IJAS Iranian Journal of Applied Animal Science

Estimating Heritabilities and Breeding Values for Real and Predicted Milk Production in Holstein Dairy Cows with Artificial Neural Network and Multiple Linear Regression Models

Research Article

M. Nosrati[1], S.H. Hafezian[1*] and M. Gholizadeh[1]

[1] Department of Animal Science, Faculty of Animal Science and Fishery, Sari Agricultural Sciences and Natural Resources University, Sari, Iran

Received on: 17 Apr 2020
Revised on: 30 Jul 2020
Accepted on: 15 Aug 2020
Online Published on: Mar 2021

*Correspondence E-mail: h.hafezian@sanra.ac.ir

Online version is available on: www.ijas.ir

ABSTRACT

The success of a dairy herd depends on milk production. Prediction of future records can reduce recording time, accelerate the computation of genetic evaluations, decrease generation interval, and increase genetic progress. Multiple linear regression (MLR) is the most common prediction method. However, artificial neural networks (ANN) can handle complex linear and non-linear functions to solve a wide range of prediction problems. In this study, MLR and ANN models were applied to the prediction of 305-day milk production in the first and second lactations of dairy cows using variables related to milk production, test-day records and estimated breeding values (EBVs). The 305-day first lactation records were also used to predict 305-day second lactation records. ANN and MLR predictions were compared in terms of accuracy and efficiency. Dairy records from 7856 dairy cows in two herds were used in this research. The best ANN model was a multilayer perceptron with a back-propagation learning algorithm. Results showed that ANN and MLR predicted values were acceptable. However, ANN prediction accuracies for 305-day milk production in the first and second lactations were higher than those of MLR. Correlation coefficients between real and predicted 305-day milk production records in the first and second lactations ranged from 0.88 to 0.96 for ANN and from 0.66 to 0.89 for MLR. Adding test-day records and EBVs for 305-day milk production in the first lactation to the set of independent variables used to predict 305-day milk production in the second lactation increased more the prediction efficiency of ANN than MLR. Thus, ANN could be used to decrease the interval between collecting records and computing animal breeding values. In addition, real data and ANN-predicted data from the first lactation were used to compute EBVs. The correlation between EBVs with real and predicted data was 0.93. Results suggested that ANN could be useful for predicting complex traits using high dimensional genomic information.

KEY WORDS artificial neural network, generation interval, milk prediction, milk production, multiple linear regression.

INTRODUCTION

Selection of the livestock with superior production traits has greatly improved livestock production. The most important economic traits in dairy cows are quantitative traits influenced by multiple genetic and environmental factors. Selection of animals for these traits in breeding programs is performed based on breeding values obtained when they reach a certain age using phenotypes from the animals themselves and their relatives. This results in medium to long generation gaps and reduces annual genetic progress (Boichard *et al*. 2015). Advances in genetic improvement programs are aimed at developing increasingly more effective procedures for genetic selection of livestock.

Milk production is an important economic trait in dairy cattle. Production of milk with desirable quality characteristics has become the primary goal of dairy cattle improvement programs. To achieve rapid genetic progress, it is important to accurately predict the animal breeding values used for genetic selection (Miglior *et al*. 2017). Linear models have been extensively used to predict milk production and estimate the genetic parameters for a single trait and for multiple traits. Accurate prediction of breeding values with linear models relies on large datasets containing correct information. Thus, small datasets or datasets with incorrect information may lead to erroneous outcomes (Visentin *et al*. 2015).

The use of artificial neural networks (ANN) has also been used for modeling in various fields of science and has had significant success over the past three decades. ANN are easy to use and have a high capacity for modeling complex functions and relationships (Ashton, 2013; Karadas *et al*. 2017). Different types of input variables are used in these networks and have the ability to model biological processes that are inherently nonlinear. If the data are influenced by a large number of ambiguous and complex mathematical factors and parametric methods cannot be correctly inferred. ANN can obviate the difficulties of linear models without specifying the input variables in advance (Khazaei and Nikosiar, 2008; Abbasi *et al*. 2016).

ANN can learn nonlinear and multidimensional relationships between independent (input) and dependent (output) variables. ANN are used in a variety of disciplines and fields (Park *et al*. 2005; Njubi *et al*. 2010; Bahreini Behzadi, 2015). Using predicted data instead of real data can play a significant role in the prediction of breeding values even in unborn and unrecorded animals, the estimation of the production records of subsequent periods, and the reduction of maintenance costs (Karadas *et al*. 2017).

Much research has been done in the animal sciences about ANN independently or in comparison with regression models. The studies on predicting the performance of milk production (Gorgulu, 2012; Shahinfar *et al*. 2012; Pour Hamidi *et al*. 2017), predicting the growth in Baluchi sheep (Bahreini Behzadi *et al*. 2010), investigating calving period with economic traits of dairy cows (Ghaderi Zefrehei *et al*. 2016), and predicting semen production in rams using phenotypic traits (Qotbi *et al*. 2010) are among the studies that represent the high capabilities of the simultaneous use of linear and nonlinear models through ANN compared to linear methods.

Boniecki *et al*. (2013) predicted the average daily milk production of dairy cows with daily and peak temperature data using ANN modeling and reported a high correlation between real and predicted data. They introduced the multilayer perceptron neural network as the most appropriate network for predicting short-term milk production. Adesh *et al*. (2007) compared MLR and ANN models to predict milk production for 305 days during the first lactation based on partial lactation records and reported a higher capability of the ANN model.

Chaturvedi *et al*. (2013) obtained acceptable predictions for cow milk production using a back-propagation neural network with two hidden layers for nonlinear relationships between independent variables and milk production. Some researchers used a hidden layer with a sigmoid activation function in ANN and different independent variables to predict milk production in sheep. Ince and Sofu (2013) reported high predictability and accuracies for sheep milk production with an artificial neural network. Njubi *et al*. (2009) found that milk production in the first lactation was better predicted by ANN than MLR.

Simultaneous use of linear and nonlinear relationships has made ANN one of the most useful modeling techniques for conducting high-precision simulations when mathematical linear models are not responsive. The increase in the use of ANN is due to its flexibility and ability to utilize linear and nonlinear modeling of systems without the use of prior knowledge (Guresen *et al*. 2011).

Therefore, it may be possible to use the capabilities of ANN for the prediction of trait records and estimation of heritabilities and breeding values with a reasonable level of confidence. This way, production and economic efficiency can be increased by removing low producing animals as well as compensating for missing parental and ancestral records.

Therefore, the objective of this study was to predict livestock production data by MLR and ANN models, compare predicted and real data, and then replace the real data with predicted data if an acceptable degree of confidence and non-significant deviation are obtained. Then, the real data, the predicted data from MLR models, and the predicted data from ANN models will be used in genetic models to estimate the hereditability of production traits, predict breeding values of animals, and compare results among the three types of data. The prediction error for dependent variables (difference between values computed by ANN and target values) was the basis for deciding whether to use ANN data to estimate these variables.

MATERIALS AND METHODS

Data

Records from Holstein dairy cows associated with the Breeding Center and the Livestock Production Improvement of Iran were used in this study. Data from the two herds with the largest number of records was used to perform statistical tests. Records from cows with at least the

first two lactations were kept and the remaining records were removed. The remaining data included 8725 dairy cows with milk production and other related records between 2001 and 2019. These records were used to predict total milk production in the first and second 305-day lactation periods using ANN and MLR models. After editing the generated data and deleting inappropriate data in the two herds, the final dataset included records from 7856 animals. The record for each animal included herd, sire, dam, age, parity, date of calving, lactation days, dry period, Holstein percentage, test-day records, and 305-day milk production.

The edited dataset was processed using R statistical software. Statistical summaries of each of the variables were examined and then used for the ANN and MLR analyses.

Before transferring data to an ANN and depending on the domain of livestock production, the input data was sorted in ascending order based on a production interval in several floors and were separated into two matrices to perform ANN calculations (matrix of input variables and matrix of output variables). Before entering the matrix of input variables to ANN, it was standardized and normalized to increase the speed and accuracy of the network. Because each parameter had a different range, normalization utilizes the range of the parameters to place numbers within a finite range to prevent over-reduction of weights (Kasy and Kummer, 2008). Relation 1 was used for normalization and the data were standardized in the range of 0.1 to 0.9.

$$N_i = 0.8 \times ((X_i - X_{min})/(X_{max} - X_{min})) + 0.1 \quad (1)$$

Where:
N_i: standardized values.
X_i: real values.
X_{min}: lowest real value and the highest real value.

Due to the large size of the input matrix, the effect of each component was calculated to reduce the size of the input matrix and prevent linear dependence between its columns. Components without a significant effect on the response variable or components with a linear dependence on the main components were removed from the input matrix to eliminate the correlation between elements of the input vector.

The data were then randomly divided into three categories: training, validation, and test. Because the placement of marginal values in the training set improves the performance of the models, these values were included in the training set to increase the performance of ANN.

Steps and states of the prediction of milk production

The main steps of the research are shown in Figure 1. These steps were repeated for each of the different models. All coding operations for the ANN and MLR models were performed in the R software environment.

The ANN and MLR models were examined using independent variables in two lactation periods, and the results of each model were compared with the real data.

In the first state, prediction of milk production in the first lactation period was performed using the independent variables of herd, sire, dam, age, parity, date of calving, lactation days, dry period and Holstein percentage, and then test-day data from the first four months of lactation was used to predict 305-day milk production in the first lactation.

In the second state, the prediction of milk production for the second lactation period was performed using independent variables of herd, sire, dam, first-calving age, second-calving age, parity, date of calving, lactation days, dry period, Holstein percentage and 305-day milk production of the first lactation and then test-day data from the first four months of lactation was used to predict 305-day milk production in the second lactation.

Subsequently, animal breeding values were estimated using the records of the first lactation period.

In the third state, the combination of the estimated breeding values and the second state were used to predict 305-day milk production in the first and second lactation periods.

Estimating animal breeding values as early as possible in animal's lives is important to speed up genetic progress. Thus, predicted first lactation records (which had the best correlation with the actual records) and real first lactation records were used to estimate genetic parameters and animal breeding values with single-trait animal models. Results from these analyses were compared to assess the ability of ANN and MLR predictive models for estimating animal breeding values.

Genetic analysis and estimation of breeding values

Wombat software was used to estimate breeding values. First, general editing of the records was performed. Records from animals whose first calving age was not in the range of 18 to 40 months and animals that had obvious contradictions in the information were removed (Hashemi and Nayebpour, 2008).

Then, analysis of variance and generalized linear model (GLM) procedure in R software were used to determine the level of significance of the effect of non-genetic factors (herd, calving season, age at calving, lactation days, dry period, etc.) on milk production.

The Wombat input data file included herd number, animal number, sire number, date of birth, date of calving, drying off date, lactation period, milk production (lactation records corrected to 305 days and two milkings per day) and recording date of each animal.

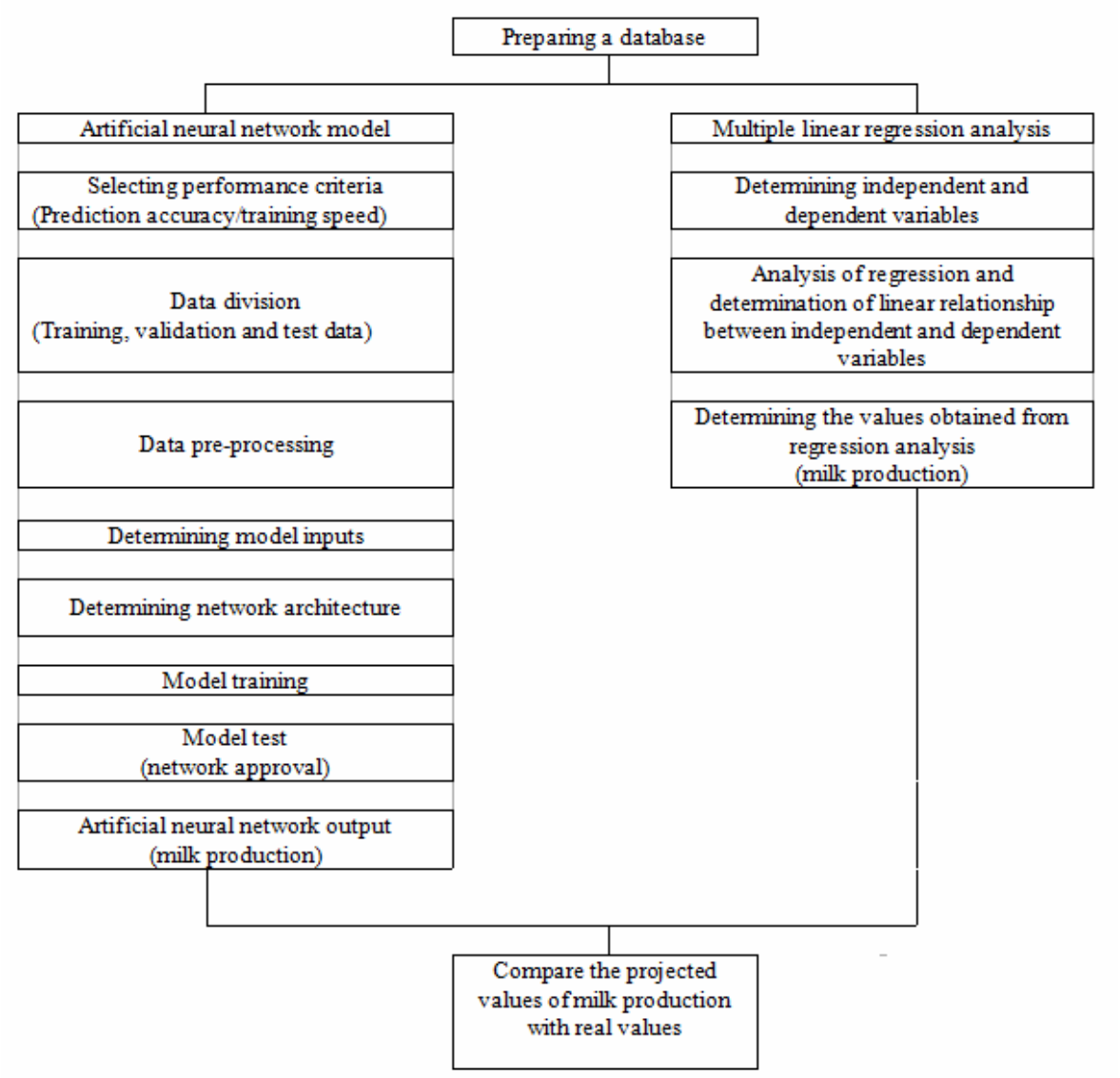

Figure 1 Steps of research

The pedigree file included the animal number, sire number, and dam number for animals from both herds. Because Wombat required the animal number to be larger than the number of its sire and dam, numbers in the pedigree file were recoded with software Pedigree 0.1 to ensure that progeny had larger numbers than their parents. The recoded pedigree file contained zeroes for unknown parents in the original pedigree. To match the actual and coded numbers, the original pedigree file was combined with the recoded pedigree file via R software.

Genetic analysis of 305-day first lactation records with a single-trait animal model

A single-trait animal model was used to estimate variance components and animal breeding values for 305-day first lactation records with the average information-restricted maximum likelihood algorithm (AI-REML) from Wombat. The single-trait animal linear mixed model in matrix notation was as follows:

$$y = Xb + Za + e \quad (2)$$

Where:

y: vector of either real or predicted 305-day first lactation records.

b: vector of fixed effects.

a: vector of random animal additive genetic effects.

e: vector residual random effects.

X: an incidence matrices relating records in vector y to fixed effects in vector b.

Z: an incidence matrix relating records in vector y to random additive genetic effects in vector a.

The mixed model equations (MME) for the single-trait animal model was as follows:

$$\begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}\alpha \end{pmatrix}\begin{pmatrix} \hat{b} \\ \hat{a} \end{pmatrix} = \begin{pmatrix} X'y \\ Z'y \end{pmatrix} \quad (3)$$

Where:

$\mathbf{A^{-1}}$: inverse of pedigree relationship matrix and $\alpha = \frac{\sigma_e^2}{\sigma_a^2} = \frac{1-h^2}{h^2}$

The single-trait animal linear mixed model in scalar notation was as follows:

$$y_{ijk} = M + HYS_i + b(age_{ij} - \overline{age}) + a_j + e_{ijk} \quad (4)$$

Where:

Y_{ijk}: k^{th} real or predicted 305-day first lactation record of the j^{th} animal in the i^{th} herd-year-season.

M: population mean.

HYS_i: i^{th} fixed herd-year-season effect.

age_{ij}: regression coefficient of 305-day first lactation records on the calving age cows of the j^{th} age in the i^{th} herd-year-season.

$\overline{age}$: mean calving age of first-lactation cows.

a_j: random additive genetic effect of the j^{th} animal.

e_{ijk}: k^{th} residual of the j^{th} animal in the i^{th} herd-year-season.

Implementation of the multiple linear regression model

A backward stepwise multiple regression approach in R was used to obtain a valid MLR model. The dependent variable in the MLR model was 305-day first lactation milk yield. The independent variables in the MLR model were mean number of lactation days, cow age at first calving, drying off date, paternal genetic group, age at first calving, record number for lactation period, minimum number lactation days, maximum number of lactation days. To predict 305-day milk production in the second lactation, the age of cow at first calving and the 305-day milk production in the first lactation and (or) the predicted breeding values for 305-day milk production in the first lactation were also included in the MLR model.

The backwards stepwise MLR kept an independent variable if it was significant and uncorrelated with other independent variables, otherwise it was removed. The regression coefficients estimated for the variables in the MLR model indicated the importance of each variable for 305-day milk production in either the first or the second lactation.

The size of the regression coefficients indicated the extent of the effect and its sign indicated the direction of its association with milk production. Positive regression coefficients indicate that an increase in the value of an independent variable will increase milk production, whereas negative regression coefficients indicate that as the value of an independent variable increases milk production will decrease.

Implementation of the proposed artificial neural network model

To create an ANN suitable for predicting milk production, the data were divided into three sets: network training, validation, and test. For better performance of the ANN model, the statistical characteristics of the training and test datasets were considered to be almost identical, and marginal values were used in the training set. Also, in the training step, a linear approach was used for preprocessing, to turn the data from the irreal domain into a domain that has more efficient neural network (Adamowski and Chan, 2011).

Due to the type of data and the high number of independent variables affecting milk production, the ANN was monitored using a supervised learning method. In supervised learning, values of dependent variables are specified and by calculating the difference between the network output and the target variable values, the prediction error is measured for each observation. Then using different repetition algorithms such as the back-propagation algorithm, network weights are adjusted to minimize prediction errors and thus the training network is given by changing weights in each iteration. Multi-layer perceptron (MLP) and radial basis function (RBF) structures were used as feed-forward or feedback with the back-propagation learning algorithm.

Independent variables of different inputs, data splitting method, initial weights, number of hidden layers, number of neurons in each hidden layer, different activation functions including hyperbolic tangent, sigmoid, linear hyperbolic tangent and various training algorithms including conjugate gradient descent algorithm, Quasi-Newton algorithm, Levenberg-Marquardt and online back-propagation for network training were examined in both hidden and output layers, as shown in Table 1. One of the factors with a great impact on model learning is the number of training data for network training, which was examined in the range of 60 to 80 percent, 15 percent of which was used as validation data. Each input must be weighed before entering the main core of the processor element, which is multiplied by an initial weight (W) and a biased value, and the result is multiplied by an activation function.

Table 1 Range of parameters used in the artificial neural network model

Artificial neural networks (ANN) parameters	Test range
Number of inputs	1-10
Number of hidden layers	1-4
Number of neurons at hidden layers	1-8
Transfer function	Linear-Sigmoid-Hyperbolic Tangent
Learning algorithm	Conjugate gradient descent algorithm-Quasi-Newton algorithm-Levenberg Marquardt-online back-propagation

The activation function was used to calculate the output layers from the network input and the network training function to optimize the weights and biased values of the network.

The best model, i.e., the model with the smallest prediction error in each of the states was obtained through trial and error. The optimal model was constructed by first determining the parameters for the ANN and then testing it using validation data.

The ANN model for the prediction of 305-day milk production in the first lactation (output variable) contained mean lactation days, cow age at calving, drying off date, paternal genetic group, record number for lactation period, minimum number of lactation days, maximum number of lactation days as independent input variables. To predict 305-day milk production in the second lactation, in addition to the input variables for the first lactation, the ANN model included the age of cow at first calving and the 305-day milk production of the first lactation period and (or) the predicted breeding values for 305-day milk production in the first lactation.

Evaluation of prediction accuracy

The predictive ability and accuracy of the ANN and MLR models were assessed using pearson's correlation coefficients between real and predicted observations, coefficients of determination, and root mean square errors (RMSE). The pearson correlation coefficient (r) was obtained with equation 5, the coefficient of determination with equation 6, and root mean square error with equation 7:

$$r = \frac{\mathrm{cov}(y_i, \hat{y}_i)}{\sigma_y . \sigma_{\hat{y}}} \quad (5)$$

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (6)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{n}} \quad (7)$$

Where:

n: number of records.

y_i: real value.

$\hat{y}_i$: predicted value.

$\bar{y}$: mean of real records.

cov: covariance.

σ: standard deviation.

The ideal model would have pearson correlation and a coefficient of determination equal to 1, and a root mean square error and a mean absolute error equal to zero.

RESULTS AND DISCUSSION

Description of data

After general data edition and proper data extraction, 7856 animals from two herds were selected for the first and second lactation periods. The remaining data were processed using R statistical software and descriptive statistics for each trait were examined. The descriptive statistics and the graphs related to the variables of Holstein percentage, dry period, lactation days, age of calving and milk production during the first and second lactation periods are shown in Tables 2 and 3.

By calculating the effect of each component, the components without a significant effect on the dependent variable and explanatory variables linearly dependent on other explanatory variables were removed from the prediction model.

Descriptive statistics for the training and test datasets

Use of 75% of the data for network training showed the best performance in ANN models, with 60% of these data used in the training set and another 15% used to validate the model. Table 4 presents the statistical analysis of the parameters for the training, validation, test and total data sets.

The statistical characteristics of the training and test datasets were highly similar, leading to a better performance of the ANN model. The skewness coefficients for the dependent variables of 305-day milk production in the first and second lactations were low.

Table 2 Characteristics of independent input variables and dependent output variables in the first lactation

Trait	Mean	Standard deviation	Lowest	Middle	Highest
Dry period	60.0	6.7	32.8	59.9	85.8
Lactation days	311.7	50.1	201	305	399
Age of calving in period (days)	753	68.3	455	740	999
Test day production 1	32.99	7.19	2.43	33.20	49.42
Test day production 2	33.92	8.48	2.51	34.14	50.81
Test day production 3	34.80	8.27	2.56	35.02	52.13
Test day production 4	29.52	8.04	2.17	29.71	44.22
305-day milk production	10170.0	2479.9	749.8	10236.3	15234.4

Table 3 Characteristics of independent input variables and dependent output variables in the second lactation

Trait	Mean	Standard deviation	Lowest	Middle	Highest
Dry period	57.9	6.9	30.8	58.0	89.7
Lactation days	302	39.3	67	300	399
Age of calving in period (days)	1140	68	844	1126	1389
305-day milk production	10914.2	2479.9	1494.6	10980.2	15978.8

Table 4 Statistical analysis of training data, validation, test, and 305-day milk production in the first and second lactations

Parameter		Number of data	Mean	Maximum	Minimum	Standard deviation	Skewness coefficient	Autocorrelation coefficient
305-day milk production first lactation	Training	4713	10050.7	15234.4	749.3	2486.3	-0.21	0.964
	Validation	1178	10126.2	14235.6	1123.6	2103.2	-0.32	0.932
	Test	1964	10487.3	12256.7	3231.6	894.3	-0.27	0.958
305-day milk production second lactation	Training	4713	10965.6	15978.8	1494.6	1524.3	-0.18	0.971
	Validation	1178	10855.3	14231.3	2561.5	1412.2	-0.23	0.948
	Test	1964	10833.2	11532.9	6548.8	854.6	-0.21	0.953

These low values of the skewness coefficient improved the performance of the ANN. The high autocorrelation coefficient of the dependent variables in the training and test sets also increased the performance of ANN. In general, Table 4 shows satisfactory statistical characteristics for the training and test sets in terms of mean values and correlation coefficients, which increased the performance of the prediction model.

Estimation of animal breeding values using records from the first lactation

A single-trait animal model was used to estimate the genetic parameters for 305-day milk production in the first lactation. The genetic variance and phenotypic variance for 305-day milk production in the first lactation were equal to 345126 kg^2 and 1380504 kg^2, resulting in an estimate of heritability of 0.265 ± 0.01. Predicted animal breeding values ranged between -4.236 and 5.516 kg of milk.

Structure of artificial neural network and multiple linear regression models

The feed-forward multi-layered perceptron artificial neural network was selected in all states, with the back-propagation learning algorithm including one input layer, two hidden layers, and one output layer as the basis for the best structure.

Increasing the number of hidden layers had no effect on increasing the efficiency of the models in any of the states.

To determine network factors, with regard to error back-propagation learning algorithm method, the calculated error values were transferred to the previous layers after the ANN output calculations, to determine the initial structure of the ANN for training. Then, the random weight matrix was assigned to all model connections, and the initial weights were corrected sequentially by the error propagation method to minimize differences between real and predicted output values. The performance of the ANN was investigated by changing the structure and specifications of the model and computing pearson correlation coefficients between real and predicted values, coefficients of determination, and mean square errors. The training was completed when the structure of the ANN model was optimized in each state. Significant independent explanatory variables for 305-day milk production in each MLR model were determined in the training dataset using a stepwise procedure until the final model and multiple linear regression coefficients were determined.

Structure of the artificial neural network model for predicting 305-day milk production in the first state

The parameters of the best ANN model for the prediction of 305-day milk production in the first lactation were determ-

ined using a test dataset and prediction errors. The independent variables in the ANN model were herd, sire, dam, age, parity, date of calving, lactation days, dry period and Holstein percentage. Nine input variables were selected for the model in two hidden layers. The number of neurons in the first hidden layer was six and the number of neurons in the second hidden layer was four.

The functions of sigmoid transmission and hyperbolic tangent, as well as the learning algorithms of Levenberg Marquardt and descending gradients, did not make a significant difference in the efficiency of the model. The descending gradient algorithm seeks a vector to optimize the weight space to minimize errors. This algorithm starts with a desired value for the weight vector and it changes the weights at each step in such a way that the error decreases in the direction of the descending slope of the curve. The output layer activation function was selected linearly.

By adding test-day data for the first four months to the independent variables and increasing the number of inputs to 13, the structure of the best predicting ANN model did not change in terms of the number of hidden layers, transmission function, and learning algorithm, and the parameters of the number of neurons in the first hidden layer and the number of neurons in the second hidden layer were changed to 7 and 4 neurons, respectively.

Structure of the artificial neural network model for predicting 305-day milk production in the second state

To predict 305-day milk production in the second lactation, the 305-day milk production in the first lactation was added to the set of input variables of the ANN model. The sigmoid transfer and hyperbolic tangent functions as well as Levenberg Marquardt and descending gradient learning algorithms provided similar levels of efficiency for ANN.

Structure of the artificial neural network and multiple linear regression models in the third state

The efficiency of the ANN model to predict 305-day milk production in the second lactation was increased by adding the animal estimated breeding values from the 305-day milk production in the first lactation to the input variables, using the hyperbolic tangent transfer function instead of the sigmoid function, and the Levenberg Marquardt learning algorithm instead of the descending gradient learning algorithm.

Network training with the Levenberg Marquardt algorithm is appropriate for prediction of 305-day milk production in the second lactation because of its speed, accuracy, and reliability, and ability to create nonlinear mathematical relationships for interpolation.

Prediction by artificial neural network and multiple linear regression models

The efficiency of ANN and MLR models for predicting the dependent variable was examined after determining their best structure for each state using pearson correlation coefficients, coefficients of determination, root-mean square errors and mean absolute error values.

Prediction by artificial neural network and multiple linear regression models in the first state

The 305-day milk production in the first lactation was predicted with the ANN model using the independent variables of herd, sire, dam, date of calving, Holstein percentage, dry period, lactation days and age at first calving, whereas the MLR model included the independent variables of Holstein percentage, dry period, lactation days, and age at first calving. The values of the statistics used to determine the efficiency of the ANN and MLR models to predict 305-day milk production in the first lactation, namely the correlation coefficient, determination coefficient, and root-mean square error are presented in Table 5. The various datasets showed positive pearson correlation coefficients between real and predicted 305-day milk production in the first lactation. However, the higher correlation coefficients and coefficient of determination as well as the lower root mean square values for ANN than MLR indicated a higher predictive ability of ANN than MLR for 305-day milk production in the first lactation. The performance of ANN and MLR increased by adding test-day records to the input variables.

Prediction by artificial neural network and multiple linear regression models in the second state

The 305-day milk production in the second lactation was predicted with the ANN model using the independent variables of herd, sire, dam, date of calving, Holstein percentage, dry period, lactation days, first calving age, second calving age, and 305-day milk production in the first lactation, whereas the MLR model included the independent of Holstein percentage, dry period, lactation days, first calving age, and second calving age. The values of the statistics used to determine the efficiency of the ANN and MLR models to predict 305-day milk production in the second lactation are shown in Table 5. The efficiency of the ANN and MLR models for predicting 305-day milk production in the second lactation was higher than in the first lactation, and the increase in efficiency was higher for ANN model than for the MLR model. Adding test-day data from the first lactation also increased the efficiency of the ANN and MLR models, although this increase was lower than in the first state (Table 5).

Table 5 Statistics used to determine the efficiency of the artificial neural network and multiple linear regression models

State	Models	Pearson correlation coefficient (r)	Coefficient of determination (R^2)	Root mean square error (RMSE)
1	Artificial neural network	0.85	0.72	0.28
	Multiple linear regression	0.66	0.56	1.38
2	Artificial neural network	0.92	0.86	0.24
	Multiple linear regression	0.81	0.66	1.02
3	Artificial neural network	0.87	0,76	0.19
	Multiple linear regression	0.77	0.59	0.24
4	Artificial neural network	0.93	0.88	0.17
	Multiple linear regression	0.81	0.65	0.22
5	Artificial neural network	0.95	0.91	0.14
	Multiple linear regression	0.87	0.75	0.18
6	Artificial neural network	0.96	0.93	0.12
	Multiple linear regression	0.89	0.80	0.17

State 1: 305-day milk production in the first lactation using independent variables; State 2: 305-day milk production in the first lactation using independent variables and test day data; State 3: 305-day milk production in the second lactation using independent variables from the first lactation and 305-day milk production in the first lactation; State 4: 305-day milk production in the second lactation using independent variables from the first lactation, 305-day milk production in the first lactation, and test-day records from the first lactation; State 5: 305-day milk production in the second lactation using independent variables from the first lactation, 305-day milk production of the first lactation, and animal breeding values for 305-day milk production in the first lactation and State 6: 305-day milk production in the second lactation using independent variables from the first lactation, 305-day milk production of the first lactation, test-day records from the first lactation, and animal breeding values for 305-day milk production in the first lactation.

Prediction by artificial neural network and multiple linear regression models in the third state

Animal estimated breeding values were used as an input variable for the ANN and MLR models to predict 305-day milk production in the second lactation, and the second state was reexamined. Table 5 presents the values of the statistics used to determine the efficiency of the ANN and MLR models to predict 305-day milk production in the second lactation with animal estimated breeding values added as an independent variable are presented in Table 5. Adding estimated breeding values to the input variables increased the ability of ANN and MLR to predict 305-day milk production in the second lactation, and this increase was higher for ANN than for MLR. Thus, the predictive accuracy of ANN was higher than that of MLR to predict 305-day milk production in all three states.

Estimation of genetic parameters and animal breeding values with predicted records

Single-trait animal models were used to assess prediction ability of ANN by comparing estimated genetic parameters and animal breeding values from real 305-day milk production in the first lactation and 305-day milk production in the first lactation predicted with the ANN model.

The heritability for 305-day milk production in the first lactation with ANN predicted data was equal to 0.271, which was slightly higher than the value of 0.265 estimated using real data. Table 6 shows the pearson correlation coefficient, coefficient of determination, and root mean square error used to compare estimated breeding values from the real and predicted 305-day milk production in the first lactation.

These statistics indicate that the ANN predicted 305-day milk production in the first lactation can be used to estimate animal breeding values.

Accurate prediction of milk production records will increase animal husbandry profits by reducing the length of the recording period and saving time, as well as reducing costs. By estimating animal breeding values with predicted records before real records are available, the generation interval is reduced and the selection intensity and genetic progress per year are increased. Estimated breeding values based on predicted data could be used for animal selection if production records are predicted with high accuracy. The type of prediction model, parameters and input variables were the most important factors affecting the accuracy of prediction evaluated in this study. Multiple linear regression, a common statistical model, and ANN, a model can be easily applied to complex functions, were the two prediction models analyzed in this study. Various parameters were examined to measure the accuracy of prediction of ANN and MLR for 305-day milk production in the first and second lactations using independent variables affecting milk production, test-day records and estimated breeding values.

Pour Hamidi *et al*. (2017) predicted the breeding value of milk production using sire, herd, twice daily milking production, season and month via a neural network model. In this study, 70% of the data were used for training, 15% for testing and 15% for validation. The Levenberg-Marquardt algorithm with 1000 iterations was selected for network training and the sigmoid tangent function for activation. pearson coefficients and root-mean-square errors showed a better predictive ability for ANN than for MLR.

Table 6 Statistics used to compare estimated breeding values computed with real and predicted 305-day milk production in the first lactation from the artificial neural network model

Model	Pearson correlation coefficient (r)	Coefficient of determination (R^2)	Root mean square error (RMSE)
Artificial neural network	0.93	0.87	0.90

Safari (2016) used various ANN structures with a supervised training method and multilayer perceptron structure with a back-propagation error algorithm to predict milk production in two breeding herds using herd, age, abdomen, and milk production records from the first to the tenth month of lactation as independent variables. In our study, a hidden layer, the Levenberg-Marquardt algorithm, and sigmoid activity and hyperbolic tangent functions were selected as the best network options by examining the pearson correlation coefficient, the coefficient of determination, the root-mean-square error, and the mean absolute error as ANN prediction criteria.

Results here showed that ANN and MLR could be highly efficient for predicting milk production, and that the ANN model had a higher correlation coefficient and a lower mean square error than the MLR model. Given that milk production is affected by multiple factors and the relationship between these factors and milk production may be either linear or nonlinear, it could be concluded that the reason for the higher accuracy of prediction of ANN was due to high compatibility, nonlinearity, generalizability, error tolerance, and the ability to solve complex problems through the learning process without taking into account any preconditions of the input data.

The multi-layer perceptron model with error back-propagation learning algorithm and sigmoid transfer and hyperbolic tangent functions was selected as the best model for predicting milk production among ANN models based on high values for independent variables affecting milk production.

The accuracy of prediction of 305-day milk production in the first and second lactations was increased by adding test-day records to the other prediction variables. Results showed that the learning ability of the ANN model increased, and because the lactation curve changes during lactation, utilization of test-day records provides ANN information on the lactation curve, hence increasing the accuracy of prediction. If the 305-day milk production in the first lactation is used together with test-day records to predict 305-day milk production in the second lactation, the learning level of the lactation curve increased further, and so does the predictive accuracy of ANN.

If the animal breeding values are estimated for 305-day milk production in the first lactation, and they are used to predict 305-day milk production in the second lactation, the learning ability and efficiency of ANN will also increase.

The predictive efficiency of the MLR model to predict 305-day milk production in the second lactation also increased with either the addition of 305-day milk production in the first lactation or the animal estimated breeding values for 305-day milk production in the first lactation to the set of independent variables. However, because dependent variables may not be linear, MLR is less efficient than ANN.

The results of this study are consistent with the results of various studies that have been conducted on a case-by-case basis in the field of neural network application to prediction of production traits. Grzesiak *et al*. (2003) reported the effectiveness of ANN for predicting 305-day milk production of dairy cows in Poland. Hosseinia *et al*. (2007) using two different types of ANN and different independent input variables predicted two milk and fat production traits in the second lactation and reported coefficients of determination ranging from 0.59 to 0.90. Gorgulu (2012) found coefficients of determination for various ANN for milk production ranging from 0.38 to 0.90 and indicated that the efficiency of ANN was higher MLR. Nobari *et al*. (2019) comparing real and predicted data reported that ANN using the activation function of the hyperbolic tangent and the Levenberg-Marquardt algorithm were the best options to optimize prediction of milk production prediction in a multi-layered perceptron structure. Ali-Jafari (2016) compared Holstein milk production predicted using test-day records with an ANN and a genetic algorithm. Ali-Jafari (2016) used an ANN with 6 to 15 input variables, 3 neurons in the hidden layer, one neuron in the output layer, and the Levenberg-Marquardt training function. Ali-Jafari (2016) reported that while both methods had the required accuracy for predicting milk production, the coefficient of determination for prediction accuracy was higher for the genetic algorithm than that for ANN.

CONCLUSION

The 305-day milk production in the second lactation can be predicted using an artificial neural network model with high accuracy utilizing several independent variables, 305-day milk production in the first lactation, test-day records from the first lactation, and estimated breeding values for 305-day milk production in the first lactation. The generation interval will decrease by reducing the time interval between collection of records and computation of estimated breeding

values for dairy animals, particularly sires. Various independent variables and test-day records can be used to increase the accuracy of prediction for 305-day milk production in the first lactation. Regarding the different research results obtained in the field of artificial network models to predict milk production, it becomes clear that the type of model, functions, number of hidden layers, number of neurons in each layer, as well as the independent input variables have a high impact on the predicted records. The importance of independent input variables will depend on cattle genetics and environmental factors, hence different sets of independent input variables will likely need to be included in artificial neural networks applied to cattle of distinct genetic characteristics under specific environmental conditions.

ACKNOWLEDGEMENT

The authors thank to Sari Agricultural Sciences and Natural Resources University, Sari, Iran for support.

REFERENCES

Abbasi A.R., Bahreini Behzadi M.R. and Talebi M.A. (2016). Prediction of some carcass characteristics from body measurements using linear regression and artificial neural network methods in Lori-Bakhtiari sheep. *J. Rumin. Res*. **3,** 1-20.

Adamowski J. and Chan H.F. (2011). A wavelet neural network conjunction model for groundwater level forecasting. J. Hydrol. **407(1),** 28-40.

Adesh K., Sharma R.K. and Kasana H.S. (2007). Prediction of first lactation 305-day milk yield in Karan Fries dairy cattle using ANN modeling. *Appl. Soft Comput*. **7,** 1112-1120.

Ali-Jafari Z. (2016). Modeling of milk production curve in Holstein dairy cows using artificial neural network and genetic algorithm. MS Thesis. Shahid Bahonar Univ., Kerman, Iran.

Ashton Q. (2013). Advances in Machine Learning Research and Application. Nova Publishers, Atlanta, Geogia.

Bahreini Behzadi M.R. (2015). Comparison of different growth models and artificial neural network to fit the growth curve of Lori-Bakhtiari sheep. *J. Rumin. Res*. **3,** 125-148.

Bahreini Behzadi M.R. and Aslaminejad A.A. (2010). A comparison of neural network and nonlinear regression predictions of sheep growth. *J. Anim. Vet. Adv*. **9,** 2128-2131.

Boichard D., Ducrocq V. and Fritz S. (2015). Sustainable dairy cattle selection in the genomic era. J. Anim. Breed. Genet. **132(2),** 135-143.

Boniecki P., Lipiński M., Koszela K. and Przyby J. (2013). Neural prediction of cows' milk yield according to environment temperature. *African J. Biotechnol*. **12,** 4707-4712.

Chaturvedi S., Yadav R.L., Gupta A.K. and Sharma A.K. (2013). Life time milk amount prediction in dairy cows using artificial neural networks. *Int. J. Res. Res. Rev*. **5,** 1-6.

Ghaderi Zefrehei M., Bahreini Behzadi M.R., Fayaz M.R. and Sharifi S. (2016). Association between calving interval and productive traits in dairy cattle over different inseminations using artificial neural network. *J. Rumin. Res*. **3(4),** 169-187.

Gorgulu O. (2012). Prediction of 305-day milk yield in Brown Swiss cattle using artificial neural networks. *South African J. Anim. Sci*. **42(3),** 280-287.

Grzesiak W., Lacroix R., Wójcik J. and Blaszczyk P. (2003). A comparison of neural network and multiple regression predictions for 305-day lactation yield using partial lactation records. *Canadian J. Anim. Sci*. **83,** 307-310.

Guresen E., Kayakutlu G. and Daim, T.U. (2011). Using artificial neural network models instock market index prediction. *Exp. Syst. Appl*. **38,** 10389-10397.

Hashemi A. and Nayebpour M. (2008). Estimation of genetic and phenotypic parameters for milk production in Holstein-Friesian cows. *Res. J. Biol. Sci*. **3(6),** 678-682.

Hosseinia P., Edris M., Edriss M.A. and Nilforooshan M.A. (2007). Prediction of second parity milk yield and fat percentage of dairy cows based on first parity information using neural network system. *J. Appl. Sci*. **7,** 3274-3279.

Ince D. and Sofu A. (2013). Estimation of lactation milk yield of Awassi sheep with artificial neural network modeling. *Small Rumin. Res*. **113,** 15-19.

Karadas K., Tariq M., Tariq M.M. and Eyduran E. (2017). Measuring predictive performance of data mining and artificial neural network algorithms for predicting lactation milk yield in indigenous Akkaraman Sheep. *Pakistan J. Zool*. **49(1),** 1-7.

Kasy M. and Kummer M. (2008). Market Entry in E-Commerce. The Networks, Electronic Commerce, and Telecommunications (NET) Institute. Available at: http://www.NETinst.org.

Khazaei J. and Nikosiar M. (2008). Approximating milk yield and milk fat and protein concentration of cows through the use of mathematical and artificial neural networks models. Pp. 123-129 in Proc. World Conf. Agric. Inf. IT/IAALD/AFITA WCCA 2008, Tokyo, Japan.

Miglior F., Fleming A., Malchiodi F., Brito L.F., Martin P. and Baes C.F. (2017). A 100-year review: Identification and genetic selection of economically important traits in dairy cattle. *J. Dairy Sci*. **100(12),** 10251-10271.

Njubi D.M., Wakhungu J.W. and Badamana M.S. (2009). Milk yield prediction in Kenyan Holstein-Friesian cattle using computer neural networks. Livest. Res. Rural Dev. Available at: https://www.lrrd.cipav.org.co/lrrd21/4/njub21046.htm.

Njubi D.M., Wakhungu J.W. and Badamana M.S. (2010). Use of test-day records to predict first lactation 305-day milk yield using artificial neural network in Kenyan Holstein–Friesian dairy cows. *Trop. Anim. Health Prod*. **42,** 639-644.

Nobari K., Bane H., Esmaeilkhanian S., Yousefi K. and Samiei R. (2019). Comparison of linear model and artificial neural network to prediction of milk yield using first recorded parity. *J. Rumin. Res*. **6(4),** 89-100.

Park S.J., Hwang C.S. and Vlek P.L.G. (2005). Comparison of adaptive techniques to predict crop yield responseunder varying soil and land management conditions. *Agric. Syst. J*. **85,** 59-81.

Pour Hamidi S., Mohammadabadi M.R., Asadi Foozi M. and Nezamabadi-Pour H. (2017). Prediction of breeding values for

the milk production trait in Iranian Holstein cows applying artificial neural networks. *J. Livest. Sci. Technol*. **5(2),** 53-61.

Qotbi A.A.A., Hossein Nia P., Seidavi A. and Ghovvati S. (2010). Predictions of semen production in ram using phenotypic traits by artificial neural network. *African J. Biotechnol*. **9(30),** 4822-4825.

Safari R. (2016). The use of artificial neural networks in predicting the production of Holstein dairy cows. MS Thesis. University of Tabriz, Tabriz, Iran.

Shahinfar S., Mehrabani-Yeganeh H., Lucas C., Kalhor A., Kazemian M., Kent A. and Weigel K.A. (2012). Prediction of breeding values for dairy cattle using artificial neural networks and neuro-fuzzy systems. *Comput. Math. Methods Med*. **2012,** 1-9.

Visentin G., McDermott A., McParland S., Berry D.P., Kenny O.A., Brodkorb A., Fenelon M.A. and De Marchi M. (2015). Prediction of bovine milk technological traits from mid-infrared spectroscopy analysis in dairy cows. *J. Dairy Sci*. **98(9),** 6620-6629.