GENERATING INVESTIGATIVE LEADS FROM FORENSIC DNA DATA:

Mapping Y-STR Profiles to Ancestral Haplogroups

Jan Paolo V. Moreno[1,2], Kevin Ansel S. Dy[1], Francis Erdey M. Capati[1], Mia Cielo G. Oliveros[1], Kristine Ann M. Carandang[1,3], Christian M. Alis[1,3]

[1]Aboitiz School of Innovation, Technology & Entrepreneurship, Asian Institute of Management [2]DNA Laboratory Division, Philippine National Police Forensic Group [3]Analytics, Computing, and Complex Systems Laboratory, Asian Institute of Management

FROM STATIC DNA EVIDENCE TO FORENSIC INTELLIGENCE

While Y-chromosome short tandem repeat (Y-STR) typing is the routine forensic analysis for individual identification, it could not provide investigative leads without a direct match from a previously populated database. This research bridges that gap by using machine learning to predict Y-chromosome single nucleotide polymorphism (Y-SNP) haplogroup, transforming what was static DNA evidence into generative intelligence for paternal ancestry and cold case leads.

PROJECT OBJECTIVES

- To predict Y-SNP haplogroups directly from forensic Y-STR profiles using machine learning.
- To demonstrate the use of SHAP (SHapley Additive exPlanations) for model interpretability and judicial transparency.

THE DATASET AND DATA PROCESSING PIPELINE

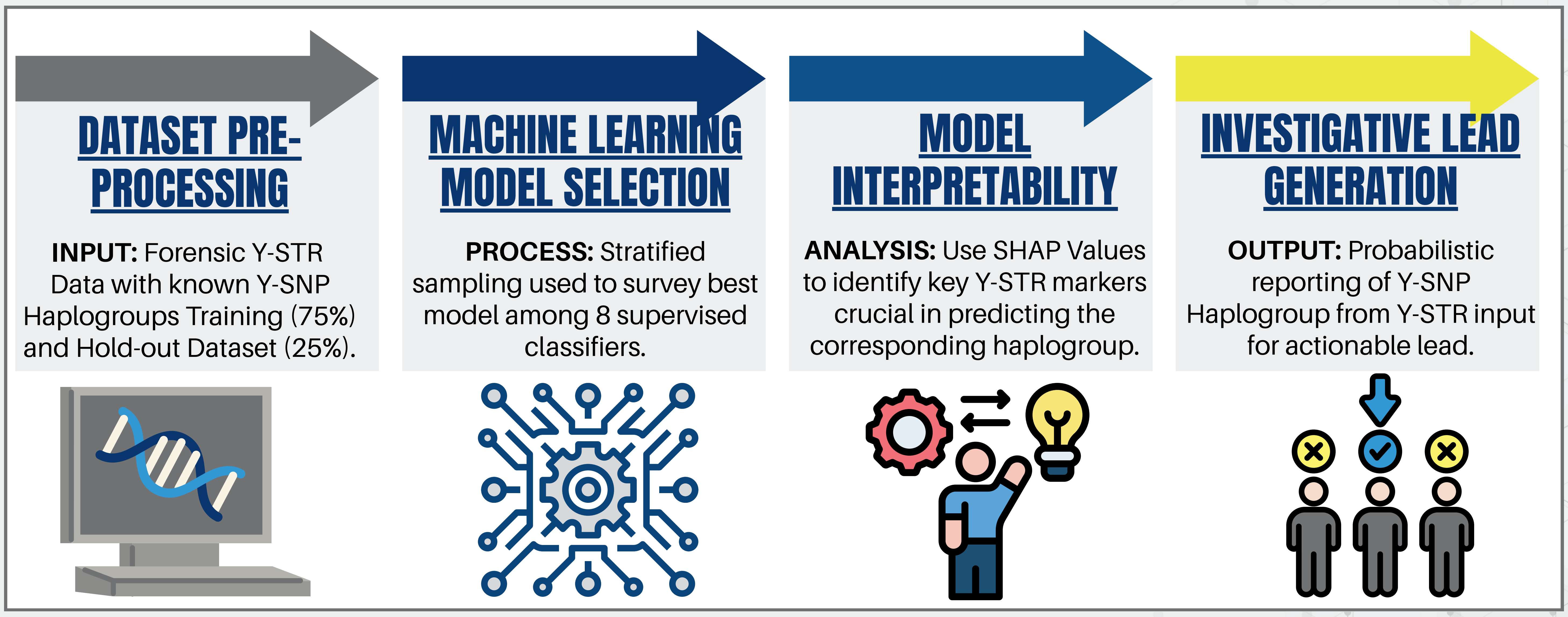
We engineered a machine learning pipeline using **stratified sampling** and **class weighting** to address extreme population imbalances and ensure high precision for minority lineages.

4,064
unique Y-STR profiles characterized by

27 Y-STR
genetic markers to associate with

13 Y-SNP Haplogroups
representing paternal lineages

sourced from the comparative study by **Song et al.** (2024) *YHP: Y-chromosome Haplogroup Predictor for predicting male lineages based on Y-STRs.*



DATASET PRE-PROCESSING

INPUT: Forensic Y-STR Data with known Y-SNP Haplogroups Training (75%) and Hold-out Dataset (25%).

MACHINE LEARNING MODEL SELECTION

PROCESS: Stratified sampling used to survey best model among 8 supervised classifiers.

MODEL INTERPRETABILITY

ANALYSIS: Use SHAP Values to identify key Y-STR markers crucial in predicting the corresponding haplogroup.

INVESTIGATIVE LEAD GENERATION

OUTPUT: Probabilistic reporting of Y-SNP Haplogroup from Y-STR input for actionable lead.

Figure 1. *Design of data processing pipeline used in the study.*

KEY FINDINGS: MODEL PERFORMANCE & INTERPRETABILITY FOR LAW ENFORCEMENT APPLICATIONS

Machine Learning	Model Accuracy
XGBoost	**0.9698**
Random Forest	0.9679
SVM	0.9623
kNN	0.9616
LDA	0.9438
Gaussian NB	0.9351
Decision Tree	0.9342
Elastic Net	0.2029

Table 1. Machine Learning Models Performance.

The optimized XGBoost framework achieved a superior **96.98%** Accuracy and Macro F1-score of **0.9810**, drastically outperforming linear models like Elastic Net (20.29%).

Haplogroup Class	Precision	Recall	F1-Score
Major (e.g. O)	0.99	0.98	0.99
Minor (e.g. L)	0.97	0.96	0.97
Overall Model	**98.24%**	**96.98%**	**0.981**

Table 2. Summary of XGBoost Model Performance.

The model maintains high fidelity across imbalanced populations, ensuring reliable investigative leads regardless of ancestral origin.

SHAP FEATURE IMPORTANCE BY CLASS

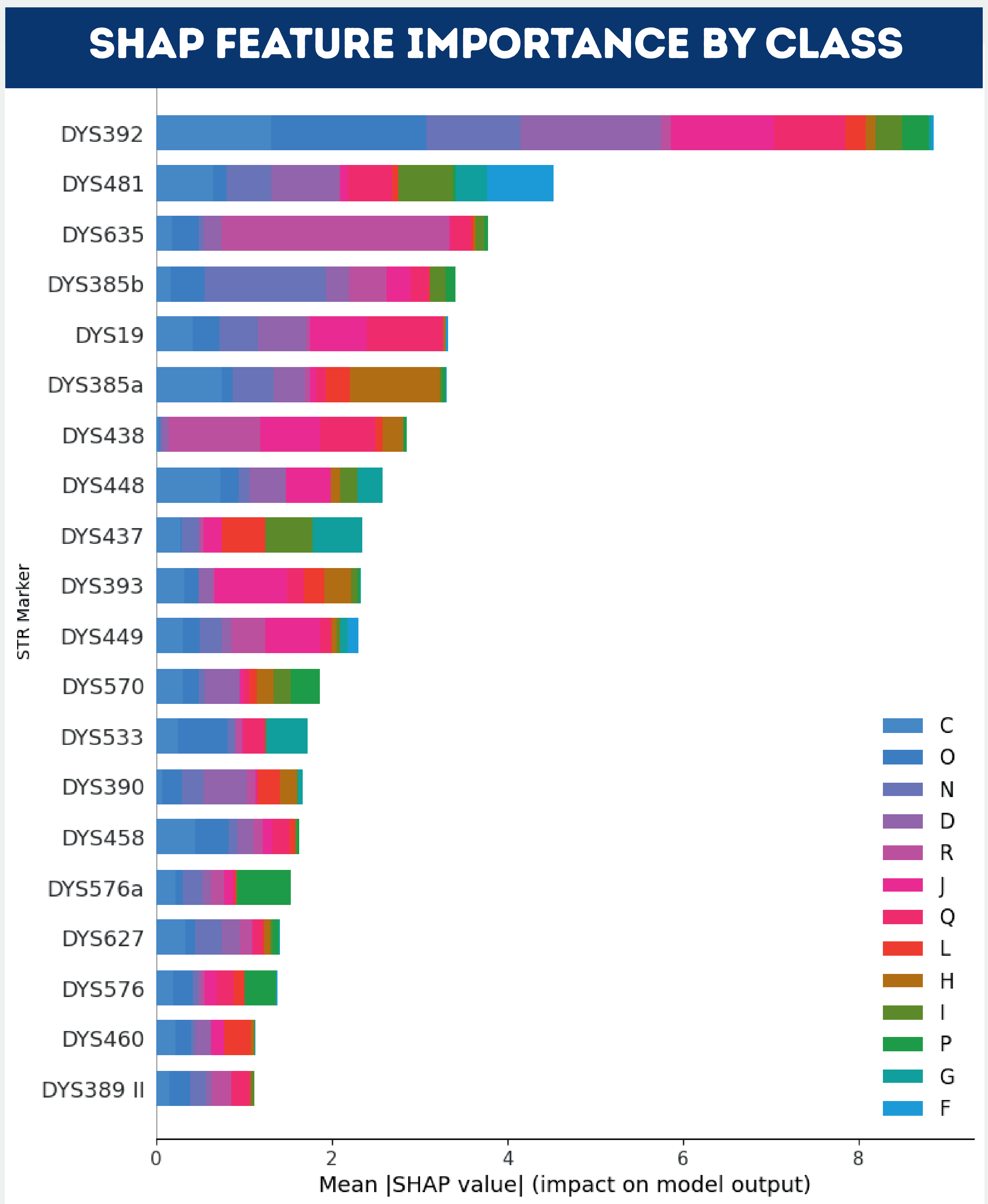


Figure 2. *Summary of SHAP Values per Y-STR genetic marker.*

STEPPING INTO THE FUTURE OF FORENSIC INVESTIGATIONS

- This research transforms static DNA data into actionable leads by mapping Y-STR profiles to Y-SNP haplogroups with high fidelity, achieving an overall F1-score of **0.9810**.
- By integrating **XGBoost with SHAP interpretability** and class weighting, the framework ensures the judicial transparency and demographic fairness essential for "Trustworthy AI" in forensic settings.
- To operationalize this tool, we recommend implementing **regional data retraining** and **independent peer-validation** of explainability results to **establish a standardized, ethical global protocol** for generative investigative intelligence.

This project was first presented to the Machine Learning Class of Professor Kenneth T. Co and first formulated in the Data Mining and Wrangling Class of Professor Jose G. Bautista. We express our sincere gratitude to the guidance of the Asian Institute of Management ACCeSsLab. We also thank the noble men and women of the Philippine National Police Forensic Group DNA Laboratory Division, forensic practitioners whose feedback helped refine the applicabiliy and explainability of this framework.

To God be the glory!