Study of Population Structure and Genetic Prediction of Buffalo from Different Provinces of Iran using Machine Learning Method

Zahra Azizi[1], Hossein Moradi Shahrbabak[2], Seyed Abbas Rafat[1,*], Mohammad Moradi Shahrbabak[2] and Jalil Shodja[1]

[1] *Department of Animal Science, Faculty of Agriculture, University of Tabriz, Tabriz, Iran*

[2] *Department of Animal Science, Faculty of Agricultural Science and Engineering, College of Agriculture and Natural Resources, University of Tehran, Iran*

Abstract: Considering breeding livestock programs to milk production and type traits based on existence two different ecotypes of Iranian's buffalo, a study carried out to investigate the population structure of Iranian buffalo and validate its classification accuracy according to different ecotypes from Iran (Azerbaijan and North) using data SNP chip 90K by means Support vector Machine (SVM), Random Forest (RF) and Discriminant Analysis Principal Component (DAPC) methods. A total of 258 buffalo were sampled and genotyped. The results of admixture, multidimensional scaling (MDS), and DAPC showed a close relationship between the animals of different provinces. Two ecotypes indicated higher accuracy of 96% that the Area Under Curve (AUC) confirmed the obtained result of the SVM approach while the DAPC and RF approach demonstrated lower accuracy of 88% and 80 %, respectively. SVM method proved high accuracy compared with DAPC and RF methods and assigned animals to their herds with more accuracy. According to these results, buffaloes distributed in two different ecotypes are one breed, and therefore the same breeding program should be used in the future. The water buffalo ecotype of the northern provinces of Iran and Azerbaijan seem to belong to the same population.

Keywords Classification, Buffalo, Machine learning, SNP Chip data.

1. INTRODUCTION

Buffalo plays an essential role in the animal husbandry and agricultural economy of many countries all over the world [1]. Moreover, it influences the rural family's economy where it is bred due to its abilities for producing milk and meat. In some studies, the role of buffalo in the Socioeconomic Development of Rural Areas has been studied [2], and the reasons buffalos failure to contribute to livestock production is include lack of financial resources and interest in the private sector [3]. Asian buffalos are divided into two subspecies, including River buffalo and Swamp buffalo, which the Iranian buffalo belongs to the River type [4]. Furthermore, there are three buffalo ecotypes in Iran consist of Azerbaijan ecotype, which is living in East-Azerbaijan, West-Azerbaijan, and Ardabil provinces, North (Northern) ecotype found in Guilan and Mazandaran provinces, and Khuzestan ecotype observed in Khuzestan province.

Geographically, widespread species often exhibit considerable genetic diversity across the populations [5]. One of the interesting subjects in large scale studies is to study the existence of genetic differences among subdivided groups ascertained from various geographical locations. The Single Nucleotide Polymorphism (SNP) markers are useful information to do research on livestock genetic diversity and population structure [6, 7]. The advent of the new large-scale genotyping and sequencing technologies have provided the evaluation of the genetic structure and the relationship between animals in the populations. Inference of population structure of genetic markers has been previously used in a variety of aspects such as communication and evolutionary studies, classification of subspecies or population connectivity [8]. The main problems related to the accuracy of GWAS studies are the subpopulation structure or population stratification [9-11], and indistinguishable individuals, which are out of genetic groups, should be eliminated from the analyses.

Machine learning techniques are divided into supervised and unsupervised procedures. The unsupervised classification deals with samples that are not class labeled aims to group samples with similar attributes together. On the other hand, supervised machine learning involves training a model based on data samples that include known class labels [12].

There are different methods to assess the population structure like model-based clustering method that uses multilocus genotype data to detect the presence of population stratification [13-15]; and the multivariate reduction analyses such as the principal component analysis (PCA) method [16, 17], the MDS method [18, 19] and DAPC which is the

[*] Address correspondence to this author at the Department of Animal Science, University of Tabriz, Tabriz, Iran; Tel: +98 -4133392029; Fax: +984133345332; E-mail: rafata@tabrizu.ac.ir

ISSN: 1927-5196 / E-ISSN: 1927-520X/20

method to describes clusters of genetically related samples [20]. In other words, DAPC identifies genetic clusters and optimizes the separation of individuals into pre-defined groups and provides group membership probabilities [21]. Clustering methods were used for the inference of population structure of human [22], Mongolian domestic camels [23], horse breeds [24], and Italian domestic pigeons [25].

The supervised learning approach is efficient when individuals are classified into pre-defined populations, particularly in quality control for large scale genome-wide association studies (Bridges *et al*. 2011). Diverse methods of machine learning such as Random Forest (RF), Support Vector Machine (SVM) [26], and DAPC tackle the classification problems.

Support Vector Machine is extensively being applied as a solution to classification issues [27], and dealing with the great dimensionality problem in a computationally flexible manner [28]. Random forest is a classification algorithm that constructs multiple decision trees, each of which is built on a bootstrap sample of the training data using a randomly selected subset of variables [29]. The random forest has excellent comparable performance to SVM in classification tasks and conducts both classification and regression precisely. Regression models and SVM were employed in the subpopulation assignment of German Warmblood horses [30]. Machine learning has been applied to proteomics tandem mass spectrometry data, classification, and biomarker identification in post-genomics biology. Also, it has been used for GWAS and genetic prediction of a discrete and complex trait [31-36].

In this research, diverse ecotypes were based on various climate conditions. Given the importance of the study population structure for decision-making in the implementation of breeding programs, the structure of the population was studied. Subsequently, the classification strategy in two ecotypes was assessed by SVM, RF, and DAPC methods.

Our hypothesis is that SVM, RF, and DAPC approaches may show a better prediction for identifying animals in distinctive native breeds and (sub) population, due to the use of prior knowledge of the membership of the populations. The aim of the research is to determine the most accurate methods of predicting animals, especially for recognizing ecotypes, subpopulation, and native breeds.

2. MATERIALS AND METHODS

2.1. Animal Samples and Genotyping

Hair and blood samples were collected from flocks that had the registration and recording system of the National Animal Breeding Centre of Iran. The selection of the sampling regions was performed according to the registered farms by the National Animal Breeding Center and Promotion of Animal Products, the ministry of agriculture. Two factors were considered to select the samples: different geographical distribution and relativity of the breed in the pedigree. Animal sampling for the Azerbaijan ecotype was performed in West Azerbaijan province (3 cities), Ardabil province (2 cities), and East Azerbaijan province (5 cities) and for the North ecotype it was conducted in Guilan province (7 cities) (Figure **1**). Totally, 262 samples were genotyped, including: 68 from East Azerbaijan, 65 from west Azerbaijan, 56 from Ardabil and 73 from Guilan provinces (Table **1**).

Table 1: The Number of Samples for Two Ecotypes

Ecotype (N)	
North (73)	Azari(189)

Genomic DNA was extracted from the hair roots [37], and whole blood by applying a salting-out protocol [38]. Samples of DNA genotyped using the Axiom® Buffalo Genotyping 90 K Array (Affymetrix), and after quality control of the genotyped data, population structure analysis was performed

2.2. Data Quality Control

SNP genotypes were extracted from raw data by using the AffyPipe workflow [52] and applying default. Primary quality control and filtering, Initial Quality Control (QC) were carried out, and genotypes exported in PLINK. In the genotyping process, 4 samples (2 samples from Ardabil province and 2 samples from Guilan province) with more than 5% missing data were excluded from further analysis. The reason for the omitted data could be related to the DNA quality, which was not so high, so that likely to show more missing data and incorrect genotype calls [39]. In total, 19 SNPs were removed due owing to unknown position, 8855 SNPs were removed due to minor allele frequency (MAF<0.01), and 336 SNPs were deleted through Hardy-Weinberg disequilibrium at the 5% level. A total of 64750 SNPs passed QC steps. Quality

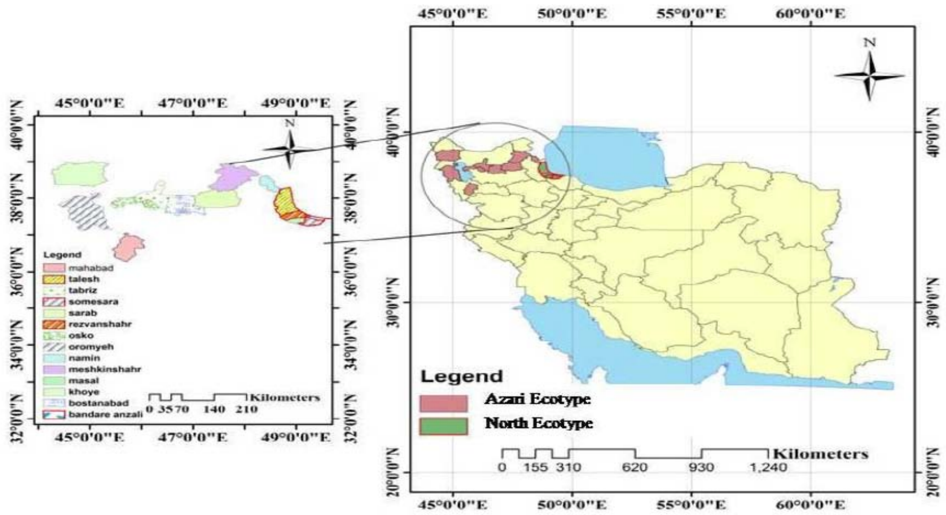

Figure 1: Geographical distributions of the animals used in this study are shown.

control was performed by PLINK for the initial data to ensure the overall quality of genotyped samples. The samples with more than 1% missing data were excluded from the analysis. Then MAF and call percentages were calculated for each SNP. The SNPs that had a call rate lower than 95% and a MAF < 1% were discarded. Deviation from Hardy-Weinberg equilibrium (p < 10^{-6}) was estimated for the remaining SNPs to identify genotyping errors [40]. The Bonferroni Correction (β=α/n) was used to address the multiple testing comparison problems [41]. The number of tests was taken to be the number of SNPs (n = 64,000), being 10^{-6} the corresponding value to $\alpha = 0.05$ experiment-wise error. We initiated the QC test with the edited Affymetrix data comprising 64750 SNPs. Then, 19 SNPs were removed due to unknown position, 7 SNPs were removed due to MAF<0.01 and 5 SNPs were not in Hardy-Weinberg equilibrium at the 5% level. Finally, a total of 64719 SNPs passed quality control steps.

2.3. Statistical analysis

2.3.1. Determining Population Structure

Admixture (Unsupervised Hierarchical Clustering)

The model-based clustering algorithm was performed in ADMIXTURE v1.23 [42] to investigate genetic structure in the combined dataset, and the genetic share of populations were plotted. ADMIXTURE is a program for estimating ancestry from a large autosomal SNP genotype dataset where the individuals are unrelated (admixture). In admixture, k is a factor involved in determining the number of ancestral populations.

Multidimensional Scaling

Multidimensional scaling [43] was conducted using PLINK on the basis of the genome-wide average proportion of Identical by State (IBS) shared alleles between every two individuals to visualize substructure and provide quantitative indices of population genetic variation and furthermore, identify outlying individuals [18]. Multidimensional scaling was performed based on the matrix with i and j elements (average proportion of IBS alleles shared by i and j individuals) using cmdscale in R Software.

Discriminant Analysis Principal Component

Discriminant analysis principal component is consists of unsupervised (as clustering method) and supervised procedures (as a predictive model). Discriminant analysis principal component analysis is employed when groups are often unknown, and there is a need for identifying genetic clusters before describing them. The number of clusters obtained by means of the *find.clusters* function and the optimal number of clusters were determined with Bayesian Information Criterion (BIC) that the rate of decrease in BIC values was visually examined to identify values of k, after which BIC values decreased only subtly Discriminant analysis principal component would be like trying k means with different ks, calculating BIC for each k and choosing the best k and defined as:

$$\text{BIC}= n\log(W(\text{X})) + g\log(n)$$

Where ***W* (X)** is the variance within groups, **g** is the number of groups, and n is the number of observations, then low BIC values are better than high ones [44].

The DAPC was performed using the *adegenet* package (function DAPC) for R software (Jombart, 2008). The supervised procedure of DAPC provides membership probabilities of each individual for the different groups that obtain indications of how clear-cut genetic clusters are.

2.3.2. Prediction Models

The supervised methods are used to predict unknown animals and determine the probability of membership populations. Hence, it requires that the data be prepared.

Data Preparation

Labeled Samples for Classification

The labeling was based on two ecotypes in which Azerbaijan ecotype with North ecotype considered as two classes and analyzed simultaneously.

Support Vector Machine Classifiers

An SVM constructs a set of hyperplanes in a high or infinite-dimensional space, which can be used for classification, regression, or other tasks. A good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class. The region bounded by two hyperplanes is called the margin, and the maximum margin hyperplane is the hyperplane that lies halfway between them. The original optimal hyperplane algorithm was a linear classifier [27]. Nonlinear classifiers were created by applying the kernel trick to maximum-margin hyperplanes that allow the algorithm to fit the maximum-margin hyperplane in high-dimensional feature space [45].

The effectiveness of SVM depends on the choice of the kernel, the kernel parameters, and a soft margin parameter C. A common choice is the Gaussian kernel (Radial Basis Function), which has a single parameter γ and RBF kernel can be defined as:

$$K(x,x') = \exp(-\gamma \| x - x' \|^2)$$

Where $\| x - x' \|^2$ is squared Euclidean distance between the two feature vectors and γ is a parameter needed for kernel. The parameters of *C* and γ is often selected by a grid search with exponentially growing sequences of *C* and γ. Typically, each combination of parameter choices is checked using cross-validation, and the parameters with best cross-validation accuracy in the training set of each fold are picked [46]. In this study, the package of e1071 was used for SVM, and the function of tune () was used to set the parameters using R Software that we tune the parameters for each training set.

Random Forest Classifiers

Random forest is an assemble learning algorithm for classification developed by Leo Breiman, which uses an ensemble of unpruned decision trees, each of which is built on a bootstrap sample of the training data using a randomly selected subset of variables. Each tree gets a vote in classifying. Individual trees are constructed as follows from data having ***n*** animals and ***m*** SNP:

1. Choose a training set by selecting ***n*** animals, with replacement, from the data.
2. At each node in the tree, randomly select ***m*** SNP from the entire set of ***m*** SNP in the data.
3. Choose the best split at that node from among the ***m*** SNP.
4. Repeat the second and third steps until the tree is fully grown.

Random forest uses bagging and random variable selection for tree buildings that result in a low correlation of the individual trees. The algorithm makes an ensemble that can achieve both low bias and low variance [29]. For finding the relationship between x and y, RF, build a model for predicting the value of y for a new value of x. A RF classification model is a collection of classification tree predictors

$$\{h(x,\Theta_k), k = 1,..\}$$

Where the Θ_k is P×N matrix that P is a (p×1) vector of animals and N is a (1×n) vector representing the genotype of each animal (0, 1, or 2) for n SNP, to which k decision trees are built. An interesting feature in the RF is "Out of Bag" **(**OOB) that has been explained below.

Out of Bag Sample

Considering a single tree from a random forest, grow the tree on a bootstrap sample (the bag). About two-thirds of the cases are in the bag, and the remaining one-third data are out-of-bag. The out-of-bag data are like a test set for this tree [29]. The out of bag data accuracy is the accuracy of the RF predictor that gives an estimate of test set accuracy.

There are two components of randomness involved in the building of the Random Forest, and they need to be tuned. First, for constructing each tree, a random subsample of the total data set was selected to grow the tree (ntree). After that, at each node of the tree, a well-performing variable from a random subset of all variables (mtry) was chosen as a splitter variable.

We ran RF in R software using the randomForest package. This implementation depends on the original Fortran code authored by Leo Breiman, the inventor of RF [47]. Different parameter configurations were considered for the values as: ntree=(200,500,1000), mtry=(150, 250, 6400) [34, 48] and nodesize (2,3,5). The best-performing configuration was selected by nested cross-validation.

2.4. Determine Predictive Performance

2.4.1. Cross-Validation

Data were randomly divided into training and testing sets. The training set was used for the statistical model construction, i.e., learning the classifier. The testing set was applied to check the accurate estimation of the classifier. In k-fold cross-validation, firstly, the training set was divided into k subsets of equal size. Sequentially one subset was tested using the classifier trained on the remaining k-1 subsets. The cross-validation procedure can prevent the overfitting problem, and estimations will be unbiased because each testing set was used only once to estimate the performance of a single classification model that was built by using training data exclusively. The accuracy was estimated as the average accuracy obtained after the k-fold cross-validation.

2.4.2. Metrics

Confusion Matrix

The confusion matrix is a method to examine the performance of classifiers. A confusion matrix contains information about actual and predicted classifications done by a given classification method. The performance of such a system is commonly evaluated using data in the matrix. Table **2** shows the confusion matrix for two-class classifiers.

Table 2: The Confusion Matrix for Two-Class Classification

	Predicted Negative	Predicted Positive
Actual Negative	TN	FN
Actual Positive	FP	TP

According to Table **2**, Sensitivity, Fall-out, Specificity, and Accuracy were calculated based on the formula:

$$Sensitivity = TPR = \frac{TP}{TP + FN}$$

$$Fall - out = FPR = \frac{FP}{FP + TN}$$

$$Specificity = \frac{TN}{FP + TN}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where TN is the number of correct negative predictions; FN is the number of incorrect positive predictions; FP is the number of incorrect negative predictions, and TP is the number of correct positive predictions.

The predictive accuracy of the classifier was estimated using sensitivity, specificity, accuracy, and phi coefficient correlation. The most important criterion for determining the performance of a classifier algorithm is its accuracy. In fact, it is the most famous and common criterion for calculating the efficiency of classifier algorithms.

Phi Coefficient Correlation

The phi coefficient correlation is identical to the Pearson that estimated for two binary variables.

In this study, we used the caret package and VCD package to calculate the confusion matrix and the phi coefficient(49, 50).

Area under the Receiver Operating Characteristic (ROC) curve

The ROC graph provides information on the performance of the classification model [51]. Moreover, it is a plot with the False Positive Rate on the X-axis (FPR) and the True Positive Rate on the Y-axis (TPR). The area under the ROC curve (AUC) is widely used for the performance measurement in classification and diagnostic rules [52]. If AUC is close to 1, the result of the test is excellent. On the contrary, the closer the AUC to 0.5, the lesser accuracy of the test result. The AUC can be used as a model comparison criterion and can be interpreted as the probability that a given classifier assigns a higher probability to a correct label when the animals are randomly picked. Individuals with a true genetic susceptibility above or below the population average were assumed positive or negative cases, respectively. Models with higher values of AUC are desirable and are considered more robust [53]. The packages of ROCR and pROC were used to calculate the area under the ROC and plot corresponding graphs [54, 55].

3. RESULT

3.1. Determining Population Structure

3.1.1. Multidimensional Scaling

The multidimensional scaling demonstrated the individual distribution of the different provinces. It presented separated clusters, but there is mixing and overlapping between individuals from the various provinces (Figure **2**).

3.1.2. Discriminant Analysis Principal Component

The DAPC method showed the distribution of individuals of the different provinces that presented separated clusters, but there is a close relationship between individuals (Figure **3**). Overlapping distributions of genetic clusters on the ordination plot indicated a low degree of genetic differentiation. Furthermore, Figure **3** illustrated the density of different provinces and the close relationship between individuals. This method used k means for finding the number of clusters. The number of clusters (k) that can be inferred from the estimation of BIC plot by means of DAPC procedure is shown in Figure **4**, and additionally, BIC had the lowest value of k=1 (Figure **4**) that displayed well the actual number of populations and confirmed that the population is a member of a group.

The results of DAPC analysis demonstrated that 250 PCs correspond to about 99% of variance exhibiting in an ordination plot with the first two axes (Figure **5**). According to the Figure **5**, explanation 90% of the variance needed more than 200 components, but retaining too many components with respect to the number of individuals can lead to overfitting in the

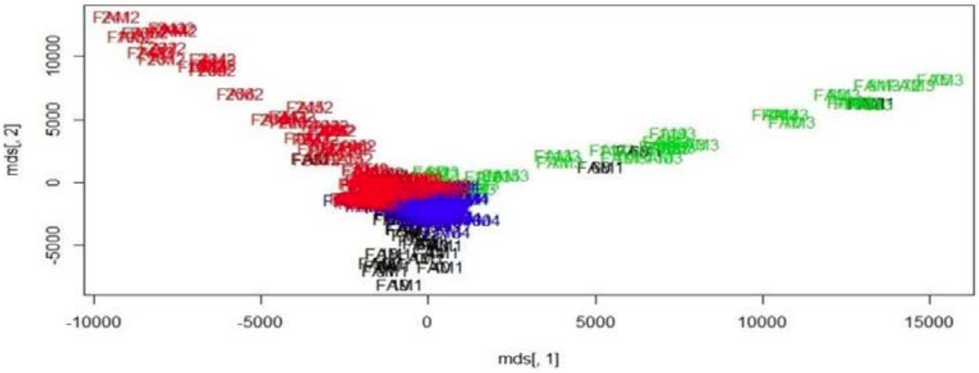

Figure 2: Plot of MDS (y-axis is related to coordinate1, and the x-axis is related to coordinate2 and FAM1, FAM2, FAM3, and FAM4 represent West Azerbaijan, Guilan, Ardabil and East Azerbaijan provinces respectively).

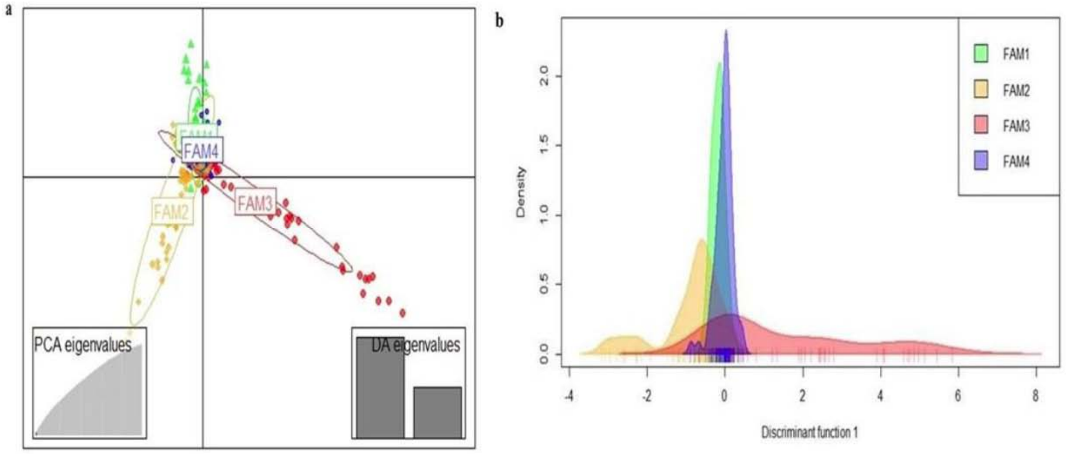

Figure 3: Ordination plot of DAPC for the four genetic clusters. Genetic clusters are shown by different colors and inertia ellipses, and dots represent individuals. The bottom-right inset shows eigenvalues of the two principal components in relative magnitude. Right plot represents the density of individuals.

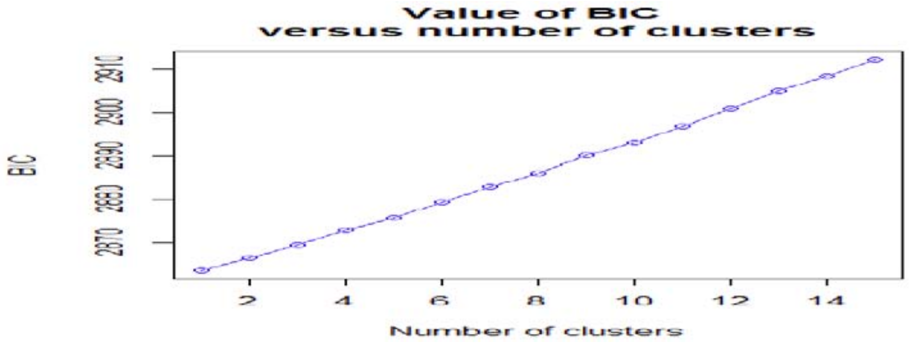

Figure 4: Inference of the number of clusters in the population of boffaloes.

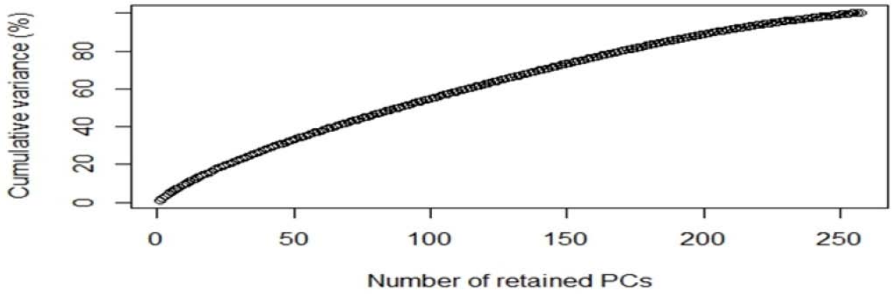

Figure 5: Variance explained by the first two PCA.

membership probabilities returned. The result of cross-validation for optimization of the trade-off between maintenance of too few and too many PCs in the model with 30 replicates indicated that the number of retaining PCA should be 60 PCs.

3.1.3. Admixture

The model-based clustering shows ancestral and mixture proportions of the individuals of different provinces. According to the results obtained from the model-based method, the animals in different provinces belonged to one breed, and they could not be distinguished from each other. The distribution of colors in Figure **6** illustrates an admixture between the individuals of different provinces.

The different methods used for studying the structure of the population demonstrated that these populations belong to one group. It means the breeding programs should be the same for these populations. In the next step, the accuracy of the prediction of individuals from two ecotypes was investigated by a supervised method.

3.2. Prediction Models

3.2.1. Parameter Regulation of SVM and RF

For SVM, the C-classification SVM with a radial kernel was applied for classification in the R package e1071 [56] with γ = 1.5×10^{-5} and a regularization (cost) parameter =1 determined by a grid search. As well, for

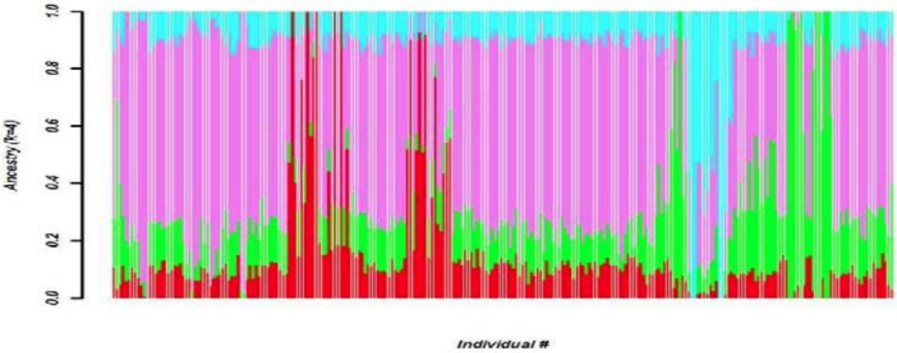

Figure 6: ADMIXTURE structure. In each plot, each cluster is represented by a different color, and each individual is represented by a vertical line divided into K colored segments with heights proportional to genotype memberships in the clusters.

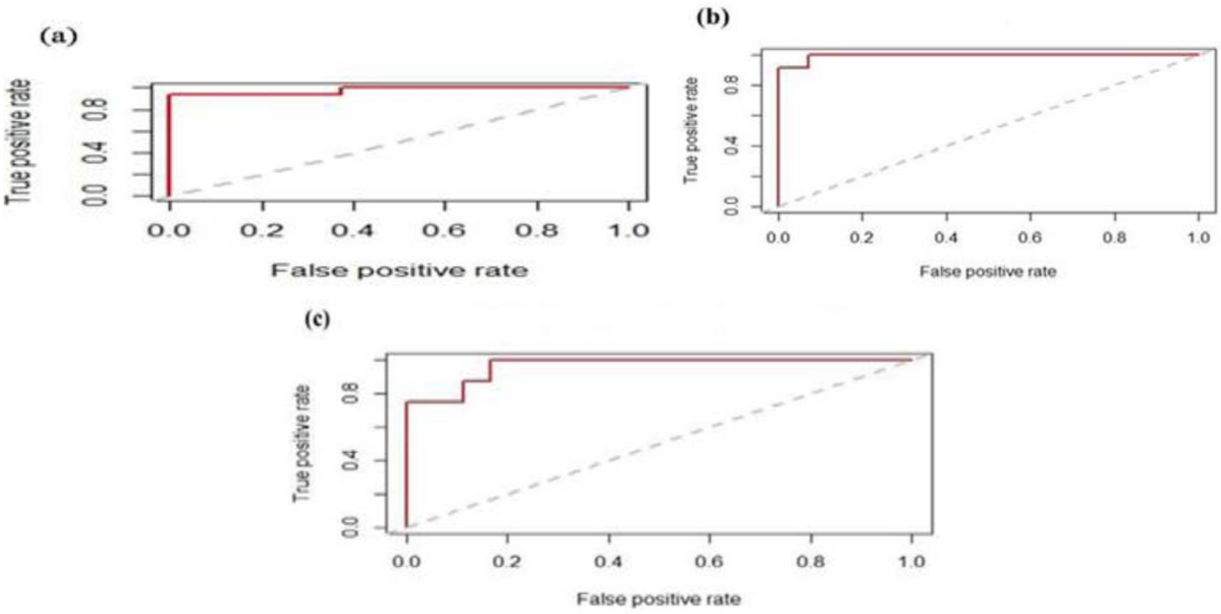

Figure 7: (**a**), (**b**), and (**c**) curve respectively represent the area under the receiving operator curve regarding the classification accuracy of an SVM, DAPC, and RF model for two classes (ecotype).

RF, the optimal tree size was determined to be at ntree= 1000, and the optimal mtry=250 along with nodesize=5 were defined by an average OOB error of 31.47%.

3.2.2. Analysis of the Two Ecotypes

In this section of the analysis, data set gained from three provinces of Azerbaijan ecotypes (members of different provinces merged) were considered as the first class, and the ecotype of Guilan was considered as the second class where two classes were analyzed together. In SVM analysis, the results of cross-validation with k=10 showed better accuracy of 98%, while DAPC and RF analysis represented an accuracy of 88% and 80%, respectively. The ROC curve displayed a good classification performance, and the area under the curve confirmed the classifier accuracy (Figure **7a**, **b**, and **c**) representing a better performance of the classifier with a test set of SVM, DAPC, and RF methods.

Table **3** shows the results of ROC curve components represented better accurate methods. SVM, DAPC, and RF methods predict the individuals of each province on the basis of their genomic data. The above procedures indicated the probabilities of each class and could provide the QC of our data sets by removing those individuals whose classifications show very low probability and affect predicted accuracy (result not shown). Hence, the individuals were

Table 3: Summarizes the Results of the Index Curve ROC, Overall Accuracy, and Kappa Coefficient SVM, RF, and DAPC Classifier Methods

DAPC	SVM	RF	Index/Methods
1.0000	0.9545	1.0000	Sensitivity
0.4000	1.0000	0.2857	Specificity
0.8750	1.0000	0.7917	PosPred Value
1.0000	0.8000	1.0000	NegPred Value
0.7000	0.9773	0.6429	Balanced Accuracy
00.8846	0.9615	0.8077	Accuracy
(0.6985, 0.9755)	0.8036, 0.999	0.6065, 0.934	95% CI
0.5185	0.866	0.3689	Kappa
0.9732	0.9812	0.9610	AUC
0.7643	0.874	0.542	Phi-Coefficient

assigned to the cluster to which they had the highest probability to belong.

The results showed a better individual classification of the two ecotypes with the SVM method. Despite the short distance between the animals of the two ecotypes, they could be predicted more accurately based on the training data set. The animals of two ecotypes and their membership probability were distinguished according to the predictions achieved by these methods. In the present study, SVM acted better than DAPC and RF. Interestingly, these methods allow for a probabilistic assignment of individuals to each group.

4. DISCUSSION

Nowadays, animal breeding researches tend towards the understanding of genome structure, mechanisms of evolution, and finding loci under selection with the increasing use of genomic information. Understanding population genetic structure is valuable for better implementation of breeding programs and, most importantly, preservation of genetic resources. Recent large-scale genotyping and sequencing technologies, e.g., next-generation sequencing, are useful for the study of genetic livestock diversity and population structure. In these large scale genome-wide association studies, it is necessary to determine whether the animals included from different herds and regions belonged to one or more different breeds and (sub) populations. Challenges related to stratification in the studying of the populations can be considered as a problem for GWAS studies [57, 58]. Hence, studying population structure is important. Multidimensional scaling and DAPC analysis of exploring the population structure of SE and SC African Bantu-speaking population showed that the populations from distant sampling localities could be clustered closely in the plot [59]. The population structure of the Korean cattle breeds was studied using a multivariate approach and model-based methods. The authors found that DAPC, PCA, and MDS result determined 20 separated clusters, and unsupervised hierarchical clustering was showed ancestry ratio and admixture of breeds [60]. The previous studies proposed that DAPC can be used as an efficient genetic clustering method [20, 61]. In this study, the result of MDS, Admixture, and DAPC showed a close relationship between the animals of the different provinces. These results suggest that it is better the breeding goals and programs to be considered for one population that can deal with decreases in the costs of the breeding programs. When the genetic admixture is high, it requires the determination of the probability of membership or unknown animals, and supervised methods are accurate classification ones to detect the individuals from each other. The most widely used classifiers are SVM and RF, which are supervised learning methods [62]. Supervised classifiers are able to recognize patterns in different features and assign individuals into one population or another. The supervised learning methods are able to classify individuals from two populations within Scotland in comparison with PCA; also, they can be used for QC in large scale genome-wide association studies [63]. The results of our study showed that the accuracy obtained from SVM and DAPC approaches were better than the RF one. In comparison with RF and SVM for microarray-based cancer classification, the results presented that SVM offers advantageous classification performance [64]. Comparing the other classification methods using seven microarray gene expression data sets, SVM, and more sophisticated classifiers such as RF showed the best performance among all methods [65].

Support Vector Machine was used to infer recent genetic ancestry of the American population, which showed 86% accuracy [66]. Prediction of the population assignments using whole-genome regression models and SVM revealed high prediction accuracy for the classification of horses into four German Warmblood breeds [30]. Discriminant analysis principal component method used for the analysis of genetically structured populations showed correct assignment rates ranging from 80% to 97% [20]. The SVM, DAPC, and RF supplied explicit probabilities for the classification of each individual. Despite the short distance between populations of the different provinces, they can be separated more accurately based on the training data set. Consequently, an individual with a specified genotype can be attributed more accurately to a breed, a region or a province that it belongs to. Furthermore, SVM, DAPC, and RF can classify populations based on a large number of markers without the necessity for strong assumptions to determine the population structure. These methods, including SVM, DAPC, and FR, can predict and assign the individuals for each group according to the determination of the probability of membership.

CONCLUSION

In fact, buffaloes from different regions belong to various ecotypes, and it related to climate conditions.

However, this study showed that different buffalo ecotypes in Iran belong to one population, and in breeding programs and GWAS, they can be analyzed in the form of one population. Although the breeds studied belonged to two groups, it seems that the utilized SNP density was unable to distinguish them completely according to their phenotype. Another reason could be the mismatch between provincial names and provincial divisions of water buffalo distributed regions. That is, there is a crossbreeding between local breeds of Azerbaijan and northern provinces. To ensure the correctness of individual grouping and prediction of new individuals, supervised methods such as SVM, RF, and DAPC were used. Among the suited methods, SVM can be used particularly for identifying animals belongs to different (sub) populations, breeds, ecotypes.

ACKNOWLEDGEMENTS

This work was financially supported by the National Animal Breeding Center, and Promotion of Animal Products, Noandish Alborz Company (TAKDANA), and Department of Animal Science (College of Agriculture and Natural Resources, University of Tehran, Karaj, Alborz Province, IRAN) and we appreciate all of them.

REFERENCE

[1] Moaeen-ud-Din M, Bilal G. Sequence diversity and molecular evolutionary rates between buffalo and cattle. J Anim Breed Genet 2015; 132(1): 74-84. https://doi.org/10.1111/jbg.12100

[2] Bibi S, Khan MF, Rehman A. Population Diversity and Role in the Socioeconomic Development of Domestic Buffaloes of Rural Areas of District Haripur, KPK Pakistan. Journal of Buffalo Science 2018; 7(3): 38-42. https://doi.org/10.6000/1927-520X.2018.07.03.1

[3] Wilson RT. The Domestic (Water) Buffalo in Africa: New and Unusual Records. Journal of Buffalo Science 2016; 5(2): 23-31. https://doi.org/10.6000/1927-520X.2016.05.02.1

[4] Naserian AA, Saremi B. Water buffalo industry in Iran. Italian Journal of Animal Science 2010; 6(2s): 1404-5. https://doi.org/10.4081/ijas.2007.s2.1404

[5] McTavish EJ, Hillis DM. A Genomic Approach for Distinguishing between Recent and Ancient Admixture as Applied to Cattle. J Hered 2014. https://doi.org/10.1093/jhered/esu001

[6] Lin BZ, Sasazaki S, Mannen H. Genetic diversity and structure in Bos taurus and Bos indicus populations analyzed by SNP markers. Anim Sci J 2010; 81(3): 281-9. https://doi.org/10.1111/j.1740-0929.2010.00744.x

[7] McKay SD, Schnabel RD, Murdoch BM, Matukumalli LK, Aerts J, Coppieters W, *et al*. An assessment of population structure in eight breeds of cattle using a whole genome SNP panel. BMC Genet 2008; 9: 37. https://doi.org/10.1186/1471-2156-9-37

[8] Epps CW, Castillo JA, Schmidt-Kuntzel A, du Preez P, Stuart-Hill G, Jago M, *et al*. Contrasting historical and recent gene flow among African buffalo herds in the Caprivi Strip of Namibia. J Hered 2013; 104(2): 172-81. https://doi.org/10.1093/jhered/ess142

[9] Lykkjen S, Dolvik NI, McCue ME, Rendahl AK, Mickelson JR, Roed KH. Genome-wide association analysis of osteochondrosis of the tibiotarsal joint in Norwegian Standardbred trotters. Anim Genet 2010; 41 Suppl 2: 111-20. https://doi.org/10.1111/j.1365-2052.2010.02117.x

[10] Tian C, Gregersen PK, Seldin MF. Accounting for ancestry: population substructure and genome-wide association studies. Hum Mol Genet 2008; 17(R2): R143-50. https://doi.org/10.1093/hmg/ddn268

[11] Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, *et al*. Demonstrating stratification in a European American population. Nat Genet 2005; 37(8): 868-72. https://doi.org/10.1038/ng1607

[12] Larranaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, *et al*. Machine learning in bioinformatics. Brief Bioinform 2006; 7(1): 86-112. https://doi.org/10.1093/bib/bbk007

[13] Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics 2000; 155(2): 945-59.

[14] Hoggart CJ, Shriver MD, Kittles RA, Clayton DG, McKeigue PM. Design and analysis of admixture mapping studies. The American Journal of Human Genetics 2004; 74(5): 965-78. https://doi.org/10.1086/420855

[15] Verdu P, Pemberton TJ, Laurent R, Kemp BM, Gonzalez-Oliver A, Gorodezky C, *et al*. Patterns of admixture and population structure in native populations of Northwest North America 2014. https://doi.org/10.1371/journal.pgen.1004530

[16] Patterson N, Price AL, Reich D. Population structure and eigenanalysis 2006. https://doi.org/10.1371/journal.pgen.0020190

[17] Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics 2006; 38(8): 904-9. https://doi.org/10.1038/ng1847

[18] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 2007; 81(3): 559-75. https://doi.org/10.1086/519795

[19] Li Q, Yu K. Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. Genetic Epidemiology 2008; 32(3): 215-26. https://doi.org/10.1002/gepi.20296

[20] Jombart T, Devillard S, Balloux F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genetics 2010; 11(1): 1. https://doi.org/10.1186/1471-2156-11-94

[21] Jombart T, Collins C. A tutorial for discriminant analysis of principal components (DAPC) using adegenet 2.0. 0 2015.

[22] Sethuraman A. On inferring and interpreting genetic population structure-applications to conservation, and the estimation of pairwise genetic relatedness 2013.

[23] Chuluunbat B, Charruau P, Silbermayr K, Khorloojav T, Burger PA. Genetic diversity and population structure of Mongolian domestic Bactrian camels (Camelus bactrianus). Anim Genet 2014; 45(4): 550-8. https://doi.org/10.1111/age.12158

[24] Felicetti M, Lopes MS, Verini-Supplizi A, Machado Ada C, Silvestrelli M, Mendonca D, *et al*. Genetic diversity in the

Maremmano horse and its relationship with other European horse breeds. Anim Genet 2010; 41 Suppl 2: 53-5. https://doi.org/10.1111/j.1365-2052.2010.02102.x

[25] Bigi D, Mucci N, Mengoni C, Baldaccini E, Randi E. Genetic investigation of Italian domestic pigeons increases knowledge about the long-bred history of Columba livia (Aves: Columbidae). Italian Journal of Zoology 2016; 83(2): 173-82. https://doi.org/10.1080/11250003.2016.1172121

[26] González-Recio O, Rosa GJ, Gianola D. Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. Livestock Science 2014; 166: 217-31. https://doi.org/10.1016/j.livsci.2014.05.036

[27] Vapnik VN, Vapnik V. Statistical learning theory: Wiley New York; 1998.

[28] Gunn SR. Support vector machines for classification and regression. ISIS technical report. 1998; 14.

[29] Breiman L. Random forests. Machine learning 2001; 45(1): 5-32. https://doi.org/10.1023/A:1010933404324

[30] Heuer C, Scheel C, Tetens J, Kühn C, Thaller G. Genomic prediction of unordered categorical traits: an application to subpopulation assignment in German Warmblood horses. Genetics Selection Evolution 2016; 48(1): 1. https://doi.org/10.1186/s12711-016-0192-2

[31] Swan AL, Mobasheri A, Allaway D, Liddell S, Bacardit J. Application of machine learning to proteomics data: classification and biomarker identification in post-genomics biology. OMICS 2013; 17(12): 595-610. https://doi.org/10.1089/omi.2013.0017

[32] Sun CS, Markey MK. Recent advances in computational analysis of mass spectrometry for proteomic profiling. J Mass Spectrom 2011; 46(5): 443-56. https://doi.org/10.1002/jms.1909

[33] Moore JH, Asselbergs FW, Williams SM. Bioinformatics challenges for genome-wide association studies. Bioinformatics 2010; 26(4): 445-55. https://doi.org/10.1093/bioinformatics/btp713

[34] Goldstein BA, Hubbard AE, Cutler A, Barcellos LF. An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. BMC genetics 2010; 11(1): 1. https://doi.org/10.1186/1471-2156-11-49

[35] González-Recio O, Forni S. Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. Genet Sel Evol 2011; 43(7): 21329522. https://doi.org/10.1186/1297-9686-43-7

[36] Long N, Gianola D, Rosa GJ, Weigel KA, Kranis A, Gonzalez-Recio O. Radial basis function regression methods for predicting quantitative traits using SNP markers. Genetics Research 2010; 92(03): 209-25. https://doi.org/10.1017/S0016672310000157

[37] Alberts CC, Ribeiro-Paes JT, Aranda-Selverio G, Cursino-Santos JR, Moreno-Cotulio VR, Oliveira AL, *et al.* DNA extraction from hair shafts of wild Brazilian felids and canids. Genet Mol Res 2010; 9(4): 2429-35. https://doi.org/10.4238/vol9-4gmr1027

[38] Grimberg J, Nawoschik S, Belluscio L, McKee R, Turck A, Eisenberg A. A simple and efficient non-organic procedure for the isolation of genomic DNA from blood. Nucleic Acids Res 1989; 17(20): 8390. https://doi.org/10.1093/nar/17.20.8390

[39] Barendse W, Harrison BE, Bunch RJ, Thomas MB, Turner LB. Genome wide signatures of positive selection: the comparison of independent samples and the identification of regions associated to traits. BMC Genomics 2009; 10: 178. https://doi.org/10.1186/1471-2164-10-178

[40] Teo YY, Fry AE, Clark TG, Tai ES, Seielstad M. On the usage of HWE for identifying genotyping errors. Ann Hum Genet 2007; 71(Pt 5): 701-3. https://doi.org/10.1111/j.1469-1809.2007.00356.x

[41] Abdi H. Bonferroni and Š idák corrections for multiple comparisons(http://www.utdallas.edu/~herve/Abdi-Bonferroni2007-pretty.pdf). In NJ Salkind (ed.). Encyclopedia of Measurement and Statistics. Encyclopedia of measurement and statistics 2007.

[42] Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 2009; 19(9): 1655-64. https://doi.org/10.1101/gr.094052.109

[43] Kruskal JB, Wish M. Multidimensional scaling: Sage; 1978. https://doi.org/10.4135/9781412985130

[44] Schwarz G. Estimating the dimension of a model. The Annals of Statistics 1978; 6(2): 461-4. https://doi.org/10.1214/aos/1176344136

[45] Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological Review 1958; 65(6): 386. https://doi.org/10.1037/h0042519

[46] Hsu C-W, Chang C-C, Lin C-J. A practical guide to support vector classification 2003.

[47] Liaw A, Wiener M. Classification and regression by randomForest. R news 2002; 2(3): 18-22.

[48] Díaz-Uriarte R, De Andres SA. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 2006; 7(1): 1. https://doi.org/10.1186/1471-2105-7-3

[49] Schaeffer L, Jamrozik J, Kistemaker G, Van Doormaal J. Experience with a test-day model. Journal of Dairy Science 2000; 83(5): 1135-44. https://doi.org/10.3168/jds.S0022-0302(00)74979-4

[50] https://cran.r-project.org/web/packages/GenABEL/index.html.

[51] Swets JA. Measuring the accuracy of diagnostic systems. Science 1988; 240(4857): 1285-93. https://doi.org/10.1126/science.3287615

[52] Hand DJ. Measuring classifier performance: a coherent alternative to the area under the ROC curve. Machine learning 2009; 77(1): 103-23. https://doi.org/10.1007/s10994-009-5119-5

[53] Gonzalez-Recio O, Forni S. Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. Genet Sel Evol 2011; 43: 7. https://doi.org/10.1186/1297-9686-43-7

[54] Schaeffer L. Application of random regression models in animal breeding. Livestock Production Science 2004; 86(1-3): 35-45. https://doi.org/10.1016/S0301-6226(03)00151-9

[55] Geetha E, Chakravarty A, Kumar KV. Estimates of genetie parameters using random regression test day model for first lactation milk yield in Murrah buffaloes. The Indian Journal of Animal Sciences 2007; 77(9).

[56] Dimitriadou E, Hornik K, Leisch F, Meyer D, Weingessel A. Misc functions of the Department of Statistics (e1071), TU Wien. R package 2008: 1.5-24.

[57] Wacholder S, Rothman N, Caporaso N. Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. Cancer Epidemiol Biomarkers Prev 2002; 11(6): 513-20.

[58] Thomas DC, Witte JS. Point: population stratification: a problem for case-control studies of candidate-gene associations? Cancer Epidemiol Biomarkers Prev 2002; 11(6): 505-12.

[59] Marks SJ, Montinaro F, Levy H, Brisighelli F, Ferri G, Bertoncini S, *et al.* Static and moving frontiers: the genetic

landscape of Southern African Bantu-speaking populations. Molecular biology and evolution 2014: msu263. https://doi.org/10.1093/molbev/msu263

[60] Sharma A, Lee S-H, Lim D, Chai H-H, Choi B-H, Cho Y. A genome-wide assessment of genetic diversity and population structure of Korean native cattle breeds. BMC Genetics 2016; 17(1): 139. https://doi.org/10.1186/s12863-016-0444-8

[61] Jemaa SB, Boussaha M, Mehdi MB, Lee JH, Lee S-H. Genome-wide insights into population structure and genetic history of Tunisian local cattle using the illumina bovinesnp50 beadchip. BMC Genomics 2015; 16(1): 1. https://doi.org/10.1186/s12864-015-1638-6

[62] Gutierrez S, Tardaguila J, Fernandez-Novales J, Diago MP. Support Vector Machine and Artificial Neural Network Models for the Classification of Grapevine Varieties Using a Portable NIR Spectrophotometer. PLoS ONE 2015; 10(11): e0143197. https://doi.org/10.1371/journal.pone.0143197

[63] Bridges M, Heron EA, O'Dushlaine C, Segurado R, Morris D, Corvin A, *et al*. Genetic classification of populations using supervised learning. PLoS One 2011; 6(5): e14802. https://doi.org/10.1371/journal.pone.0014802

[64] Statnikov A, Wang L, Aliferis CF. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics 2008; 9(1): 319. https://doi.org/10.1186/1471-2105-9-319

[65] Lee JW, Lee JB, Park M, Song SH. An extensive comparison of recent classification tools applied to microarray data. Computational Statistics & Data Analysis 2005; 48(4): 869-85. https://doi.org/10.1016/j.csda.2004.03.017

[66] Haasl RJ, McCarty CA, Payseur BA. Genetic ancestry inference using support vector machines, and the active emergence of a unique American population. European Journal of Human Genetics 2013; 21(5): 554-62. https://doi.org/10.1038/ejhg.2012.258

Received on 02-12-2019 Accepted on 30-04-2020 Published on 05-06-2020

DOI: https://doi.org/10.6000/1927-520X.2020.09.07