Journal of Intelligent Learning Systems and Applications, 2022, 14, 1-13
https://www.scirp.org/journal/jilsa
ISSN Online: 2150-8410
ISSN Print: 2150-8402

Toward an Intelligent System for Taurine Cattle Recognition

Fulbert Bembamba[1], Frédéric T. Ouédraogo[1], Soudré Albert[2], Amadou Traoré[3]

[1]Université Norbert Zongo, Laboratoire Mathématiques, Informatique et Applications (LAMIA), Koudougou, Burkina Faso
[2]Université Norbert Zongo, Koudougou, Burkina Faso
[3]Institut de l'Environnement et de Recherches Agricoles (INERA), Ouagadougou, Burkina Faso
Email: bembaplus@gmail.com, frederic.ouedraogo@unz.bf, asoudre@yahoo.fr, traore_pa@yahoo.fr

How to cite this paper: Bembamba, F., Ouédraogo, F.T., Albert, S. and Traoré, A. (2022) Toward an Intelligent System for Taurine Cattle Recognition. *Journal of Intelligent Learning Systems and Applications*, **14**, 1-13.
https://doi.org/10.4236/jilsa.2022.141001

Received: February 1, 2022
Accepted: February 21, 2022
Published: February 28, 2022

Abstract

Unlike zebus, taurine cattle have the natural ability to resist trypanosomosis, a parasitic disease endemic to the humid areas of West Africa. However, repeated crossbreeding between zebus and taurine cattle is jeopardizing the genetic heritage of the Taurines and their ability to resist trypanosomosis. To strengthen protection and conservation efforts, it is essential to accurately distinguish purebred taurines from crossbreds. In this study, five Machine Learning models were built using morphological data collected from 1968 cattle. These models were trained to determine whether a given individual is purebred taurine or not. The classifiers yielded promising results. The random forest model and RBF Kernel SVM performed the best with up to 86% and 85% accuracy respectively. Moreover, the study of the correlation coefficients and the feature importance scores allowed us to define the most discriminating morphological traits.

Keywords

Machine Learning, Trypanosomosis, Purebred Taurine, Accuracy, Model

1. Introduction

In a world literally drowned in data, Artificial Intelligence (AI) is becoming an increasingly important part of our lives. This new science at the junction of algebra, statistics, probability and computer science has diversified to meet the needs. Among the different branches of AI is Machine Learning (ML), which is used when it is difficult or impossible to define explicit instructions to give to a computer to solve a problem, but we have many illustrative examples at hand. We can oppose a classical program which uses a procedure and the data it receives (input) to produce answers (output), to a Machine Learning program, which

uses the data and the answers in order to produce the procedure which makes it possible to obtain the latter from the first [1].

AI, in general, and machine learning in particular, are progressively becoming strategic research axis for decision support solutions in several fields such as finance, marketing, security, etc. AI has also popped into agriculture and livestock, especially by contributing to improving the health and production of animals [2] [3], but also in the field of genetic improvement and conservation [4]. This is the case of the West African taurine cattle, also known as *Lobi* or *Baoulé*. Taurine cattle are tolerant to trypanosomosis disease though smaller in size and with lower productivity compared to most zebu-type cattle [5]. Trypanosomosis is the main parasitic disease of ruminants in wetlands, causing enormous economic losses to producers. However, for the Sahel region, these wetlands are the most suitable places for livestock production because of the abundance of fodder and pasture. The effects of climate change are accelerating the phenomenon of zebu migration to these areas that were once known as taurine sanctuaries. Uncontrolled and indiscriminate crossbreeding among local cattle types is thus taking place, leading to the dilution of trypano-tolerance ability and threats to the genetic integrity of West African taurine cattle types [5]. Therefore, empirical methods of distinguishing the two species, formerly based on visual differences in morphological traits (size, presence of hump, etc.) no longer work. An efficient yet very costly method is the laboratory analysis of blood samples. Our study aims at proposing a low-cost method inspired by machine learning techniques to easily make this distinction. In the long run, it is planned to integrate the results achieved here with image processing applications to identify purebred taurines using their images.

This paper is structured as follows. In Section 2, we present the context of the problem we have to address. In Section 3, we give an overview of related work. Section 4 will provide definitions and background. In Section 5 and Section 6, we will respectively unveil some results and conduct discussions. Finally, Section 7 concludes the paper.

2. Context

There are two subspecies of cattle: zebus and taurines. The taurine cattle live in the wetlands. This fodder-rich region is unfortunately infested with tsetse flies, a vector for the spread of an endemic parasitic disease called trypanosomosis, that causes enormous losses to livestock:

- direct economic losses due to morbidity;
- stunted growth of young animals;
- weight loss;
- low milk production;
- infertility;
- abortion of cows;
- etc.

The taurines are special in that they have a natural resistance to trypanosomo-

sis. Unfortunately, this genetic faculty is undermined by repeated cross-breeding over several generations with zebus due to the seasonal transhumance of the latter towards the wetlands and deliberate actions of breeders who seek larger animals through these crossings.

In order to preserve this type of cattle, it is necessary to find out whether a given individual is pure taurine or not. The empirical segregation methods are less and less accurate because of the massive crossings. The only formal method is a genetic analysis which is too costly in time and resources. Therefore, artificial intelligence is used for this characterization.

A conservation project working in the Sahel that focuses on the preservation of bulls has made several scientific productions on the topic, though in the field of natural and social sciences. For our study, we have in hand the data collected by this research project. Phenotypic data were measured on several thousand cattle in accordance with the 2012 FAO guidelines [6] for the phenotypic characterization of animal genetic resources (**Table 1**). Blood samples were also taken for laboratory analysis. These analyses allowed, among other things, to determine formally if an individual is a purebred taurine (with full trypano-tolerance capacity), pure zebu (no trypano-tolerance capacity) or a crossbred (some percentage of trypano-tolerance capacity). In the present work, we use the first dataset of 1968 individuals (taurines, zebus, crossbreds) in which six traits have been assessed: height at withers, chest girth, body length, weight, sex and age.

3. Related Work

Animal species identification is an important issue for the modernization of livestock. The scientific literature reveals different techniques that replace direct observation methods. These techniques are mainly based on body measurements, images or biological markers. One important issue is how to obtain the body features. Traditional direct measurement of animals consumes time and effort. For instance, the use of scales for live weight measurement requires a vehicle, some qualified personnel and special facilities. To overcome this difficulty, [7] and [6] used barymetric equations to estimate the weight applicable to Niger Azawak and Burkina Faso taurine cattle. This technique main drawback is that it provides low accuracy with adult animals because of the possible fattening or the

Table 1. List of quantitative traits.

Head measurements	Body measurements
cranial length, head width, head length, cranial width, facial length, facial width, muzzle circumference, distance between horn tips, distance between horn bases, horn length, earn length	height at withers, thoracic perimeter, height at sacrum, body length, length of scapula ischium, hip width, ischium width, tail length, chest depth, shoulder width, chest width, teat length, weight

physiological state of females.

Rudendko [8] derived cows' weight using artificial neural network algorithms. This is achieved in two steps: firstly, a convolution neural network(CNN) is used to detect cows in the picture, and the stereopsis method allows the system to obtain their size measurements such as wither height, hipheight, body length and hip width via photogrammetry; secondly, these measurements are used to determine the cow live weight.

References [9] and [10] trained CNN classifiers to classify images of dogs to the appropriate class out of 120 breeds of dogs. The problem is tackled as an image classification problem using a deep convolutional neural network. In [10], the image is divided into numerous lattices and the extracted descriptors serve as input for the CNNs that are trained to identify dog species.

Reference [11] implemented an effective breed identification system using genetic markers single nucleotide polymorphisms (SNPs) genotyped from pigmeat products. Six machine learning methods were trained to make this identification task. SVM yielded the most accurate performance.

The identification methods outlined above are based on costly techniques in computational resources as well as material resources. Our approach, which also offers good accuracy, is based on Machine Learning, using phenotypic data collected from hundreds of cows to predict their sub-species. This method can be integrated into a lite and affordable intelligent system for breed recognition in the Sahel social and economic context.

4. Methods

4.1. Conceptual Framework

The problem that we have to deal with is to decide whether a designated bovine individual is pure taurine or not. For this purpose, we dispose of its morphological measurements. To train our model, we also have at our disposal the measurements of thousands of other individuals (examples) with their label: the "pure" character. The problem is, therefore, a supervised learning mater. According to [12], supervised learning is the machine learning task of learning a function that maps an input to an output based on input-output pairs.

For the sake of simplicity, we will restrict our attention in this phase, to determining whether the individual is pure or not, regardless of the inter-breeding rate. So, the space of labels is binary: {pure, notpure}. We are thus reduced to a binary classification problem.

4.2. Selection of Algorithms

Machine Learning relies on different algorithms to solve data problems. Choosing an appropriate classification algorithm for a particular problem task requires practice and experience [13]. At this stage of our study, we have chosen a limited number of the most commonly used algorithms, making sure that they are as representative as possible of the different types of algorithms: linear, non-linear,

instance-based, bayesian, and ensemble methods.

After a model is trained, we evaluate its performance on the test set to guarantee that future measurements in similar situations are sufficiently accurate. To compare the two models, we can compare their accuracy, precision, or recall values. Reference [14] recommends that AUC (Area Under the Curves) be used in preference to overall accuracy for single number evaluation of machine learning algorithms.

4.3. Overview of a Few ML Algorithms

In the following lines, we will briefly describe the general principles of the machine learning algorithms that we will implement in this study. We will consider the following notations:

- n: the number of examples;
- m: the number of features of an example;
- $x^{(i)}$: $i \in [1, \cdots, n]$: the *ith* example;
- $x_j^{(i)}$; $j \in [1, \cdots, m]$ or simply x_j if there is no ambiguity: the *jth* feature of the *ith* example
- $y^{(i)}$ and $\hat{y}^{(i)}$ respectively the true class label and the predicted class label of the *ith* training example
- β_j: the *jth* model weight.

4.3.1. Random Forest (RF)

Decision Trees are considered to be one of the most popular approaches for representing classifiers [15]. However, they are known to have high variance and so, tend to overfit. Random forest is an ensemble method that allows to combining several trees in order to avoid overfitting. Ensemble methods apply the "wisdom of crowd" concept. This concept is based on the idea that combining many weak learners results in a performance that is far beyond the individual performance of those learners, because their errors compensate for each other. [16] proposes to build the individual trees of the forest using different variables. So at each node a number p of variables smaller than the total number is selected before applying the splitting criteria.

4.3.2. Logistic Regression (LR)

It's a fast model to learn and effective on binary classification problems. It's one of the most widely used algorithms for classification in the industry [17]. The basic principle is like linear regression, where the hypothesis space consists of a linear combination of the variables:

$$h_\theta(x) = \sum_{j=0}^{m} \beta_j x_j \tag{1}$$

This linear hypothesis can yield very high values as well as very low values (below zero). Logistic regression transforms this output using the sigmoid function to return a probability value: between 0 and 1. Concretely, we apply the sigmoid function:

$$\phi(z) = \frac{1}{1+e^{-z}} \tag{2}$$

Therefore, we have:

$$h_\theta(x) = \phi(z) \tag{3}$$

with

$$z = \sum_{j=0}^{m} \beta_j x_j \tag{4}$$

The model is trained by minimizing the cost function $J(\theta)$, using the descending gradient technique:

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left[-y^{(i)}\log\left(\phi\left(z^{(i)}\right)\right) - \left(1-y^{(i)}\right)\log\left(1-\phi\left(z^{(i)}\right)\right)\right] \tag{5}$$

4.3.3. Naïve Bayes (NB)

The model is comprised of two types of probabilities that can be calculated directly from your training data:

- The prior probability of each class.
- The conditional probability for each class given each x value.

$$P\left(y|x_1,\cdots,x_n\right) = \frac{P(y)P\left(x_1,\cdots,x_n|y\right)}{P\left(x_1,\cdots,x_n\right)} \tag{6}$$

where $P\left(y|x_1,\cdots,x_n\right)$ is the posterior probability,
$P(y)$ is the class prior probability,
$P\left(x_1,\cdots,x_n|y\right)$ is the likelihood,
$P\left(x_1,\cdots,x_n\right)$ is the predictor prior probability.

The predictor prior probability term is constant with regard to the class values. Therefore, we can write:

$$P\left(y|x^{(i)}\right) \propto P(y)P\left(x^{(i)}|y\right) \tag{7}$$

The different Bayes classifiers differ mainly by the assumptions they make regarding the distribution of $P\left(x^{(i)}|y\right)$ [18].

With the naive conditional independence assumption for example, this expression becomes:

$$P\left(x^{(i)}|y\right) = \prod_{j=1}^{m} P\left(x_j^{(i)}|y\right) \tag{8}$$

4.3.4. K-Nearest Neighbors (KNN)

Nearest neighbors algorithm is one of the simplest predictive models there is [13]. Predictions are made for a new data point by searching through the entire training set for the K most similar instances (the neighbors) and summarizing the output variable for those instances. We select in advance the number (K) of neighbors to consider and the notion of distance to apply. KNN is called "lazy" not because of its apparent simplicity, but because it doesn't learn a discriminative function from the training data but memorizes the training dataset instead

[17]. KNN belongs to a subcategory of non-parametric models that are described as instance-based learning.

4.3.5. Support Vector Machine Kernel (SVMk)

SVM might be one of the most powerful and widely used classifiers and can be considered an extension of the perceptron [17]. In SVM, our optimization objective is to maximize the margin that we define as the distance between the separating hyperplane (decision boundary) and the support vectors. The support vectors are the training examples that are closest to this hyperplane. The optimal hyperplane can be set as:

$$\omega x^{\mathrm{T}} + b = 0 \tag{9}$$

where w is the weight vector, x the input vector and b, the bias. For all elements of the training set, w and b should verify [19]:

$$\begin{aligned} \omega\left(x^{(i)}\right)^{\mathrm{T}} + b \geq +1 \text{ if } y^{(i)} = 1 \\ \omega\left(x^{(i)}\right)^{\mathrm{T}} + b \leq -1 \text{ if } y^{(i)} = -1 \end{aligned} \tag{10}$$

Support vectors are those $x^{(i)}$ for which

$$\left|y^{(i)}\right|\left(\omega\left(x^{(i)}\right)^{\mathrm{T}} + b\right) = 1 \tag{11}$$

The training objective is to find the right parameters (ω and b) so that the hyperplane separates the data and maximizes the margin $\frac{1}{\|\omega\|^2}$. Which is equivalent to minimizing $\|\omega\|^2$.

SVM is also popular because it can be kernelized to solve nonlinear classification problems. In practice, we use a mapping function to transform the training data into a higher dimensional feature space. We now train a linear SVM to classify the data in this new feature space. The "kernel trick" allows saving expensive cost of calculations. We define a kernel function as:

$$k\left(x^{(i)}, x^{(j)}\right) = \phi\left(x^{(i)}\right)^{\mathrm{T}} \phi\left(x^{(j)}\right) \tag{12}$$

One popular kernel function is called Radial Basis Function (RBF) which can be written as:

$$k\left(x^{(i)}, x^{(j)}\right) = \exp\left(-\gamma\left\|x^{(i)} - x^{(j)}\right\|^2\right) \tag{13}$$

Here, γ is a free parameter to be optimized k can be interpreted as a similarity score, ranging from 0 (very dissimilar examples) to 1 (exactlysimilar examples).

4.4. Data Preparation

The quality of the data and the amount of useful information that it contains are key factors that determine how well a machine learning algorithm can learn [17]. Data preparation is the process of transforming raw data so that they can be run

through machine learning algorithms. This involves handling categorical data and missing values, rescaling data, etc. Supervised machine learning techniques require splitting data into multiple parts for training and testing steps. However, if we are dividing a dataset, we have to keep in mind that we are withholding valuable information that the learning algorithm could benefit from. At the same time, the smaller the test set, the more inaccurate the estimation of the generalization error. Therefore, dividing a dataset into training and test sets is all about balancing this tradeoff [20]. Within the framework of this work, we used the "hold out" method. Basically, we split the dataset into two chunks: the calibration sample and the test sample. The default proportions of 70% - 30% were used.

5. Results

In the data preparation process, we cleaned the data by discarding entries that contained missing values or outliers. These inconsistent data represented 8.6% of the entire dataset. Finally, 1797 observations were validated for the study.

Data mining allowed us to visualize the shape of the feature distributions. We used pair plots to assess the correlation between the features. Graphs plotting features one against the other on the one hand and one against the label, on the other hand, showed that the interdependence is not negligible. In particular, a strong correlation was noted between weight and chest girth with a correlation coefficient of 0.948463 (**Figure 1**).

Correlations between the different descriptors and the "pure" trait were also analyzed. Height at withers has the highest coefficient with the label: −0.472985. This is corroborated by the feature importance graphic (**Figure 3**).

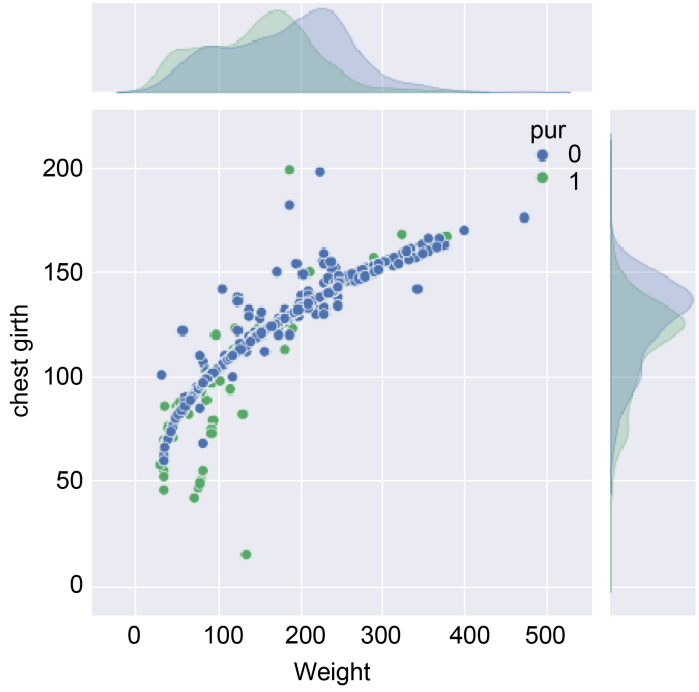

Figure 1. Correlation weight-girth width.

We trained the five algorithms presented earlier on the calibration set. The resulting predictive models were tested on the test sample to measure the generalization ability. Hyper parameters were adjusted to yield the best performances. The results obtained are shown in Table 2 and Figure 2.

Table 2. Performances measures.

	Accuracy	Precision	Recall
RF	86%	0.87	0.88
NB	64%	0.62	0.58
LR	81%	0.79	0.80
SVMk	85%	0.83	0.85
KNN	83%	0.83	0.78

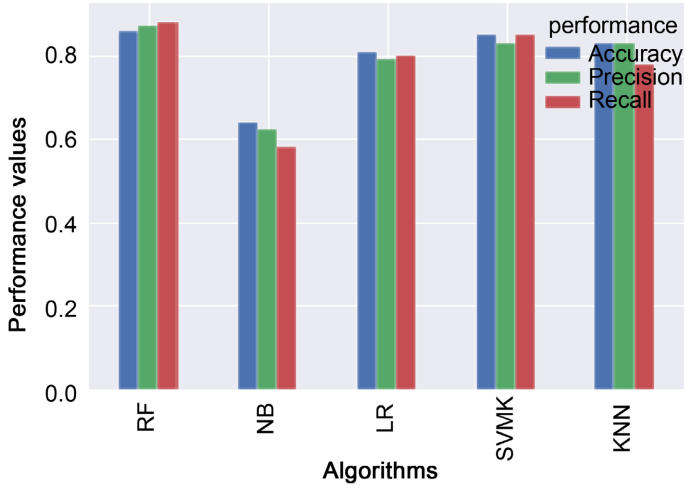

Figure 2. Performance metrics.

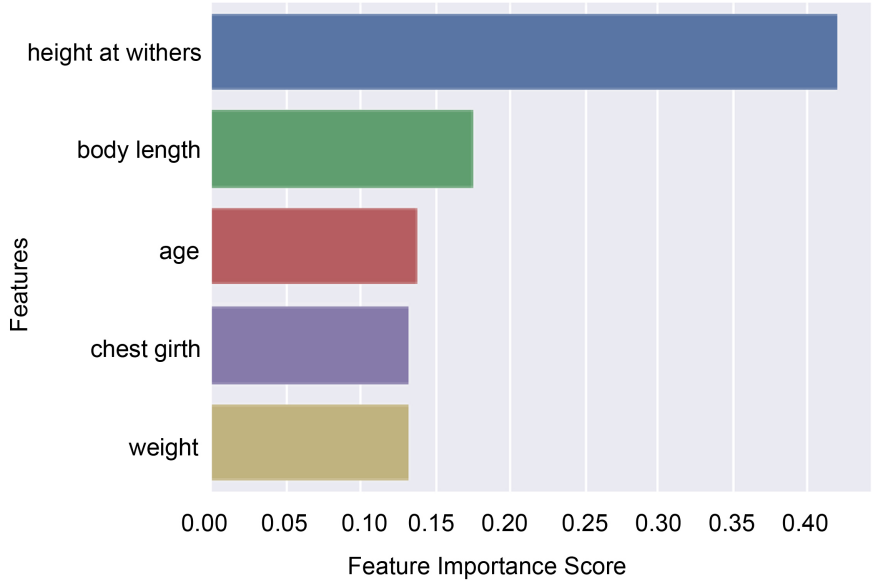

Figure 3. Feature importance.

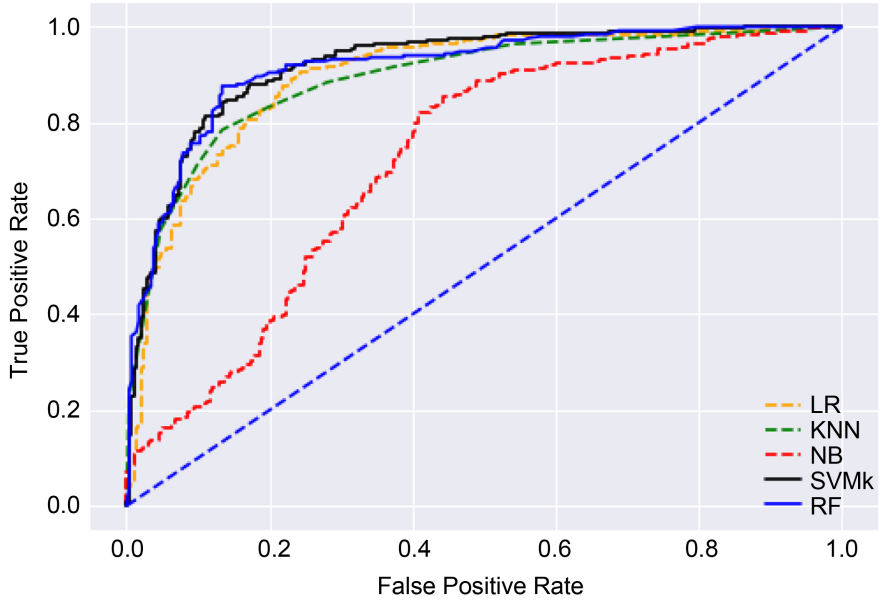

Figure 4. ROC curves.

ROC curves (Receiver Operating Characteristic) were also drawn (see **Figure 4**). As per [14] recommendations, AUC (Area Under the Curve)values have been calculated in order to compare the performances of the different methods.

From the AUC perspective, it appears that nonlinear Kernel SVM is the most efficient algorithm (AUC = 0.9202), followed by Random Forest (AUC = 0.9161). K-Nearest Neighbors (K = 10) and Logistic Regression have very similar performances. Their ROC curves overlap at some cut-off points and their AUC are very closed: 0.8963324783059543 and 0.8919186268624134 respectively. Naive Bayes yields the worst result. The Random Forest model remains the most accurate as far as accuracy (86%), precision (87%) and recall (88%) are concerned.

6. Discussion

Naive Bayes yields the lowest performance: 64% of accuracy. This is due to the algorithm's strong independence assumption. It is clearly difficult to completely decouple age from chest girth or weight for example. Reference [1] demonstrated that the Naive Bayes model reaches its best performance in two opposite cases: completely independent features and functionally dependent features. In the present case, the level of feature dependence is in between.

Furthermore, the analysis of the coefficients of the regression model confirms that size (height at withers) is the most significant discriminant variable among the two species. The negative sign of the coefficients indicates inverse proportionality. This reinforces the general view that taurines are smaller than zebus. Feature importance plot (**Figure 3**) supports this since height at withers and body length score the most.

The strong correlation between weight and girth width can be explained by the measurement technique used in the field. Indeed, technicians used a weigh

band, a tool that deduces the weight from the chest width, that is measured directly on the subject [6].

The performance ranking showed that nonlinear models provide better results. Random Forest gives an accuracy of 86%, kernel SVM and KNN performed 85% and 83% respectively. These algorithms often lead to models with high variance. There is therefore a non-negligible risk of overfitting.

7. Conclusions

Trypanosomosis, which is prevalent in humid areas of West Africa, leads to a drop in livestock production and higher operating costs. The taurines, unlike the zebu species, have an innate ability to resist this disease. Unfortunately, uncontrolled crossbreeding between those two species of cattle leads to the dilution of this resistance capacity and threatens the genetic heritage of the taurines. Innovative means such as machine learning applications are needed to contribute to the preservation of the taurine species and its precious trypanotolerance capacity. To achieve this, it is crucial to distinguish purebred taurines from others. In this study, we applied five machine learning algorithms to train supervised models in order to make this identification. Random Forest performed the best with up to 86% accuracy, 88% recall and 0.9161 of AUC. The study confirmed that height at withers is the most discriminating descriptor among the six descriptors analyzed.

To obtain better results, it is important to continue the study by including the other morphological variables (**Table 1**). As this preliminary study confirmed, nonlinear methods seem to be more efficient. This trend could be further explored by the implementation of some cutting-edge models like Artificial Neural Network, XGBoost, etc. Moreover, the generalization capacity of the models trained here can be improved by associating other sampling methods such as bootstrapping.

Acknowledgements

We wish to express our sincere gratitude to all the partners who provided us with data for this study, in particular the Local Cattle Breed of Burkina Faso (LoCaBreed) team. We would like to thank Austrian Partnership Programme in Higher Education and Research for Development (APPEAR project 120). We are also grateful to Ministère de l'Enseignement Supérieurde la Recherche Scientifique et de l'Innovation (MESRSI) of Burkina Faso.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Rish, I., *et al.* (2001) An Empirical Study of the Naive Bayes Classifier. *IJCAI* 2001

Workshop on Empirical Methods in Artificial Intelligence, Seattle, 4-6 August 2001, 41-46.

[2] Wanga, H.P., Ghani, N. and Kalegele, K. (2015) Designing a Machine Learning-Based Framework for Enhancing Performance of Livestock Mobile Application System. *American Journal of Software Engineering and Applications*, **4**, 56. https://doi.org/10.11648/j.ajsea.20150403.13

[3] Olasehinde, O. (2021) Infrared Thermography and Machine Learning in Livestock Production. *International Journal of Advanced Research and Review*, **6**, 38-57.

[4] Libbrecht, M.W. and Noble, W.S. (2015) Machine Learning Applications in Genetics and Genomics. *Nature Reviews Genetics*, **16**, 321-332. https://doi.org/10.1038/nrg3920

[5] Ouedraogo, D., Ouedraogo-Kone, S., Yougbare, B., *et al*. (2021) Population Structure, Inbreeding and Admixture in Local Cattle Populations Managed by Community-Based Breeding Programs in Burkina Faso. *Journal of Animal Breeding and Genetics*, **138**, 379-388. https://doi.org/10.1111/jbg.12529

[6] Yougbare, B., Soudre, A., Ouedraogo, D., *et al*. (2021) Genome-Wide Association Study of Trypanosome Prevalence and Morphometric Traits in Purebred and Crossbred Baoulé Cattle of Burkina Faso. *PLOS ONE*, **16**, e0255089. https://doi.org/10.1371/journal.pone.0255089

[7] Dodo, K., Pandey, V.S. and Illiassou, M.S. (2001) Utilisation de labarymetrie pour l'estimation du poids chez le zebu Azawak au Niger. *Revue d'élevage et de médecine vétérinaire des pays tropicaux*, **54**, 63-68. https://doi.org/10.19182/remvt.9808

[8] Rudenko, O., Megel, Y., Bezsonov, O., *et al*. (2020) Cattle Breed Identification and Live Weight Evaluation on the Basis of Machine Learning and Computer Vision. *Proceedings of the Third International Workshop on Computer Modeling and Intelligent Systems* (*CMIS*-2020), Zaporizhzhia, 27 April-1 May, 2020, 939-954. https://doi.org/10.32782/cmis/2608-70

[9] Raduly, Z., Sulyok, C., *et al*. (2018) Dog Breed Identification Using Deep Learning. *IEEE* 16*th International Symposium on Intelligent Systems and Informatics* (*SISY*), Subotica, 13-15 September 2018, 271-276. https://doi.org/10.1109/SISY.2018.8524715

[10] Kumar, R., Sharma, M., Dhawale, K., *et al*. (2019) Identification of Dog Breeds Using Deep Learning. 2019 *IEEE* 9*th International Conference on Advanced Computing* (*IACC*), Tiruchirappalli, 13-14 December 2019, 193-198. https://doi.org/10.1109/IACC48062.2019.8971604

[11] Xu, Z.T., Diao, S.Q., Teng, J.Y., *et al*. (2021) Breed Identification of Meat Using Machine Learning and Breed Tag SNPs. *Food Control*, **125**, Article ID: 107971. https://doi.org/10.1016/j.foodcont.2021.107971

[12] Mahesh, B. (2020) Machine Learning Algorithms—A Review. *International Journal of Science and Research* (*IJSR*), **9**, 381-386.

[13] Grus, J. (2015) Data Science from Scratch: First Principles with Python. O'Reilly, Sebastopol.

[14] Bradley, A.P. (1997) The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognition*, **30**, 1145-1159. https://doi.org/10.1016/S0031-3203(96)00142-2

[15] Rokach, L. and Maimon, O. (2009) Classification Trees. In: *Data Mining and Knowledge Discovery Handbook*, Springer, Berlin, 149-174. https://doi.org/10.1007/978-0-387-09823-4_9

[16] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32.

https://doi.org/10.1023/A:1010933404324

[17] Raschka, S. and Mirjalili, V. (2019) Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-Learn and TensorFlow 2. 3rd Edition, Packt Publishing, Birmingham.

[18] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011) Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825-2830.

[19] Huang, K.X., Xiao, C., Glass, L.M., *et al*. (2021) Machine Learning Applications for Therapeutic Tasks with Genomics Data. *Patterns*, **2**, Article ID: 100328.

[20] Hawkins, D.M., Basak, S.C. and Mills, D. (2003) Assessing Model Fit by Cross-Validation. *The Journal for Chemical Information and Computer Scientists*, **43**, 579-586. https://doi.org/10.1021/ci025626i