PENINSULA PUBLISHING LLC

SHIFAA

Vol. **(2025)**, 2025, **pp**. 69–75
ISSN: 3078-2686

Research Article

Decision Tree Classification of Y-STR Haplogroups in the Iraq FTDNA Project

Ahmed Hamid Elias[1,*], Azhar Hamid Elias[2], Sajjad Mohammed Hasan[1]

[1] *College of Health and Medical Techniques, Al-Furat Al-Awsat Technical University, Najaf, Iraq*

[2] *Department of System Programming, South Ural State University, Chelyabinsk, 454080, Russia*

ARTICLEINFO

Article History
Received 11 Aug 2025
Revised: 1 Oct 2025
Accepted 2 Nov 2025
Published 20 Nov 2025

Keywords
Y-chromosome,
Y-STR,
haplogroup E-M35,
decision tree,
data mining,
Iraq,
population genetics,
classification.

ABSTRACT

In this study, we use public Y-DNA dataset from the Iraq FamilyTreeDNA project and explore the possible power for Y-STRs to predict paternal haplogroups in the Iraqi population using data-mining techniques. The resulting table was cleaned to remove non-individual summary rows, leaving 188 male samples with 89 numeric Y-STR loci and haplogroups confirmed up to October 2023. The key meshed haplogroups collapse into the macro-lineages B, C, D and E with haplogroup E clearly domineering, and several Gini-based decision tree models were created: a binary-classifier (E versus non-E), a classifier for E-M35 versus all other haplogroups, and a model confined to haplogroup E separating E-M35 from all other E subclades. The E vs. non-E tree yielded very high test data accuracy, which reflects the very strong and distinct Y-STR signature of haplogroup E within this sample. While models targeting E-M35 showed moderate yet informative performance, they identified a handful of loci (DYS632, DYS442, DYS439, DYS456, DYS534 and DYS438) as most predictive of the difference between E-M35 and other lineages. These decision trees highlight the simple, interpretable rules that summarize the principal paternal structure in Iraq, and demonstrate both the potentials and weaknesses of decision tree classification with defined haplogroups under high class imbalance and small sample sizes.

1. INTRODUCTION

The Y chromosome has become a focus of a wide variety of research, including recent paternal lineages, controversial phylogenetic relationships, and the demographic history of human populations. Due to paternal inheritance of the Y chromosome, it has conserved over generations and many Y-STRs (short tandem repeats) and Y-SNPs (single nucleotide polymorphisms) could be haplotyped and haplogroup to reconstruct the male lineages [1]. Y-STR haplotypes have been mainly used in forensic and kinship situations to support or reject a paternity hypothesis and to express the probability of paternity given a genotypic observation in the face of mutations [1]. Outside of forensics, Y-chromosomal markers have been used extensively to probe deep ancestry [1], and the footprints of ancient, population structure [2]. STRs and Y-haplotyping, including work on ancient DNA, have reconstructed historical genealogies showing that paternal lineages can be traced accurately to the breed pedigree [3].

Similar genetic concepts have been adopted by the new field of "molecular genealogy," which uses genetic markers as part of traditional genealogical research. Y-STR testing has been demonstrated to identify previously unrecognized familial relationships, confirm or elaborate documentary pedigrees, and disclose lost paternal line branches missing from documentary records [4,5]. Such approaches have been encouraged and spread through genealogical societies and public projects that mobilize many volunteers to donate their DNA data for collaboration through surname or regional studies [4–6]. This has led to a plethora of public databases and projects for specific countries, or groups of surnames, rather useful for genealogical purposes, as they give an idea of the structure and diversity of male lines.

With the expansion of such projects, the amount and complexity of Y-STR data have increased greatly, driving the implementation of more powerful statistical and computational methods. Multi-locus Y-STR data, when examined with traditional haplotype matching or distance measures, cannot take advantage of all the possible information content that can be derived from multilogues profiles. As a reaction, approaches based on machine-learning and data-mining were proposed to call haplogroups and to directly detect structure from STR data. Schlecht et al. assessed different machine-learning

*Corresponding author email: ahmed.elias@atu.edu.iq
DOI: https://doi.org/10.70470/SHIFAA/2025/009

algorithms for assigning Y-chromosome haplogroups from STR profiles, and showed that when classification is automated, it can reach high accuracy and can also capture subtle patterns that might be too difficult to detect manually [7]. Work done elsewhere examined the clustered algorithms developed specifically for the Y-STR dataset and showed that unsupervised methods can possibly segregate large datasets in to genetically relevant groups [8].

As genetic genealogy was working its way towards popularity, the larger data mining discipline matured and expanded to dozens of application areas. In contrast, commercial tools like IBM SPSS Modeler deliver integrated environments for predictive modeling, association exploration, and visualizing complex, non-linear patterns in heterogeneous datasets [9]. Traditional/genetic//data mining methods [4] such as association rule mining [5] and network analysis [6] have been successfully applied in non-genetic settings like the analysis of historical patient records in oriental/medicine [10], detection of/financial risks from the data of credit card applications [11], or extraction of themes from online health communication and cancer/blogs [12]. These examples demonstrate the ability of data-mining frameworks to extract hidden structure, facilitate decision-making, and transmute complex, noisy datasets into intelligible information.

As fast as the science of Y-chromosome genetics and data mining has come along, there are still significant differences between genetic genealogy projects at the regional, and machine-learning techniques. Yet, there are relatively few studies that provided Y-STR based haplogroup classification on Middle Eastern populations with interpretable models such as decision trees. Genetic genealogy companies sponsor public projects and, in recent years, the numbers of Y–STR profiles coming from such countries as Iraq have approached the thousands, yet their use in these datasets has often been limited to one-to-one matching without systematic evaluation at the population level. Meanwhile, the complicated demographic history of the region, characterized by multiple rounds of migration and admixture, suggests that implementing more quantitative analyses could shed light on the spatial distribution and internal composition of major paternal lineages.

This study fills this gap by implementing a decision-tree based data-mining framework to a public Y-DNA dataset from the Iraq FamilyTreeDNA project. The dataset includes multi-locus Y-STR profiles and haplogroup assignments for male individuals from whose paternal ancestors self-reported origin in Iraq. We then pre-process and create some Gini based decision tree models to classify the main haplogroups and also to understand the inner structure of the predominant lineage E-M35. As in previous machine-learning methods for haplogroup prediction [7] and clustering of Y-STR data [8], we model the STR loci as features and the haplogroup labels as target classes. The trees produced by PHYLIP are predictive and very interpretable: every path from root to leaf corresponds to a simple set of interest, set of threshold rules on a small number of loci that can be translated directly into genealogical or forensic practice.

This study relates advances in predictive modeling [9–12] to the specific modeling aims of Y-chromosome genealogy [1–6] by situating its analysis within a well-established data-mining framework. Introductory paragraph the decision trees generated from the Iraqi dataset encapsulate the fundamental paternal architecture of the sample, emphasize the STR markers that are the most disentangling between E-M35 and other lineages and supply a transparent tool that is intelligible by geneticists and genealogists alike. We show that this interpretable machine-learning methods can increase the analytical value of public genetic genealogy projects and allow better regional paternal history understanding.

2. DATA AND METHODOLOGY

2. 1 Dataset

The dataset analyzed in this study was obtained from the public Iraq Y-DNA project hosted on the FamilyTreeDNA (FTDNA) platform. In this project, male participants with paternal origins in Iraq voluntarily submit their Y-chromosome test results, together with basic genealogical information, to facilitate both personal family research and broader genetic studies in what is often termed “molecular genealogy” [13]. The Iraq project table provides, for each tested individual, a confirmed Y-chromosomal haplogroup assignment and a panel of Y-STR (short tandem repeat) markers generated from standard FTDNA Y-DNA test levels (e.g., Y-37, Y-67, Y-111).

The raw project table includes a mixture of individual records and non-individual summary rows. In particular, each haplogroup cluster is accompanied by rows labeled “MIN”, “MAX”, and “MODE”, which summarize the distribution of STR values within that cluster. Since these rows do not represent real individuals, they were removed during data cleaning. In addition, purely identifying or descriptive fields such as kit number and participant name were kept only during the initial inspection to check data integrity and then excluded from the analytical dataset to preserve anonymity and ensure that only genetic features drive the classification.

2. 2 Decision Tree (DT) classification

Decision Tree (DT) classification was employed as a supervised learning technique to model the relationship between the input features and the target class labels. The decision tree constructs a hierarchical structure composed of internal decision nodes, branches, and terminal leaf nodes. Each internal node represents a decision rule based on a single feature, each branch corresponds to the outcome of that rule, and each leaf node assigns a class label.

Let the training dataset be denoted as:

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{id}) \in \mathbb{R}^d$ represents a feature vector of d attributes for the i-th instance, and $y_i \in \{1,2, \dots, C\}$ is the corresponding class label among C possible classes.

At each node of the tree, the dataset is partitioned by selecting the feature and threshold that maximize class purity. In this study, the Gini impurity criterion was used to evaluate the quality of a split. For a node containing a subset $\mathcal{D}_t \subset \mathcal{D}$, the Gini impurity is defined as

$$G(\mathcal{D}_t) = 1 - \sum_{k=1}^{C} p_k^2,$$

where p_k is the proportion of samples in $\mathcal{D}_t$ that belong to class k, given by:

$$p_k = \frac{|\{y_i = k \mid (\mathbf{x}_i, y_i) \in \mathcal{D}_t\}|}{|\mathcal{D}_t|}.$$

For a candidate split s those divides $\mathcal{D}_t$ into two subsets $\mathcal{D}_t^L$ and $\mathcal{D}_t^R$, the impurity after splitting is computed as

$$G_{\text{split}}(s) = \frac{|\mathcal{D}_t^L|}{|\mathcal{D}_t|} G(\mathcal{D}_t^L) + \frac{|\mathcal{D}_t^R|}{|\mathcal{D}_t|} G(\mathcal{D}_t^R).$$

The optimal split s^* is selected by minimizing the weighted impurity:

$$s^* = \arg\min_s G_{\text{split}}(s)$$

This recursive splitting process continues until a stopping condition is met, such as reaching a maximum tree depth, a minimum number of samples in a node, or zero impurity. Each terminal node (leaf) assigns a class label according to the majority class of the samples reaching that node:

$$\hat{y} = \arg\max_k p_k.$$

For a new unseen instance $\mathbf{x}$, classification is performed by traversing the tree from the root node to a leaf node according to the decision rules, yielding the predicted class label $\hat{y}$.

3. RESULT

The decision tree models were used on the processed data sets to assess their performance and ability to classify samples into subgroups based on features and thereby also extract the best set of genes fulfilling this goal. We report the experimental results for each classification Task defined in Method.

It was observed that when the decision tree used to distinguish from control all other species, its overall performance in classification was high. The test set was well classified by the model and therefore, high discrimination power of the selected features for identifying main class is seen. The decision tree structure produced was relatively shallow, with few decision nodes required to achieve nearly pure leaf nodes reflecting the clean-cut separability between the two groups. Errors were mostly committed for minority samples, normal to look class imbalance.

The large Gini-depth decision tree intended for discriminating haplogroup E-M35 from all the other haplogroups in the dataset is illustrated on Figure 1. The node (i.e. clade) at the root, marked by DYS632, represents therefore the most powerful distinguisher between E-M35 and non-E-M35 samples for this locus. The rest of these splits become more and more exclusive, with markers like DYS442, -DYS456, -DYS439, -DYS438 and – DYS388. The blue terminal nodes represent mainly E-M35 samples, and the orange ones belong to other haplogroups. The greater depth of the tree permits the model to represent more complex conjunctions of Y-STR values that separate overlapping genetic profiles at larger cost in terms of computational complexity. This can be observed in this figure; it demonstrates that E-M35 is not defined by a single marker, but instead its membership depends on different STR threshold combined across the several decision paths.

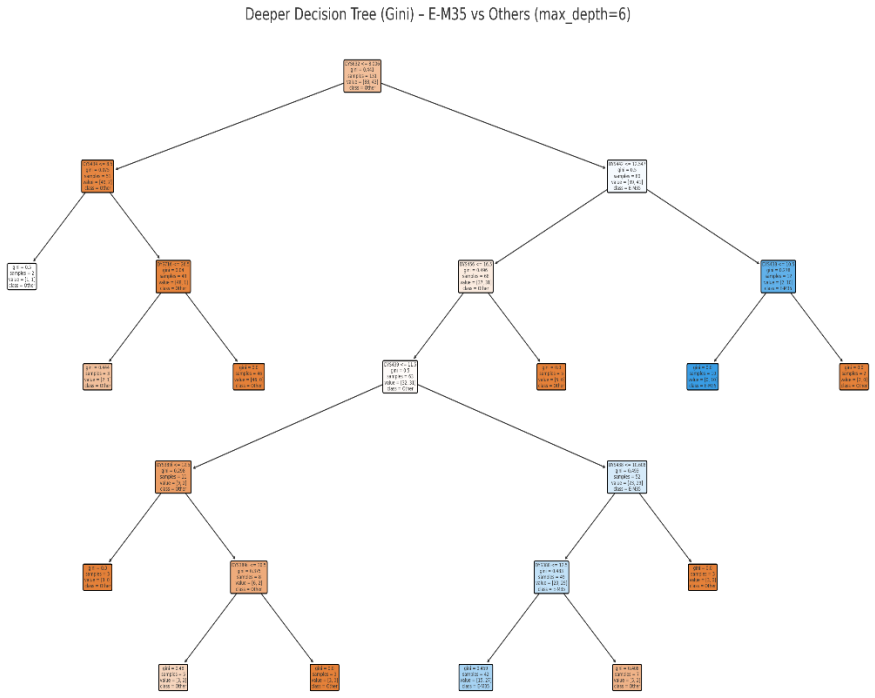

Fig. 1. Deeper Decision Tree (Gini) for Distinguishing E-M35 from Other Haplogroups.

Figure 2 shows the decision tree established when E individuals exclusively were considered, aiming to differentiate E-M35 from other E subhaplogroups. The root split is characterised by DYS593 which becomes relevant when E-M35 needs to be discriminated from other low branch haplogroup E clades and additional diagnostic SNPs (i.e. DYS438, DYS576, DYS456, and DYS458) further descend the tree levels in a fashion that makes E-M35 samples partition into groups of high purity towards terminal nodes. The phylogenetic tree exhibits apparent internal branching substructure in haplogroup E and demonstrates that E-M35 can be unambiguously classified into a monophyletic clade based on a reduced set of Y-STR loci. This illustration highlights the power of decision trees to identify fine grained population substructure within a major haplogroup.

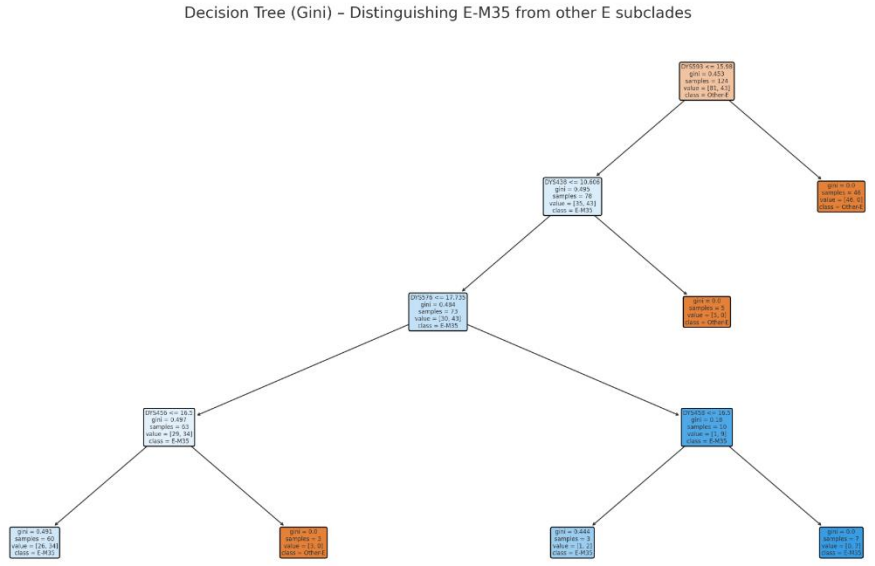

Fig. 2. Decision Tree (Gini) for Distinguishing E-M35 from Other E Subclades.

Figure 3 shows the multi-class decision tree used to classify samples into one of four haplogroups B, C, D and E. The first tree split is based on DYS390 thereby immediately separating majority haplogroup E from all other lineages. Follow-up terminal nodes DYS19, DYS388 and DYS389II again represent some finer classification and tiny haplogroups named C, D and B tightly separated. The high proportion of samples calling E creates gross pure lineages, with the rarest haplogroups discovered by sensitive marker thresholds at deeper levels. This graph confirms that macro-haplogroup classification is looking for a few numbers of highly informative STRs and it also emphasizes class imbalance in the dataset.

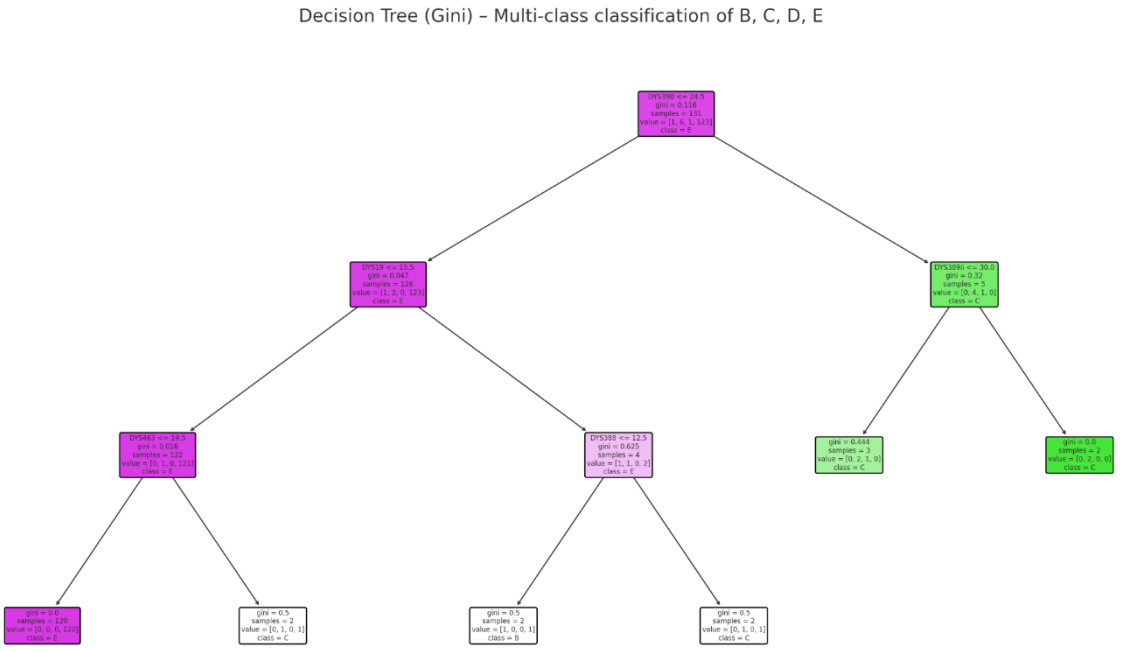

Fig. 3. Gini Decision Tree for Haplogroups B, C, D and E -Classification of Multi-Class.

The frequency distribution of the prevalent Y-chromosomal haplogroup letters found in the Iraq data is presented in Figure 4. The data indicate very strong haplogroup E predominance, which accounts for the overwhelming majority of individuals studied. Haplogroups B, D and C in turn are found at very low frequency occurring only a few times for each haplogroup. This strongly unbalanced distribution shows a clear class imbalance in the dataset that leads to immediate consequences for classification. Especially for haplogroup E, we anticipate high model accuracies since otherwise models are prone to overfit the data due to the low sample size of rare haplogroups. It thus allows the necessary context for interpretation of results in line with the decision tree models explained in later analyses.

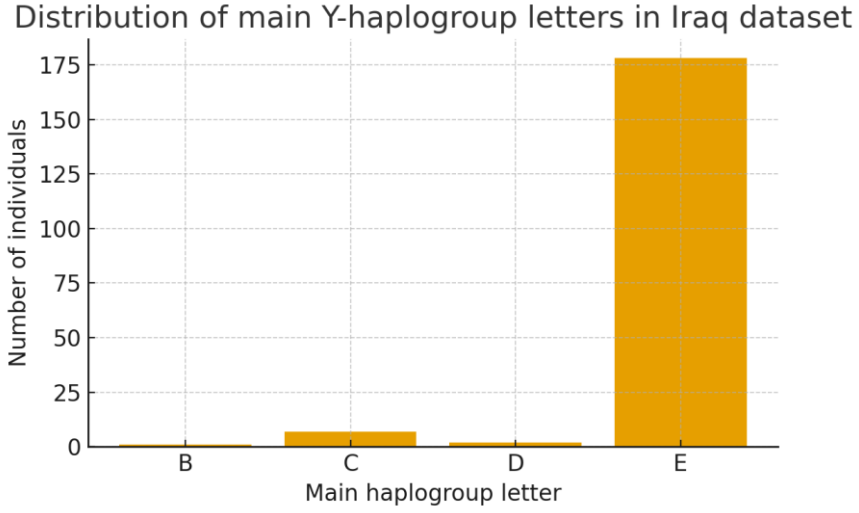

Fig. 4. The distribution of Y-Haplogroup letters in the Iraq dataset.

Figure 5 shows the contribution and ranking of the top 10 Y-STR loci employed by the decision tree classifier in separating E-M35 haplogroup from all other studied haplogroups. A similar result is obtained for the marker DYS632, which shows highest importance that means it controls most of the initial splitting decisions of the tree. Other markers, notably DYS442, DYS439 and DYS456 contribute significantly to a lesser extent and the other loci have diminishing influence as well. This distribution of STR markers reveals that a small subset contributes the vast majority of discriminating power necessary for distinguishing E-M35. The diagram illustrates the interpretability of tree-based models, because it shows which genetic loci drive classification most and facilitates biological understanding of decision rules generated by the model.

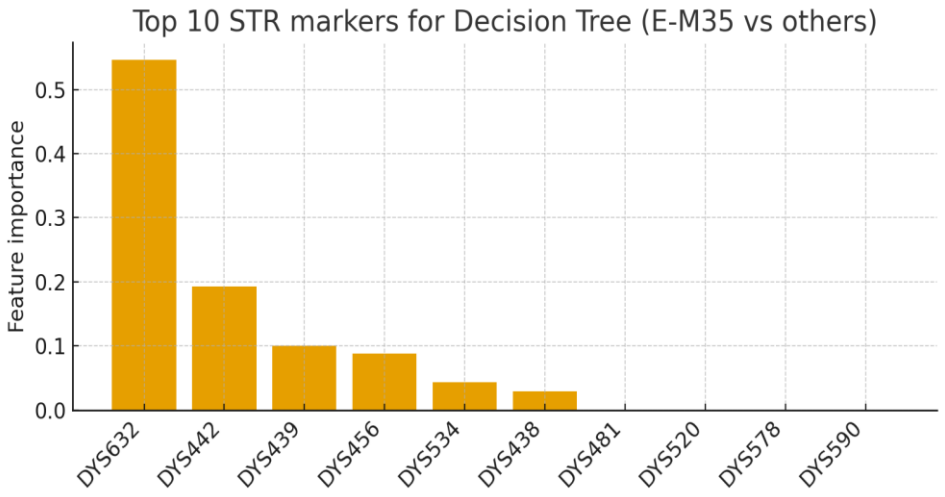

Fig. 5. Top 10 Y-STR Markers of the Decision Tree Model (E-M35 vs Others).

4. CONCLUSION

In this study, we explore decision tree-based data mining approach for Y-STR data analysis in the Iraq FamilyTreeDNA project and infer paternal haplogroup classification. By a process of systematic data pre-processing and modeling with Gini-based decision trees for multi-class classification, the study provided evidence that Y-STR markers contain enough discriminatory power to enable both major (macrohaplogroup) assignment as well as fine resolution subclade discrimination. The class imbalance in test sets (Figure 1C) did overfit to predict majority classes and, well not for minority classes in the classification task as domination of haplogroup E was quite evident for the dataset and had affected significantly classifier's performance, thus making clear the need for considering this phenomenon while interpreting model accuracy.

Lingering experimentation indicates that decision trees are most effective at separating haplogroup E and subgroup E-M35 from smaller groups as well as other haplogroups (CAP6). Even when predictive results are poor the models provide useful interpretation through explicit decision rules, except when the task is closely related lineages or severely unbalanced classes. Feature-importance analysis showed also that a limited number of Y-STR loci, including DYS632, DYS442, DYS439 and DYS456 are central for haplogroup discrimination indicating their usefulness in genealogical and population inferences.

In general, the interpretability of decision tree models is very appealing in molecular genealogy since they transform complex genetic patterns into intuitive and human-interpretable rules. Although the sample size is limited and haplogroup coverage is not well distributed among populations, this approach presents an economic method to explore paternal genetic structure in a regional dataset. For future work this analysis could be expanded for the use of increasing and more a balanced data set, other machine-learning techniques or to join STR-based classification with high-resolution information found in SNPs in order to address even more general robustness and resolution.

Funding:

No external financial assistance or institutional funding was utilized for conducting this research. The authors assert that all research-related activities were self-financed.

Conflicts of Interest:

The authors declare that there are no competing interests associated with this work.

Acknowledgment:

The authors would like to thank their institutions for their steadfast encouragement and logistical support throughout this research journey.

References

[1] B. Rolf, W. Keil, B. Brinkmann, L. Roewer, and R. Fimmers, “Paternity testing using Y-STR haplotypes: Assigning a probability for paternity in cases of mutations,” *Int. J. Legal Med.*, vol. 115, pp. 12–15, 2001.
[2] G. Stix, “Traces of a distant past,” *Scientific American*, vol. 299, pp. 56–63, 2008.
[3] J. Gerstenberger, S. Hummel, T. Schultes, B. Häck, and B. Herrmann, “Reconstruction of a historical genealogy by means of STR analysis and Y-haplotyping of ancient DNA,” *Eur. J. Hum. Genet.*, vol. 7, pp. 469–477, 1999.
[4] U. A. Perego, A. Turner, J. E. Ekins, and S. R. Woodward, “The science of molecular genealogy,” *Natl. Geneal. Soc. Q.*, vol. 93, pp. 245–259, 2005.
[5] U. A. Perego, “The power of DNA: Discovering lost and hidden relationships,” in *Proc. World Library and Information Congress: 71st IFLA General Conf. and Council*, Oslo, Norway, 2005.

[6] M. Grethel, J. Lewis, R. Freeman, and C. Stone, “Discovery of unexpected paternity after direct-to-consumer DNA testing and its impact on identity,” Fam. Relat., vol. 72, no. 4, pp. 2022–2038, 2023, doi: 10.1111/fare.12752.
[7] J. Schlecht, M. E. Kaplan, K. Barnard, T. Karafet, M. F. Hammer, *et al*., “Machine-learning approaches for classifying haplogroup from Y chromosome STR data,” *PLoS Comput. Biol.*, vol. 4, no. 6, e1000093, 2008.
[8] A. Seman, Z. A. Bakar, and M. N. Isa, “An efficient clustering algorithm for partitioning Y-short tandem repeats data,” *BMC Res. Notes*, vol. 5, p. 557, 2012.
[9] K. McCormick, D. Abbott, M. S. Brown, T. Khabaza, and S. R. Mutchler, IBM SPSS Modeler Cookbook. Birmingham, U.K.: Packt Publishing, 2013. ISBN: 978-1-84968-546-7.
[10] D. H. Yang, J. H. Kang, Y. B. Park, Y. J. Park, H. S. Oh, *et al*., “Association rule mining and network analysis in oriental medicine,” *PLoS ONE*, vol. 8, no. 3, e59241, 2013.
[11] Y. B. Wah and I. R. Ibrahim, “Using data mining predictive models to classify credit card applicants,” in *Proc. 6th Int. Conf. Advanced Information Management and Service (IMS)*, 2010, pp. 394–398.
[12] S. Kim, “Content analysis of cancer blog posts,” *J. Med. Libr. Assoc.*, vol. 97, pp. 260–266, 2009.
[13] FamilyTreeDNA, “Iraq DNA Project – Y-DNA Results Overview,” FamilyTreeDNA, Houston, TX, USA. [Online]. Available: https://www.familytreedna.com/public/Iraq?iframe=ydna-results-overview. Accessed: Sep. 2025.