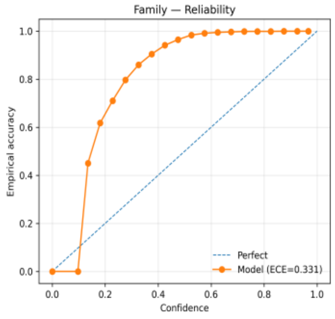
[6] Stage 2 Metrics - Family


Family — Confusion (top 40)

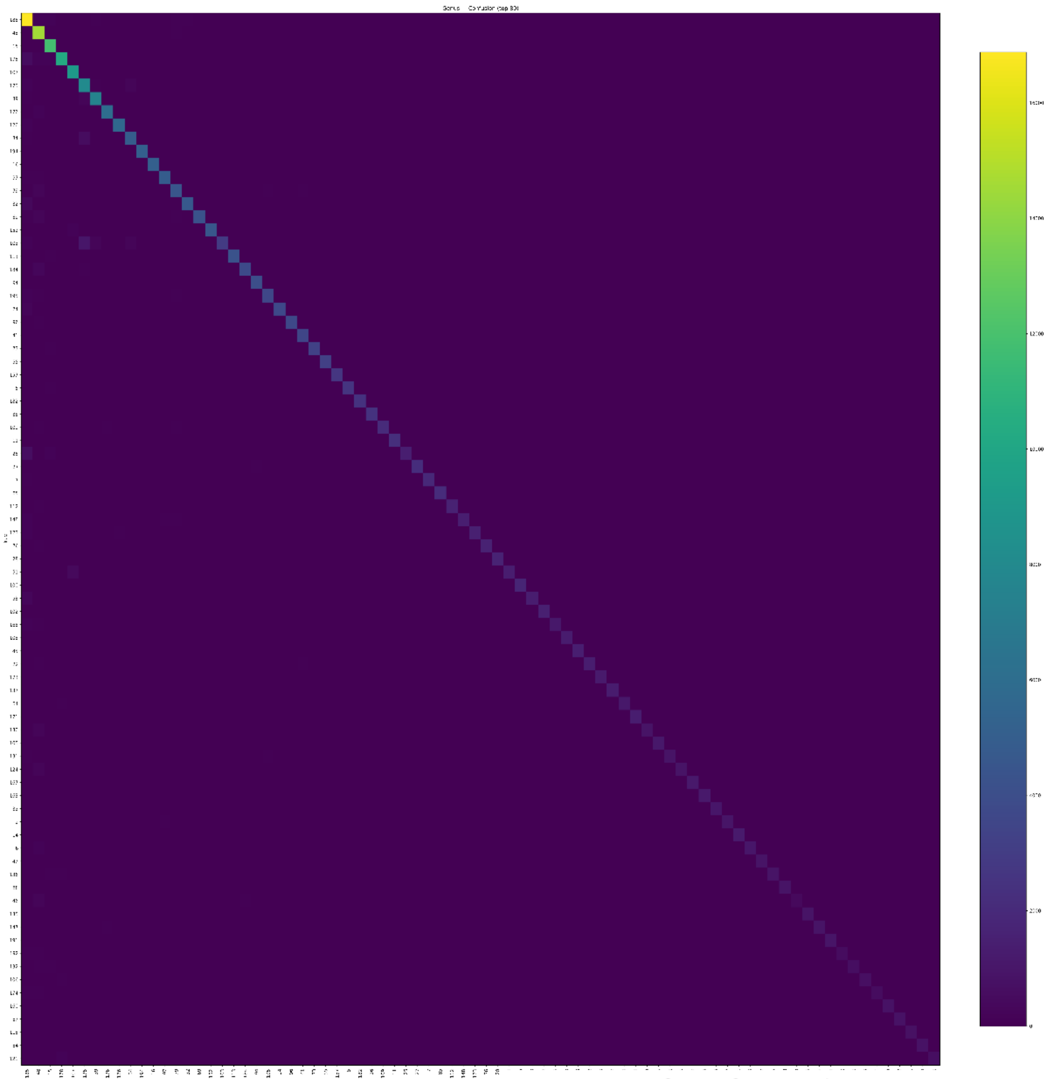

Family — Per-class accuracy vs support

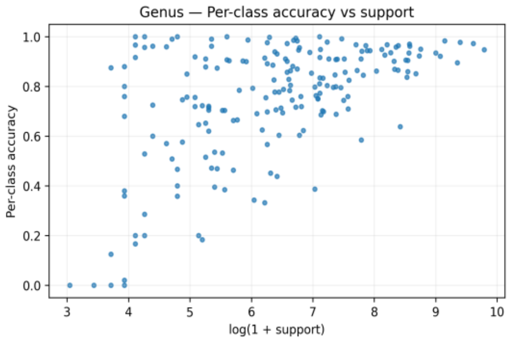

Family — Coverage vs Accuracy

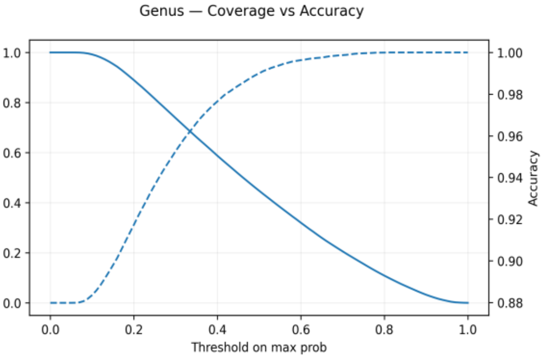

Family — Confidence (correct vs incorrect)

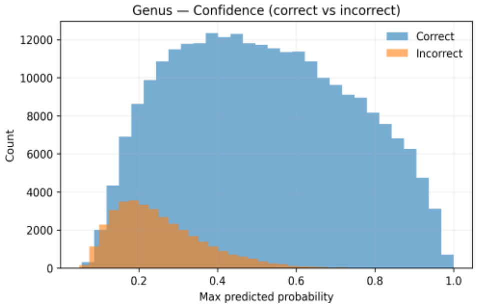

Family — Reliability

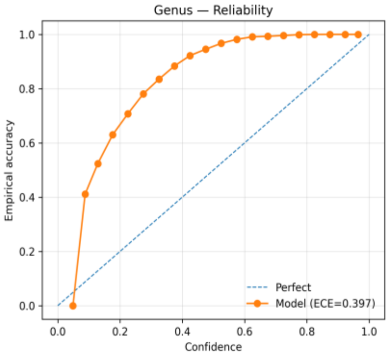
[7] Stage 2 Metrics - Genus



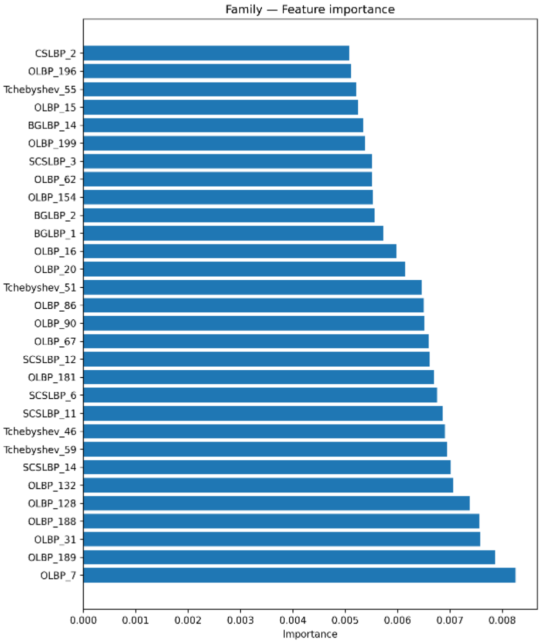


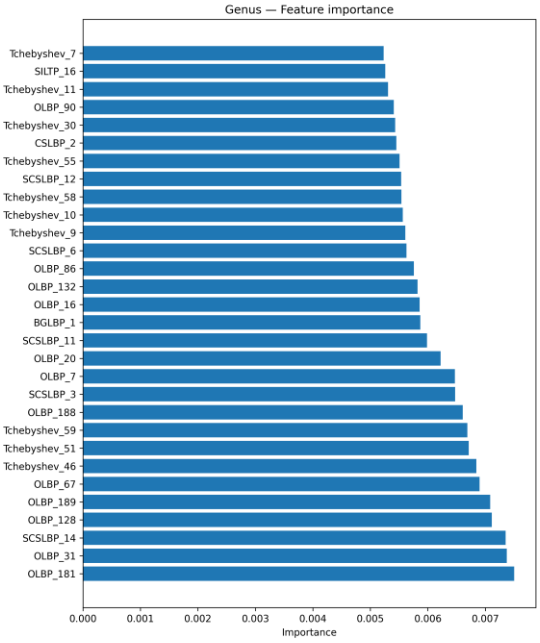






[8] – Stage 2 Feature Importance





[9] – Stage 3: