RESEARCH ARTICLE

Prediction of Frozen Semen Doses Production in Dairy Studs using Machine Learning Algorithm

Chandrashekharaiah Jeevan[1]*, Sarojini M.K. Karthickeyan[1], Alagappan Gopinathan[1], Gopalan K. Tirumurugaan[2], Subbiah Vairamuthu[3]

Abstract

The ability to predict the frozen semen doses produced per ejaculate would be of considerable benefit for the management of skill, human resource, capital and time. The new computing paradigm called machine learning involves in predicting dependent variable by learning complex and non-linear relationship among independent variables. The purpose of this study is to develop prediction model using one of the conventional and machine learning modelling techniques called Multiple Linear Regression (MLR) and Artificial Neural Network (ANN), respectively. A Total of 1,57,532 ejaculates data were used for modelling. The modelling involved prediction of frozen semen doses produced per ejaculate using independent variables namely volume of ejaculate, ejaculate number, sperm concentration, initial motility and post thaw motility. Various combinations of architectural parameters were employed to explore optimum configuration for each model. The ANN (R^2=90.66) modelling was observed to be efficient over MLR (R^2=73.52). The root mean squared error (RMSE) value was found to be lower in ANN (33.89) when compared to MLR (57.31). Hence, the ANN modelling approach is efficient to predict frozen semen doses that could be produced per ejaculate.

Keywords: Artificial neural network, Dairy studs, Frozen semen production, Machine learning, Multiple linear regression, Seminal traits.

Ind J Vet Sci and Biotech (2022): 10.21887/ijvsbt.18.3.15

Introduction

Machine learning techniques namely artificial neural networks (ANN), support vector machines and random forest are being used and applied in various field of modern world for predicting the future events. These modern statistical approaches eclipse conventional approach like multiple linear regressions for the fact that former approach considers non-linear complexity of independent variables. At present, ANN or connectionist model, a paradigm branch is gaining momentum in solving real life problems (Yang *et al*., 1999 : Fang *et al*., 2000). This model is inspired by biological neurons, comprised of network of nodes like neurons connected by dendrites. The input nodes receive data of independent variables and signals to several layers of the network. The strength of connection between the nodes are altered by learning process to give an output data or dependent variable (Sharma *et al.,* 2007). This approach surpasses limitations involved in predicting the complex relationship among and between independent variables by learning itself and applying the same on unknown data for predicting dependent variable. The ability of ANN to learn data, parallel processing and generalization make it distinguishable from other methods. These characteristics facilitate the advantage of speed, efficient and error tolerant modelling (Mammadova and Keskin, 2015). ANN models are being used and tested in many fields of livestock and dairying, especially in predicting mastitis (Panchal *et al.,* 2017), conception success (Hempstalk *et al.,* 2015), milk yield (Sharma *et al.,* 2007; Murphy *et al.,* 2014) and breeding values (Shahinfar *et al.,* 2012).

One of the keys to propel India as a leading milk producer is crossbreeding programme. Indigenous cows are bred with the semen of progeny selected bulls through artificial insemination technique. For successful insemination, fertility of the bull is crucial as few bulls are used against large number of females. The demand for semen from best progeny tested

[1]Department of Animal Genetics and Breeding, Madras Veterinary College, Tamil Nadu Veterinary and Animal Sciences University, Chennai - 600 007, Tamil Nadu, India

[2]Zoonoses Research Laboratory, Madhavaram Milk Colony, Tamil Nadu Veterinary and Animal Sciences University, Chennai - 600 007, Tamil Nadu, India

[3]Centralized Clinical Laboratory, Madras Veterinary College, Tamil Nadu Veterinary and Animal Sciences University, Chennai - 600 007, Tamil Nadu, India

Corresponding Author: Chandrashekharaiah Jeevan, Department of Animal Genetics and Breeding, Madras Veterinary College, Tamil Nadu Veterinary and Animal Sciences University, Chennai - 600 007, Tamil Nadu, India, e-mail: drjeegene@gmail.com

How to cite this article: Jeevan, C., Karthickeyan, S.M.K., Gopinathan, A., Tirumurugaan, G.K., Vairamuthu, S. (2022). Prediction of Frozen Semen Doses Production in Dairy Studs using Machine Learning Algorithm. Ind J Vet Sci and Biotech. 18(3), 67-70.

Source of support: Nil

Conflict of interest: None.

Submitted: 20/12/2022 **Accepted:** 25/03/2022 **Published:** 10/07/2022

bull is growing exponentially and necessitates for production and dissemination of large quantity of cryopreserved semen doses. Moreover, many genetic and non-genetic factors contribute to the total number of doses produced per ejaculate. However, prediction of the total frozen semen doses produced per ejaculate aids in capital, skill, human resource and time management effectively. The studies pertaining to predict the total frozen semen doses produced by using basic initial parameters are scanty. Hence, this study involves in predicting the frozen semen doses (FSD) production per ejaculate by using initially obtained variables like ejaculate number, volume of ejaculate, initial motility and post-thaw motility by ANN modelling method compared with the conventional multiple linear regression (MLR) modelling.

Materials and Methods

Data Collection and Classification

Data for the study were collected from Exotic Cattle Breeding Farm, Eachankottai, Tamil Nadu, India. The data included ejaculate parameters namely, volume of ejaculate, ejaculate number, sperm concentration, initial motility, post-thaw motility (PTM) and number of frozen semen doses (FSD) produced per ejaculate. These data were collected from various genetic groups namely Murrah buffalo, Jersey, Umblachery and Jersey crossbred cattle breeds. The data were spread over 11 years (2008 - 2019) accounting to 1,57,532 ejaculates.

Statistical Analyses

The compiled data were divided into training and test data randomly in different ratios at every experimental run for multiple linear regression and artificial neural network. All statistical analyses were performed by using R statistical software, version 3.6.2 (R Core Team, 2019). The models were designed and tested to predict the number of FSD produced. A series of combinations of variables were used and tested for the better prediction accuracy using both models separately.

Multiple Linear Regression (MLR)

The prediction model used to predict the FSD produced per ejaculate by multiple linear regressions was as follow:

$$Y_{ij} = b_o + b_1X_1 + b_2X_2 + \ldots + b_iX_{ij} + e_{ij}$$

where,

Y_{ij}=Observation on no. of doses produced per ejaculate
b_o=Intercept
b_i= Partial regression coefficients for X_{ij} variables
X_{ij}= j^{th} observation under i^{th} variable
e_{ij}= Random error

Artificial Neural Network (ANN)

The ANN model operates based on the principle of neural network of brain (Haykin, 1998). It is a non-linear statistical tool where it adapts its network topology for modelling complex relationship between the variables. Feed forward resilient back propagation (rprop+) multilayer perceptron algorithm was used in this study (Riedmiller and Braun 1993). The neural network consists of one input layer, one output layer and varied number of hidden layers. These nodes contain logistic activation functions as search grid (Hagan *et al.,* 1997). The logistic active function with input layers were fed with independent variables, *viz.*, ejaculate number, volume, sperm concentration, initial motility and post-thaw motility; whilst single output node is the dependent variable that needs to be predicted, *i.e.,* FSD produced per ejaculate.

The experiments for ANN were run for the given set of data by sub-setting into training and test data. The ratio of train and test data were altered in different ratios to fit model for best prediction of the dependent variable. Other parameters namely, number of hidden layers, number of neurons in each hidden layer, activation function, learning rate, stepmax and threshold were altered until the model was best fit for prediction.

Model Evaluation Parameters

The prediction performance of MLR and ANN were evaluated in terms of root mean squared error (RMSE), coefficient of determination (R^2) and relative percent error (RPE). The greater R^2 values for a model suggest better prediction capabilities of model. Lower RMSE and RPE shows the lesser deviation of the predicted values and better prediction accuracy. The formulae used were as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_l)^2}$$

$$R^2 = \frac{\Sigma_i(y_l - \bar{y})^2}{\Sigma_i(y_i - \bar{y})^2}$$

$$RPE\ \% = \sum_{i=1}^{n}\frac{\hat{y}_l - y_i}{y_i - \bar{y}_l} X\ 100$$

Results and Discussion

The overall means for volume of ejaculate, concentration of spermatozoa in various genetic groups are presented in Table 1. The overall mean for the volume of ejaculate was found to be 3.59 mL, while the overall mean concentration of spermatozoa was found to be 1181.2 million per mL. It was observed that the mean semen volume was more in Jersey (4 mL) than in Murrah buffalo (2.66 mL). Concentration of spermatozoa was higher in Murrah buffalo (1267.29 million per mL) and lower in Umblachery cattle (1010.10 million per mL). When traits were compared over ejaculate number; volume, sperm concentration and FSD produced per ejaculate were observed to be greater in first ejaculate than in the second ejaculate.

Multiple Linear Regression

The experimental runs to predict genetic groupwise production of FSD projected poor model evaluation

Table 1: Breed wise means (± SE) of semen production traits

Genetic group	*Volume (ml)*	*Sperm concentration (10^6/ml)*	*Initial motility (%)*	*Post-thaw motility (%)*	*FSD produced (per ejaculate)*
Overall mean	3.59 ± 0.04	1181.89 ± 1.42	72.47 ± 1.42	51.38 ± 0.01	172.27 ± 0.29
Murrah	2.66 ± 1.19^a	1267.52 ± 556.69^c	72.67 ± 6.5	51.72 ± 6.21	136.85 ± 89.15^b
CBJY	3.91 ± 1.55^c	1155.16 ± 552.29^b	72.35 ± 6.5	50.82 ± 7.30	181.64 ± 118.04^c
Jersey	4.00 ± 1.44^c	1148.07 ± 470.61^b	72.51 ± 4.0	51.88 ± 5.82	192.85 ± 108.72^c
Umblachery	3.09 ± 1.16^b	1010.10 ± 373.98^a	71.14 ± 2.7	51.44 ± 7.70	128.06 ± 75.18^a

Means with at least one common superscript within classes do not differ significantly

Table 2: Comparison of model evaluation parameter

	R^2 (%)	*RMSE (%)*	*RPE (%)*
Multiple linear regression	73.52	57.31	37.42
Artificial neural network	90.66	33.89	24.71

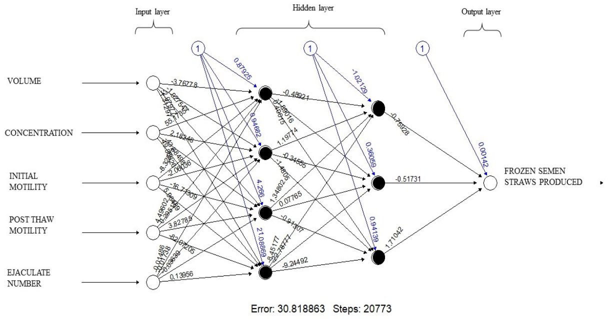

Fig. 1: Schematic representation of artificial neural network model with input and output variables considered in the study

parameters due to paucity in adequate number of data points requiring to train and test the prediction model. However, all genetic groups put-together predicted the target variable efficiently. The composite data set was divided into test data and train data randomly in different ratios for every run. It was observed that the model R^2 was the best when test and train data were in ratio of 70:30 and this ratio was maintained in further analyses. The MLR attained prediction accuracy with R^2, RMSE and RPE values as 73.52, 57.31 and 45.02 per cent, respectively are presented in Table 2.

Artificial Neural Network

The artificial neural networks were employed for the prediction of FSD produced per ejaculate as output variable. The training experiments with different combinations of the internal parameters were conducted to arrive at the best predictive model. The ANN model involved two hidden layers with 4 and 3 nodes in each layer. The learning rate of 0.001, threshold of 0.1 and stepmax value of 1,00,000 with 5 repetitions yielded the best model. The ratio of 70:30 of randomly selected train and test data showed the model with R^2 value of 90.66%. Fig. 1 shows the schematic diagram of ANN model plot with error in the end of the run reaching maximum stepmax. The ANN model exhibited model parameter values of 33.89 and 24.71 per cent for RMSE and RPE respectively (Table 2).

Neural Network Vs Multiple Linear Regression

The R^2 value was observed to be 90.66 in ANN, while 73.52 in MLR indicating that the ANN has better prediction capability. The RMSE being the absolute measure of fit, value for MLR was 57.31 which is greater than ANN (33.89), hence the ANN model is a better fit. The relative percent error was lower in ANN (24.71) thus proving that ANN architecture is superior in prediction. Similar study was conducted by Deb *et al*. (2105) to compare the effectiveness of MLR and ANN for prediction of post-thaw motility based on number of ejaculates, volume and concentration of sperm. Deb *et al*. (2105) reported the values of R^2 and RMSE for MLR 32.04 and 8.61 respectively, while 34.87 and 8.43 respectively, for ANN reiterating the ability of ANN in prediction of variables precisely. Upon comparing the model evaluation parameters in the present study, ANN takes upper hand than MLR in predicting the number of frozen semen doses produced per ejaculate. The outcome of study deduces the pragmatic potential of ANN to predict the frozen semen doses produced over the conventional statistical modelling methods.

Aknwoledgement

The authors wish to thank Department of Animal Husbandry and Veterinary Sciences, Govt. of Tamil Nadu for providing the data pertaining to semen production in the state for this study.

References

Deb, Rajib,. Singh, U., Raja, T.V., Kumar, S., Tyagi, S., Alyethodi, R.R., Alex, R., Sengar, G., & Sharma, S. (2015). Designing of an artificial neural network model to evaluatethe association of three combined Y-specific microsatelliteloci on the actual and predicted postthaw motility incrossbred bull semen. *Theriogenology, 83*(9), 1145-1450.

Fang, Q., Hanna, M.A., Haque, E., & Spillman, C.K. (2000). Neural network modeling of energy requirements for size reduction of wheat. *Transactions of the ASAE*, *43*(4), 947.

Hagan, M.T., Demuth, H.B., & Beale, M. (1997). *Neural Network Design,* 2nd Edn. PWS Publishing Co., Okhlama state university, Still water, United States of America.

Haykin, S. (1998). *Neural Networks: A Comprehensive Foundation*. 2nd edn. Prentice Hall PTR, USA.

Hempstalk, K., McParland, S., & Berry, D.P. (2015). Machine learning algorithms for the prediction of conception success to a given insemination in lactating dairy cows. *Journal of Dairy Science*, *98*(8), 5262-5273.

Mammadova, N.M., & Keskin, I. (2015). Application of neural network and adaptive neuro-fuzzy inference system to predict subclinical mastitis in dairy cattle. *Indian Journal of Animal Research*, *49*(5), 671-679.

Murphy, M.D., O'Mahony, M.J., Shalloo, L., French, P., & Upton, J. (2014). Comparison of modelling techniques for milk-production forecasting. *Journal of Dairy Science*, *97*(6), 3352-3363.

Panchal, I., Sawhney, I.K., Sharma, A.K., Garg, M.K., & Dang, A.K. (2017). Mastitis detection in Murrah buffaloes with intelligent models based upon electro-chemical and quality parameters of milk. *Indian Journal of Animal Research*, *51*(5), 922-926.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Accessed at http://www.R-project.org/.

Riedmiller, M., & Braun, H. (1993). A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In: IEEE *International conference on neural networks*, Baden-Württemberg, Germany, 586-591.

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

Sharma, A.K., Jain, D.K., Chakravarty, A.K., Malhotra, R., & Ruhil, A.P. (2013). Predicting economic traits in Murrah buffaloes with connectionist models. *Journal of the Indian Society of Agricultural Statistics*, *67*(1), 1-11.

Sharma, A.K., Sharma, R.K., & Kasana, H.S. (2007). Prediction of first lactation 305-day milk yield in Karan Fries dairy cattle using ANN modelling. *Applied Soft Computing*, ***7***(3), 1112-1120.

Yang, X.Z., Lacroix, R., & Wade, K.M. (1999). Neural detection of mastitis from dairy herd improvement records. *Transactions of the ASAE*, *42*(4), 1063.

ANNOUNCEMENT: SVSBT-NS-2022

IX Annual Convention and National Seminar of SVSBT

The ***IX Annual Convention*** and ***National Seminar*** of The Society for Veterinary Science & Biotechnology ***(SVSBT) on "Recent Biotechnological Advances in Health and Management to Augment Productivity of Livestock and Poultry"*** will be **organized at Ramayanpatti, Tirunelveli - 627 358, Tamil Nadu, during September 22-24, 2022** (Thursday, Friday & Saturday) by Veterinary College & Research Institute, Tirunelveli - 627 358, TANUVAS, (TN). The detailed Brochure cum Invitation showing Theme Areas/ Sessions, Registration Fee, Bank Details for online payment and deadlines, etc. has been floated on the Whats Apps and e-mails. Accordingly, the organizing committee of ***SVSBT NS-2022 invites abstracts*** of original and quality research work on theme areas of seminar limited to 250 words by e-mail on svsbttnns2022@gmail.com or mopandian69@gmail.com latest by 30th August, 2022 for inclusion in the Souvenir cum Compendium to be published on the occasion.

For Further details, please contact:

Dr. M. Chennapandian

Organizing Secretary cum Professor and Head

Department of Animal Nutrition, Veterinary College & Research Institute, TANUVAS, Ramayanpatti, Tirunelveli - 627 358 (Tamil Nadu), India

E-mail: svsbttnns2022@gmail.com; mopandian69@gmail.com; annvcritni@tanuvas.org.in mobile +91 94423 29003, 88256 79231