Communications for Statistical Applications and Methods
2024, Vol. 31, No. 3, 279–289

https://doi.org/10.29220/CSAM.2024.31.3.279
Print ISSN 2287-7843 / Online ISSN 2383-4757

Genetic classification of various familial relationships using the stacking ensemble machine learning approaches

Su Jin Jeong[a], Hyo-Jung Lee[b], Soong Deok Lee[c], Ji Eun Park[d], Jae Won Lee[1,d]

[a]Statistics Support Part, Medical Science Research Institute, Kyung Hee University Medical Center, Korea; [b]Product Development HQ, Dong-A ST, Korea;
[c]Department of Forensic Medicine, College of Medicine, Seoul National University, Korea;
[d]Department of Statistics, Korea University, Korea

Abstract

Familial searching is a useful technique in a forensic investigation. Using genetic information, it is possible to identify individuals, determine familial relationships, and obtain racial/ethnic information. The total number of shared alleles (TNSA) and likelihood ratio (LR) methods have traditionally been used, and novel data-mining classification methods have recently been applied here as well. However, it is difficult to apply these methods to identify familial relationships above the third degree (e.g., uncle-nephew and first cousins). Therefore, we propose to apply a stacking ensemble machine learning algorithm to improve the accuracy of familial relationship identification. Using real data analysis, we obtain superior relationship identification results when applying meta-classifiers with a stacking algorithm rather than applying traditional TNSA or LR methods and data mining techniques.

Keywords: familial relationships, Korean family, STR marker, likelihood ratio, genetic classification, machine learning, stacking ensemble model

1. Introduction

The genes used for identification, including paternity identification and racial/ethnic identification, are parts of an entire genetic sequence. Such genes have a partly repeated nucleotide sequence or include a genetic mutation with a single or more base pair difference in the genetic sequence. Usually, these nucleotide sequences used for gene identification are called markers. By comparing and analyzing different DNA polymorphisms, individual identification, familial relationships, and racial/ethnic identification can be performed (Butler and Hill, 2012).

Research has sought to develop DNA markers that can accurately characterize an individual from among numerous others, and this continues to this day. Among the types of DNA markers explored, DNA nucleotide sequences that appear repeatedly at a specific position exhibit differences in the number of repetitions in each individual (Schneider, 2007). As with other genes, this factor is passed down from parent to offspring following the Mendelian inheritance rules. A short tandem repeat (STR) is a unit, normally 2–5 base pairs long, that repeats multiple times at a specific point on a chromosome (Evett and Weir, 1998). Each individual has two alleles of this repetition, each expressed as the number of repeats in one STR marker. Where alleles have identical repeated times, the code is

[1]Corresponding author: Department of Statistics, Korea University, 145, Anam-ro, Seongbuk-gu, Seoul 02841, Korea. E-mail: jael@korea.ac.kr

Published 31 May 2024 / journal homepage: http://csam.or.kr

called homozygous, and if the number of repetitions is different, it is called heterozygous (Evett and Weir, 1998).

These unique genes can identify individuals in the same way that fingerprints can, and criminals are being identified through this type of forensic investigation at increasing rates. It is possible to extract genetic information from evidence such as bloodstains or cigarette butts, making it possible to identify individuals and possibly criminal suspects. Beyond individual identification, these genes can be used to identify familial relationships, as they are inherited half from each parent, also making it possible to identify siblings (Butler and Hill, 2012). It is now possible to trace a criminal offender using only a tiny blood stain at the scene of the crime and to check parent-child relationships, as well as even identifying the ethnicity of a criminal (Bieber *et al.*, 2006). In addition, methods of matching the genetic information of a suspect with other information in a database have emerged, in particular, the establishment of a criminal database that stores criminals' genetic information, can be used to identify the criminal who matches stored data (Myers *et al.*, 2011).

With such a database, individual suspects can be identified, and moreover, familial searching technique can be used, in which families of blood relatives are investigated when a DNA profile that matches genetic evidence can not be found (Myers *et al.*, 2011). A representative application of this cutting-edge genetic investigation technique was the Grim Sleeper case, in which a killer terrorized Los Angeles for 25 years (Myers *et al.*, 2011). In this case, an arrested young man's father was revealed to be a serial killer through the familial relationship between them, as ascertained by genetic evidence collected at the crime scene of a murder (Myers *et al.*, 2011).

Familial searching has been used in the investigation of various crimes, and it has produced successful apprehension of suspects (Schneider, 2007). Research to apply its use to a broader range of investigations is ongoing in many countries. However, it is difficult to apply it universally, due to the limited range of identifications and because the standard criteria for quantification is ambiguous (Lee *et al.*, 2007; Jeong *et al.*, 2016). In particular, it is necessary to develop a new method of analysis that goes beyond the current limits in familial relationship identification, whose power decreases at third and fourth cousins. In this study, we propose to apply a stacking ensemble machine learning algorithm, employing various statistical models that can improve the identification accuracy of familial relationship analysis in relatives from the first to fourth degrees (i.e., parent-child, full siblings, grandparent-grandchild, uncle-nephew, first cousins, etc.).

2. Materials and methods

2.1. Data

In this study, an analysis of familial data was conducted among 896 people (118 households) performed under the auspices of the Department of Forensic Medicine at Seoul National University and the National Bank of Korea (NBK) at Korea National Institute of Health. The data contained 790 parent-child pairs (first-degree relatives), 522 siblings pairs (second-degree relatives), 221 grandparent-grandchild pairs (second-degree relatives), 859 uncle-nephew pairs (third-degree relatives), and 468 first cousin pairs (fourth-degree relatives). In addition, 1,000 pairs were formed from individuals who were not related by blood. The experiment was conducted at the Department of Forensic Medicine at Seoul National University, and the genetic data included information on allele and genotype frequencies (Jeong *et al.*, 2016; Jeong *et al.*, 2019).

2.2. STR marker

The STR markers used in the analysis were the CODIS 13 markers (D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, CSF1PO, FGA, TH01, TPOX, vWA) used by the US Federal Bureau of Investigation and 9 markers commonly used in forensic analysis (D2S1338, D19S433, PentaE, PentaD, D1S1656, D2S441, D10S1248, D12S391, D22S1045), for a total of 22 markers (Yang *et al.*, 2013; Budowle *et al.*, 2001).

2.3. Statistical classification methods

2.3.1. Total number of shared Alleles analysis

The familial relationship analysis method using the total number of shared alleles (TNSA) as indicated by identical by state (IBS) corresponds to the most fundamental method of estimation suggested by Evett and Weir (1998), which measures the number of alleles between two people who are thought to have a familial relationship, and when the number of alleles surpasses a certain critical value, they are considered to be related by blood (Cowen and Thomson, 2008).

2.3.2. Likelihood ratio method

The likelihood ratio (LR) method is commonly used to calculate the likelihood ratio of the probability that a gene is observed in the case of a familial relationship to the probability that a gene is observed in the absence of a familial relationship. Using this likelihood ratio value, the probability of a familial relation can be determined when the likelihood ratio value is greater than or equal to a pre-specified critical value (Bieber *et al.*, 2006).

2.3.3. Shared Allele count and likelihood ratio analysis

The total number of shared alleles and likelihood ratio value (TNSA & LR) are considered at the same time in kinship identification. Where the number of shared alleles and likelihood ratio value exhibit at least an appropriate critical value, the sources of the samples are considered to be related (Jeong *et al.*, 2016). The relevant probability is estimated with the use of a logistic regression model, and when the estimated value exceeds an appropriate critical value, it is concluded that the pair have a familial relationship (Jeong *et al.*, 2016).

2.3.4. Data mining methods

Using previous familial relationship analysis methods in the identification of third-degree relatives does not yield a good result. This is because the distribution of unrelated groups overlaps the distribution of related groups, making it difficult to distinguish familial relationships through a comparison with the foregoing analysis method. Therefore, to evaluate the identification of familial relationships above the third degree, statistical methods can be applied to the classification analysis. These methods include linear discriminant analysis (LDA) (Friedman, 1988), shrinkage discriminant analysis (SDA) (Friedman, 1988), regularized discriminant analysis (RDA) (Friedman, 1988), classification and regression trees (CART) (Quinlan, 2007), C5.0 (Jansson, 2016), random forest (RF) (Breiman, 2001), extreme gradient boosting (Xgboost) (Chen and Guestrin, 2016), logistic regression (Cox, 1958), K-nearest neighbor (KNN) (Altman, 1992), Naive Bayes (Bickel and Levina, 2004), penalized multivariate analysis (PMA) (Witten *et al.*, 2009), support vector machine (SVM) (Menon, 2009), and artificial neural networks (ANN) (Basheer and Hajmeer, 2000).

LDA seeks to maximize separation between classes and to reduce dimensions while maintaining

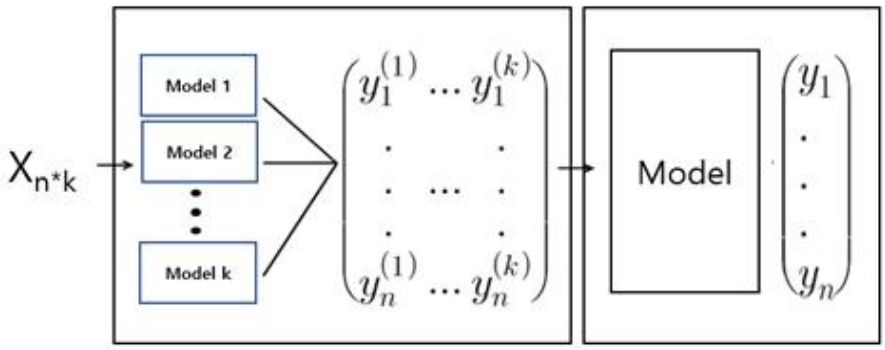

Figure 1: *Stacking ensemble machine learning algorithm.*

the information between classes as much as possible (Friedman, 1988). That is, LDA maximizes between-class scatter and reduces within-class scatter by projecting data from W-dimensional feature space into $W'(W > W')$ dimensional space (Friedman, 1988). SDA is a simplified discriminant analysis that replaces the non-diagonal elements of a sample covariance matrix with 0 in the LDA reduction model (Friedman, 1988). Doing so optimizes the model through classification errors as an ideal threshold for minimization based on a higher criticism threshold (HCT) or false non-discovery rates (FNDR) (Friedman, 1988). RDA optimizes parametric estimates to reduce the variance of estimates by using convenience estimates instead of sample-based inconvenience estimates when they are unstable and have large variances (Friedman, 1988).

CART is binary decision trees that use entropy or a Gini index as an impurity function and return a decision tree in an operation that best distinguishes a given data point through a separator that maximizes the value of the impurity (Quinlan, 2007). C5.0 is an algorithm that uses entropy to perform multiple separations on a classification basis with reduced entropy (Jansson, 2016). To determine optimal segmentation using entropy, segmentation efficiency can be measured through InfoGain (F). This efficiency is calculated as the difference between the entropy S1 of the segment before the division and entropy S2 after the division (Jansson, 2016). RF is a type of ensemble learning method that is defined by means of classification or regression analysis results in multiple decision trees during training (Breiman, 2001). Following this method maximizes confidence, with information gain being used as a measurement criterion (Breiman, 2001). Xgboost is a type of gradient tree boosting algorithm that allows the use of multiple decision trees in parallel (Chen and Guestrin, 2016). Xgboost generates optimized models to control the complexity of trees to minimize training loss and prevent overfitting (Chen and Guestrin, 2016). Logistic regression does not require data normality or the assumption that the covariance structures of the two groups are the same; it is a statistical method that is used to predict the likelihood that an event uses a linear combination of independent variables (Cox, 1958). Here, the odds represent the probability of success divided by the probability of failure, and the logistic model is achieved by performing a logit transformation of the odds (Cox, 1958).

In machine learning algorithms, KNN is a type of nonparametric method that is used for classification or regression. It is trained to select k classes from among the closest set within the feature space (Altman, 1992). KNN uses a condensed neighbor (CNN) to reduce the dataset (Altman, 1992). It selects a set of prototypes through training data, which classifies the first one-nearest neighbor and continues to repeatedly perform CNNs from this process (Altman, 1992). Naïve Bayes is a con-

Table 1: Result for identification of first-degree relatives (parent-child pairs) by classification models

Classification models		Sensitivity (n = 158)	Specificity (n = 158)	Accuracy (n = 316)
Traditional method	TNSA[a]	98.23 ± 0.85	99.63 ± 0.40	98.93 ± 0.45
	LR[b]	97.39 ± 1.01	99.03 ± 0.66	98.21 ± 0.61
	TNSA & LR	100.00 ± 0.00	98.80 ± 0.69	99.40 ± 0.34
One-step	LDA	99.42 ± 0.58	99.22 ± 0.61	99.32 ± 0.40
	SDA	98.80 ± 0.75	99.32 ± 0.51	99.06 ± 0.46
	RDA	50.37 ± 4.15	98.86 ± 0.83	74.61 ± 2.05
	CART	99.68 ± 0.58	99.89 ± 0.24	99.78 ± 0.29
	C5.0	99.84 ± 0.35	99.86 ± 0.26	99.85 ± 0.20
	RF	99.84 ± 0.33	99.84 ± 0.29	99.84 ± 0.21
	Xgboost	**99.85 ± 0.35**	**99.86 ± 0.26**	**99.85 ± 0.20**
	Logistic	98.58 ± 1.05	98.97 ± 0.80	98.78 ± 0.63
	kNN	99.22 ± 0.62	99.16 ± 0.60	99.19 ± 0.46
	NB	90.83 ± 2.27	98.99 ± 0.77	94.91 ± 1.18
	PAM	98.79 ± 0.73	99.34 ± 0.50	99.06 ± 0.43
	SVM	99.15 ± 0.63	99.25 ± 0.61	99.20 ± 0.44
	ANN	99.09 ± 0.88	99.36 ± 0.61	99.23 ± 0.53
Two-step (stacking)	**Xgboost**	**99.86 ± 0.33**	**99.86 ± 0.26**	**99.86 ± 0.20**

a. The cut-off value of TNSA is 23
b. The cut-off value of LR is 4.36

ditional probability model that applies the Bayes theorem, which assumes independence within a characteristic (Bickel and Levina, 2004). It is one of the simplest Bayesian network models, but it can be combined with a kernel density estimation to achieve high accuracy. The parameter estimation for the Naïve Bayes model uses the maximum-likelihood method (Bickel and Levina, 2004). PMA applies a penalized matrix decomposition that imposes the Lasso penalty on raw material approximation through singular value decomposition (Witten *et al.*, 2009). SVM separates data using a non-probabilistic binary linear classification model and identifies a hyperplane parameterized by a normal vector w that balances the objectives that maximize the margin (Menon, 2009). A hyperplane of a higher-dimensional space is defined as an internal operation of a set of points and a constant vector. The vectors defined in the hyperplane are selected and linearly combined with the image vector parameters that appear in the database (Menon, 2009). An ANN is built by simulating neurons, which are the basic structures of perception and of the real neural networks in the brain. This produces a multi-layered neural network model consisting of an input layer, a hidden layer, and an output layer (Basheer and Hajmeer, 2000). This allows the model to be optimized with error back propagation with learning based on the gradient descent method (Basheer and Hajmeer, 2000).

2.3.5. Stacking ensemble machine learning algorithm

The stacking model refers to an ensemble model that generates a meta-model by re-learning predictions generated through two or more algorithms with learning data, as shown in Figure 1 (Rho and Kim, 2009). Because each individual model was developed under the assumption that it was independent, it was optimized to have the smallest misclassification rate in response to outliers (Rokach, 2010). The stacking model is largely developed out in two stages.

As described in Figure 1, in the first step, various basic models are selected and learned. The posterior probability is derived from this learned model, and a new learning dataset is generated by collecting the derived posterior probability (Rokach, 2010). In the learning data X_{n^*k}, k refers to the k^{th} classification model, n refers to the number of learning data. This type of algorithms includes LDA,

Table 2: Results for identification of second-degree relatives (siblings and grandparent-grandchild pairs) by classification models

Classification models		Siblings			Grandparent-grandchild		
		Sensitivity (n = 104)	Specificity (n = 104)	Accuracy (n = 208)	Sensitivity (n = 44)	Specificity (n = 44)	Accuracy (n = 88)
Traditional method	TNSA[a]	94.99 ± 2.02	98.48 ± 1.08	96.73 ± 1.12	83.47 ± 5.02	85.78 ± 4.27	84.62 ± 3.48
	LR[b]	95.80 ± 1.64	99.12 ± 0.80	97.46 ± 0.87	88.44 ± 4.09	90.56 ± 3.87	89.50 ± 2.96
	TNSA & LR	99.52 ± 0.60	98.48 ± 1.08	99.00 ± 0.62	91.24 ± 3.68	77.40 ± 5.89	84.32 ± 3.60
One-step	LDA	95.80 ± 1.89	99.13 ± 0.84	97.47 ± 0.98	84.51 ± 5.21	94.69 ± 3.31	89.60 ± 2.81
	SDA	70.70 ± 14.13	99.86 ± 0.34	85.28 ± 7.08	84.22 ± 5.90	94.73 ± 3.23	89.48 ± 3.27
	RDA	59.49 ± 10.50	97.81 ± 1.46	78.65 ± 5.08	44.84 ± 10.29	95.73 ± 2.89	70.29 ± 5.40
	CART	98.11 ± 1.98	99.06 ± 0.98	98.59 ± 1.06	89.51 ± 4.79	93.51 ± 4.63	91.51 ± 2.85
	C5.0	98.82 ± 0.97	99.48 ± 0.79	99.15 ± 0.57	91.36 ± 4.07	94.73 ± 3.55	93.04 ± 2.63
	RF	**99.00 ± 0.94**	**99.30 ± 0.76**	**99.15 ± 0.55**	**92.42 ± 3.45**	**95.69 ± 3.47**	**94.06 ± 2.24**
	Xgboost	98.97 ± 0.98	99.14 ± 0.90	99.06 ± 0.59	91.20 ± 3.71	95.09 ± 3.02	93.14 ± 2.23
	Logistic	96.50 ± 2.01	96.50 ± 2.24	96.50 ± 1.41	89.09 ± 4.88	89.91 ± 4.80	89.50 ± 3.42
	kNN	95.60 ± 2.01	98.78 ± 1.07	97.19 ± 1.12	86.11 ± 4.70	90.16 ± 4.36	88.13 ± 2.92
	NB	90.88 ± 2.49	96.02 ± 1.99	93.45 ± 1.48	69.80 ± 7.99	96.40 ± 2.60	83.10 ± 4.33
	PAM	96.54 ± 1.54	99.40 ± 0.74	97.97 ± 0.80	87.36 ± 4.70	94.04 ± 3.40	90.70 ± 2.67
	SVM	98.05 ± 1.21	98.09 ± 1.18	98.07 ± 0.77	88.49 ± 4.35	93.09 ± 3.87	90.79 ± 3.02
	ANN	98.42 ± 1.26	98.02 ± 1.57	98.22 ± 0.97	90.02 ± 4.35	90.07 ± 4.59	90.04 ± 2.95
Two-step (stacking)	**RF**	**99.50 ± 0.61**	**99.33 ± 0.76**	**99.41 ± 0.49**	**93.49 ± 3.13**	**97.31 ± 2.13**	**95.40 ± 1.74**

a. The cut-off value of TNSA is 21, 18 in siblings and grandparent-grandchild.
b. The cut-off value of LR is 0.15, 2.93 in siblings and grandparent-grandchild.

SDA, RDA, CART, C5.0, RF, Xgboost, Logistic Regression, KNN, Naïve Bayes, PMA, SVM, and ANN, as mentioned in 2.3.4 (Rokach, 2010; Park *et al.*, 2019). $y_i^{(k)}$is the i^{th} output value from the k^{th} classification model (Rokach, 2010; Park *et al.*, 2019). The second step refers to the optimization of the classification model through learning with another independent classification model using a new set of learning data that were generated in the previous step as input (Rokach, 2010). The output data obtained through the first step become the input data in the second step, and the final output classification for y_i is optimally estimated through the selected model (Rokach, 2010). Stacking models can be used with several different model types and are classified as cumulative generalizations rather than as winner-take-all approaches (Park *et al.*, 2019).

The analysis was carried out using the statistical programs SAS 9.4 (SAS Institute, Inc., Cary, NC) and R3.5.1; the packages 'lda', 'sda', 'klaR', 'tree', 'C50', 'varSelRF', 'xgboost', 'foreign', 'class', 'klaR', 'pamr', 'e1071', and 'nnet' were used.

3. Results

The method of the basic analysis of the familial relationship relies on calculating the TNSA and LR for each of the 22 STR gene loci. It is then compared to unrelated pairs, and an optimal cut-off value can be obtained to determine the existence of familial relationships. Taking into account sensitivity (the probability that a test correctly identifies that people are related), specificity (the probability that a test correctly identifies that people are unrelated), and accuracy (the ratio of the number of correctly classified divided by the total number analyzed), the set sharing the highest sensitivity and specificity results was defined as the optimal cut-off. In the stacking model, a new learning dataset was generated that used various data mining models as a basic model (step 1), and the method of optimizing the classification model was performed using it as input data for study with an appropriate classification model (step 2).

Table 3: Results for identification of third-degree relatives (uncle-nephew pairs) by classification models

Classification models		Sensitivity (n = 171)	Specificity (n = 171)	Accuracy (n = 342)
Traditional method	TNSA[a]	73.99 ± 2.90	86.86 ± 2.24	80.43 ± 1.90
	LR[b]	79.07 ± 2.42	91.02 ± 1.59	85.05 ± 1.51
	TNSA & LR	94.67 ± 1.45	83.54 ± 2.36	89.10 ± 1.47
One-step	LDA	84.99 ± 2.51	86.33 ± 2.47	85.66 ± 1.81
	SDA	85.38±2.56	85.17±2.71	85.28±1.90
	RDA	36.33±4.56	96.23±1.47	66.28±2.26
	CART	90.80±2.57	90.37±2.60	90.58±1.48
	C5.0	**91.51±2.15**	**92.37±1.82**	**91.94±1.47**
	RF	91.00±2.48	91.44±2.02	91.22±1.63
	Xgboost	91.52±2.07	92.09±2.02	91.81±1.44
	Logistic	85.45±2.58	86.08±2.47	85.76±1.79
	kNN	80.12±2.91	88.87±2.37	84.49±1.64
	NB	50.67±4.78	94.96±1.60	72.82±2.30
	PAM	85.21±2.49	84.21±2.31	84.71±1.70
	SVM	88.74±2.20	89.05±2.29	88.90±1.65
	ANN	87.63±2.87	87.26±2.58	87.45±1.88
Two-step (stacking)	**C5.0**	**94.32±1.61**	**92.99±1.77**	**93.65±1.22**

a. The cut-off value of TNSA is 18
b. The cut-off value of LR is −0.13

3.1. Parent-child (first-degree relatives)

Analysis was conducted on 790 parent-child pairs and 790 randomly selected unrelated pairs. To avoid overfitting, a sample was allocated from among the total 1,580 pairs, as follows; train:validation:test = 6:2:2 (948 pairs:316 pairs:316 pairs). The results are obtained with 100 replicates and are presented as means ± standard deviations. The optimal cut-off TNSA value for distinguishing parent-child pairs is 23, and the cut-off value of the log-likelihood ratio is 4.36. Among conventional analysis methods (TNSA, LR, TNSA&LR), TNSA&LR had the highest accuracy (99.40 ± 0.34%), and among classification models (LDA, SDA, RDA, CART, C5.0, RF, Xgboost, logistic, KNN, NB, PMA, SVM, and ANN), the Xgboost model had the highest value (99.85 ± 0.20%). In addition, the estimate based on the Xgboost model, which had the highest accuracy in the classification model, was applied to the stacking model. Thus, a highest accuracy of 99.86 ± 0.20% was obtained (cf. Table 1).

3.2. Siblings (second-degree relatives)

Analysis was based on 522 pairs of siblings and 522 pairs randomly selected from among unrelated pairs. To prevent overfitting, the sample of 1,044 pairs was distributed as train:validation:test = 6:2:2 (628 pairs:208 pairs:208 pairs). The analysis was iterated 100 times. Then, the results are presented as the means ± standard deviations. The optimal cut-off value for TNSA in distinguishing siblings is 21, and the cut-off value of the log-likelihood ratio is 0.15. In the conventional analysis method, TNSA&LR had the highest accuracy (99.00 ± 0.62%), and among classification models, the RF model had the best performance (99.15 ± 0.55%). In addition, the estimate based on the RF model, which had the highest accuracy among the classification models, was applied to the stacking model, with the highest accuracy, of 99.41 ± 0.49% (cf. Table 2).

3.3. Grandparent-grandchild (second-degree relatives)

A set of 221 grandparent-grandchild pairs and 221 randomly selected unrelated pairs were analyzed. To avoid overfitting, the sample of 442 pairs was assigned as train:validation:test = 6:2:2 (266 pairs:88

Table 4: Results for identification of fourth-degree relatives (first cousins) by classification models

Classification models		Sensitivity (n = 93)	Specificity (n = 93)	Accuracy (n = 186)
Traditional method	TNSA[a]	71.51±3.78	63.90±3.48	67.71±2.49
	LR[b]	73.12±4.42	68.94±4.22	71.03±3.26
	TNSA & LR	90.31±2.75	56.91±4.05	73.61±2.35
One-step	LDA	69.66±4.95	72.55±4.00	71.11±2.91
	SDA	71.07±4.56	72.29±4.33	71.68±2.73
	RDA	24.11±6.29	95.96±2.13	60.03±3.03
	CART	73.96±8.96	71.65±8.67	72.80±3.45
	C5.0	**80.55±4.12**	**78.00±4.26**	**79.28±2.93**
	RF	79.67±4.49	73.98±4.85	76.82±2.89
	Xgboost	80.00±3.75	76.91±4.70	78.46±3.06
	Logistic	68.88±4.84	72.13±4.12	70.51±2.96
	kNN	57.85±5.25	70.48±5.00	64.16±3.58
	NB	32.72±5.25	91.36±3.53	62.04±2.70
	PAM	71.83±4.35	69.06±4.29	70.45±2.60
	SVM	75.38±4.65	70.79±4.67	73.09±2.71
	ANN	68.82±5.63	70.29±5.24	69.55±3.77
Two-step (stacking)	**C5.0**	**88.05±3.22**	**76.15±4.70**	**82.10±2.97**

a. The cut-off value of TNSA is 16
b. The cut-off value of LR is −0.11

pairs:88 pairs), and the analysis was replicated 100 times. The results are provided as means ± standard deviations. The optimal cut-off value of the TNSA for distinguishing grandparent-grandchild was 18, and the cut-off value of the log-likelihood ratio was 2.93. The LR showed the highest value in terms of accuracy (89.50 ± 2.96%), and the RF model had the highest (94.06 ± 2.24%) value among the classification models. In addition, the estimate based on the RF model was applied to the stacking model. The results showed the highest accuracy, at 95.40 ± 1.74% (cf. Table 2).

3.4. Uncle-nephew (third-degree relatives)

In all, 859 uncle-nephew pairs and 859 randomly selected pairs were assessed. To avoid overfitting, the sample of 1,718 pairs was assigned as train:validation:test = 6:2:2 (1034 pairs:342 pairs:342 pairs), and the analysis was replicated 100 times. The analysis was performed with 100 replicates, the results are given as means ± standard deviations. The optimal cut-off value of TNSA for identifying uncle-nephew relationships is 18, and the cut-off value of the log-likelihood ratio is −0.13. In the case of the earlier analysis method, TNSA&LR had the highest accuracy (89.10 ± 1.47%), and among classification models, the C5.0 model had the highest accuracy (91.94 ± 1.47%). An estimate of the C5.0 model was applied to the stacking model, which it showed the highest accuracy with 93.65 ± 1.22% (cf. Table 3).

3.5. First cousins (fourth-degree relatives)

The analysis was performed based on 468 pairs of first cousins and 468 randomly selected from among unrelated pairs. The total size of 936 pairs was analyzed, a sample allocated as train:validation:test = 6:2:2 (564 pairs:186 pairs:186 pairs) to avoid overfitting, and the analysis was performed with 100 replicates. The results are presented as means ± standard deviations. The cut-off value of the TNSA for discriminating the fourth-degree relatives is 16, and the cut-off value of the log-likelihood ratio is −0.11. In previous analysis method, TNSA&LR showed the highest accuracy (73.61 ± 2.35%), and the C5.0 model had the best performance among classification models (79.28 ± 2.93%). As in the

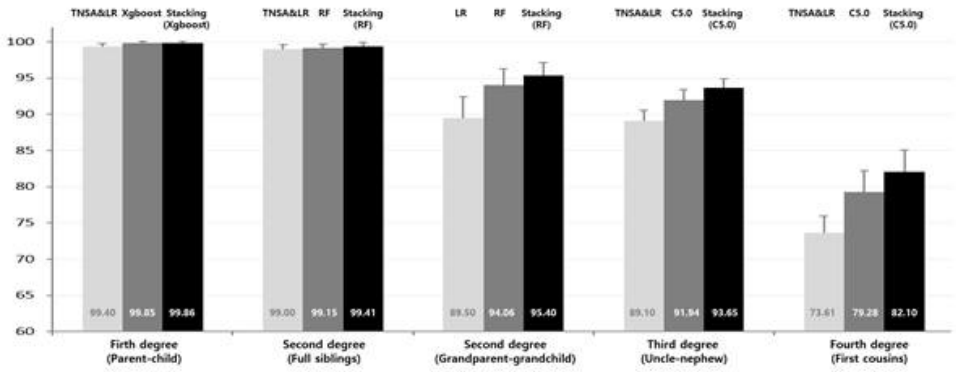

Figure 2: *Stacking ensemble machine learning algorithm.*

cousin case, the estimate of C5.0 model, which had the highest accuracy in the classification model, was applied to the stacking model and showed the highest accuracy, with $82.10 \pm 2.97\%$ (cf. Table 4).

4. Discussion

The methods that are most commonly used for the identification of familial relationships can distinguish related from unrelated pairs by calculating LR through STR markers consisting of repeats of specific nucleotide sequences. Therefore, using traditional methods, such as the number of shared alleles (TNSA) or likelihood ratio (LR), a cut-off value that minimizes the classification error rate was deduced, and familial relations were identified based on the cut-off values (Jeong *et al.*, 2019; Yang *et al.*, 2013). However, the cut-off values obtained using the above method may differ in relation to the sample size or the size or type of family members included in the sample, especially among third-degree relatives or even in higher degrees of relation, so the number of shared alleles or distribution of log-likelihood ratio shows significant overlap with unrelated groups. Thus, in the existing method, there is a limitation to calculating a cut-off value that can be accurately distinguished between the related and the unrelated.

Therefore, a more effective classification method is needed to identify familial relationships beyond the third cousin, and for this purpose, various data mining classification techniques have been applied (Yang *et al.*, 2013). Thus, the accuracy of the classification was somewhat higher than that of traditional methods, such as TNSA or LR, but limitations were seen in the cases of third- or higher-degree relatives. To address these issues, we combined several types of data mining analysis models through stacking ensemble algorithm to identify familial relationships using cumulative generalization methods rather than winner-only approaches.

The classification result obtained by applying the TNSA and LR methods, which have been the main methods used in the past, is as follows. For fourth-degree relatives, the accuracy was $73.61 \pm 2.35\%$, but a specificity of only about $56.91 \pm 4.05\%$ was predicted, and the misclassification rate was high. Therefore, to increase accuracy, it was necessary to produce and apply a new classification method rather than using existing methods, and this result can be shown in Figure 2. Our proposed stacking ensemble machine learning algorithm allowed us to find that the classification of third and

fourth-degree relatives, which had been difficult to identify in the past, could be improved to reach a maximum of 93.65 ± 1.22% (uncle-nephew) and 82.10 ± 2.97% (first cousins).

If two people do not know which blood relationship is, we do not apply all the above five models at the same time but apply in step by step from the first-degree modeling. In other words, if it is classified as the first-degree relatives (parent-child pairs) by applying the first-degree modeling we presented (stacking model of xgboost), the blood relationship is predicted as the parent-child pairs. If it is not classified as the first-degree relatives (parent-child pairs), the next step is to apply our modeling in the order of second-degree relatives (full siblings pairs, grandparent-grandchild pairs), third-degree relatives (uncle-nephew pairs), and fourth-degree relatives (first cousins pairs) to determine the relationship sequentially.

This result can be applied to the calculation of the familial relationship distance between distant relatives of more than the fourth degree in future work. From this, we judge that this work can become important material for the development of more accurate and quantitative familial relationship determinations in various fields that require genetic identification through familial searching using a system of databases storing DNA profile.

Acknowledgements

This research was supported and funded by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00208882) and the Korean National Police Agency [Project Name: Advancing the Appraisal Techniques of Forensic Entomology / Project Number: PR10-04-000-22].

References

Altman NS (1992). An introduction to kernel and nearest-neighbor nonparametric regression, *The American Statistician*, **46**, 175–185.

Basheer IA and Hajmeer M (2000). Artificial neural networks: Fundamentals, computing, design, and application, *Journal of Microbiological Methods*, **43**, 3–31.

Bickel PJ and Levina E (2004). Some theory for Fisher's linear discriminant function, "Naive Bayes", and some alternatives when there are many more variables than observations, *Bernoulli*, **10**, 989–1010.

Bieber FR, Brenner CH, and Lazer D (2006). Human genetics: Finding criminals through DNA of their relatives, *Science*, **312**, 1315–1316.

Breiman L (2001). Random forests, *Machine Learning*, **45**, 5–32.

Budowle B, Shea B, Niezgoda S, and Chakraborty R (2001). CODIS STR loci data from 41 sample populations, *Journal of Forensic Sciences*, **46**, 453–489.

Butler JM and Hill CR (2012). Biology and genetics of new autosomal STR loci useful for forensic DNA analysis, *Forensic Science Review*, **24**, 15–26.

Chen T and Guestrin C (2016). XGBoost: A scalable tree boosting system, In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Cowen S and Thomson J (2008). A likelihood ratio approach to familial searching of large DNA databases, *Forensic Science International Genetics Supplement*, **1**, 643–645.

Cox DR (1958). The regression analysis of binary sequences, *Journal of the Royal Statistical Society. Series B (Methodological)*, **20**, 215–242.

Evett IW and Weir BS (1998). *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*, Sinauer Associates, Inc, Seunderland.

Friedman JH (1988). Regularized discriminant analysis, *Journal of the American Statistical Association*, **84**, 165–175.

Jansson J (2016). Decision tree classification of products using C5.0 and prediction of workload using time series analysis, TRITA-EE, Sweden.

Jeong SJ, Lee HJ, Lee SD, Lee SH, Park SJ, Kim JS, and Lee JW (2019). Classification of common relationships based on short tandem repeat profiles using data mining, *Korean Journal of Legal Medicine*, **43**, 97–105.

Jeong SJ, Lee JW, Lee SD, Lee SH, Park SJ, Kim JS, and Lee HJ (2016). Statistical evaluation of common relationships using STR markers in Korean population, *The Korean Academy of Scientific Criminal Investigation*, **10**, 107–115.

Lee JW, Lee HS, and Lee HJ (2007). Statistical evaluation of sibling relationship, *Communications for Statistical Applications and Methods*, **14**, 541–549.

Menon AK (2009). Large-scale Support Vector Machines: Algorithms and Theory, *Research Exam*, University of California, San Diego, CA.

Myers SP, Timken MD, and Piucci ML (2011). Searching for first-degree familial relationships in California's offender DNA database: Validation of a likelihood ratio-based approach, *Forensic Science International: Genetic*, **5**, 493–500.

Park SJ, Kim YM, and Ahn JJ (2019). Development of product recommender system using collaborative filtering and stacking model, *Journal of Convergence for Information Technology*, **9**, 83–90.

Quinlan JR (2007). Induction of Decision Trees, *New South Wales Institute of Technology*, Sydney, Australia.

Rho DS and Kim E (2009). A study on the voltage regulation method based on artificial neural networks for distribution systems interconnected with distributed generation, *Journal of Korea Academia-Industrial Cooperation Society*, **10**, 3130–3136.

Rokach L (2010). Ensemble-based classifiers, *Artificial Intelligence Review*, **33**, 1–39.

Schneider PM (2007). Scientific standards for studies in forensic genetics, *Forensic Science International*, **165**, 238–243.

Witten DM, Tibshirani R, and Hastie T (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, *Biostatistics*, **10**, 515–534.

Yang IS, Lee HY, and Park SJ (2013). Analysis of kinship index distributions in Koreans using simulated autosomal STR profiles, *Korean Journal of Legal Medicine*, **37**, 57–65.

Received May 24, 2023; Revised January 12, 2024; Accepted January 12, 2024