Turkish Journal of Engineering

https://dergipark.org.tr/en/pub/tuje

e-ISSN 2587-1366

Application of artificial intelligence methods for bovine gender prediction

Ali Öztürk[*1,2], **Novruz Allahverdi**[1], **Fatih Saday**[3]

[1]*KTO Karatay University, Engineering Faculty, Computer Engineering Department, Konya, Turkey*
[2]*Havelsan Inc., Ankara, Turkey*
[3]*Organized Industrial Zone Directorate, Konya, Turkey*

Keywords
Bovine gender prediction
Fuzzy logic
Artificial neural networks
Support vector machines
Random forests

ABSTRACT

This study investigates determining the gender of calves using some artificial intelligence (AI) techniques. Gender identification is important in animal breeding, focusing on the desired outcome and planning. The data used to determine the gender of calves were the speed, magnitude, and density of the bull's semen. The analysis of the related studies showed that there was not a study on gender prediction of bovine with the application of AI methods. In this study, fuzzy logic (FL), artificial neural networks (ANN), support vector machines (SVM), and random forests (RF) were used. The efficiency of these approaches was verified by statistical analysis parameters such as accuracy, specificity, sensitivity (recall), precision, and F-score. The FL, ANN, SVM, and RF models had 84%, 96%, 97%, 99% accuracy, 93.75%, 96.88%, 100%, 100% sensitivity, 66.66%, 94.44%, 92.31%, 97.30% specificity, 83.33%, 96.88%, 95.31%, 98.44% precision results, respectively. Application of these AI techniques for prediction bovine gender proves that these methods may be used by semen breeders as supporting information tools. In particular, it was observed that the RF method yielded the highest accuracy results.

1. INTRODUCTION

Biotechnological progress is being exploited to improve herd fertility. One of the last points in biotechnology improvement is the development of a method for determining the features of bovine sperm to determine the offspring sex (Seidel, 2003). Bovine genders are largely determined by the bulls' semen. In cattle, the gender of breeders is shaped during fertilization.

Gender identification enables the planning of production strategies and biotechnological study programs of enterprises that produce milk or meat. Today, alternative breeding systems are being studied in terms of calf production in cattle breeding (Erten and Yilmaz, 2012).

In cattle breeding, methods, such as centrifugation, electrophoresis, sedimentation, filtration, pH changes in the preservation medium, immunological techniques, and motility criteria, are used in the detection of the X and Y chromosomes in sperm. However, the practical use of the mentioned techniques is not very reliable because of the significant differences in the gender-determined sperm rates obtained as a result of these methods (Anderson, 1997; Johnson et al. 1994; Niemann and Meinecke, 1993). In contrast, the gender selection of offspring of cows in cattle currently represents a great perspective for genetic improvement and for meeting market demand. A new proposal for determining the proportion of the X- and Y-bearing cells in a bovine sperm sample was settled using actual polymerase chain reaction (Parati et al. 2006).

The forecast of male fertility with sperm quality parameters in vitro remains a problem for the bull industry. Fluorescein staining furthermore computer semen analysis (CASA) provides kinetically correct and functionally objective results to improve sperm control parameters. Therefore, Inanc et al. (2018) sought to study the kinetic parameters of the CASA and fluorescein staining of cryopreserved bull semen. They concluded that various kinetic parameters obtained with the help of algorithms of the CASA software system and fluorescein dyes can be related to fertility. However, further research is needed to establish a more accurate relationship with fertility.

[*] **Corresponding Author**

[*](aliozturk2002@gmail.com) ORCID ID 0000-0002-1797-2039
(novruz.allahverdi@karatay.edu.tr) ORCID ID 0000-0001-9807-884X
(fatihsaday@gmail.com) ORCID ID 0000-0001-7496-2796

Cite this article

Ozturk A, Allahverdi N & Saday F (2022). Application of artificial intelligence methods for bovine gender prediction. Turkish Journal of Engineering, 6(1), 54-62

Research Article / DOI: 10.31127/tuje.807019

Received: 07/10/2020; Accepted: 07/12/2020

A review by Sendag et al. (2005) about forecasting prenatal embryonic or fetal sex uses a variety of methods to apply sperm, embryo, or fetus. The review provides detailed information on these methods.

A Fuzzy Inference System (FES) for determining the productivity of livestock (milk and meat) was described by Vásquez et al. (2019). Using FL in terms of modeling the variables affecting livestock productivity, one can benefit from the knowledge and experience gained by producers, as well as from what they have learned from many years of observation and practice. Consequently, these results can be shared with agricultural producers and technicians to increase livestock productivity. The accuracy of the designed expert system was 86.67%.

RF method was used (Nicolas et al. 2016) to improve the downscaling of Gridded Livestock of the World database and provided better results than the stratified regression models. To identify the significant associations between single nucleotide polymorphism (SNP) and residual feed intake in dairy cattle, RF algorithm was used by Yao et al. (2013). The obtained results of RF could be used to identify large additive or epistatic SNP and informative quantitative trait loci. Multiple logistic regression, Naïve Bayes, and RF were compared to predict individual survival to the second lactation in dairy heifers (Heide et al. 2019). RF had the highest Area Under Curve (AUC) among the methods after first calving.

Mikail and Keskin (2013) evaluated SVM to assess its performance in detection of the mastitis in dairy cows. They achieved a sensitivity of 89% and specificity of 92% on the prediction of somatic cell counts in milk samples. SVM was also used by Miekley et al. (2013) to investigate its application for the mastitis detection in dairy cows. They obtained a sensitivity of 84.6% and specificity of 78.3%, and concluded that SVM could principally be applied for disease detection. Martiskainen et al. (2009) constructed SVM classification models based on nine features corresponding to different cow behaviors. The data were obtained using a three-dimensional accelerometer to investigate cow behavior pattern recognition. They concluded that SVM proved to be useful in the classification of measured behavior patterns. Huma and Iqbal (2019) used traditional linear models, regression trees, SVM, and RF methods to predict the bodyweight of farm animals. They found that RF had the best results among the methods for both the training and test datasets.

Allahverdi and Saday (2018) investigated the gender prediction of the bovine subject in a preliminary form, where they described the use of the ANN in predicting gender offspring. For the same goal, SVM and RF techniques were used in this study. The accuracy, sensitivity, specificity, precision, and F-score were determined for all these approaches. The problem of determining the gender of descendants in the animal herd with the methods of artificial intelligence described here is particularly lacking in the literature. In this study, bull sperm cells' features were used to predict the sex of descendants in the animal herd with the above-mentioned methods and their performances were compared in terms of various statistical analysis parameters.

The rest of the paper is organized as follows: Section 2 discusses the procedure for obtaining semen signs and the dataset used; describes the artificial intelligence methods (FL, ANN, SVM, RF); gives the definitions on the accuracy, sensitivity, and specificity determinations; Section 3 gives details of the experimental study with the obtained results; and Section 4 concludes the study.

2. METHOD

2.1. Preparing the Dataset

Sperm processing was performed by the commercial company Super Genetics Ltd. Sti. via cryobiology method. To ensure the classification of X and Y sperm cells, semen was collected from sexually mature bulls using an artificial vagina. Sperm with mobility greater than 60% were separated. It was diluted in egg diluent with egg yolk which was determined as 4% for sperm freezing process. It was cooled at 4 ° C for 90 minutes. The sperm were transferred by machine to 0.25 ml tubes and the sexless sperm were frozen in a programmable freezer as described. At the end of the process, the sperm were immersed in nitrogen.

A total of 100 individual cells derived from the animals were marked using the labeling function of the analysis software. The original image was manually divided into segments by digital zooming. Sperm measurements were then performed. The data set, as a result of these measurements, includes speed, size, density, and gender type. As for the knowledge base, 100 samples described in the study (Allahverdi and Saday, 2018) were used. There were 64 males and 36 females in the knowledge base. The input parameters in the system were speed (µ/s), magnitude (µ), and density (µg/µm^3). For the input parameters, the mean values were 45.61, 62.99, 49.76, the minimum values were 37.82, 52.99, 42.34 and the maximum values were 54.28, 72.62, 62.21, respectively. The output parameter was a numerical outcome which made a basis for female or male estimation depending on the predefined threshold value. The classification (male or female) was neither inferred by Super Genetic LTD tool, but the real genders of the animals were eventually determined and shared by the company in the dataset.

2.2. Fuzzy Logic Method

FL was introduced by Lotfi A. Zadeh (1965) to manage inaccurate and vague knowledge. If in the classical theory of sets elements either belong to a set or not, then in the fuzzy theory of sets elements may belong to a set to some extent. More formally, let X be a set of elements called a reference set. A fuzzy subset A of X is defined by a membership function µA(x), or simply A(x), which assigns a value to any x ϵ X within a real number range between 0 and 1. As in the classical case, 0 means no membership and 1 full membership, but now the value between 0 and 1 represents the extent to which x can be considered as an element X (Bobillo and Straccia, 2008).

The FL approach for applications is generally used as an FES, where instead of a strict knowledge base in an expert system, a fuzzy knowledge base is used. When the input data are entered into the system, one or a few rules can be activated, and an inference mechanism is used to calculate the correct fuzzy answer (Allahverdi, 2002; Allahverdi, 2020). The FES implementation details which are specific to this study are given in Section 3.1.

2.3. Artificial Neural Networks Method

ANN is a robust method against errors in training data for approximating real, discrete, or vector-valued functions (Adeli and Hung, 1995; Oztemel, 2016). They learn the input-output mapping given by the training data with a highly parallel and distributed process through automated weight tuning. Every bounded continuous function can be approximated with an arbitrarily small error by ANN with one hidden layer. For the ANN to be capable of representing nonlinear functions, the output of the neurons must be calculated with a differentiable nonlinear transfer function. One such function is the sigmoid defined as σ(net) = 1 / (1 + e^{-net}) that squeezes the output between 0 and 1. The input to the sigmoid is defined on neuron x_j as $net_j = \sum_{i=0}^{n} w_{ij} x_i$, where w_{ij} is the weight between the neurons x_i and x_j. The derivative of the sigmoid function was used to calculate the error value for the neurons and defined as dσ(y) / dy = σ(y)(1 – σ(y)).

The backpropagation is based on the stochastic gradient descent where the initial random weights are updated for each training example. The error between the target and computed output values is iteratively minimized until the termination condition is met. The termination condition can be either the reduction of the total network error to a predefined level or reaching to a predefined number of training steps. The error E is computed as follows:

$$E = (1/2)\sum_{j}(x_{o} - x_{t})^2 \tag{1}$$

Where x_o is the output value, and x_t is the actual (target) value.

The weights are updated according to the following formula:

$$\Delta W_{ji}(n + 1) = \eta\delta_{pj}O_{pi} + \alpha\Delta W_{ji}(n) \tag{2}$$

Where η is the learning rate; δ_{pj} is the error value for the neuron on L^{th} layer, and α is the momentum coefficient introduced to escape from the local minima during training.

δ_{pj} is calculated for the output layer neurons as

$$\delta_{pj} = \left(O_{t_{pj}} - O_{pj}\right)O_{pj}(1 - O_{pj}) \tag{3}$$

For the hidden layer neurons:

$$\delta_{pj} = O_{pj}(1 - O_{pj})\sum_{k}\delta_{pk}w_{kj} \tag{4}$$

2.4. Artificial Neural Networks Method

The SVM algorithm was first proposed by Vapnik (1995) for solving classification problems using a nonlinear function which maps an input dataset X into a high dimensional feature space F. The estimation function for the SVM is

$$f(x) = \left(w \times \emptyset(x)\right) + b \tag{5}$$

Where w and b are the estimated coefficients from the dataset, and $\emptyset(x)$ is the non-linear function used in feature space.

The risk function to be minimized is

$$R(w,\xi^*) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \tag{6}$$

And,

$$d_i - w\emptyset(x_i) - b_i \le \varepsilon + \xi_i \tag{7}$$

$$\left(w\emptyset(x)\right) + b - d_i \le \varepsilon + \xi_i^* \tag{8}$$

Where $\xi_i, \xi_i^* > 0$.

Vapnik (1999) introduced ε-insensitive loss function as an extension to the SVM to solve regression problems as well. The Support Vector Regression estimation function is

$$f(x) = \sum_{i=1}^{NSV}(\alpha_i - \alpha_i^*)K(X,X_i) + b \tag{9}$$

Where α_i and α_i^* are Lagrange multipliers and NSV is the number of support vectors. The kernel function $K\left(X_i,X_j\right) = \emptyset(X_i)\emptyset(X_j)$ is used in feature space to perform computation in input space.

2.5. Random Forest Method

Random Forests are formed by a decision tree classifiers set (DTCS) as in the following.

$$DTCS = \left\{h\left(x,\theta(k)\right),\ k = 1,2,3,\dots,K\right\} \tag{10}$$

Where x is the input vector and Θ(k) are random vectors which independently determine the growth of a single tree. Each random tree in the set casts a unit vote for determining the output of the forest (Breiman, 2001). In the case of regression, the random trees take on numerical values rather than discrete labels and the output is obtained by taking the average over each random tree.

If A is continuous attribute and m is the sample subsets on a node, then the RF algorithm is defined as in the following:

- The samples are sorted in ascending order on the continuous attribute A using corresponding discrete sequence {A_1, A_2,…, A_m}.
- On the sequence, m-1 division points are generated. The division point j(0<j<m) is adjusted by the formula

$$\frac{W_1 - (A_j + A_{j+1})}{2} \quad (11)$$

Then the sample set is divided into subsets as

$$\{s \mid s \in S, A(s) \leq W_j\} \text{ and } \{s \mid s \in S, A(s) > W_j\} \quad (12)$$

- The Gini coefficients of m-1 divided points are calculated as in Eq. (13) and the points having minimum Gini coefficients are selected to divide the sample set.

$$Gini(S) = 1 - \sum_{n=1}^{n} p_i^2 \quad (13)$$

Here S is the sample set and |S| is the total number of samples. The number of samples in class C_i is $|C_i|$ and the probability p_i is $\frac{|Ci|}{|S|}$ (Xu, 2017).

2.6. Performance measure analysis

The accuracy, sensitivity, specificity, precision and F-score (Vapnik and Vapnik, 1998) are calculated for each artificial intelligence method. For this, the following definitions are used:

Male: positive for male; Female: negative for male

True positive (TP) = the number of statuses correctly identified as male; True negative (TN) = the number of statuses correctly identified as female; False positive (FP) = the number of statuses incorrectly identified as male; False negative (FN) = the number of statuses incorrectly identified as female.

Accuracy: Accuracy is about how close you are to the right results. In our case, the accuracy gives how correctly the male and female statuses are differentiated. The accuracy is defined as

$$Accuracy = ((TP + TN) / (TP + TN + FP + FN)) \quad (14)$$

Sensitivity or **Recall:** The sensitivity of a test is its ability to correctly determine the male statuses. For its estimation, the proportion of the true positive in the male statuses was used. The sensitivity is defined as

$$Sensitivity = (TP / (TP + FN)) \quad (15)$$

Specificity: The specificity of a test is its ability to correctly determine the female statuses. For its estimation, the proportion of the true negative in the female statuses was used. The specificity is defined as

$$Specificity = (\mathrm{TN} / (TN + FP)) \quad (16)$$

Precision: Precision is about getting the same results in the same way. For its estimation, the proportion of the true positive in the male statuses was used. The precision is defined as

$$Precision = (TP / (TP + FP)) \quad (17)$$

F-score or **F-measure** is a weighted average of accuracy and recall. Therefore, this indicator takes into account both false positives and false negatives. If there is an irregular distribution of classes then using F score is usually more useful than accuracy. The F-score is defined as

$$F - score = 2 \times TP / (2 \times TP + FP + FN) \quad (18)$$

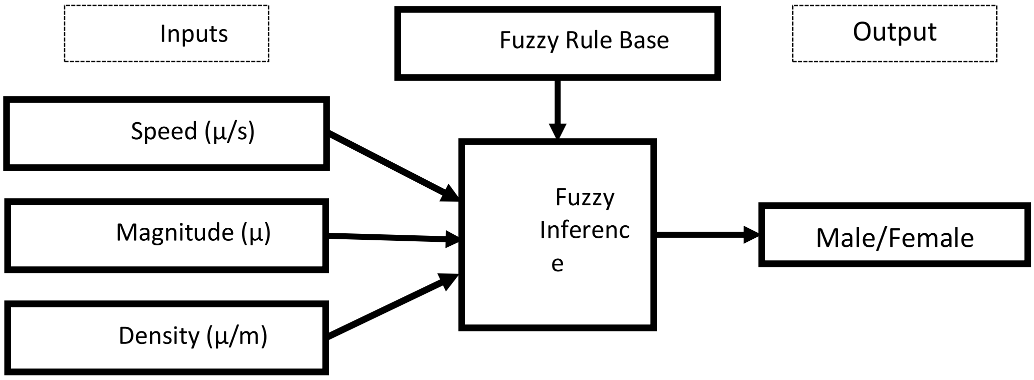

Figure 1. Structure of the designed FES

3. EXPERIMENTAL STUDY

In determining the sex of bovine animals, the speed, size, and density characteristics of the semen cell were the input elements used for gender estimation. An accurate analysis of these elements and the prediction of gender in this context are possible. After the cells were examined under the microscope, the analysis results were transferred to the artificial intelligence methods.

3.1. Fuzzy Expert System Implementation

The FES structure designed in this study is given in Figure 1. The Mamdani inference approach was used herein because it is widely preferred due to its simplicity. The centroid defuzzification method was chosen to obtain more strict results.

Table 1 show the selected min and max value ranges of the input and output parameters for the designed FES.

Five fuzzy sets were selected for the fuzzification of the input parameters “speed” and “magnitude.” Three fuzzy sets were selected for the fuzzification of the input parameter “density.” Two fuzzy sets were used for the output parameter “gender” (male/female). The ranges for the fuzzy sets and the fuzzy rules were determined by the agreement of the domain experts and the fuzzy system designer. As an example, Figure 2 shows the speed fuzzy set. Its fuzzy formulas are presented by expressions (19)–(23).

Table 1. Min and max value range of the input and output parameters of FES

Input/Output	Fuzzy values	Fuzzy Sets				
Input	Speed (μ/s)	Very Slow (30-37)	Slow (35-41)	Medium (38-46)	Fast (45-55)	Very Fast (> 50)
Input	Magnitude (μ)	Very Small (45-56)	Small (52-62)	Medium (64-68)	Big (64-74)	Very Big (> 70)
Input	Density (μg/μm^3)	Less Dense (35-48)	Dense (42-58)	Much Dense (> 55)		
Output	Gender (%)	Male (< 35)	Female (< 35)			

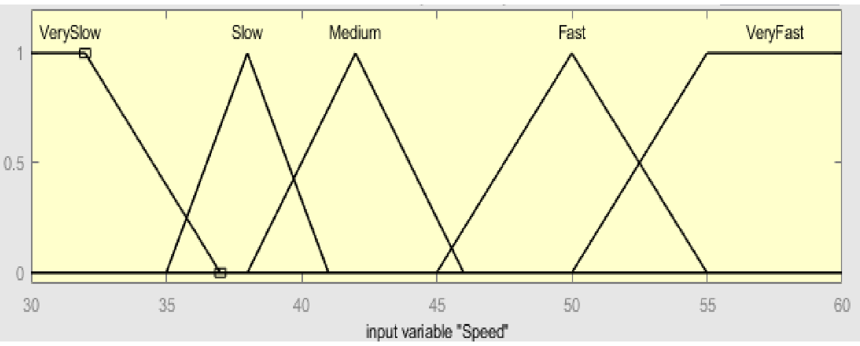

Figure 2. Fuzzification of the speed input value

$$\mu_{VS}(x) = \begin{cases} 1 & \text{if } x < 32 \\ (37 - x) \big/ (37 - 32) & \text{if } 32 < x < 37 \end{cases} \quad (19)$$

$$\mu_{S}(x) = \begin{cases} (x - 35) \big/ (37 - 35) & \text{if } 35 < x \leq 37 \\ (41 - x) \big/ (41 - 37) & \text{if } 37 < x < 41 \end{cases} \quad (20)$$

$$\mu_{M}(x) = \begin{cases} (x - 37) \big/ (42 - 37) & \text{if } 37 < x \leq 42 \\ (46 - x) \big/ (46 - 45) & \text{if } 45 < x < 46 \end{cases} \quad (21)$$

$$\mu_{F}(x) = \begin{cases} (x - 45) \big/ (50 - 45) & \text{if } 45 < x \leq 50 \\ (55 - x) \big/ (55 - 50) & \text{if } 50 < x < 55 \end{cases} \quad (22)$$

$$\mu_{VF}(x) = \begin{cases} (x - 50) \big/ (55 - 50) & \text{if } 50 < x \leq 55 \\ 1 & \text{if } 55 < x \end{cases} \quad (23)$$

Note that the output parameter "gender" was determined as "Female% = 100% − Male%" or "Male% = 100% − Female%." In addition, the value of 35% will be an undetermined answer when calculating the calf sex (Figure 3 and formulas (24-25)).

Some of the fuzzy sets of sperm speed will be described as the next expressions:

$$\mu_{VS}(x) = 1 / 30 + 1 / 32 + 0.3 / 36 + 0 / 37$$

$$\mu_{M}(x) = 0 / 38 + 0.45 / 39 + 1 / 42 + 0.1 / 45.5 + 0 / 46$$

$$\mu_{VF}(x) = 0 / 50 + 0.5 / 52.5 + 1 / 55 + 1 / 60$$

$$\mu_{F}(x) = \begin{cases} 1 & if \quad x < 30 \\ (40 - x) \big/ (40 - 30) & if \ 30 < x < 40 \end{cases} \quad (24)$$

$$\mu_{M}(x) = \begin{cases} (x - 30) \big/ (40 - 30) & if \ 30 < x \leq 40 \\ 1 & if \ 40 < x \end{cases} \quad (25)$$

Some of the fuzzy sets (Figure 3) of gender (output value) will be described as the next expressions:

$$\mu_{F}(x) = 1 / 0 + 1 / 30 + 0.5 / 35 + 0 / 40$$

$$\mu_{M}(x) = 0 / 30 + 0.5 / 35 + 1 / 40 + 1 / 100$$

The number of fuzzy rules is determined by the multiplication of the number of input fuzzy sets. In our case, it will be 5 × 5 × 3 = 75 rules. The dataset was applied to the FES containing 75 rules to obtain the outputs. Some fuzzy rules are also used in the designed FES:

R5: If (speed is very low) and (magnitude is small) and (density is dense), then gender is female.

R36: If speed is medium and magnitude is small and density is much dense then output is female.

R67: If (speed is very fast) and (magnitude is medium) and (density is less dense), then gender is male.

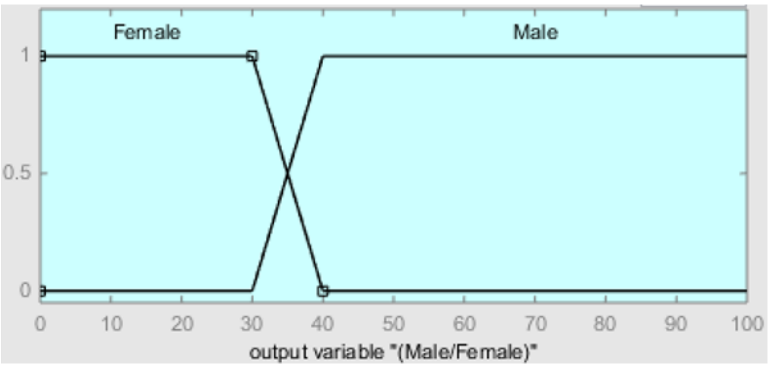

Figure 3 Fuzzification of the gender output value

3.2. ANN implementation

The Weka (Frank et al. 2016) machine learning package was used to implement the ANN. Various ANN structures were evaluated with a different number of hidden neurons beginning from 2 up to 10. Furthermore, various momentum and learning rate values were applied for comparison purposes. The ANN with a hidden layer of five neurons trained with a momentum of 0.3 and a learning rate of 0.2 gave the best performance in terms of sensitivity, specificity and accuracy. The ANN structure with five hidden neurons is shown in Figure 4. The ANN was evaluated herein with a 10 fold cross-validation to obtain the results for the data set. The ANN was trained with 500 numbers of steps.

3.3. SVM implementation

In this study, the Pearson VII function based kernel (PUK) proposed by Ustun et al (2006) was used. Because, it gave better results than the Radial Basis Function (RBF) kernel and poly kernel functions. The default values were chosen for the Ω and ∑ parameters of the PUK function which were 1.0 and 1.0, respectively.

The Sequential Minimal Optimization (SMO) algorithm was proposed by Smola and Schölkopf (1998) as an extension of the original SMO algorithm for solving regression problems.

Shevade et al. (2000) suggested the use of two threshold parameters instead of one and devised two variants of the original SMO Regression (SMOReg) algorithm. These variant algorithms are much more efficient than the original SMOReg. In this study, the first variant of the SMOReg algorithm was applied with the complexity parameter, the round-off error parameter, the ε-insensitive loss function parameter, the tolerance parameter for checking the stopping criterion were set as 1.0, 1.0e-12, 0.01 and 0.01, respectively. The SVM was evaluated with a 10 fold cross-validation.

3.4. RF implementation

Random forest algorithm combines bagging with random feature selection because bagging increases accuracy when random features are used while growing the trees. Depending on the bag size, some of the samples in the dataset are not used for constructing the random trees. If, for example, the bag size percent is 80 then 20% of the training samples are out-of-bag and they are not used for tree growing. In the random forest algorithm, it is possible to include an out-of-bag error in the generalization error estimate while building the forest. The out-of-bag error is estimated by aggregating the votes for each (x,y) training sample only over the trees those were grown by bootstrapped training sets Tk not containing (x,y).

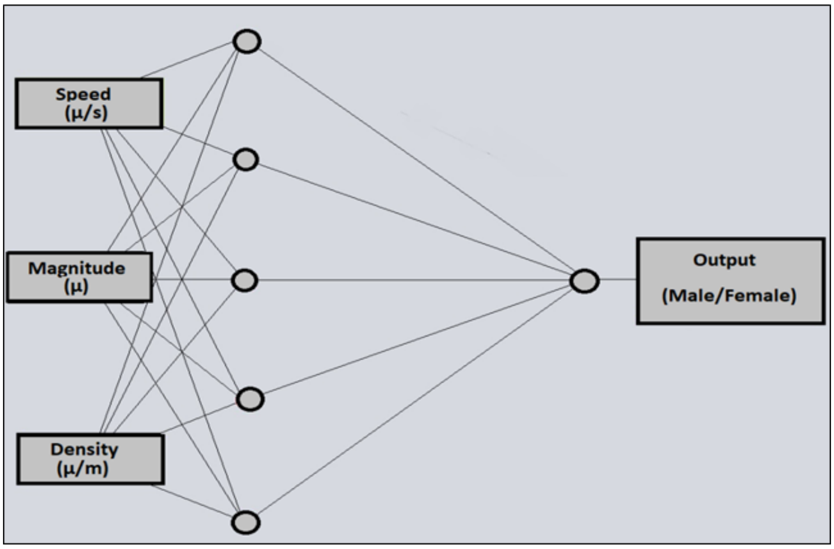

Figure 4. Artificial neural network structure

In this study, the bag size percentage was set as 100% which means that every training sample was used to construct the random forest. The number of attributes in feature selection and the maximum depth of the growing trees were not limited. The out-of-bag error was not included in the generalization error estimation of the random forest. The maximum number of iterations to build the random forests was set as 100. The RF was evaluated with a 10 fold cross-validation to obtain the results.

3.5. Performance Comparison of Artificial Intelligence Methods

Table 2 presents the values of the TP, TN, FP, and FN status for each method. The overall results were obtained by evaluating the performances of the ANN, SVM and RF on each 10-fold test set.

Table 2 TP, TN, FP, and FN values.

Status	FL	ANN	SVM	RF
TP	60	62	61	63
TN	24	34	36	36
FP	12	2	3	1
FN	4	2	0	0
Total	100	100	100	100

Table 3. Comparison of FL, ANN, SVM and RF in terms of performance measures.

Methods	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F-score
FL	84	93.75	66.66	83.33	0.88
ANN	96	96.88	94.44	96.88	0.97
SVM	97	100	92.31	95.31	0.98
RF	99	100	97.30	98.44	0.99

Table 3 shows a comparison of the results of the FL, ANN, SVM, and RF approaches in terms of the accuracy, sensitivity (recall), precision, specificity and F-score. The results showed that the sensitivity values of SVM and RF approaches were 100% indicating that both of these methods were good at predicting the male gender. However, the RF specificity was significantly higher than the other methods denoting that RF was much more successful in predicting female gender.

The RF approach also outperformed FL and ANN methods in terms of overall accuracy. The outputs (predicting bovine gender) produced by these approaches for different test inputs (100 data) are investigated.

Table 4 provides ten samples. The FL approach incorrectly predicted the data in row 7 as male, although the answer must be female. In row 8, the ANN failed, but FL succeeded. In row 9, both methods were unsuccessful in the prediction. The tenth-row dataset was successful for all approaches, besides SVM techniques.

So, the research generally shows that the RF forecast bovine gender is more accurate. The output threshold to classify an instance as female or male was determined as 35, depending on the suggestions of Super Genetic LTD. This assumption was made for the prediction outputs of all the artificial intelligence methods as in FES. Although the application of the FL method results for samples 3, 5, 8, and 10 in Table 4 indicate male gender, the results are at the intersection of the fuzzy functions of the output parameter (i.e., between 30 and 40). As can also be seen from the Table 4, 7th, and 9th rows for the FL method, as well as in the rows of the 8th and 9th rows, for the ANN method did not correctly guess the gender of the calf.

4. CONCLUSIONS AND FUTURE WORKS

This study used some bull sperm cell features to compare the fuzzy logic and machine learning approaches in automatically determining the bovine offspring gender. The results of this study have shown that farmers may well use the proposed methods, where the RF method gave the best results (prediction accuracy = 99% and precision = 98.44%) among the methods. The prediction accuracy and precision of FL, ANN and SVM were 84% and 83.33%, 96% and 96.88, 97% and 95.31%, respectively. The F-scores for FL, ANN, SVM and RF were 0.88, 0.97, 0.98, and 0.99, respectively. This means that the RF method achieves approximately the best prediction result, where the maximum value of F-score can be 1.

To further check the statistical significance, a two-sample assuming equal variances t-test was performed for the RF method which gave the best results. The null hypothesis was that the actual and predicted values come from normal distributions with the same variance. The P-value was found as 0.94 which indicated that the null hypothesis could not be rejected at 5% level of significance. The total number of observations used in t-test was 100 and the t-stat value was found as 0.07.

Since the data kindly provided by the company Super Genetics Ltd. was used, the accuracy and other values which were calculated in this study reflect the state of these data. In the future, for use in practice, the results of the actual use of these data in the insemination of cows will be obtained.

Some variables such as body temperature, semen concentration and extraction temperature, semen quality during freezing processes, quantity as well as morphology and % acrosome, can also be used in calculations as input data. However, such data were not considered by the company in the dataset. Despite this, promising performances were obtained in bovine gender prediction by using the provided dataset of the limited number of variables.

Table 4. Comparison of some of the outputs for the FL, ANN; SVM and RF methods

Sample No	Inputs			Outputs (Actual)		Calculated Outputs							
						ANN		FL		SVM		RF	
	Speed (μ/s)	Magnitude (μ)	Density (μg/μm^3)	Female (%)	Male (%)	Female (%)	Male (%)	Female (%)	Male (%)	Female (%)	Male (%)	Female (%)	Male (%)
1	47.07	68.60	61.08	**81.89**	18.11	**82.54**	17.46	**82.95**	17.05	**81.83**	18.17	**83.58**	16.42
2	40.60	56.03	44.76	16.06	**83.94**	21.42	**78.58**	48.22	**51.78**	16.71	**83.29**	16.71	**83.29**
3	45.18	60.66	45.96	10.04	**89.96**	18.02	**81.98**	34.51	**65.49**	8.16	**91.84**	14.11	**85.89**
4	43.78	62.02	61.00	**86.25**	13.75	**97.50**	2.50	**83.63**	16.37	**84.03**	15.97	**79.27**	20.73
5	49.35	64.77	47.96	10.65	**89.35**	3.05	**96.95**	33.53	**66.47**	9.40	**90.60**	7.42	**92.58**
6	43.87	63.94	56.93	**76.32**	23.68	**70.46**	29.54	**81.95**	18.05	**78.60**	21.40	**84.32**	15.68
7	44.74	62.02	50.98	**80.64**	19.36	**74.82**	25.18	32.91	**67.09**	**46.12**	53.88	**77.83**	22.17
8	50.99	69.42	48.20	14.27	**85.73**	**79.47**	20.53	33.39	**66.61**	20.48	**79.72**	37.83	**62.17**
9	45.37	64.28	51.93	**86.27**	13.73	42.91	**57.09**	46.05	**53.95**	**85.65**	14.35	**83.66**	16.34
10	43.78	63.55	49.15	9.66	**90.34**	27.91	**72.09**	33.72	**66.28**	43.70	**56.30**	16.56	**83.44**

ACKNOWLEDGEMENT

The research reported here is supported by KTO Karatay University. The calculations are performed at the Computer Engineering Laboratory at the Department of Computer Engineering, KTO Karatay University. The authors also thank the Company Super Genetics Ltd. in Konya for their sharing the data.

Author Contributions

Ali Öztürk: Random Forest method and Support Vector Machines method. **Novruz Allahverdi:** Fuzzy Logic method. **Fatih Saday:** Artificial Neural Networks method. The rest of the paper was prepared together.

Conflict of Interest

There is no conflict of interest between the authors.

REFERENCES

Adeli H & Hung S L (1995). Machine learning - neural networks, genetic algorithms and fuzzy systems. John Wiley & Sons Inc. ISBN: 9780471016335.

Allahverdi N & Saday F (2018). An artificial neural network study for predicting sex in bulls. 7th International Conference on Advanced Technologies (ICAT'18), 727-731, Antalya, Turkey.

Allahverdi N (2002). Uzman Sistemler. Atlas, Istanbul, Turkey (in Turkish). ISBN: 975-6574-11-9.

Allahverdi N (2020). Bulanık Mantık ve Tıptaki Uygulamaları. KTO Karatay Üniversitesi Yayınları, Konya, Turkey (in Turkish). ISBN:9786056934636.

Anderson G B (1997). Identification of embryonic sex by detection of H-Y antigens. Theriogenology, 27, 81-97.

Bobillo F & Straccia U (2008). Towards a Crisp Representation of Fuzzy Description Logics under Łukasiewicz Semantics. International Symposium on Methodologies for Intelligent Systems (ISMIS 2008), 309-318, Toronto, Canada.

Breiman L (2001). Random forests. Machine Learning, 45 (1), 5-32.

Erten O & Yılmaz O (2012). Techniques of sex-selected calf production in dairy cattle breeding. Van, Yüzüncü Yil Üniversitesi Journal of Veterinary Faculty, 23 (3), 155-157 (in Turkish).

Frank E, Hall M A & Witten I H (2016). The WEKA Workbench Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques. 4th ed. San Francisco, CA, USA: Morgan Kaufmann. ISBN:9780128042915.

Heide, E.M.M., Veerkamp, R.F., Pelt, M.L., Kamphuis, C., Athanasiadis, I. et al., (2019), Comparing regression, naive Bayes, and random forest methods in the prediction of individual survival to second lactation in Holstein cattle, Journal of Dairy Science, 102 (10), 9409-9421.

Huma ZE & Iqbal F (2019). Predicting the body weight of Balochi sheep using a machine learning approach. Turkish Journal of Veterinary and Animal Sciences, 43, 500-506.

Inanc M E, Çil B, Tekin K & Alemdar H (2018). The combination of CASA kinetic parameters and fluorescein staining as a fertility tool in cryopreserved bull semen. Turkish Journal of Veterinary and Animal Sciences, 42, 452-458.

Johnson L A, Cran D G & Polge C (1994). Recent advances in sex preselection of cattle: Flow cytometric sorting of X-Y-chromosome bearing sperm based on DNA to progeny. Theriogenology, 4, 51-56.

Martiskainen P, Jarvinen M, Skön J P, Tiirikainen J, Kolehmainen M, et al. (2009). Cow behaviour pattern recognition using a three-dimensional accelerometer and support vector machines. Applied Animal Behaviour Science, 119 (1-2), 32-38.

Miekley B, Traulsen I & Krieter J (2013). Mastitis detection in dairy cows: the application of support vector machines. The Journal of Agricultural Science, 151 (6), 889-897.

Mikail N & Keskin I (2013). Application of the support vector machine to predict subclinical mastitis in dairy cattle. The Scientific World Journal, 1: 603897.

Nicolas G, Robinson TP, Wint W & Conchedda G (2016). Using Random Forest to Improve the Downscaling of Global Livestock Census Data. PLoS ONE, 11 (3), e0150424.

Niemann H & Meinecke B (1993). Embryo transfer und assoziierte biotechniken bei landwirtschaftlichen nutztieren. Ferdinand Enke Verlag, Stuttgart (In German). ISBN: 9783432254715.

Oztemel E (2016). Yapay Sinir Ağları. Papatya Yayınları. Istanbul, Turkey (in Turkish). ISBN: 9789756797396.

Parati K, Bongioni G, Aleandri R & Galli A (2006). Sex ratio determination in bovine semen: A new approach by quantitative real-time PCR. Theriogenology, 66, 2202–2209.

Seidel G EJ (2003). Economics of selecting for sex: the most important genetic trait. Theriogenology, 59, 585-598.

Sendag S, Aydin I & Celik HA (2005). Prenatal embryonic or fetal sex determination in cows. J Fac Vet Med., Univ. Erciyes, 2 (1), 39-44 (in Turkish).

Shevade S K, Keerthi SS, Bhattacharyya C & Murthy K R K (2000). Improvements to SMO algorithm for SVM regression. IEEE Transactions on Neural Networks, 11(5), 1188–1193.

Smola A J & Schölkopf B (1998). A tutorial on support vector regression. NeuroCOLT Technical Report TR 1998-030, Royal Holloway College, London, UK.

Ustun B, Melssen WJ, Buydens LMC (2006). Facilitating the application of support vector regression by using a universal Pearson VII function-based kernel. Chemometrics and Intelligent Laboratory Systems, 81, 29-20.

Vapnik V (1995). The Nature of Statistical Learning Theory. Springer-Verlag, New York. ISBN:9781475724400.

Vapnik V (1999). An overview of statistical learning theory. IEEE Transactions on Neural Networks, 10(5), 988–999.

Vapnik VN, Vapnik V (1998). Statistical Learning Theory. New York, USA: Wiley. ISBN: 9780471030034.

Vásquez R P, Anguilar-Lasserre A A, Lopez-Segura M V, Rivero LC, Rodriguez-Duran AA & Rojaz-Luna AA (2019). Expert system based on a fuzzy logic model for the analysis of the sustainable livestock production dynamic system. Computers and Electronics in Agriculture, 161, 104-120.

Xu Y (2017). Research and implementation of improved random forest algorithm based on Spark. IEEE 2nd International Conference on Big Data Analysis, 499–503, Beijing, China.

Yao C, Spurlock DM, Armentano LE, Page Jr C D, VandeHaar MJ et al. (2013). Random Forests approach for identifying additive and epistatic single nucleotide polymorphisms associated with residual feed intake in dairy cattle. Journal of Dairy Science, 96 (10), 6716-6729.

Zadeh L (1965). Fuzzy sets. Information and Control, 8, 338-353.