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 [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

Abstract

Genetic markers, particularly Y-chromosome short tandem repeats (Y-STRs), play a critical role in forensic investigations. Since Y-STRs are inherited strictly along the paternal line, it can help differentiate male lineages. However, its forensic value depends heavily on the availability of reference profiles in population databases. When no corresponding entry exists, a generated Y-STR profile cannot be used for direct identification. Acknowledging this gap, this study aims to investigate the possible features critical in developing machine learning framework for predicting Y-chromosome Single Nucleotide Polymorphism (Y-SNP) haplogroups from standard Y-STR profiles. Prediction of haplogroups provides information on paternal lineage ancestry, enabling the generation of intelligence leads useful for police investigations. Through comprehensive evaluation of multiple supervised classifiers on a dataset of 4,064 Y-STR profiles, the optimized XGBoost (Extreme Gradient Boosting) classifier was selected for its superior raw predictive power, achieving the highest overall accuracy of 96.98% and a Macro F1-score of 0.9810. Critically, the framework employs stratified sampling and class weighting to ensure fairness across demographically underrepresented ancestral groups Evaluation of the model incorporates stratified sampling and class weighting to mitigate inherent demographic data imbalance, ensuring fairness across minority ancestral groups. Furthermore, the integration of SHAP (SHapley Additive exPlanations) provides the necessary model interpretability to guide ethical and legal requirements for deployment in police investigations, thus advancing the paradigm of trustworthy AI in law enforcement.

Code — https://github.com/kibindy/LT8_final_project/blob/master/Mapping\%20Forensic\%20Y-STR\%20to\%20Haplogroups.ipynb
Datasets — https://github.com/cissy123/YHP-Y-Haplogroup-Predictor-

Introduction

Deoxyribonucleic acid (DNA) fingerprinting or short term forensic repeat (STR) typing has fundamentally reshaped the criminal justice system, establishing the gold standard for human identification in criminal cases (Kayser 2017; Ndomondo et al. 2025). Being cost-effective, Y-STR typing is currently one of the routine procedures performed in a forensic laboratory.

Y-STRs are short repeating DNA segments on the Y chromosome that are inherited strictly along the paternal line, providing high variability for individual discrimination but possessing limited ancestral information due to their high mutation rate (Rolf et al. 2001; Ndomondo et al. 2025). However, its use to resolve cases is highly dependent on the existence of a reference profile within a DNA database such as the Combined DNA Index System (CODIS) or a Y chromosome Haplotype Reference Database (YHRD) (Ndomondo et al. 2025; Costa et al. 2025). When a profile from a crime scene fails to produce a match, the evidence becomes static, often including the case in the unsolved cold case backlog (National Institute of Justice 2002).

In such cases, another procedure can be performed, the single nucleotide polymorphism (SNP) array analysis, in which broader ancestral information can be derived. This utilizes Y-SNP (Y chromosome SNP) to determine the haplogroup to which a sample belongs. Y-SNPs mutate slowly, defining distinct phylogenetic clades known as haplogroups. This is strongly correlated to deep human migratory history and biogeographical ancestry. By understanding these haplogroups, law enforcers can generate leads pertaining to a suspect's likely continental or regional origin, thereby significantly narrowing the initial suspect pool and prioritizing resources. However, this process is resource-intensive in terms of cost, time, and sample quantity, leading to its exclusion from initial forensic casework (National Institute of Justice 2002; Li et al. 2025).

To address these challenges and facilitate lead generation for law enforcement investigations without relying on database comparison, our work focuses on classifying DNA samples using supervised machine learning into 13 haplogroups based on their Y-STR profiles, each containing allele scores for 27 Y-STR markers. We also demonstrate how interpretability analysis through SHapley Additive exPlanations (SHAP) can facilitate transparency and inform policy decisions, contributing to trustworthy use of artificial intelligence (AI) in law enforcement.

Related Work

Generating Forensic Intelligence from DNA Profiles. The role of DNA analysis in criminal justice has evolved significantly since the initial adoption of Restriction Fragment Length Polymorphism and the subsequent shift to STR markers. While the primary function remains unique for individual identification, the lack of matches in a high percentage of cases has pushed the field toward forensic lead generation. This area encompasses techniques such as forensic phenotyping (predicting physical appearance) and inferring biogeographical ancestry (BGA). Early BGA tools relied heavily on Autosomal SNPs (Kayser 2017), but Y-chromosome analysis, due to its paternal-lineage specificity, offers a distinct and powerful investigative tool, particularly in sexual assault and patrilineal missing person cases. Current computational approaches however have limitations. Y-STR typing requires an established reference databases with autosearch capabilities while Y-SNP array analysis involves resource-intensive procedures with different laboratory equipment and computing capacity (Li et al. 2025; Ndomondo et al. 2025). Accordingly, we use Y-STR profiles of individuals to predict the haplogroup to which a profile belongs to, to facilitate lead generation despite the absence of a reference database.

Haplogroup Prediction. Prior attempts to predict Y-SNP haplogroups from Y-STR data originated in population genetics and statistical models. These approaches typically relied on frequency databases, using maximum likelihood or Bayesian statistics to calculate the probability of a haplotype belonging to a specific haplogroup (Willuweit and Roewer 2015; Babić Jordamović et al. 2021).

The field has seen a recent shift toward supervised machine learning to overcome the limitations of statistical models (Song et al. 2024; Yin et al. 2022; Fan et al. 2023). Song et al. (2024) have used multiple algorithms such as Random Forest, Support Vector Machines, and Neural Networks to model the non-linear mapping between Y-STRs and Y-SNPs. These models demonstrate superior predictive accuracy and generalization compared to their statistical counterparts. However, a critical gap remains. Much of the existing machine learning literature prioritizes raw accuracy over the necessary legal and ethical requirements of interpretability, transparency, and fairness. Without explicit methodological interventions to address demographic imbalance and the "black box" nature of deep learning or ensemble methods, these highly accurate tools cannot be held accountable once integrated into the justice system (Mittelstadt 2019). Apart from developing machine learning models, we perform model interpretability analysis on the best performing model to enhance explainability of model results.

Method

We first describe the dataset we use to demonstrate how we can use machine learning and interpretability analysis in lead generation. We then elaborate on the approaches used in model training and evaluation.

Dataset

We use the Y Haplogroup Predictor (YHP) dataset from the study of Song et al. (2024), which contains 4,064 unique samples, each characterized by 27 Y-STR markers and their corresponding confirmed Y-SNP haplogroup assignments. Each Y-STR profile contains 27 Y-STR markers. Each Y-STR marker represents a distinct Y-STR locus genotyped by Yfiler Plus kit represented by the allele score (i.e., the number of repeats of a short DNA sequence in the locus). The allele scores vary per individual; hence, the distributions of the allele scores likewise vary as shown in Figure 1. We use these variations in the Y-STR profile to predict which of the 13 major Y-SNP haplogroup present in the dataset a sample belongs to.

In terms of major haplogroup assignment, we note imbalance in the distribution of samples (see Figure 2), which is typical of most available forensic population data. Specific haplogroups, such as Haplogroup O (predominant in East Asian populations), constitute a large majority, while Haplogroups G, H, or L represent minimal fractions of the dataset. With this imbalance, a model trained naively can be expected to achieve high overall accuracy by defaulting to the majority class, which may lead to high rates of misclassification for minority ancestral groups. In the justice setting, partiality would be unacceptable as it would result in an unequal opportunity for lead generation based on a suspect's lineage.

Model Training and Evaluation

We describe here the models experimented on, the evaluation metrics we monitored and the cross-validation process implemented to select the best model for predicting haplogroup assignment from Y-STR profiles.

Models. Several machine learning models for classification were trained and evaluated. These include Extreme Gradient Boosting (XGBoost), K-Nearest Neighbors (kNN), Decision Tree, Support Vector Machine (SVM), Random Forest, Gaussian Naive Bayes, Linear Discriminant Analysis (LDA) and Elastic Net.

Evaluation Metrics. Considering the class imbalance of the haplogroup assignment, macro-F1 score as the primary evaluation metric while we also report and monitor accuracy and precision scores. The macro-F1 score is the unweighted average of the F1 scores for each of the 13 classes, ensuring that the performance of the smallest minority class carries the same weight as the largest majority class. Achieving a high macro-F1 score would be the primary ethical consideration, confirming that the model's predictive reliability is consistent and equitable across all ancestral demographics, aligning directly with the principles of trustworthy AI in public safety (O'Neil, 2016).

Model Selection. We performed hyperparameter tuning using stratified cross-validation (CV) strategy with $k = 5$ to select the best performing model with its corresponding optimal hyperparameters. We split the dataset of 4,064 samples into a training-validation set (75%) and a holdout set (n=1,016 or 25%) using stratified sampling. This critical step ensured that the proportional representation of all 13 hap-

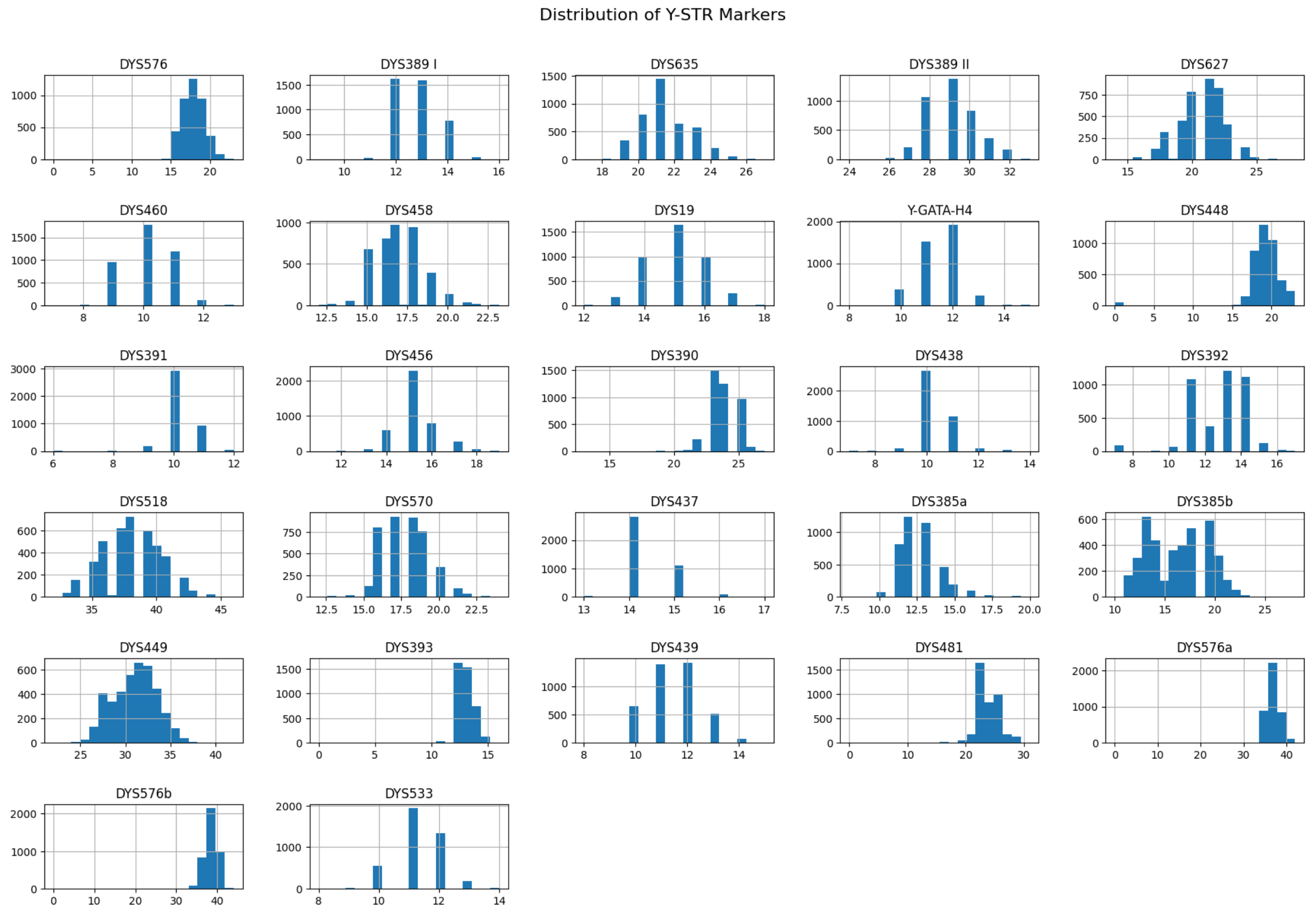

Figure 1: **Distribution of the Y-STR Allele Scores Across the Genetic Markers.** Most Y-STR markers show clear, unimodal distributions with tight clustering around common allele values (e.g., DYS576, DYS449), indicating strong population-level consistency. A few markers exhibit broader or multi-modal patterns (e.g., DYS389II, DYS456), suggesting higher variability and greater discriminatory power.

logroups was maintained in both the training and testing partitions, preventing the model from encountering unseen minority classes during final evaluation. Stratified CV ensures that each fold maintains the original class distribution, providing a far more stable and reliable estimate of a model's true performance compared to standard CV, which is highly vulnerable to imbalanced data.

Model Interpretability Analysis

We use SHAP (SHapley ADditive exPlanations) on our best performing machine learning model to enable law enforcers gain insights into the Y-STR markers that influence the classification of the sample into a certain haplogroup. This enhances the transparency and trustworthiness of the prediction process. Results of this could likewise inform policy decisions.

Results and Discussion

We show that we can classify DNA samples into their respective haplogroups using machine learning, and utilize model interpretability analyses to inform decisions on how the predicted classification should be used and demonstrate increased transparency in using machine learning models.

Model Performance and Robustness

As shown in Table 1, the XGBoost model achieved the highest performance metrics, justifying its selection as a considerable model for forensic intelligence generation. The model achieved an overall accuracy of 97%. More significantly, the macro-F1 score of 0.91 confirms the success of the fairness interventions, demonstrating strong performance even in the most resource-scarce classes (e.g., F1 scores for Haplogroups G and L were substantially higher than non-weighted models). Poor performance of the Elastic Net model can be attributed to the highly non-linear nature of relationship between Y-STR markers and Y-SNP haplogroups. Y-STR data constitutes natural evolutionary nuances such as homoplasy, locus interactions, and other threshold effects inherent in genetic data. As a result, the model would default majority classes under severe class im-

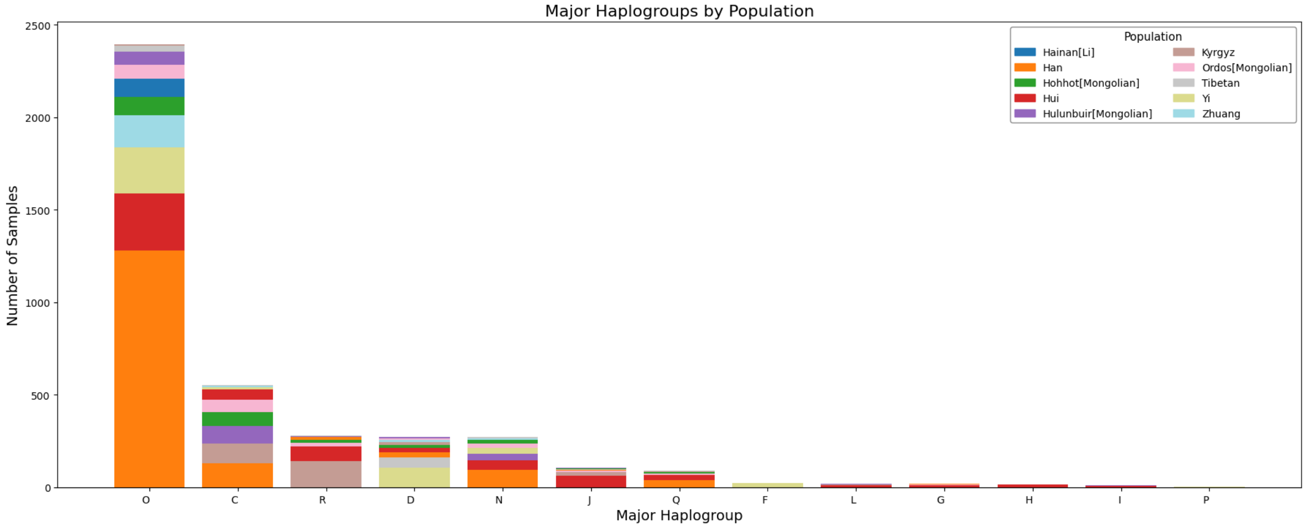

Figure 2: **Distribution of Major Haplogroups.** The plot shows a highly imbalanced distribution, with haplogroup O contributing the vast majority of samples, while all other haplogroups are sparsely represented. This imbalance means models trained on these data may become biased toward haplogroup O and struggle to learn meaningful patterns for minority haplogroups.

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 Model Performance.** XGBoost achieved the strongest performance among all models, reaching an overall accuracy of 0.97. However, despite this high accuracy, variability in macro-level metrics indicates that minority haplogroups remain more difficult to classify, reflecting the effects of dataset imbalance.

balance, leading to very low predictive performance despite regularization.

The key weaknesses of the model were exposed in the classification of minority classes, demonstrating the difficulty of classifying rare genetic profiles. Table 3 below summarizes the detailed per-haplogroup performance on the holdout set.

The disparity between rare haplogroups (e.g., the perfect performance of F, G, and H versus the low recall of L and I) is attributed to interactions within the class inherent to the nature of the data. Haplogroups like F, G, and H, despite being minority classes in the dataset, likely possess highly distinct Y-STR signatures that do not overlap with majority groups (like O or R). This 'genetic distance' allows the model to draw clear decision boundaries even with limited data. Conversely, Haplogroup L exhibits low recall (0.20)

Haplogroup	Accuracy	Precision	Recall	F1-Score
O	0.97	0.97	0.99	0.98
C	0.99	0.97	0.96	0.96
R	1.00	1.00	1.00	1.00
D	0.99	0.97	0.94	0.96
N	0.99	0.97	0.94	0.96
Q	0.99	0.91	0.91	0.91
J	0.99	1.00	0.96	0.98
F	1.00	1.00	1.00	1.00
L	0.99	1.00	0.20	0.33
G	1.00	1.00	1.00	1.00
H	1.00	1.00	1.00	1.00
I	0.99	1.00	0.67	0.80

Table 2: **Performance of Best-Performing Model per Haplogroup.** Summary of the model's consistent high performance across most haplogroups, with F1-scores near or equal to 1.00 for the majority classes. Conversely, performance drops sharply for haplogroups with extremely low sample counts highlighting the direct impact of class imbalance on minority-lineage predictability.

because its Y-STR profiles may closely resemble those of a majority class. When the model encounters a profile that could be either 'Common Group O' or 'Rare Group L,' it defaults to the majority class to minimize overall loss, leading to high precision but poor recall.

The per-class accuracy is the overall accuracy of the model when treating that specific class as the positive class, and all other classes as the negative class. This metric is important because it shows the model is generally excellent at distinguishing a specific haplogroup from all others.

The decision to maximize Weighted F1 (0.9810) and Accuracy is an intentional operational choice based on the ethical requirements of the justice system. In qualifying investigative leads or forensic intelligence, the error profile would ultimately influence the fate of a human being. To contextualize these metrics in the justice system, a false Positive, or an incorrect prediction that falsely associates a crime scene profile with a specific haplogroup translates to wrongly accusing an innocent individual just because he belongs to the group of interest. On the other hand, failure to correctly classify a profile in the case of a false negative could ultimately result in acquitting the guilty and/or losing the only investigative lead that could drive the case entirely.

The Need for Legal Transparency

The core tenet of Western law is the Blackstone Ratio, which holds that it is "better that ten guilty persons escape than that one innocent party suffer." Our model's objective is to satisfy this ratio by aggressively minimizing False Positives. The high Weighted F1 Score confirms the model's reliability on the most common haplogroups thereby avoiding spurious leads and satisfying the mandate to protect the innocent.

In the justice system, the output of any computational model must be transparent, hence "black box" algorithms are functionally inadmissible due to the fundamental concerns on accountability and the right to confrontation (Mittelstadt 2019). The ability to audit the model's decision-making process is mandatory for maintaining judicial trust and public acceptance (Lundberg and Lee 2017). To meet this requirement, SHAP (SHapley Additive exPlanations) analysis were implemented on the validated Random Forest model.

SHAP values provide both a global overview of feature influence and a local, per-prediction explanation. This technique quantifies the contribution of each of the 27 Y-STR markers to the final haplogroup prediction.

Figure 4 shows a summary of the global SHAP values of the genetic loci used in predicting the haplogroup. This revealed that Y-STR markers such as DYS392, DYS481 and DYS635 were the most influential in the model's overall classification logic. This means that these are the markers that had the largest overall impact, or highest average absolute SHAP value, on the model's predictions across the entire dataset, giving them the most weight when deciding on the haplogroup.

Furthermore, analyzing markers with intermediate global importance, such as DYS392, DYS481, and DYS635, reveals that while they may not be the primary discriminators, they possess crucial power for distinguishing specific haplogroups. DYS392 and DYS481 often exhibit allele length variance that strongly correlates with minor clades, and DYS635 is known for its high heterozygosity. ((Willuweit and Roewer 2015)).

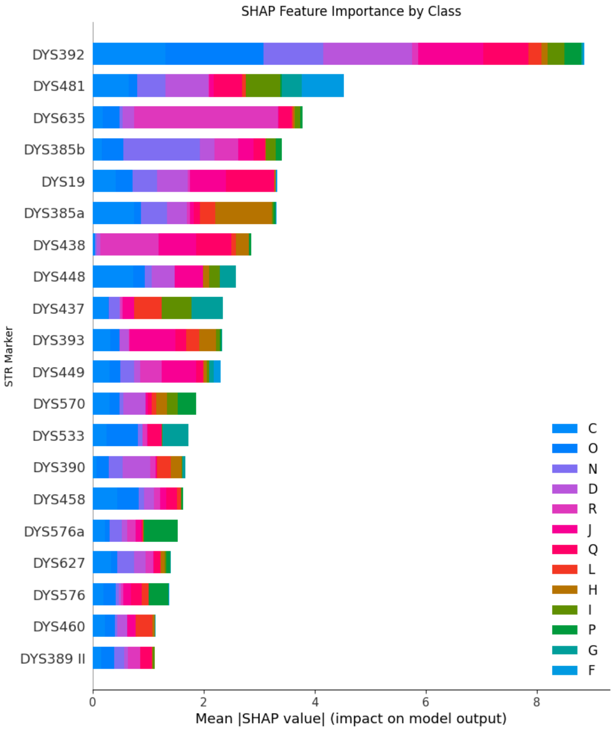

Figure 3: **Summary of Mean SHAP Values per Y-STR marker.** For the most important marker is DYS392, almost doubled compared to the next marker, DYS481. This highlights the magnitude of dataset imbalance. In the justice setting, interpreting such important predictors could qualify the credibility of an investigative lead.

Generating a local SHAP plot for every novel Y-STR crime scene profile could describe the exact positive and negative contribution of each of the 27 markers to the predicted haplogroup. This output would allow a forensic expert to articulate the basis of the investigative lead which can be operationalized through an addendum through the Y-STR DNA profile. The ability to articulate that the model predicted that a certain STR profile belonged to a specific haplogroup because the allele length for the loci was this score could strongly push the probability away from another Haplogroup. This level of documented transparency is critical for internal audits, external review, and maintaining judicial and public trust in an AI-driven forensic intelligence framework.

Conclusion

The current operational challenges in forensic investigations necessitate a strategic shift toward AI-driven intelligence generation. This research successfully develops a machine learning framework that accurately maps Y-STR profiles to ancestral haplogroups, achieving high reliability (97% accuracy). Most significantly, the methodology is engineered to consider specific interventions for fairness (class weighting to protect minority groups) and accountability (SHAP analysis for total transparency). By adhering to the principles of trustworthy AI, this framework not only provides law enforcement with a powerful, demographically sensitive investigative tool but also offers a model for responsible AI integration across the broader public safety and justice domain.

Limitations and Future Work

The present model was implemented using the YHP dataset, which has a primary geographical representation (Song et al. 2024). This introduces a necessary caution regarding geographical bias; the model may perform suboptimally when presented with data from populations that are genetically distant from the training set.

Future deployment of the model must include rigorous retraining and validation using local reference data for any new geopolitical region to prevent systemic bias in investigative lead generation. Furthermore, clear legal and policy frameworks must be established to define the ethical boundaries for the use and retention of haplogroup information generated, ensuring it remains strictly an investigative lead and not a piece of individualizing evidence.

Predicting major haplogroups could provide a substantial investigative lead, especially for populations of diverse ethnicity groups such as the Philippines. However, the pursuit to finer resolution prediction will always be desired. The model can be extended to predict sub-clades (more granular haplogroups), which would further narrow the suspect pool. This technical advancement necessitates incorporating a larger feature set that includes highly discriminating Y-SNP markers.

A critical limitation in high-stakes environments is the lack of explicit uncertainty of quantification. The current model provides a point of prediction and a probability score, but future work must integrate rigorous uncertainty measures possibly through Bayesian machine learning or Conformal Prediction to provide a statistically sound confidence interval alongside the haplogroup prediction. This would be critical for investigators to appropriately weigh the evidentiary value of the generative lead in their overall strategy.

Acknowledgments

The authors sincerely thank Professor Christopher Monterola, School Head of the Aboitiz School of Innovation, Tehcnology and Entrepreneurship of Asian Institute of Management, for his leadership, guidance, and steadfast support of interdisciplinary research at the intersection of data science and public service. His encouragement and institutional support were instrumental in the completion of this work.

The authors are also grateful to Professor Kenneth T. Co and Jose Miguel G. Bautista for guiding the student authors through the formulation of the research topic and the completion of the initial project submitted for their Master of Science in Data Science course in Machine Learning I. The authors also acknowledge the Master of Science in Data Science 2026 Cohort for the shared learning journey, intellectual exchange, and mutual support that fostered growth and perseverance throughout this endeavor.

The authors respectfully recognize Police Colonel Pierre Paul F Carpio, Officer-in-Charge of the PNP Forensic Group; Police Lieutenant Colonel Edmar A Dela Torre, Chief of the DNA Laboratory Division; Police Lieutenant Colonel Jasper P Magana, Assistant Chief, DLD, Police Major Rey-Ann L Valencia; and Police Lieutenant Colonel Eleazar P Macaraeg for their valuable support, professional insights, and commitment to advancing forensic science and evidence-based policing, which significantly informed the practical contributions of this work in the field. The same commendation is extended to the men and women of the Philippine National Police Forensic Group DNA Laboratory Division.

Finally, this work is dedicated to the primary author's wife, Mrs. Mary Claire Leoncio Moreno, RPm, and daughter, Ms. Emma Haraya Celyne Leoncio Moreno whose love and patience fueled the completion of this work.

References

Babić Jordamović, N.; Kojović, T.; Dogan, S.; Bešić, L.; Salihefendić, L.; Konjhodžić, R.; Škaro, V.; Projić, P.; Hadžiavdić, V.; Ašić, A.; et al. 2021. Haplogroup prediction using Y-chromosomal short tandem repeats in the general population of Bosnia and Herzegovina. *Frontiers in genetics*, 12: 671467.

Costa, R.; Fadoni, J.; Amorim, A.; and Cainé, L. 2025. Y-STR Databases—Application in Sexual Crimes. *Genes*, 16(5): 484.

Fan, G.-Y.; Jiang, D.-Z.; Jiang, Y.-H.; Song, W.; He, Y.-Y.; and Wuo, N. A. 2023. Phylogenetic analyses of 41 Y-STRs and machine learning-based haplogroup prediction in the Qingdao Han population from Shandong Province, Eastern China. *Annals of Human Biology*, 50(1): 35–41.

Kayser, M. 2017. Forensic use of Y-chromosome DNA: a general overview. *Human genetics*, 136(5): 621–635.

Li, Y.; Jiang, L.; Liu, Q.; Chen, B.; Zhuang, B.; Zhao, L.; Han, J.; and Li, C. 2025. Integrated Microfluidic System for Rapid 89-Plex Y-SNP Profiling: Development and Forensic Validation. *Electrophoresis*, 46(20): 1534–1547.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Mittelstadt, B. 2019. Principles alone cannot guarantee ethical AI. *Nature machine intelligence*, 1(11): 501–507.

National Institute of Justice, N. 2002. Using DNA to Solve Cold Cases: Special Report.

Ndomondo, S.; Sharma, P.; Mevada, V.; et al. 2025. Forensic applications of rapidly mutating Y-STR markers: current status and future perspectives. *Forensic Science, Medicine and Pathology*, 1–16.

Rolf, B.; Keil, W.; Brinkmann, B.; Roewer, L.; and Fimmers, R. 2001. Paternity testing using Y-STR haplotypes: assigning a probability for paternity in cases of mutations. *International journal of legal medicine*, 115(1): 12–15.

Song, M.; Zhou, Y.; Zhao, C.; Song, F.; and Hou, Y. 2024. YHP: Y-chromosome Haplogroup Predictor for predicting male lineages based on Y-STRs. *Forensic Science International*, 361: 112113.

Willuweit, S.; and Roewer, L. 2015. The new Y chromosome haplotype reference database. *Forensic Science International: Genetics*, 15: 43–48.

Yin, C.; He, Z.; Wang, Y.; He, X.; Zhang, X.; Xia, M.; Zhai, D.; Chang, K.; Chen, X.; Chen, X.; et al. 2022. Improving the regional Y-STR haplotype resolution utilizing haplogroup-determining Y-SNPs and the application of machine learning in Y-SNP haplogroup prediction in a forensic Y-STR database: A pilot study on male Chinese Yunnan Zhaoyang Han population. *Forensic Science International: Genetics*, 57: 102659.