Proceedings of ICBioMed 2025 Symposium: AI for Healthcare: Advanced Medical Data Analytics and Smart Rehabilitation
DOI: 10.54254/2753-8818/2025.AU27524

A Study on Hybrid STR Profile Analysis Model Based on Distance-Weighted Self-Attention Mechanism

Zijin Su

School of Mathematics, Nankai University, Tianjin, China
suzijin040318@126.com

Abstract. Forensic DNA analysis is a critical component of judicial practice, where identifying individuals from mixed DNA samples poses significant challenges. Short tandem repeat (STR) analysis serves as the core technique in forensic DNA identification. However, peak height overlap, allele loss, and background noise in mixed samples lead to insufficient identifiability and cumulative errors in traditional probability models during subject identification and proportion estimation. This study aims to construct a deep learning-based cascade analysis framework to systematically enhance the resolution accuracy and robustness of mixed STR profiles. This paper proposes a multimodal fusion algorithm based on the Distance-Weighted Self-Attention (DWSA) mechanism. To address the problem of inferring the number of contributors, a dual-branch network architecture is designed. By integrating sequence features and macro-statistical features, the DWSA mechanism captures spatial correlations among alleles. For contributor proportion estimation, a nonlinear distance-decaying self-attention network is constructed. Jensen-Shannon divergence is employed as the loss function, and multi-locus outputs are integrated in logit space. Experimental results demonstrate that the proposed model achieves 100% accuracy in predicting the number of individuals across 50 test samples. The overall average ensemble MAE for proportion prediction is 0.0809, significantly outperforming traditional methods. This study provides a novel solution for analyzing forensic DNA admixture samples, holding significant theoretical implications and practical application value.

Keywords: Hybrid STR analysis, Distance-weighted self-attention, Multi-model fusion

1. Introduction

Forensic DNA analysis plays a pivotal role in crime scene investigations, with Short Tandem Repeat (STR) analysis at its core [1]. STRs are referred to as the Deoxyribonucleic acid(DNA) fingerprint of cells due to the individual polymorphism in the number of repeats within their core sequence. Within STR profiles, allele size and peak height respectively reflect DNA fragment length and abundance. Each diploid individual carries two alleles at autosomal loci, conferring individual specificity to genotypes [2].

When multiple individuals' DNA is present in biological evidence, mixed STR profiles exhibit complex allele superposition patterns accompanied by interference such as random losses, random gains, and peak height imbalances [3]. Traditional probabilistic models face challenges of insufficient resolvability and error accumulation when handling these complex scenarios, necessitating the development of new computational methods to enhance the analytical capability of mixed STR profiles. Recent studies have explored machine learning approaches to improve mixture interpretation, yet challenges remain in handling high-order mixtures and low-template samples [4]. Deep learning frameworks, particularly those incorporating biological constraints, show promising potential for robust forensic DNA profile analysis [5].

This study proposes a deep learning model that integrates biological prior knowledge to tackle the above problem. The main contributions of this study are: 1) Integrating biological prior knowledge with deep learning techniques to design an attention mechanism tailored to STR profile characteristics; 2) Proposing effective multimodal fusion and ensemble learning strategies to enhance model performance in complex scenarios; 3) Providing novel technical approaches and solutions for forensic DNA mixture analysis.

2. Materials and methods

2.1. Data sources and preprocessing

This study utilized mixed STR profile data measured in forensic laboratories, encompassing detection results for 16 autosomal STR loci (D3S1358, D13S317, D21S11, vWA, etc.). Sample types encompassed DNA mixtures from 2 to 5 individuals. Raw data for each sample included key characteristics such as fragment size (Size, unit: bp) and peak height (Height, unit: RFU) for each allele.

Data preprocessing includes the following steps:

a. Dynamic threshold filtering and valid peak identification: Employing dynamic threshold analysis

$$\mathrm{T}_h = \max\left(20\mathrm{RFU},\ \mathrm{Percentile}(\{h_\mathrm{k}\}, 8)\right) \tag{1}$$

Distinguish between genuine allele signals and background noise.

Only peaks exceeding the specified height are considered “valid peaks.” Participate in subsequent analysis and feature calculation.

b. Normalization: To eliminate variations in amplification efficiency and detection sensitivity across different loci, all sequence features (Size, Height) and numerical auxiliary features undergo minimum-maximum scaling, independently normalized to the range [0, 1].

$$\overline{\mathrm{x}} = \frac{\mathrm{x}-\mathrm{x}_{\mathrm{min}}}{\mathrm{x}_{\mathrm{max}}-\mathrm{x}_{\mathrm{min}}} \tag{2}$$

c.Sequence Padding and Mask Construction: Since the number of effective alleles detected

$$\mathrm{x}_{\mathrm{ij}} = \left\{\left(\overline{\mathrm{s}}^{\mathrm{k}}_{\mathrm{ij}}, \overline{h}^{\mathrm{k}}_{\mathrm{ij}}\right)\right\}_{\mathrm{k}=1}^{\mathrm{n}_{\mathrm{ij}}^{\mathrm{eff}}} \tag{3}$$

At the same locus varies across samples, each sample's (Size, Height) pairs are padded to a uniform maximum length using zero vectors. Concurrently, a Boolean mask sequence is generated to explicitly distinguish between effective data and padding positions within the sequence.

2.2. Model architecture

2.2.1. Distance-Weighted Self-Attention mechanism (DWSA)

Self-attention mechanisms can capture global dependencies within sequences, but their standard form does not account for biological prior knowledge—namely, that the strength of associations between alleles in STR maps is closely correlated with their physical distance. To address this, this paper proposes the Distance-Weighted Self-Attention Mechanism (DWSA).

First, the input height features are projected onto the query, key, and value spaces via a linear transformation:

$$Q = X_h W_Q(4),\ K = X_h W_k(5),\ V = X_h W_V \tag{4}$$

Here, X_h denotes the height feature column within the input sequence, while W_Q , W_k , W_V represent learnable weight matrices.

Subsequently, a distance penalty term based on normalized size is introduced during attention weight computation:

$$\mathrm{Logits}_{kl} = \frac{q_k \cdot k_l^T}{\sqrt{d_k}} - \lambda \left| \bar{s}_k - \bar{s}_l \right| \tag{5}$$

Where q_k and k_l are the query and key vectors for the k-th and l-th alleles, respectively, d_k denotes the dimension of the key vector, $\bar{s}_k$ and $\bar{s}_l$ represent the normalized fragment sizes of the k-th and l-th alleles, respectively, and λ is a learnable or predefined distance decay coefficient that controls the strength of the distance penalty.

Finally, the attention weights α_{kl} and the output O are computed as follows:

$$\alpha_{kl} = \frac{\exp(\mathrm{Logits}_{kl})}{\sum_{m=1}^{L_{max}} \exp(\mathrm{Logits}_{km})} \tag{6}$$

$$\mathrm{Output} = \sum_{l=1}^{L_{max}} \alpha_{kl} v_l \tag{7}$$

The biological significance of this mechanism lies in the fact that alleles in close physical proximity (such as the main peak and its stutter peak) are more likely to originate from the same biological process. DWSA explicitly reinforces this local association through its penalty term, thereby reducing interference from unrelated distant peaks and enabling the model to identify key patterns with greater precision.

2.2.2. Multimodal dual-branch model architecture

For contributor inference, the model employs a dual-branch architecture to fuse multimodal information:

Sequence Branch: Processes normalized and padded (Size, Height) sequence data $X_{ij} \in R^{L_{max} \times 2}$. Its core component is the DWSA layer, designed to capture complex spatial interaction patterns between alleles.

Auxiliary Branch: Processes normalized macro-statistical feature vectors $Z_{ij} \in R^{D_{aux}}$. These features include effective allele size, total peak height, mean peak height, peak height standard deviation, peak height skewness, kurtosis, size difference between adjacent alleles, and peak height ratio, describing sample characteristics from a global perspective.

Fusion Strategy: Features from both branches undergo independent high-level feature extraction via multi-layer perceptrons (MLPs). The extracted feature vectors are concatenated, and a fully connected layer outputs the prediction (theoretical effective allele number).

For contributor proportion estimation, the model input is expanded into four modalities: normalized peak height sequence, raw Size sequence (retaining absolute scale), validity mask, and the number of contributors N provided by Problem 1 (injected as a conditional input). The core mechanism employs an improved nonlinear distance-decaying self-attention mechanism (using the exponential decay function $\exp\left(-\lambda D_{ij}\right)$, while also integrating macro-statistical features. The final output is a vector of logits without Softmax normalization.

2.3. Model training and evaluation

2.3.1. Training details

Contributor Number Inference: The learning objective is to predict the theoretical number of alleles $2\times\ p_{true}$. The loss function employs Mean Absolute Error (MAE):

$$\mathscr{L}_{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \widehat{y}_i\right| \tag{8}$$

The optimizer uses Adam with an early stopping strategy to prevent overfitting.

Contributor Proportion Estimation: The learning objective is to predict the proportion distribution of contributors. The loss function employs Jensen-Shannon Divergence (JSD) to stably measure differences between probability distributions:

$$JSD(P||Q) = \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M)(11),\ M = \frac{1}{2}(P+Q) \tag{9}$$

The loss function implementation includes steps such as dynamically resolving the true population size N, computing only the first N valid proportions, and ensuring numerical stability.

The model trains a separate submodel for each locus to fully leverage locus-specific information.

2.3.2. Evaluation metrics

Contributor Number Inference: Primarily assesses the accuracy of the final number prediction. Additionally, records the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) of each submodel's predicted allele counts to analyze performance variations across loci.

Contributor Proportion Estimation: Primarily evaluates the Mean Absolute Error (MAE) between predicted and actual proportions. Simultaneously analyzes MAE variations across different contributor counts and compares ensemble model performance against independent models.

2.3.3. Ensemble strategy

Estimating the number of contributors: The final predicted allele count $\widehat{n}_{final}$ for a sample is obtained by taking the weighted average of predictions from all locus models. The weight w_j is the reciprocal of the MAE of the corresponding model: $w_j = \frac{1}{MAE_j+\epsilon}$ (10), where ϵ is a small constant added to prevent the denominator from becoming zero. The final number of contributors is obtained as $\widehat{\mathrm{p}} = \left\lceil \frac{\widehat{\mathrm{n}}_{\mathrm{final}}}{2} \right\rceil$ (11).

Contributor proportion estimation: For test samples, the ensemble Logits vector $\mathrm{L}_{\mathrm{ensemble}} = \frac{1}{\mathrm{M}} \sum_{\mathrm{m}=1}^{\mathrm{M}} \mathrm{L}_{\mathrm{m}}(12)$ is computed as the arithmetic mean of Logits vectors from all 16 locus models. The first N elements are then truncated based on the known true population size N, and the final predicted proportion is obtained via Softmax. This ensemble strategy in logit space effectively smooths noise from individual models.

3. Results

3.1. Contributor count prediction results

The model achieved perfect performance on the test set (50 samples). As shown in Table 1, all mixed samples of 2, 3, 4, and 5 contributors were correctly classified, achieving an overall accuracy of 100%.

Table 1. Contributor count prediction confusion matrix

Actual	Prediction Category			
	2	3	4	5
2	15	0	0	0
3	0	16	0	0
4	0	0	11	0
5	0	0	0	9

The performance of individual locus submodels showed significant variation (Table 2). Highly polymorphic loci such as D13S317, D3S1358, and D21S11 demonstrated high prediction accuracy (100%, 98.04%, 98.04%) and low MAE (~0.13–0.16). In contrast, low-polymorphism or highly complex loci like TPOX, D7S820, and D5S818 exhibited relatively lower prediction accuracy (54.9%, 54.9%, 58.8%) and higher MAE (~0.46–0.53). This confirms that different loci contribute varying degrees of information to population size estimation and underscores the necessity of an ensemble strategy.

Table 2. Performance examples of models for different locuses

Marker	MAE	RMSE	Accuracy
D13S317	0.1613	0.2028	1.0000
D3S1358	0.1265	0.1792	0.9804
D21S11	0.1388	0.1833	0.9804
vWA	0.1290	0.1964	0.9608
TH01	0.1810	0.2417	0.9216
AMEL	0.2448	0.2978	0.9020
D8S1179	0.2886	0.3437	0.8824
FGA	0.2526	0.3207	0.8824
D18S51	0.2203	0.2964	0.8627
D16S539	0.2409	0.3241	0.8431
D19S433	0.3272	0.3648	0.8235
CSF1PO	0.3530	0.4194	0.7843
D2S1338	0.3883	0.4450	0.6863
D5S818	0.5032	0.5721	0.5882
D7S820	0.4633	0.5318	0.5490
TPOX	0.5346	0.6190	0.5490

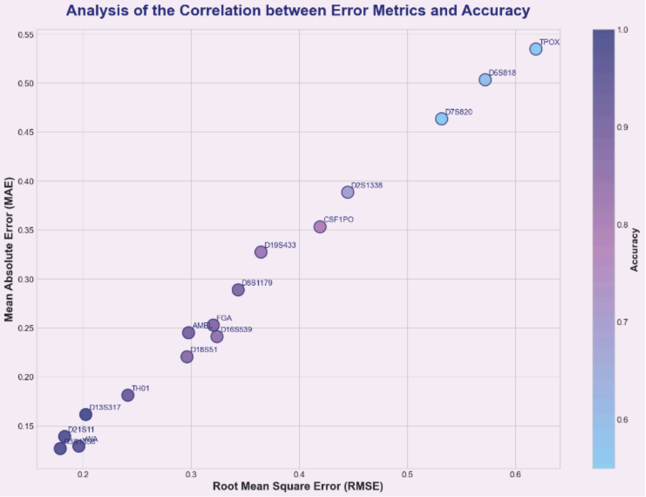

Figure 1. Analysis of the correlation between error metrics and accuracy

The weighted ensemble strategy effectively leverages reliable information from high-performance loci while suppressing noise from low-performance loci, ultimately achieving 100% accuracy. Case studies demonstrate that even for highly challenging samples involving five individuals, the model accurately predicts a theoretical allele count of 10 by integrating predictions from multiple loci, thereby correctly inferring the number of individuals as five, as shown in Figure 1.

3.2. Contributor proportion prediction results

The proportion prediction model also demonstrated excellent and robust performance. The overall average ensemble MAE across all 50 test samples was 0.0809, indicating an average absolute deviation of approximately 8.09% between predicted and actual proportions.

Model performance did not decline monotonically with increasing sample size (Table 3). For the more complex 4-person and 5-person mixed samples, the average MAE (0.0768, 0.0673) was lower than that for 2-person and 3-person samples (0.0898, 0.1002), with smaller standard deviations. This indicates that the ensemble strategy demonstrates greater advantages when handling high-complexity mixtures, effectively integrating multi-locus information to resist noise.

Table 3. Average ensemble MAE for different contributor numbers

Contributors(N)	Samplesize	AverageEnsembleMAE	MAEstandarddeviation
2	6	0.0898	0.0934
3	12	0.1002	0.0533
4	16	0.0768	0.0469
5	16	0.0673	0.0298

The performance of the ensemble model (MAE=0.0809) significantly outperformed all 16 independent locus models (whose MAEs ranged from 0.0886 to 0.0988) (Table 4). This demonstrates that multi-locus integration in Logits space effectively smooths biases arising from data quality issues and allele loss in individual loci, thereby enhancing prediction accuracy and robustness.

Table 4. Example of MAE comparison between independent and integrated models

Model	Average MAE	Model	Average MAE
MAE(AMEL)	0.0925	MAE(D5S818)	0.0922
MAE(CSF1PO)	0.0988	MAE(D7S820)	0.0886
MAE(D13S317)	0.0940	MAE(D8S1179)	0.0970
MAE(D16S539)	0.0986	MAE(FGA)	0.0954
MAE(D18S51)	0.0985	MAE(TH01)	0.0976
MAE(D19S433)	0.0962	MAE(TPOX)	0.0932
MAE(D21S11)	0.0924	MAE(vWA)	0.0937
MAE(D2S1338)	0.0975	Integrated Model	0.0809
MAE(D3S1358)	0.0955		

The predicted proportions [0.1008, 0.4747, 0.4245] for the typical sample (ID:9, 3-person mixed) closely match the actual proportions [0.1111, 0.4444, 0.4444], with an ensemble MAE as low as 0.0202, demonstrating the model's high accuracy under ideal conditions.

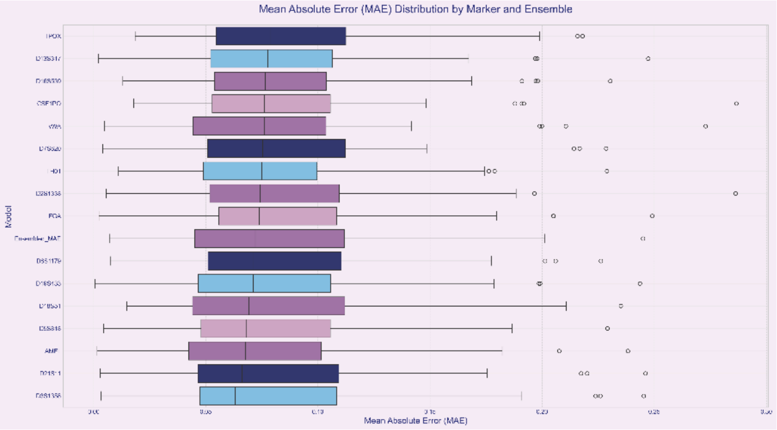

Figure 2. Mean absolute error distribution by marker and ensemble

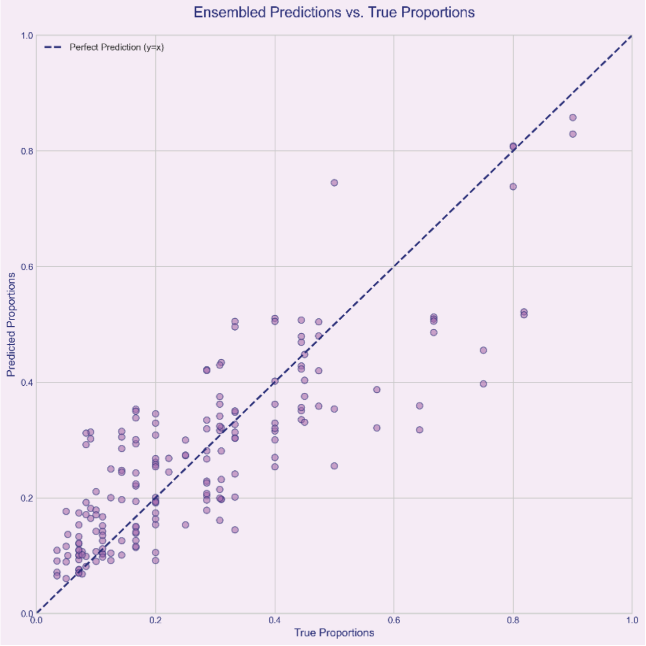

Figure 3. Ensembled predictions vs. true proportions

This result provides strong evidence that our ensemble strategy is successful. Ensemble learning effectively smooths out prediction biases in individual models caused by data noise, allele loss, or other random factors by synthesizing the "opinions" of multiple models, as shown in Figure 2. Even when data quality at a single locus leads to inaccurate predictions in its independent model, information from other loci can correct this bias. Consequently, the final ensemble result more closely approximates the true value and exhibits greater robustnes, as shown in Figure 3.

4. Model analysis

The cascade model constructed in this study demonstrates significant advantages in hybrid STR profile analysis:

First, the successful application of the distance-weighted self-attention mechanism (DWSA) represents one of the core innovations enabling the model's high accuracy. This mechanism successfully encodes the key biological prior that physically proximate alleles exhibit stronger associations into the model. By introducing distance penalty terms (linear or nonlinear decay), the model precisely captures local biological features such as stutter peaks, significantly enhancing its ability to recognize complex spatial interaction patterns between alleles. This attention mechanism design grounded in biological prior offers new insights for deep learning applications in specialized domains.

Second, the multimodal fusion design enables the model to simultaneously leverage micro-level sequence patterns (Size, Height sequences and their interactions) and macro-level statistical features (peak height distribution, effective gene number, etc.). This fusion provides a more comprehensive, multidimensional information view, enhancing the model's understanding and robustness toward complex, variable data. Particularly in applications like forensic DNA analysis where data quality varies significantly, the complementarity of multi-source information is crucial.

Third, the ingenious application of ensemble learning strategies effectively addresses the disparity in information content and reliability across loci. Weighted ensemble (for individual prediction) and logit-space averaging (for proportion prediction) allow the model to synthesize multiple information sources, automatically assigning higher weights to reliable loci. This approach effectively mitigates the impact of low-quality data or noise interference, ensuring the stability and accuracy of final decisions. This strategy demonstrates significant advantages when handling challenging samples.

Finally, the “number→proportion” cascade analysis framework aligns with forensic experts' logical interpretation process. The accurate number prior provided in Problem 1 establishes the foundation for precise proportion prediction in Problem 2, forming a systematic solution. This cascade design not only enhances analytical efficiency but also improves result interpretability.

Future research directions include: 1) developing more robust feature extraction methods, particularly for low-polymorphism loci; 2) exploring end-to-end training strategies to reduce error accumulation in cascading frameworks; 3) incorporating additional prior biological knowledge, such as stutter ratios and allele frequencies across different loci; 4) expanding dataset scale, especially by including rare proportion and highly degraded samples, to enhance model generalization.

5. Conclusion

This paper addresses key challenges in forensic DNA mixture analysis by proposing a mixed STR profile parsing model based on distance-weighted self-attention (DWSA). Through multimodal feature fusion and ensemble learning strategies, it systematically resolves the problems of contributor number inference and proportion estimation.

Experimental results demonstrate that the proposed model achieves 100% accuracy in number prediction and an average MAE of 0.0809 in proportion estimation, significantly outperforming traditional methods. The DWSA mechanism effectively captures local associations among alleles, multimodal fusion provides more comprehensive feature representations, and the ensemble learning strategy ensures model robustness.

However, this study still has some limitations. First, model performance is highly dependent on the accuracy of the initial number prediction. Any misjudgment of the number will lead to cumulative errors in subsequent proportion estimation. Second, although the ensemble strategy mitigates the impact of low-polymorphic loci to some extent, the prediction performance for loci such as TPOX and D7S820 still has significant room for improvement. Additionally, the model's performance in handling extreme proportion mixtures (e.g., where the major contributor's proportion exceeds 90%) requires further validation.

Future work will focus on optimizing the model architecture, incorporating additional prior knowledge, and expanding the dataset scale to further enhance the model's performance and practical value in complex scenarios.

References

[1] Butler JM. Advanced topics in forensic DNA typing: methodology. Academic Press; 2011.

[2] Gill P, Sparkes R, Pinchin R, Clayton T, Whitaker J, Buckleton J. Interpreting simple STR mixtures using allele peak areas. Forensic Science International. 1998; 91(1): 41-53.

[3] Taylor D, Bright JA, Buckleton J. The interpretation of single source and mixed DNA profiles. Forensic Science International: Genetics. 2013; 7(5): 516-528.

[4] Škaro V, Primorac D, Projić P, et al. Towards advanced interpretation of STR profiles: Applications of machine learning for the identification of the number of contributors in a DNA sample. Forensic Science International: Genetics. 2021; 54: 102557.

[5] Cheng K, Jin X, Li H, et al. Deep learning model for estimating the number of contributors in DNA mixtures using a hierarchical attention mechanism. Briefings in Bioinformatics. 2022; 23(6): bbac440.