UNIVERSIDADE DE SÃO PAULO

FACULDADE DE MEDICINA DE RIBEIRÃO PRETO

LUARA AFONSO DE FREITAS JANUÁRIO

Otimização da resistência a nematódeos gastrointestinais em ovinos Santa Inês: uma abordagem de seleção genômica, *machine learning* e análise de imagens

Ribeirão Preto

2023

LUARA AFONSO DE FREITAS JANUÁRIO

Otimização da resistência a nematódeos gastrointestinais em ovinos Santa Inês: uma abordagem de seleção genômica, *machine learning* e análise de imagens

Tese apresentada ao programa de Pós-graduação em Genética da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo, para obtenção do título de Doutor(a) em Ciências.
Area de Concentração: Genética
Orientadora: Profa. Dra. Claudia Cristina Paro de Paz
Coorientador: Dr. Rodrigo Pelicioni Savegnago

Ribeirão Preto

2023

“Versão corrigida. A versão original encontra-se disponível tanto na Biblioteca da Unidade que aloja o Programa, quanto na Biblioteca Digital de Teses e Dissertações da USP (BDTD)”

Januário, Luara Afonso de Freitas

Otimização da resistência a nematódeos gastrointestinais em ovinos Santa Inês: uma abordagem de seleção genômica, *machine learning* e análise de imagens. Ribeirão Preto, 2023.
117 p.: il.; 30 cm

Tese de Doutorado, apresentada à Faculdade de Medicina de Ribeirão Preto/USP. Área de concentração: Genética.
Orientador: Cristina Paro de Paz, Claudia.
Coorientador: Pelicioni Savegnago, Rodrigo.

1. Nematódeos gastrointestinais. 2. *Ovis aries*. 3. *Machine Learning*. 4. Análise de Imagem. 5. Seleção Genômica.

Universidade de São Paulo

ATA DE DEFESA

Aluno: 17135 - 7238085 - 1 / Página 1 de 1

Ata de defesa de Tese do(a) Senhor(a) Luara Afonso de Freitas Januário no Programa: Ciências Biológicas (Genética), do(a) Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo.

Aos 27 dias do mês de março de 2023, no(a) FMRP-USP realizou-se a Defesa da Tese do(a) Senhor(a) Luara Afonso de Freitas Januário, apresentada para a obtenção do título de Doutora intitulada:

"Otimização da resistência a nematódeos gastrointestinais em ovinos Santa Inês: uma abordagem de seleção genômica, *machine learning* e análise de imagens"

Após declarada aberta a sessão, o(a) Sr(a) Presidente passa a palavra ao candidato para exposição e a seguir aos examinadores para as devidas arguições que se desenvolvem nos termos regimentais. Em seguida, a Comissão Julgadora proclama o resultado:

Nome dos Participantes da Banca	Função	Sigla da CPG	Resultado
Claudia Cristina Paro de Paz	Presidente	SAABSP(FMRP)	Não Votante
Ricardo Vieira Ventura	Titular	FMVZ - USP	APROVADA
Lucia Regina Martelli	Suplente	FMRP - USP	APROVADA
Salvador Boccaletti Ramos	Suplente	UF - Externo	APROVADA

Resultado Final: APROVADA

Parecer da Comissão Julgadora *

Eu, Ísis Sawasaki Silveira Leme Ísis Sawasaki Silveira Leme, lavrei a presente ata, que assino juntamente com os(as) Senhores(as). Ribeirão Preto, aos 27 dias do mês de março de 2023.

Ricardo Vieira Ventura

Lucia Regina Martelli

Salvador Boccaletti Ramos

Claudia Cristina Paro de Paz
Presidente da Comissão Julgadora

* Obs: Se o candidato for reprovado por algum dos membros, o preenchimento do parecer é obrigatório.

A defesa foi homologada pela Comissão de Pós-Graduação em 04/04/2023 e, portanto, o(a) aluno(a) jus ao título de Doutora em Ciências obtido no Programa Ciências Biológicas (Genética) - Área de concentração: Genética.

Presidente da Comissão de Pós-Graduação

Profª Drª Cristina Marta Del Ben
Presidente da Comissão de Pós-Graduação
FMRP-USP

DADOS CURRICULARES DO AUTOR

LUARA AFONSO DE FREITAS JANUÁRIO – Filha de Carlos Alberto de Paula Freitas e Valéria de Souza Afonso Freitas, nasceu em Ribeirão Preto, SP, no dia 3 de março de 1993. Iniciou o curso de Zootecnia em março de 2011, na Faculdade de Ciências Agrárias e Veterinárias, UNESP, Campus de Jaboticabal. Durante o ano de 2013, foi bolsista de Iniciação Científica de Aperfeiçoamento de Estágio de Pesquisa de Nivel Superior e nos anos de 2014 e 2015 foi bolsista de Iniciação Científica da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), ambas na mesma instituição de ensino, sob orientação do Prof. Dr. Danísio Prado Munari. Em março de 2016, iniciou o Curso de Mestrado do Programa de Genética e Melhoramento Animal na Faculdade de Ciências Agrárias e Veterinárias, UNESP, campus de Jaboticabal, sob orientação do Prof. Dr. Danísio Prado Munari e coorientações da Dra. Claudia Cristina Paro de Paz e do Dr. Rodrigo Pelicioni Savegnago. Foi bolsista Capes e FAPESP. No período de agosto a outubro de 2017 realizou estágio de pesquisa de Mestrado no Instituto Nacional de Investigación y Tecnología Agraria e Alimentaria (INIA), Madrid, Espanha, sob a supervisão da Dra. María Jesús Carabaño Luengo e Dra. Magdalena Serrano Noreña. Em abril de 2018, iniciou o doutorado em genética na Universidade de São Paulo, Faculdade de Medicina de Ribeirão Preto, sob a orientação da Profa. Dra. Claudia Cristina Paro de Paz e coorientação do Dr. Rodrigo Pelicioni Savegnago. No período de agosto de 2021 a Julho de 2022, realizou estágio de pesquisa na University of Wisconsin-Madison, USA, sob a supervisão do Prof. Dr. Guilherme Jordão de Magalhães Rosa. Foi bolsista Capes e FAPESP.

Como é grande a tua bondade, que reservaste para aqueles que te temem,
e que, à vista dos homens, concedes àqueles que se refugiam em ti!
Bendito seja o Senhor, pois mostrou o seu maravilhoso amor para
comigo.
Salmos 31:19,21

"Pois tudo é Dele, por Ele e para Ele. A Ele seja dada a glória para
sempre! Amém." (Romanos 11:36)

Dedico

A Deus,

Ao meu amado marido Marcelo Januário,

Aos meus Pais Valéria e Carlos,

Razão de tudo.

AGRADECIMENTOS

A **Deus** pelo dom da vida. Tudo que conquistei foi para a Tua glória. Eu te agradeço por tudo. Te amo, Deus, tua graça nunca falha.

Ao meu amado marido **Marcelo Januário**, pela paciência, apoio, companheirismo, compreensão, amor e carinho. Por não ter soltado minha mão durante todo o processo e ter sido meu porto seguro. Você foi apoio essencial nessa jornada. Te amo muito.

Aos meus **pais (Valéria e Carlos)**, pelo apoio, incentivo, carinho e amor. Agradeço imensamente a oportunidade que vocês sempre me proporcionaram para que eu pudesse estudar e fazer o que eu escolhesse. Amo vocês.

A minha orientadora **Dra. Claudia Cristina Paro de Paz**, pelos ensinamentos, incentivo, compreensão e por toda a ajuda na elaboração deste trabalho. Sua contribuição para minha formação profissional e pessoal foi muito importante.

Ao meu coorientador e amigo, **Dr. Rodrigo Pelicioni Savegnago**, por todos os ensinamentos, paciência e principalmente por sempre acreditar no meu trabalho e me incentivar. Obrigada por tudo.

À **Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP)** pela concessão da bolsa de estudos de Doutorado (Processo n° 2018/01540-2), bolsa de Estágio de Pesquisa no Exterior – BEPE (Processo n° 2020/03575-8), e projeto temático (Processo n° 2016/14522-7).

Ao prof. **Dr. Guilherme Jordão de Magalhães Rosa** da Universidade de Wisconsin-Madison, EUA, pela incrível oportunidade de estágio no exterior. Agradeço imensamente todo o aprendizado, você me ensinou que questionar e entender o processo são uma das partes mais importantes dessa jornada de pesquisa. Meus sinceros agradecimentos.

As **minhas amigas (Stephani, Ryalva e Meire) e a minha família**, sou muito grata a Deus por ter vocês em minha vida, todo o apoio e palavras de conforto que vocês sempre me disseram me motivaram ainda mais a chegar aos meus objetivos. Muito

obrigada. Amo vocês.

A **equipe do projeto temático**, que colaboraram com as coletas e análises de dados, sou muito grata pela ajuda de cada um de vocês.

Aos **proprietários das fazendas**, por terem fornecido seus animais para esta pesquisa.

To my friends and colleagues in Madison I would like to thank you all for being part of my life and for providing me such a wonderful experience in Madison.

Aos **membros da Banca Examinadora de Doutorado** Dra. Lucia Regina Martelli, Dr. Ricardo Vieira Ventura e Dr. Salvador Boccaletti Ramos. Obrigada pela disponibilidade e todas as sugestões e correções.

Ao **Programa de Pós-Graduação em Genética** da Universidade de São Paulo, Faculdade de Medicina de Ribeirão Preto.

A todos os **professores** que se dedicaram ao meu ensinamento e compartilharam muito de seus conhecimentos.

Aos **funcionários da USP** sempre dispostos e prontos a me ajudar.

A **todos** que, de alguma maneira, contribuíram para que eu chegasse até aqui, meus sinceros agradecimentos.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

RESUMO

Januário, L. A. F. Otimização da resistência a nematódeos gastrointestinais em ovinos Santa Inês: uma abordagem de seleção genômica, *machine learning* e análise de imagens. 2023. Tese (Doutorado) – Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2023.

A infecção por nematoides gastrointestinais representa uma grande ameaça à saúde e produtividade das populações de ovinos, sendo o *Haemonchus contortus* a espécie mais patogênica. Este estudo analisou uma população de ovinos Santa Inês e foi composto por cinco capítulos com os seguintes objetivos: Capítulo 1) Revisão da literatura; Capítulo 2) Avaliar a viabilidade de usar características fenotípicas facilmente mensuráveis, a fim de predizer ovinos susceptíveis a nematoides gastrointestinais por meio do uso de métodos de “Machine Learning”; Capítulo 3) Analisar imagens da conjuntiva ocular para classificar os níveis de anemia com base nos escores de Famacha© (FAM); Capítulo 4) Examinar o padrão genético aditivo de valores genéticos estimados (EBVs) para características indicadoras de resistência a nematoides gastrointestinais; Capítulo 5) Avaliar a acurácia de modelos paramétricos (GBLUP, BayesA, BayesB e Lasso Bayesiano – BLASSO) e redes neurais artificiais (BRANN) na predição genômica de características indicadoras de resistência a nematoides gastrointestinais. No Capítulo 2, os animais foram classificados em resistentes, resilientes e suscetíveis de acordo com a contagem de ovos nas fezes (OPG) e volume globular (VG), e os métodos de classificação foram ajustados usando as informações de classe de idade, mês de registro, fazenda, sexo, FAM, peso corporal e escore de condição corporal como preditores. No Capítulo 3, um modelo “Random Forest” (RF) foi usado para segmentar as imagens. Após a segmentação, os quantis de intensidade de cor (1, 10, 20, 30, 40, 50, 60, 70, 80, 90 e 99%) em cada canal de imagem (vermelho, azul e verde) foram determinados e usados como variáveis explanatórias nos modelos de classificação, sendo o FAM 1 a 5 as classes a serem previstas. No Capítulo 4, os EBVs para FAM, VG e OPG foram estimados por inferência bayesiana em um modelo animal uni-característico. Em seguida, análises de agrupamento foram realizadas usando os EBVs para FAM, VG e OPG para identificar animais resistentes, resilientes e suscetíveis a nematoides gastrointestinais.

No Capítulo 5, a acurácia e o erro de predição foram obtidos para VG, OPG e FAM usando modelos paramétricos e redes neurais artificiais. Os resultados sugerem que o uso de características facilmente mensuráveis pode fornecer informações úteis para apoiar decisões de manejo a nível de fazenda. Os resultados das análises de imagem indicam que é possível prever com sucesso o FAM, especialmente para escores 2 a 4, em ovinos por meio de análise de imagem e modelo de RF usando imagens da conjuntiva ocular coletadas em condições de fazenda. O agrupamento dos animais resistente apresentou EBVs positivos para VG e negativos para FAM e OPG, sendo os animais mais desejáveis para serem usados como candidatos a seleção para melhorar geneticamente a resistência à nematoides gastrointestinais. Em resumo, os modelos paramétricos são adequados para a predição de valores genéticos genômicos de VG, OPG e FAM em ovinos, devido à similaridade da acurácia encontradas entre eles. Apesar disso, o uso do modelo GBLUP é recomendado devido ao seu menor custo computacional e à possibilidade de incorporar animais não genotipados na análise usando procedimentos “Single-step”.

Palavras chave: Nematódeos gastrointestinais. *Ovis aries*. *Machine Learning*. Análise de Imagem. Seleção Genômica.

ABSTRACT

Januário, L. A. F. Optimization of resistance to gastrointestinal nematodes in Santa Inês sheep: a genomic selection, machine learning and image analysis approach. 2023. Tese (Doutorado) – Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2023.

Gastrointestinal nematode infection represents a major threat to the health and productivity of sheep populations, and the *Haemonchus contortus* is the most pathogenic species. This study analyzed a population of Santa Ines sheep and it was composed of five chapters with the following objectives: Chapter 1) Literature review; Chapter 2) To evaluate the feasibility of using easily-measured phenotypic traits in order to predict the susceptibility of sheep to gastrointestinal nematodes through the use of machine learning methods; Chapter 3) To analyze ocular conjunctiva images to classify anemia levels based on Famacha© scores (FAM); Chapter 4) To examine the additive-genetic patterns of estimated breeding values (EBVs) for indicator traits of resistance to gastrointestinal nematodes; Chapter 5) To assess the accuracy of parametric models (GBLUP, BayesA, BayesB e Bayesian Lasso – BLASSO) and artificial neural networks (BRANN) in genomic predictions of indicator traits of resistance to gastrointestinal nematodes. In the Chapter 2, the animals were classified into resistant, resilient, and susceptible according to fecal egg count (FEC) and packed cell volume (PCV), and the methods were fitted using the information of age class, the month of record, farm, sex, FAM, body weight, and body condition score as predictors. In the Chapter 3, a random forest model (RF) was used to segment the images. After segmentation, the quantiles of color intensity (1, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 99%) in each image channel (red, blue, and green) were determined and used as explanatory variables in the classification models, and the FAM 1 to 5 were the target classes to be predicted. In the Chapter 4, The EBVs for FAM, PCV, and FEC were estimated by Bayesian inference in a single-trait animal model. After, cluster analyses were done using the EBVs for FAM, PCV, and FEC in order to identify animals that are resistant, resilient, and susceptible to gastrointestinal nematodes. In the Chapter 5, the prediction accuracy and mean squared errors were obtained for PCV, FEC, and FAM using parametric models and artificial neural network. The results suggest that the use of easily measurable traits may provide useful information for supporting management decisions at the farm level. The image analysis results indicate that is possible to successfully predict Famacha© score, especially for scores 2 to 4, in sheep via image analysis and RF model using ocular conjunctiva images collected in farm conditions. The resistant cluster presented positive EBVs for PCV and negative for FAM and FEC, being consisted of the most desirable animals to be used as selection candidates in order to genetically improve resistance to gastrointestinal nematodes. In summary, parametric models are suitable for genome-enabled prediction of PCV, FEC and FAM in sheep, due to the small differences in accuracy found between them. Despite this, the use of the GBLUP model is recommended due to its lower computational costs and the possibility of incorporating non-genotyped animals into the analysis using single-step procedures.

Keywords: Gastrointestinal nematodes. *Ovis aries*. Machine learning. Image analysis. Genomic selection.

TABLE OF CONTENTS

CHAPTER 1. GENERAL CONSIDERATIONS

CHAPTER 2. CLASSIFICATION PERFORMANCE OF MACHINE LEARNING METHODS FOR IDENTIFYING RESISTANCE, RESILIENCE, AND SUSCEPTIBILITY TO *HAEMONCHUS CONTORTUS* INFECTIONS IN SHEEP

CHAPTER 3. IMAGE ANALYSIS TO AUTOMATICALLY CLASSIFY ANEMIA BASED ON FAMACHA© SCORE IN SHEEP INFECTED WITH *HAEMONCHUS CONTORTUS* USING OCULAR CONJUNCTIVA IMAGES

CHAPTER 4. CLUSTER ANALYSIS TO EXPLORE ADDITIVE-GENETIC PATTERNS FOR THE IDENTIFICATION OF SHEEP RESISTANT, RESILIENT AND SUSCEPTIBLE TO GASTROINTESTINAL NEMATODES

CHAPTER 1 – GENERAL CONSIDERATION

1. INTRODUCTION

Brazil has approximately 20 million sheep, with about 70% located in the Northeast region (IBGE, 2021). The adaptation of these animals to the region, along with the favorable farming characteristics and technological advancements, make sheep farming an attractive option for producers in various regions, especially in the Northeast (IBGE, 2014). However, Brazilian sheep farming, like in other countries with a tropical climate, faces a significant challenge in the form of infections caused by gastrointestinal nematodes (GIN). The primary cause of GIN is the highly pathogenic parasite *Haemonchus contortus*, which results in significant economic losses due to decreased productivity and animal death, as well as expenses related to the use of anthelmintics for control (ADDUCI et al., 2022). Moreover, the overuse of anthelmintics has resulted in the development of parasite resistance (KUISEU et al., 2021; SZEWC et al., 2021).

To reduce the problems associated with GIN, alternatives such as selection based on phenotypic traits and genetic/genomic selection can be implemented. The latter option involves identifying sheep that are resistant to gastrointestinal nematodes and incorporating them into breeding programs for genetic improvement of sheep populations. One breed of sheep known to be resistant to gastrointestinal nematodes is the Santa Inês, a breed derived from crosses of Morada Nova, Bergamasca, and the native Crioula breed (AMARANTE et al., 2004).

Studies have evaluated the feasibility of making decisions based on phenotypic and genetic selection using measures such as fecal egg count (FEC), packed cell volume (PCV), and the Famacha© method (FAM), which have been found to be highly correlated with resistance to GIN (VAN WYK and BATH, 2002; SELVAM, 2021, 2022; FREITAS et al., 2022, 2023). However, there are few genomic selection studies that focus on indicator traits for GIN resistance in sheep.

Sheep infected with gastrointestinal nematodes can be classified into three categories based on their response to infection: resistant, resilient, and susceptible (WOOLASTON and BAKER, 1996). Currently, this classification is done using phenotypic traits such as fecal egg count and packed cell volume (POLLOT et al., 2004; BISHOP, 2012; MCMANUS et al., 2014; MARQUES et al., 2018; FREITAS et al., 2023). However, these parameters require laboratory resources, which increases the cost for farmers. To reduce these costs, easily-measured phenotypic such as body weight, body condition score, age class, month of record and sex could be explored to identify susceptible animals and support management decisions at the farm level.

The Famacha© method is a useful tool for identifying parasitized animals through the diagnosis of anemia, providing scores based on visual assessment of ocular conjunctiva coloration (VAN WYK and BATH, 2002). However, this method is prone to error as it requires a subjective evaluation from the professional performing the analysis. Developing high-throughput measurements of traits such as diagnostic image analysis can help improve accuracy and eliminate human subjectivity in assessment.

Genetic improvement can also play a role in controlling GIN infections by selecting animals with genetic resistance to gastrointestinal nematodes. Genetic variability is crucial for achieving genetic progress, and the estimation of genetic parameters combined with multivariate analyses can group animals based on their estimated breeding values, helping to identify the most suitable candidates for selection based on additive-genetic patterns for indicator traits of gastrointestinal nematode resistance (FREITAS et al., 2022).

With the advent of genomic information, molecular information has provided a shift in animal breeding. High-density single nucleotide polymorphism (SNP) panels have expanded the horizons for the study of complex traits controlled by multiple genes with small effects and improved marker-assisted selection in animal and plant breeding programs (XU et al., 2020). Including molecular marker information in genomic prediction models has also increased the accuracy of breeding value prediction in genomic selection (MEUWISSEN et al., 2001). However, the performance of genomic selection models depends on numerous factors, including the methodology used to estimate genomic breeding values, so the choice of statistical method must be carefully considered before implementing genomic selection in breeding programs.

2. GENERAL OBJECTIVES

The objectives of this study are as follows:

1. To evaluate the feasibility of using readily available, easily-measured phenotypic traits in order to predict the susceptibility of sheep to gastrointestinal nematodes through the use of machine learning methods.
2. To analyze ocular conjunctiva images to automatically classify anemia levels based on Famacha© scores.
3. To examine the pattern of estimated breeding values for indicator traits of resistance to gastrointestinal nematodes, in terms of genetic additivity.
4. To assess the accuracy of parametric models and artificial neural networks in predicting indicator traits of resistance to gastrointestinal nematodes in Santa Inês sheep through genomic predictions.

2.1 Specific objectives

For Objective 1:

- Compare the performance of different methods for classifying sheep as resistant, resilient, or susceptible to gastrointestinal nematodes using multinomial logistic regression (MLR), random forest (RF), linear discriminant analysis (LDA), and artificial neural network (ANN).
- Assess the suitability of the top-performing classification model for each farm.

For objective 2:

- To compare the performance of two classification models (MLR and RF).
- To assess the effectiveness of the top-performing classification model in three sheep farms.

For objective 3:

- To classify sheep into resistant, resilient, and susceptible groups based on their additive-genetic patterns for resistance to gastrointestinal nematodes.
- To determine the best candidates for selection by analyzing their genetic patterns.

3. LITERATURE REVIEW

3.1 Brazilian Santa Inês breed

The Santa Inês breed is a hair sheep variety that originated from Northeast Brazil, resulting from crossbreeding of Morada Nova, Bergamasca, and the indigenous Crioula sheep with a rough wool coat (Figure 1). This breed has become the fastestgrowing in Brazil, with approximately 39% of the country's sheep herds consisting of Santa Inês sheep, distributed across 1,185 counties (MCMANUS et al., 2014).

Figure 1. Ram of the Santa Ines breed.
Source: Personal Archive.

One reason for its widespread presence is its high adaptability to different climates in Brazil. In addition, the Santa Inês sheep have desirable characteristics such as good meat production, rusticity, low nutritional requirements, strong maternal ability, and low susceptibility to endoparasites and ectoparasites (PAIM et al., 2013). Studies have shown that Santa Inês sheep are more resistant to infections from gastrointestinal nematodes than the Suffolk and Ile de France breeds (AMARANTE et al., 2004; BRICARELLO et al., 2005; ROCHA et al., 2005). Female Santa Inês sheep are also unique in that they have a low reproductive seasonality (BALARO et al., 2015), and show estrus with lambs at their feet, leading to a faster production cycle.

3.2 Impact of gastrointestinal nematodes on sheep farming

The presence of gastrointestinal nematodes is a major concern for the growth of sheep farming, with *Haemonchus contortus* being the primary parasite. Its high prevalence due to climatic conditions, high pathogenicity, and high prolificity (a female can lay up to 5,000 eggs per day) make it a highly harmful parasite (EMERY et al., 2016). The impact of parasitic infections on animal health can result in untreated individuals becoming ill (ALAM et al., 2020). In terms of production, the consequences include slowed growth, decreased production of wool, meat, and milk, low reproductive efficiency, low resistance to diseases, and high mortality rates (JARDIM, 1974; COBON and O'SULLIVAN, 1992).

Sheep infected with gastrointestinal nematodes can be classified into three categories based on their response to the parasitic infection: resistant, resilient, and susceptible. Animals that are resistant are less likely to be infected, as their immune response helps to prevent the establishment of the parasite, without any negative impact on production (ALBERS et al., 1987). Resilience refers to the animal's ability to maintain its production performance in the face of a disease challenge (BISHOP, 2012; ALBERS et al., 1987). On the other hand, susceptible animals are those that become infected with the parasites and experience a high degree of anemia and impaired production. To prevent the spread of the disease to other populations, it's essential to identify and select animals that are resistant to the disease.

For a long time, controlling gastrointestinal nematodes in small ruminants has been achieved through the use of anthelmintics and pasture management (RODRIGO, 2017). However, due to the indiscriminate use of anthelmintics, which has led to parasite resistance, and the complexities of pasture management, along with the long lifespan of infective larvae in the environment, it has become increasingly urgent to develop alternative control techniques to prevent production losses.

3.3 Sheep resistance to gastrointestinal nematodes

An alternative to controlling gastrointestinal parasites is selecting for resistant animals (LI et al., 2001; FREITAS et al., 2022). These animals are able to eliminate infective larvae or inhibit the development of immature parasites (BALIC et al., 2000). Herds that consist of resistant animals and implement efficient pasture management are a sustainable alternative for preventing parasite infections, leading to increased production (GREEFF and KARLSSON, 2020).

Several studies have concluded that some breeds are naturally more resistant to parasitic infections than others (AMARANTE et al., 2004; ROCHA et al., 2005; BRICARELLO et al., 2005). In Brazil, the Santa Inês breed has been considered the most resistant (AMARANTE et al., 2004; ROCHA et al., 2005; BRICARELLO et al., 2005).

The packed cell volume (PCV), fecal egg count (FEC), and Famacha© method (FAM) are the most commonly used methods for diagnosing the degree of infection with gastrointestinal nematodes (POLLOT et al., 2004; BISHOP, 2012; MCMANUS et al., 2014; MARQUES et al., 2018; FREITAS et al., 2023). The FEC is used to count eggs of parasites present in the gastrointestinal tract of small ruminants (ALBERS et al., 1987), and animals with a zero or close to zero egg count per gram of feces are considered resistant. PCV measurements are closely related to resistance to gastrointestinal nematodes as they indicate the intensity of anemia and hypoproteinemia. The Famacha© method is widely used to detect sheep that are unable to cope with *H. contortus* infection through visual assessment of ocular conjunctiva coloration (OLIVEIRA et al., 2018; FREITAS et al., 2022; 2023). *H. contortus* causes anemia, which results in changes in the color of the ocular conjunctiva from deep red in healthy sheep to shades of pink or white in non-healthy sheep (SINGH and SWARNKAR, 2012; EL-ASHRAM et al., 2017).

So far, resistance based on phenotype has been evaluated for management purposes at the farm level, with resistant animals not needing treatment and susceptible animals requiring treatment and attention. Additionally, genetic resistance to gastrointestinal parasites is a sustainable alternative for reducing losses due to GIN infections and forming future resistant herds by selecting for genetically resistant animals.

Amarante (2004) have shown that resistance or susceptibility to parasitism is genetically determined. The traits used to assess resistance, such FEC, PCV, and FAM, have adequate genetic variability to respond to selection in the Santa Inês breed. According to Berton et al. (2017), the FAM method is expected to yield a higher genetic gain compared to the other methods. The studies by Berton et al. (2017), Oliveira et al. (2018), and Freitas et al. (2022) have demonstrated that genetic progress in research and commercial herds can be achieved through selection for resistance to nematode gastrointestinal infections.

3.4 Analysis to identify animals parasitized by gastrointestinal nematodes

3.4.1 Machine learning for classification

Machine learning (ML) models have been widely used for classification purposes in various fields. These models are generally modifications or extensions of linear models, which allow for nonlinear relationships between variables and handling of collinearity and high-dimensional data (GOLDSTEIN et al., 2017; MULLAINATHAN and SPIESS, 2017). ML algorithms have been utilized in different aspects of the livestock industry, such as detecting estrus in dairy cows (ANDRADE et al., 2021), classifying pathologies in animal necropsy reports (BOLLIG et al., 2020), and categorizing dairy cattle breeds (MOAWED et al., 2017). The best method for a particular task depends on the problem being studied. The results, in conjunction with farm-level management strategies, can lead to reduced costs and more effective interventions, such as reducing deworming in GIN infected herds. Furthermore, predicting the likelihood of a particular outcome can aid in formulating, evaluating, and directing animal management and production practices.

Other techniques have been developed to address nonlinear relationships between variables. One example is the Multilayer Perceptron (MPL), which creates layers of neurons, mathematical processing units that receive inputs and produce outputs according to different response functions. Another is the Random Forest, a decision tree methodology that outputs classifications or estimates based on the average of all trees. The Random Forest is currently the most widely used ML method due to its ability to handle both linear and nonlinear variables and its low variance, which results in higher prediction accuracy compared to logistic regression and simple decision trees (SOUZA, 2017).

Combining multivariate techniques with ML concepts can also enhance model performance, such as linear discriminant analysis (a supervised learning method for classification). Linear discriminant analysis is a generalization of Fisher's linear discriminant and is used for pattern recognition and ML to find a linear combination of features that separates or characterizes different classes of objects or events. The resulting combination can be used as a linear classifier or to reduce dimensionality before further classification. It is expected that the use of ML models will lead to improved performance with a practical level of accuracy for application at the farm level, allowing for more precise classification of animals as susceptible and resistant.

3.4.2 Image analysis

Several computer vision techniques have been developed due to advancements in image capture and the growing demand for more accurate information (SZELISKI, 2010). Computer vision simulates human vision, in which there is an image input, and the output is an interpretation as a whole or partially. It has medium-level processes such as segmentation or classification operations, and high-level processes related to cognition tasks associated with human vision.

An image can be defined as a two-dimensional function, *f*(x,y), where x and y are spatial coordinates and the amplitude of *f* represents the intensity or gray level of the image at a given point (pixel). Computer vision uses a set of machine learning techniques, such as random forests and artificial neural networks, to extract information from images (LEE et al., 2015). In this context, the texture of objects plays a prominent role, as it can be easily highlighted by our interpretation of black-and-white images. In addition, colors provide additional details to complement our inferences about the objects in question, serving as a more detailed classifier (KHODASKAR and LADHAKE, 2014).

The next important step in image processing is segmentation, which consists of partitioning the image into regions of interest to extract features from these regions. To achieve efficient pattern recognition, accurate and robust segmentation is crucial (CHAOU et al., 2015). After the feature extraction step, the identification and recognition of these features are followed by their classification.

In this study, we aim to automatically classify anemia based on the Famacha© method in sheep infected with *Haemonchus contortus* using ocular conjunctiva images. The Famacha© method is based on the Famacha© card (Figure 2) and has been created to detect sheep unable to cope with *H. contortus* infection by determining clinical anemia based on the visual assessment of ocular conjunctiva coloration (VAN WYK and BATH, 2002). For this task, there are no reports in the literature of the use of image analysis to automatically classify Famacha© score using ocular conjunctiva images of sheep.

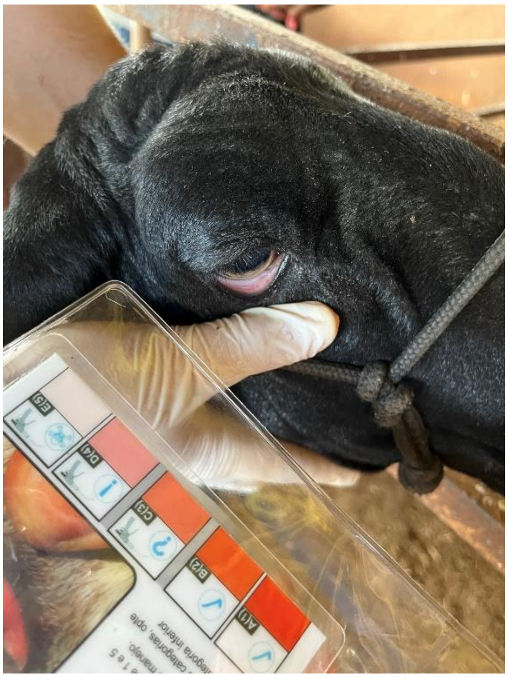

Figure 2. Application of the Famacha© method.
Source: Personal Archive.

3.5 Analysis to identify animals genetically resistant to gastrointestinal nematodes

3.5.1 Estimation of genetic parameters and cluster analysis

The selection of the best individuals for indicator traits of resistance to gastrointestinal nematodes (FEC, PCV, and FAM) can have a positive impact on the production chain. However, simply selecting individuals based on their phenotypes is not enough to guarantee the selection of animals with higher breeding value (BOUJENANE and DIALLO, 2017). Effective selection requires genetic evaluations and knowledge of the relevant genetic parameters (SAFARI et al., 2005).

To improve indicator traits of resistance to gastrointestinal nematodes, it is crucial to accurately estimate genetic parameters. This will ensure accurate predictions of breeding values and allow for effective selection. Many sheep breeding programs aim to select animals that are resistant to gastrointestinal nematodes based on these indicator traits (ZVINOROVA et al., 2016; BERTON et al., 2017; OLIVEIRA et al., 2018; HAYWARD, 2022; FREITAS et al., 2022). However, because genetic parameters can vary between populations and environments, it is important to estimate them in the population where selection will be applied.

Traits with moderate to high estimated heritability will respond more efficiently to selection compared to traits with low heritability (DOVALE et al., 2013). The indicator traits of resistance to gastrointestinal nematodes have a low to moderate magnitude of heritability. Despite this, investment in the selection of genetically superior animals for these traits is justified because the genetic gains obtained through breeding will be permanent.

In breeding programs, selection must be balanced and take into account the primary economic traits of interest. Selecting for multiple traits using weighted selection can reduce the risk of harm to any one trait due to genetic correlations. Thus, combining estimates of genetic parameters with cluster analysis can help in the decision-making process for selecting animals that will have the greatest positive impact on desired traits in the population.

The Santa Inês breed presents adequate genetic variability for the selection of indicator traits of resistance to gastrointestinal nematodes (BERTON et al., 2017). Therefore, cluster analysis can be a useful alternative for finding and grouping animals that are genetically similar based on traits of interest.

3.5.2 Genomic Selection

Genomic prediction is a method for evaluating the genetic potential of individuals for selection, based on the genotypes of single nucleotide polymorphisms (SNPs). When applied to animal breeding, it is known as genomic selection. The impact of SNPs on genomic selection is estimated using data from molecular markers and phenotypic records (MEUWISSEN et al.; 2001). The foundation of this methodology is the linkage disequilibrium between SNPs and quantitative trait loci (QTLs). Genomic prediction through regression models can be performed in two steps. The first step uses a reference population with both phenotypic and genotypic information to estimate the impact of the markers. The second step uses only marker information from candidate animals (VANRADEN et al., 2009).

In animal breeding, several genome-enabled prediction methods, such as Bayesian Alphabet, often make use of shrinkage or regularization processes to impose prior assumptions about the genetic architecture of complex traits (DE LOS CAMPOS et al., 2013). The GBLUP method replaces the traditional relationship matrix with a genomic relationship matrix based on markers (G). BayesA, as described by Meuwissen et al. (2001), allows for different variances across segments of the genome. It uses a scaled-t density as the prior distribution for SNP effects. BayesB also allows for different genetic variances explained by each locus and uses a two-component mixture prior, with a point of mass at zero and non-null effects following a scaled-t distribution with a prior probability π.

Artificial neural networks (ANNs) are mathematical information processing systems that mimic the human brain. Like the human brain, ANNs process input information (such as marker data) through interconnected artificial neurons, which can learn complex relationships between predictor variables and the target in an adaptive way through appropriate learning algorithms (BISHOP, 2006). Neural network methods have also been proposed for predicting genomic breeding values. ANNs are non-linear and can model complex functions. There are several architectures for ANNs, including the number of hidden layers and the number of neurons in each layer, as well as the type of activation function performed at each neuron. One of the most commonly used in genome-enabled prediction is the Multilayer Perceptron (MLP). MLP neural networks are of the feed-forward multilayer type, where learning is performed using a dataset (Figure 3).

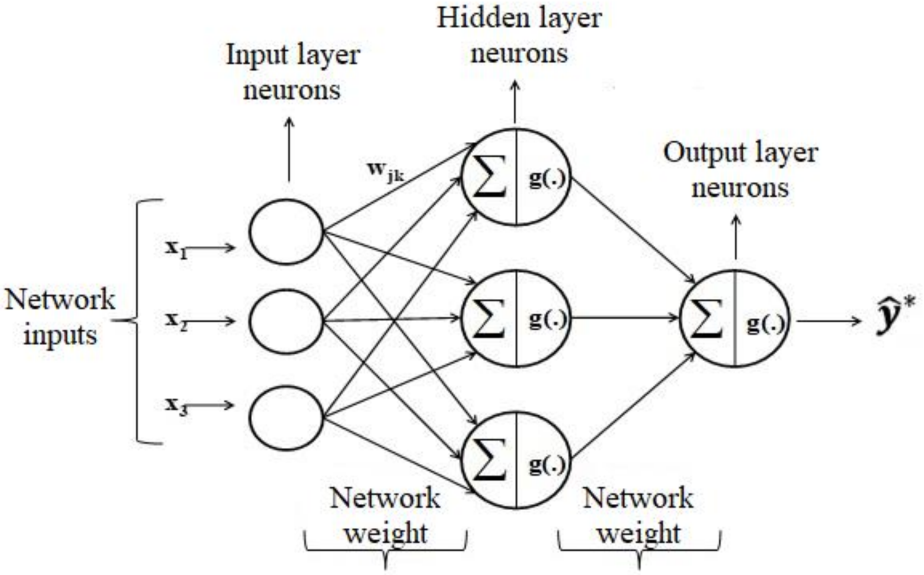

Figure 3. Topology of a multilayer perceptron (MLP) neural network with three neurons in the input layer, three in the intermediate layer and one in the output layer. The input information is x_1, x_2 and x_3. The synaptic weights between neurons in adjacent layers are represented by w_{jk}, where *j* are the neurons in the input layer and *k* are the neurons

in the intermediate layer; $\sum$ is the weighted sum given by $\sum_{j=1}^{n} x_k w_{jk}$, g(.) is the activation function and $\hat{\boldsymbol{y}}^*$ is the predicted value.

For genome-enabled prediction, a particular advantage is that ML methods are model-free, in other words, there is no necessity to impose a specific genetic structure, so that, no assumptions are required about the genetic architecture of the target traits. However, results considering empirical applications of machine learning algorithms on real data still have been few explored, particularly for Santa Ines sheep.

In this scenario, using genomic prediction methods for the identification of genetically resistance to GIN becomes a promising strategy, allowing to achieve higher genetic gains by the improvement of prediction accuracy and decrease on generation intervals (MEUWISSEN et al. 2013). Notably, the success of genomic-assisted selection depends directly on the prediction accuracy, which is associated, among other factors, with the used statistical method and the genetic architecture of the interest traits.

4. REFERENCES

ADDUCI, I. et al. Haemonchosis in Sheep and Goats, Control Strategies and Development of Vaccines against *Haemonchus contortus.* **Animals**, v. 12, n. 18, p. 2339, 2022.

ALAM, T. M. R. *Heamonchus Contortus* infection in Sheep and Goats: alterations in haematological, biochemical, immunological, trace element and oxidative stress markers, **Journal of Applied Animal Research**, v. 48, n. 1, p. 357-364, 2020.

ALBERS, G. A. A. et al. The genetics of resistance and resilience to *Haemonchus contortus* infection in young Merino sheep. **International Journal for Parasitology**, v. 17, n. 7, p. 1355-1363, 1987.

AMARANTE, A. F. et al. Resistance of Santa Ines, Suffolk and Ile de France sheep to naturally acquired gastrointestinal nematode infections. **Veterinary Parasitology**, v. 120, n. 1-2, p. 91–106, 2004.

ANDRADE, V. V. et al. Estrus prediction models for dairy gyr heifers. **Animals**, v. 11, n. 11, p. 3103, 2021.

BALARO, A. M. et al. Reproductive performance, metabolic and hormonal profiles of Santa Inês ewes in Winter and summer under tropical conditions. **Tropical Animal Health and Production**, v. 47, n. 3, p. 627-631, 2015.

BALIC, A. et al. The immunobiology of gastrointestinal nematode infections in ruminants. **Advances in Parasitology**, v.45, p.181-241, 2000.

BERTON, M. P. et al. Genomic regions and pathways associated with gastrointestinal parasites resistance in Santa Inês breed adapted to tropical climate. **Journal Animal Science Biotechnology,** v.8, p.73, 2017.

BISHOP, C. M. **Pattern Recognition and Machine learning**. New York: Springer, 2006, 738p.

BISHOP, S.C. A consideration of resistance and tolerance for ruminant nematode infections. **Frontiers in Genetics**, v. 3, p. 168, 2012.

BOLLIG, N. et al. Machine learning for syndromic surveillance using veterinary necropsy reports. **PLoS ONE,** v. 15, n. 2, p. e0228105, 2020.

BOUJENANE, I.; DIALLO, I.T. Estimates of genetic parameters and genetic trends for pre-weaning growth traits in Sardi sheep. **Small Ruminant Research**, v. 146, p. 61-68, 2017.

BRICARELLO, P. A. et al. Influence of dietary protein supply on resistance to experimental infections with *Haemonchus contortus* in Ile de France an Santa Ines lambs. **Veterinary Parasitology**, v.134, n. 1-2, p. 99-109, 2005.

CHAOU, A. K.; MEKHALDI, A.; TEGUAR, M. Elaboration of novel image processing algorithm for arcing discharges recognition on hv polluted insulator model. **IEEE Transactions on Dielectrics and Electrical Insulation**, v. 22, n. 2, p. 990–999, 2015.

COBON, D. H.; O'SULLIVAN, B. M. Effect of *Haemonchus contortus* on Productivity of Ewes, Lambs and Weaners in a Semi-Arid Environment. **The Journal of Agricultural Science**, v. 118, n. 2, p. 245–248, 1992.

DE LOS CAMPOS, G. Whole-genome regression and prediction methods applied to plant and animal breeding. **Genetics**, v. 193, n. 2, p. 327–45, 2013.

DOVALE et al. Genetic responses of traits relationship to components of nitrogen and phosphorus use efficiency in maize. **Acta Scientiarum. Agronomy**, v. 35, n. 1, p. 31-38, 2013.

EL-ASHRAM, S., et al. *Haemonchus contortus* and Ovine Host: a Retrospective Review. **International Journal of Advanced Research**, v. 5, p. 972–999, 2017.

EMERY, D. L. et al. *Haemonchus contortus*: the then and now, and where to from here? **International Journal for Parasitology**, v. 46, n.12, p. 755-769, 2016.

FREITAS, L. A. et al. Cluster analysis to explore additive-genetic patterns for the identification of sheep resistant, resilient and susceptible to gastrointestinal nematodes. **Veterinary Parasitology**, v. 301, p. 109640, 2022.

FREITAS, L. A. et al. Classification performance of machine learning methods for identifying resistance, resilience, and susceptibility to gastrointestinal nematode infections in sheep. **Animals**, v. 13, n. 3, p. 374, 2023.

GOLDSTEIN, B. A. et al. Moving beyond regression techniques in cardiovascular risk prediction: Applying machine learning to address analytic challenges. **European Heart Journal**, v. 38, n. 23, p. 1805–1814, 2017.

GREEFF J. C.; KARLSSON L. J. E. Production benefits of breeding for worm resistance in Merino sheep. **Animal Production Science**, v. 60, n. 13, p. 1643-1653, 2020.

HAYWARD, A. D. Genetic parameters for resistance to gastrointestinal nematodes in sheep: a meta-analysis, **International Journal for Parasitology**, v. 52, n. 13–14, p. 843-853, 2022.

IBGE, 2014. **Instituto Brasileiro de Geografia e Estatística**. Disponível em: <http://www.ibge.gov.br/home/estatistica/economia/ppm/2013/default.shtm>

IBGE, 2021. **Instituto brasileiro de Geografia e Estatística.** Disponivel em: <https://www.ibge.gov.br/explica/producao-agropecuaria/ovino/br>

JARDIM, W. R. **Criação de caprinos**. 11. ed. São Paulo: Nobel, 1987. 239 p.

KUISEU J. et al. Prevalence, effects and alternative control methods of *Haemonchus contortus* in small ruminants: A review. **Journal of Veterinary Medicine and Animal Health** v. 13, p. 84-97, 2021.

KHODASKAR, A. A.; LADHAKE, S. A. Pattern recognition: Advanced development, techniques and application for image retrieval. In: Communication and Network Technologies (ICCNT), Sivakasi. **Anais…** Sivakasi: India, 2014.

LEE, W. P. et al. Recognition of fish based on generalized color fourier descriptor. In: Science and Information Conference, London. **Anais…** London: United Kingdom, 2015.

LI, Y. et al. Epidemiological observations and heterosis analysis of gastrointestinal nematode parasitism in Suffolk, Gulf Coast Native, and crossbred lambs. **Veterinary Parasitology**, v. 98, n. 4, p. 273-283, 2001.

MARQUES, C. A. T. et al. The use of targeted selective treatments on controlling gastrointestinal nematodes in different sheep categories under grazing system. **Pesquisa Veterinária Brasileira**, v. 38, n. 3, p. 470–476, 2018.

MCMANUS, C. et al. Selection methods for resistance to and tolerance of helminths in livestock. **Parasite**, v. 21, p. 56, 2014.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome wide dense marker maps. **Genetics**, v. 157, n. 4, p. 1819-1829, 2001.

MEUWISSEN T. H. E.; HAYES B. J.; GODDARD M.E. Accelerating improvement of livestock with genomic selection. **Annual Review of Animal Biosciences**, v. 1, p. 221–37, 2013.

MOAWED, S.A.; OSMAN, M. The Robustness of Binary Logistic Regression and Linear Discriminant Analysis for the Classification and Differentiation between Dairy Cows and Buffaloes. **Journal of Applied Probability and Statistics**, v. 7, n. 6, p. 304–310, 2017.

MULLAINATHAN, S.; SPIESS, J. Machine learning: An applied econometric approach. **Journal of Economic Perspectives**, v. 31, n. 2, p. 87–106, 2017.

PAIM, T. P. et al. Performance, survivability and carcass traits of crossbred lambs from five paternal breeds with local hair breed Santa Inês ewes. **Small Ruminant Research**, v. 112, n. 1-3, p. 28-34, 2013.

POLLOT, G. E. et al. Genetic parameters for indicators of host resistance to parasites from weaning to hogget age in Merino sheep. **Journal of Animal Science**, v. 82, n. 10, p. 2852–2864, 2004.

ROCHA, R. A.; Amarante, A. F. T.; Bricarello, P. A. Resistance of Santa Inês andIle de France suckling lambs to gastrointestinal nematode infections. **Revista Brasileira de Parasitologia Veterinária**, v.14, n. 1, p.17-20, 2005.

RODRIGO, L. **Human Helminthiasis**. London: IntechOpen, 2017. 158 p.

SAFARI, E. et al. A review of genetic parameter estimates for wool, growth, meat and reproduction traits in sheep. **Livestock Production Science**, v. 92, n. 3, p. 271-289, 2005.

SELVAM, R. Estimation of genetic parameters of traits for evaluation of resistance to gastrointestinal nematode infection in Vembur sheep. **The Pharma Innovation Journal,** v. 10, n. 7, p. 686-688, 2021.

SELVAM, R. Genetic Parameter Analysis of Indicator Traits of Gastrointestinal Nematode Infection in Sheep Breeds. **International Journal of Bio-resource and Stress**, v. 13, p. 1, 17-21, 2022.

SINGH, D.; SWARNKAR, C. P. Evaluation of targeted selective treatment strategy in sheep farm of Rajasthan. **Indian Journal of Animal Research**, v. 82, n. 7, p. 679-686, 2012.

SOUZA, L. A. M. de. **Aplicação de aprendizado de máquina para predição de prioridade em gestão de incidentes**. 2017. 61p. Monografia (Graduação em Sistemas de Informação) – Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, 2017.

SZELISKI, R. Computer Vision: Algorithms and Applications. 1st ed. London: Springer, 2010, 812 p.

SZEWC, M.; WAAL, T.; ZINTL, A. Biological methods for the control of gastrointestinal nematodes. **The Veterinary Journal**, v. 268, p. 105602, 2021.

VAN WYK J. A.; BATH G. F. The FAMACHA system for managing haemonchosis in sheep and goats by clinically identifying individual animals for treatment. **Veterinary**

Research, v. 33, n. 5, p. 509-29, 2002.

VANRADEN, P. M. Efficient methods to compute genomic predictions. **Journal of Dairy Science** v. 91, n. 11, p. 4414–4423, 2008.

XU, Y. et al. Enhancing Genetic Gain through Genomic Selection: From Livestock to Plants. **Plant Communications**, v. 1, n. 1, p. 100005, 2020.

WOOLASTON, R. R.; BAKER, R. L. Prospects of breeding small ruminants for resistance to internal parasites. **International Journal for Parasitology**, v. 26, n. 8-9, p. 845–855, 1996.

ZVINOROVA, P. I. et al. Breeding for resistance to gastrointestinal nematodes - the potential in low-input/output small ruminant production systems. **Veterinary Parasitology**, v. 225, p. 19-28, 2016.

CHAPTER 2. CLASSIFICATION PERFORMANCE OF MACHINE LEARNING METHODS FOR IDENTIFYING RESISTANCE, RESILIENCE, AND SUSCEPTIBILITY TO *HAEMONCHUS CONTORTUS* INFECTIONS IN SHEEP

Manuscript published in the journal of Animals (Online) as part of the results obtained in this Ph.D. study. DOI: https://doi.org/10.3390/ani13030374

Received: 16 December 2022; Revised: 16 January 2023; Accepted: 20 January 2023; Published: 21 January 2023.

Following is the manuscript published open access according to the Journal of Animals sharing policies that allow the deposited manuscript in repositories.

animals

MDPI

Article

Classification Performance of Machine Learning Methods for Identifying Resistance, Resilience, and Susceptibility to *Haemonchus contortus* Infections in Sheep

Luara A. Freitas [1,2], Rodrigo P. Savegnago [3], Anderson A. C. Alves [2], Ricardo L. D. Costa [4], Danisio P. Munari [5], Nedenia B. Stafuzza [6], Guilherme J. M. Rosa [2] and Claudia C. P. Paz [1,6,*]

[1] Department of Genetics, University of Sao Paulo, Ribeirão Preto 14049-900, SP, Brazil
[2] Department of Animal and Dairy Sciences, University of Wisconsin, Madison, WI 53706, USA
[3] Department of Animal Science, Michigan State University, East Lansing, MI 48864, USA
[4] São Paulo Agency of Agribusiness and Technology, Animal Science Institute, Nova Odessa 13380-011, SP, Brazil
[5] School of Agricultural and Veterinary Sciences, São Paulo State University, Jaboticabal 14884-900, SP, Brazil
[6] Sustainable Livestock Research Center, Animal Science Institute, São José do Rio Preto 15130-000, SP, Brazil
* Correspondence: claudiacristinaparopaz@gmail.com

Simple Summary: Infection by gastrointestinal nematodes is a major sanitary issue in sheep production. Therefore, improvements in the animal's health are important to reduce losses and improve animal welfare. Thus, this study investigated the feasibility of using easy-to-measure phenotypic traits to predict sheep resistant, resilient, and susceptible to gastrointestinal nematodes, compared the classification performance of different methods, and evaluated the applicability of the best classification model on each farm. The results revealed the multinomial logistic regression and linear discriminant analysis models presented the best classification performances for the susceptible and resistant animals. The results suggest that the use of readily available records and easily measurable traits may contribute to the identification of susceptible animals, supporting management decisions at the farm level and potentially reducing the economic losses due to parasitic infection. The animals identified as resistant can also be incorporated as selection candidates into breeding programs for the genetic improvement of sheep populations.

Abstract: This study investigated the feasibility of using easy-to-measure phenotypic traits to predict sheep resistant, resilient, and susceptible to gastrointestinal nematodes, compared the classification performance of multinomial logistic regression (MLR), linear discriminant analysis (LDA), random forest (RF), and artificial neural network (ANN) methods, and evaluated the applicability of the best classification model on each farm. The database comprised 3654 records of 1250 Santa Inês sheep from 6 farms. The animals were classified into resistant (2605 records), resilient (939 records), and susceptible (110 records) according to fecal egg count and packed cell volume. A random oversampling method was performed to balance the dataset. The classification methods were fitted using the information of age class, the month of record, farm, sex, Famacha© degree, body weight, and body condition score as predictors, and the resistance, resilience, and susceptibility to gastrointestinal nematodes as the target classes to be predicted considering data from all farms randomly. An additional leave-one-farm-out cross-validation technique was used to assess prediction quality across farms. The MLR and LDA models presented good performances in predicting susceptible and resistant animals. The results suggest that the use of readily available records and easily measurable traits may provide useful information for supporting management decisions at the farm level.

Keywords: multinomial logistic regression; *Ovis aries*; precision; sensitivity

check for updates

Citation: Freitas, L.A.; Savegnago, R.P.; Alves, A.A.C.; Costa, R.L.D.; Munari, D.P.; Stafuzza, N.B.; Rosa, G.J.M.; Paz, C.C.P. Classification Performance of Machine Learning Methods for Identifying Resistance, Resilience, and Susceptibility to *Haemonchus contortus* Infections in Sheep. *Animals* **2023**, *13*, 374. https://doi.org/10.3390/ani13030374

Academic Editor: Ilektra A. Fragkou

Received: 16 December 2022
Revised: 16 January 2023
Accepted: 20 January 2023
Published: 21 January 2023

1. Introduction

Infection by gastrointestinal nematodes is a great sanitary concern for the sheep meat industry, causing, directly and indirectly, economic impacts due to the compromised health conditions of infected animals [1]. The negative impacts of gastrointestinal nematodes include subclinical weight loss and lethal pathologies such as anemia, severe protein loss, reduced feed intake due to anorexia, and in cases of massive infection, high mortality rates [2–4]. The most pathogenic blood-feeding parasitic nematode in sheep is *Haemonchus contortus*, which is responsible for causing anemia and great economic losses [5]. The indiscriminate use of anthelmintics to reduce infections by gastrointestinal nematodes resulted in serious problems of parasite resistance to anthelmintics and increased production costs [6]. The combined annual cost of the helminth infections in 18 European and neighbor countries was estimated at €1.8 billion, with 81% percent of this cost due to lost production and 19% attributed to treatment costs [7].

An alternative to reduce the problems associated with anthelmintics is the detection of animals according to their response to parasitic infection and consequently the breeding of sheep for resistance to nematodes, i.e., sheep with the immunological capacity to control parasitic infection [8,9]. The Santa Inês is a breed of hair sheep derived from crosses of Morada Nova, Bergamasca, and the native coarse-wooled Crioula, which present a higher resistance to gastrointestinal nematodes than Suffolk and Ile de France sheep [10].

Sheep can be classified into three classes according to their response to parasitic infection: resistant, resilient, and susceptible [11]. Resistant animals are less prone to be infected because their infection response acts to avoid the establishment of the parasite, without damage to production [12]. Resilience is related to the animal's ability to keep production performance in the face of a disease challenge [12,13]. Susceptible animals are those that host the parasites and have a high degree of anemia followed by impaired production. Currently, infection level can be quantified by fecal egg count and packed cell volume [13–16]. However, these parameters depend on blood and feces collection, which are difficult to perform and require laboratory resources that increase costs for farmers. Cheaper, non-invasive, and non-laborious alternatives to classify the animals' infections response must be explored in order to reduce productions costs.

Machine learning model (ML) theory represents a branch of artificial intelligence that combines statistics, computer science, and data mining principles aiming to find and learn inherent patterns and to classify (or predict) interest outcomes. MLs have been used for classification in many areas. Most of them are generalizations or extensions of linear models to make them more flexible to accommodate nonlinear relationships between variables, dealing with collinearity and high dimensional data [17,18]. These classification algorithms have been increasingly used in different tasks in the livestock industry, such as for identifying the estrus occurrence in dairy cows [19], detection of different pathologies in animal necropsy reports [20], and classification of dairy cattle breeds [21], with the most suitable method depending on the problem studied.

Although there are several studies using those models for classification, to the best of our knowledge, the potential of using ML to classify resistance, resilience, and susceptibility to gastrointestinal nematode infections in sheep populations has not been investigated yet. Hence, the objectives of this study were (1) to investigate the feasibility of using easy-to-measure phenotypic traits to predict sheep resistant, resilient, and susceptible to gastrointestinal nematodes, (2) to compare the classification performance of different methods (multinomial logistic regression—MLR, linear discriminant analysis—LDA, random forest—RF, and artificial neural network—ANN), and (3) evaluate the applicability of the best classification model on each farm.

2. Materials and Methods

2.1. Data Set

The dataset comprised 3654 phenotypic samples of 1250 Santa Inês sheep recorded in two periods: from 2013 to 2014 and from 2018 to 2020. The animals came from six farms: Cravinhos (latitude: 21°20′25″ S; longitude: 47°43′46″ W), Jardinopolis (latitude: 21°01′04″ S; longitude: 47°45′50″ W), Nova Odessa (latitude: 22°46′39″ S, longitude: 47°17′45″ W), Pontal (latitude: 21°01′21″ S; longitude: 48°02′14″ W), Serrana (latitude: 21°12′41″ S; longitude: 47°35′44″ W), and Ventania (latitude: 24°14′45″ S; longitude: 50°14′34″ W). The experimental procedures were conducted following the recommendations of the Institutional Animal Care and Use Committee of the Animal Science Institute, Nova Odessa, São Paulo, Brazil (protocol code CEUA Nº. 267-18, 3 October 2018).

The feces were collected directly from the rectal ampoule and individually subjected to fecal egg count using the modified McMaster technique of Gordon and Whitlock [22]. A pool of feces samples was separated for the preparation of larvae culture to establish the genera of nematodes prevalent in the herd [23]. The results showed that the main genus of gastrointestinal nematode found during the study was *Haemonchus* (63%) followed by *Trichostrongylus* (24%), *Cooperia* (7%), and *Oesophagostomum* (6%) [24].

Blood samples were collected in 5 mL vacutainer tubes containing EDTAK3 by puncture of the jugular vein. The packed cell volume was measured by the microhematocrit centrifugation technique [25]. Resistant animals (2605 records) had fecal egg counts equal or less than 1000 eggs/g and packed cell volume equal or higher than 22%. The resilient animals (939 records) had fecal egg counts over 1000 eggs/g and packed cell volume higher or equal to 22%, and the susceptible animals (110 records) had packed cell volume less than 22%.

Body weight was measured with an electronic scale. Body condition score was obtained by visual inspection and by palpation of the dorsal lumbar region of the spine, constituting scores from 1 to 5 and their intermediates (1.5, 2.5, 3.5, and 4.5), which represent a (1) cachectic animal; (2) thin animal; (3) moderate animal; (4) fat animal; and (5) obese animal [26,27]. The missing observations for body weight and body condition score were imputed considering the animal's most recent valid measure.

The Famacha© diagnosis was performed by trained technicians using the Famacha© card, which compares the different shades of the ocular conjunctiva on a five-point scale corresponding to the colors robust red (non-anemic—1), red/pink (non-anemic—2), pink (mildly anemic—3), pink/white (anemic—4), and white (severely anemic—5) [28].

2.2. Sampling Technique

The different classes (resistant, resilient, and susceptible) were not equally represented in our data sets (Table S1). As it is well known that prediction models can have poor prediction performance with highly unbalanced datasets, the method used to overcome the unbalanced dataset issue was the random over-sampling method, which randomly replicates instances in the minority class. The random oversampling method was performed through the R software using the caret package (R Core Team, 2018).

2.3. Classification Models

Multinomial logistic regression (MLR), linear discriminant analysis (LDA), random forest (RF), and artificial neural network (ANN) methods were used to classify the animals according to their response to the parasitic infection. The hyperparameter values were set to defaults for the MLR, LDA, and RF models. Different architectures were tested for the ANN model, and the architecture with the best performance in our training set was that with a single layer and 7 hidden neurons, 200 epochs, the sigmoidal activation function in the hidden layer, and the softmax function in the output layer. The analyses were performed with the R software (R Core Team, 2018), using the randomForest package for random forest analyses [29], the MASS package for linear discriminant analyses, and the nnet package for artificial neural networks and multinomial logistic regression analyses [30].

The response variable was the animal class (resistant, resilient, and susceptible, coded as 0, 1, 2, respectively). The exploratory variables were body weight (kg), body condition score (scores 1 to 5 and their intermediates), farm, sex (3224 female records and 430 male records), age class (194 records in class 1: 1 to 6 months of age, 588 records in class 2: 6 to

18 months of age, and 2872 records in class 3: over 18 months of age), month of record and Famacha© degree (from 1 to 5; [28]).

2.4. Classification Performance Metrics

For comparison purposes, the whole dataset was randomly split into training and testing sets, which contained 80% and 20% from the original number of observations, respectively. Classification metrics were calculated using 5-fold cross-validation with 20 replicates. Precision and sensitivity (%) were used for comparing the classification methods for each class. The classification metrics were calculated as follows:

$$\text{Precision } = \frac{\text{TP}}{(\text{TP} + \text{FP})} \times 100\%$$

$$\text{Sensitivity } = \frac{\text{TP}}{(\text{TP} + \text{FN})} \times 100\%$$

in which TP = true positive rate, FP = false positive rate, and FN = false negative rate.

Sensitivity was used to select the best model for classifying susceptible animals. Sensitivity is important when it is concerned with identifying positive outcomes (parasitized animals) and the cost of a false positive is low (animals classified as susceptible when they are not), as long as the model identifies as many actual positives as possible. Precision was used to select the best models to classify resistant and resilient animals. Precision looks at the ratio of true positives to the predicted positives. This metric is most often used when there is a high cost for having false positives (in this case, classification of resistant or resilient animals when they indeed are not).

In addition, the accuracy and AUC of the models studied were calculated. The AUC corresponds to the area under the ROC curve, whose maximum value is 1.0, indicating that the model is 100% sensitive and 100% specific [31]. The AUC was obtained using the auc function with multcap object of the R program (R Development Core Team, 2018).

2.5. Extrapolating Classification across Farms

After finding the best classification model, a leave-one-farm-out cross-validation technique was performed by excluding all the data from a target farm in the training set and using it in the testing set. We did this for exploring the classification performance of the models to predict animal classes in a specific farm that was not used to build the model. Therefore, out of the six farms in the dataset, five farms were included in the model training, whereas the data from one farm was held out as testing set in each run of the cross-validation, and the process was then repeated for each farm.

3. Results

3.1. Prediction Models

The different classes were not equally represented in the data, with the percentage of animals classified as resistant, resilient, and susceptible being 71.3%, 25.7%, and 3.0%, respectively. After balancing the classes by the random over-sampling method, we tested four different models to classify sheep according to their level of gastrointestinal parasite infection.

Among the four models tested, both MLR and LDA provided the best performances in predicting susceptible animals, as indicated by the high values of sensitivity (75.5% and 76.7%, respectively) compared to the RF and ANN models (Figure 1). The poorest performance for this infection response class was achieved with the RF, which identified only 50.7% of the susceptible animals (Figure 1). The precision of the MLR was 82.6% for resistant animals, whereas this metric ranged from 36.6 to 42.4% for resilient animals, suggesting poor performance for this specific class (Figure 1).

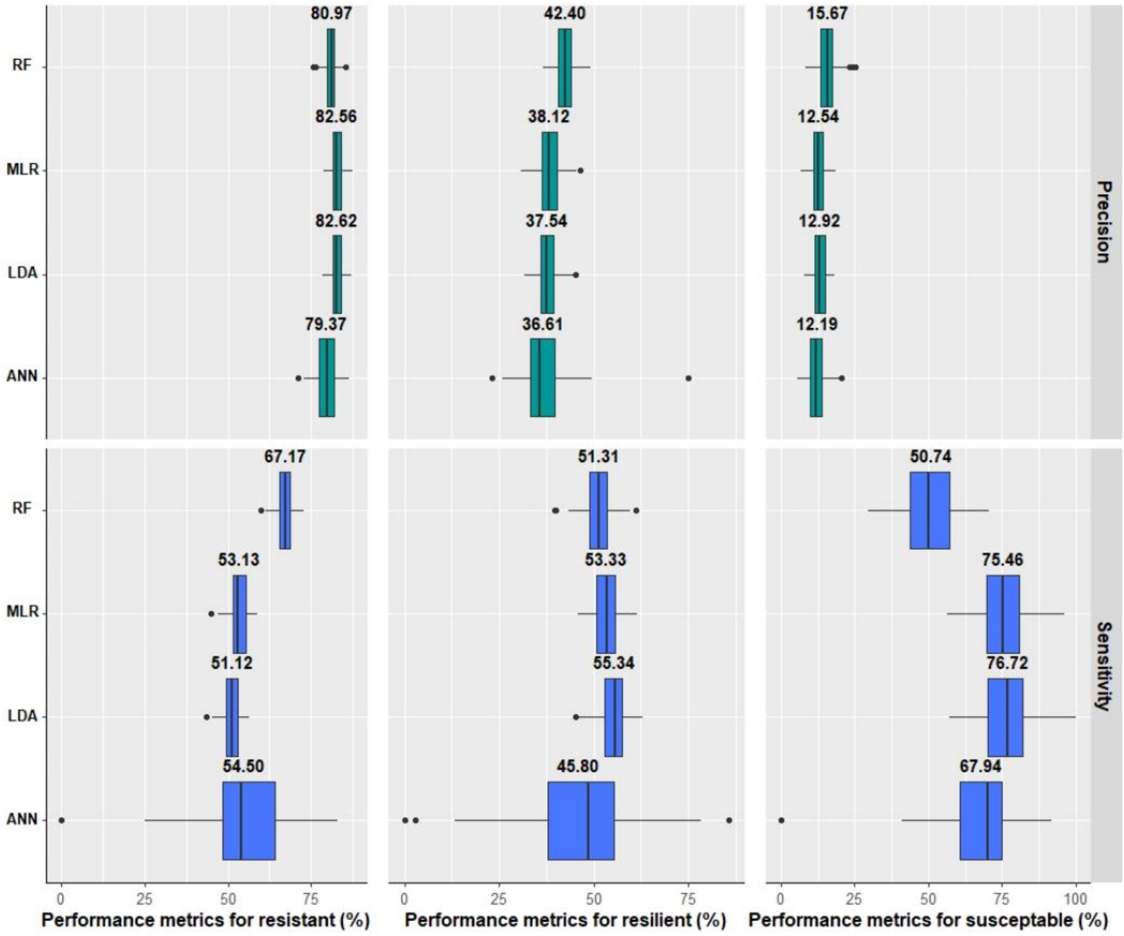

Figure 1. Precision and sensitivity for the classification of parasitic infection response (resistant, resilient, and susceptible) in sheep according to different methods (random forest—RF, multinomial logistic regression—MLR, linear discriminant analysis—LDA, and artificial neural network—ANN) trained with easy-to-measure explanatory variables. Performance metrics were computed using a 5-fold cross-validation with 20 random replicates. The left and right of the box represent first and third quartiles, respectively; the vertical line denotes the median; the value above the box is the mean; the whiskers correspond to 1.5 × interquartile distance; and the dark dots are outliers.

The mean accuracy ranged between 59% (ANN) and 63% (MLR and RF) for resistant animals, between 58% (ANN) and 63% (RF) for resilient animals, and between 71% (RF) and 80% (MLR and LDA) for susceptible animals (Table 1). In addition to the performance metrics (sensitivity and precision) and accuracy, the AUC was also used to assess the performance of the tested models (Table 1). The best performance (AUC = 0.78) was obtained for the MLR and LDA.

3.2. Classification Performance of Leave-One-Farm-Out Cross-Validation Technique

According to the results, the MLR was one of the most suitable methods for classifying resistant and susceptible animals (Figure 1). We further investigated the performance of this model across farms by alternating the data from one specific farm as the testing set and the remaining as training set. The sensitivity for the MLR model considering different farms ranged from 20 to 100% for the susceptible class, presenting better results in farms A (89.3%), D (100.0%), and F (83.0%) (Figure 2). While the precision for resistance ranged from 60.0 to 85.8% with the best results for farm B (71.0%), farm C (78.8%), farm E (85.8%), and farm F (77.0%) (Figure 2), the identification of resilient animals showed low precision in all farms, ranging from 0 (Farm E) to 53.3% (Farm D).

Table 1. Mean and standard deviation of the performance metrics of the multinomial logistic regression (MLR), linear discriminant analysis (LDA), random forest (RF), and artificial neural network (ANN) methods using sex, age class, month of record, body weight, body condition score, and Famacha© degree as explanatory variables and response classes of the parasitic infection (resistant, resilient, and susceptible) as the response variable.

Performance Metrics		Model [2]			
		MLR	LDA	RF	ANN
AUC [1]		0.78 (0.02)	0.78 (0.02)	0.77 (0.02)	0.73 (0.04)
Accuracy	Resistant	0.63 (0.02)	0.62 (0.02)	0.63 (0.02)	0.59 (0.03)
	Resilient	0.62 (0.02)	0.62 (0.02)	0.63 (0.02)	0.58(0.03)
	Susceptible	0.80 (0.04)	0.80 (0.04)	0.71 (0.05)	0.76 (0.05)

[1] AUC = area under the ROC curve. [2] Multinomial logistic regression (MLR), linear discriminant analysis (LDA), random forest (RF), and artificial neural network (ANN) methods. The value of AUC for a multiple class response by averaging pairwise comparisons is the mean of 5 repeats of cross-validation with 20 random replicates. The value of accuracy is the mean of 5 repeats of cross-validation with 20 random replicates.

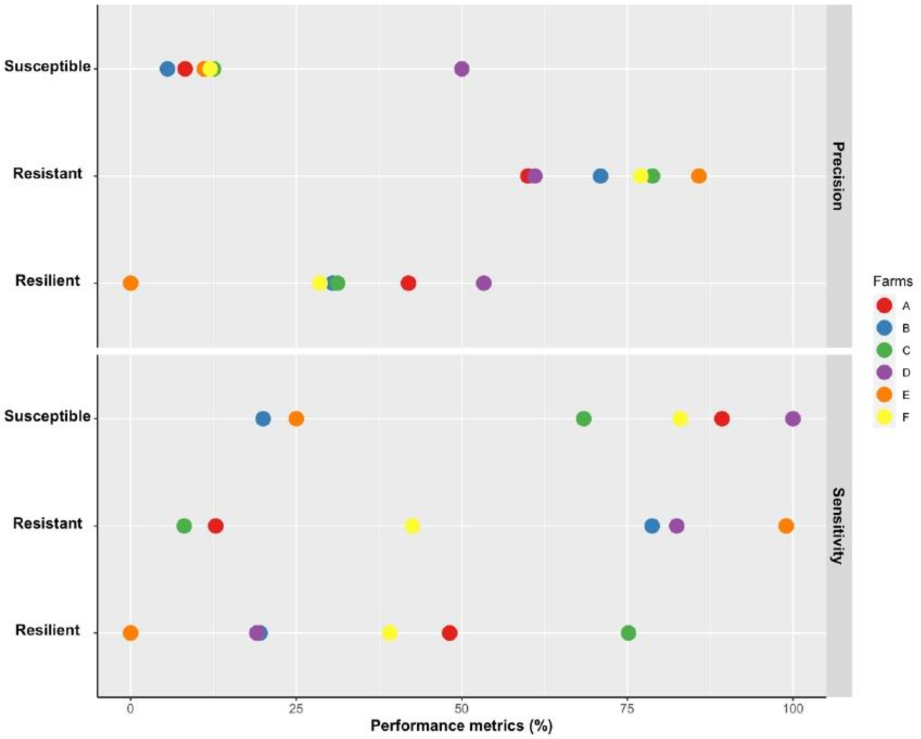

Figure 2. Precision and sensitivity of multinomial logistic regression model using sex, age class, month of record, body weight, body condition score, and Famacha© degree as explanatory variables and response classes of the parasitic infection (resistant, resilient, and susceptible) as the response variable for each farm. Performance metrics were calculated leave-one-farm-out cross-validation technique, alternating the data from one specific farm as the testing set and the remaining as training sets.

4. Discussion

In this study, sensitivity was used to select the best model to classify susceptible animals. This metric indicates how good a given model is at predicting the sick animals (susceptible animals) when the cost of a false positive is low (animals classified as susceptible when they are not) as long as the model identifies as many parasitized animals as possible. The cost of a high rate of false positives becomes low considering that the farmer treats all animals identified as susceptible. As treatment of all animals in the herd is a common management in many sheep farms, and false positives in this case are not a big issue in the production system.

Precision was used to select the best models to classify resistant and resilient animals. This metric is most often used when there is a high cost for having false positives (in this case, the classification of susceptible animals as resistant or resilient). A high precision rate for the classification of resistant animals reduces the risk of not treating real susceptible animals that are incorrectly identified as resistant or resilient, contributing to increased contamination of pastures and a higher cost with treatments.

Poor precision values were observed for the minority class (susceptible animals), and this was mostly because this metric is sensitive to the class distribution. Additionally, by performing a simple random oversampling for a minority class with too few datapoints, all observations and explanatory variables tend to appear repeatedly in the training data. It happens because this method just samples with replacement from the original data, which can contribute to an overfitting during the training process.

Therefore, considering the criteria used to compare the classification performance across models, the most suitable model combines higher sensitivity for susceptible animals and the highest precision for resistant and resilient animals. This specific criterion was chosen by considering the balance among productivity, cost reduction, and treating the parasitized animals in the herd in order to reduce the contamination of pasture and other animals while increasing the profitability of the production system and making it more sustainable. It was found that among the four models tested, the MLR model presented one of the best classification performances for the susceptible and resistant animals. The precision for resilient animals was low in all models, so it was not possible to correctly identify the best model for this class. Still, the precision average for the investigated models was higher than the expected random guess threshold for resilient class (roughly 26%), suggesting that the models are capturing relevant information from the data.

The MLR and LDA models presented higher AUC (0.78). According to Hosmer and Lemeshow [32], an AUC of 0.7 to 0.8 is considered acceptable. Thus, considering a model that combines higher sensitivity for susceptible animals and higher precision for resistant and resilient animals and presents an acceptable AUC, MLR and LDA can be classified as good for predicting classes of parasitic infection.

Excluding animals or part of the records for the same animal from the training data set are strategies generally employed for validating the model predictive ability [33,34]. Shetty et al. [34] reported that when multiple observations related to the same cow are only partially omitted from the training set, the model predictions tend to be more optimistic. Dorea et al. [33] investigated this topic hypothesizing that even when excluding an animal from the training data set, conditions related to external factors such as weather, diet, season, management, and others would inflate the prediction quality of other animals in the same circumstances. These authors concluded that inflation in the prediction ability occurred when animals from the same trial were kept in both the training and validation datasets. In the present study, we investigated this issue by hypothesizing that conditions related to the farm would inflate the prediction quality assessment. Clearly, inflation in the classification performance occurred when sheep from all farms were split randomly and the repeated measures of the same animal were present in both training and validation datasets. Our approach to validate the models by excluding an entire farm from the training dataset is closer to reality because the developed models are supposed to be used to predict new datasets coming from farms with different external factors, such as diets, weather, and management practices.

The applicability of the present study in sheep farming was validated by analyzing the classification performance by farm. This approach revealed that the MLR model was able to predict with good performance both classes (resistant and susceptible) in two farms (C and F) and at least one of the classes in four farms (susceptible class in Farms A and D or resistant class in Farms B and E). The proposed approach could help attenuating the negative impacts related to infections caused by gastrointestinal nematodes, contributing to design deworming strategies that take into account the risk of an animal being contaminated, consequently reducing the costs with anthelmintic administration and laboratory analyses based on blood or fecal samples.

The identification of resistant and susceptible animals predicted through easily measurable, non-invasive, and cost-reduced variables can be used to support animal/herd management decisions in which animals identified as resistant to gastrointestinal nematodes may be incorporated as selection candidates in breeding programs, whereas animals classified as susceptible can receive adequate treatment.

Easy-to-measure variables were used in the models due to their association with gastrointestinal parasite infection. Several studies reported that gastrointestinal nematode infections had a negative effect on meat production and that performance of sheep infected with nematodes was 85% of the performance in uninfected individuals for weight gain [1]. According to Cornelius et al. [35], body condition score showed promise as a selection index under commercial farming conditions. One survey showed that body condition score was the best criterion to detect infected sheep, with only 1.1% of false negatives [36]. Another extremely useful tool for identifying parasitized animals through the diagnosis of anemia is the Famacha© method [28]. Sheep infected with blood-sucking *H. contortus* may show anemia, eosinophilia, and hypoproteinemia, resulting in pale mucous membranes and submandibular edema [5], which seems to be the case in the present study considering the higher prevalence of *H. contortus* (63%) in the herd. Additionally, procedures requiring blood sampling and laboratory services of individuals is more expensive in terms of time and resources, which makes them less attractive for monitoring purposes. Therefore, cheaper alternatives must be studied to reduce the cost of production.

The use of models with easy-to-measure phenotypic traits for identification with high sensitivity of the susceptible animals allows treatment of only those animals instead of the whole herd. This brings many advantages to the farmer, such as reducing the cost with laboratory resources and with the use of anthelminthics, considered an expensive drug and, in most cases, only partially effective [37]. Furthermore, overuse and frequent use of anthelmintics have resulted in substantial and widespread problems with anthelmintic resistance in nematode populations [38,39]. Another advantage is the reduction in the number of agricultural residues deposited in meat and soil, increasing the environmental sustainability of the farm [40]. Therefore, there are large economic and environmental gains by increasing the control of the main parasitic diseases in sheep.

Overall, the use of prediction algorithms in sheep farming is still incipient. There are no reports in the literature on the use of machine learning models for classifying resistance, resilience, and susceptibility to gastrointestinal nematodes in sheep. This reinforces the importance of the present work, which certainly opens the way for further studies and future applicability at the field level.

5. Conclusions

Multinomial logistic regression and linear discriminant analysis achieved the best performances for classifying susceptible and resistant animals. The results suggest that the use of readily available records and easy-to-measure variables such as body weight, body condition score, farm, sex, age class, record month, and Famacha© degree may contribute to the identification of susceptible animals, supporting management decisions at the farm level and potentially reducing the economic losses due to parasitic infection with higher prevalence of *H. contortus* parasites. The animals identified as resistant can also be incorporated as selection candidates into breeding programs for genetic improvement of sheep populations.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ani13030374/s1, Table S1: Number of records for each farm and each class of parasitic infection (resistant, resilient, and susceptible).

Author Contributions: Conceptualization, L.A.F.; methodology, L.A.F., R.P.S., A.A.C.A. and G.J.M.R.; formal analysis, L.A.F.; investigation, A.A.C.A. and G.J.M.R.; data curation, C.C.P.P.; writing—original draft preparation, L.A.F.; writing—review and editing, R.P.S., A.A.C.A., R.L.D.C., D.P.M., N.B.S., G.J.M.R. and C.C.P.P.; visualization, R.L.D.C., D.P.M., N.B.S., G.J.M.R. and C.C.P.P.; supervision, G.J.M.R. and C.C.P.P.; project administration, C.C.P.P.; funding acquisition, C.C.P.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by São Paulo Research Foundation (FAPESP), grant numbers #2012/15982-0, #2016/14522-7 and #2017/50339-5. The authors also thank the FAPESP for the fellowships granted to L.A.F. (#2018/01540-2 and #2020/03575-8). C.C.P.P., N.B.S., and D.P.M. received a productivity fellowship from CNPq. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001.

Institutional Review Board Statement: The animal study protocol was approved by the Institutional Animal Care and Use Committee of the Animal Science Institute, Nova Odessa, São Paulo, Brazil (protocol code CEUA N°. 267-18, 3 October 2018).

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions, and the data that support the findings of this study are available in the supplementary material of this article.

Acknowledgments: The authors would like to thank all researchers and students that contributed to phenotype collection.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mavrot, F.; Hertzberg, H.; Torgerson, P. Effect of gastro-intestinal nematode infection on sheep performance: A systematic review and meta-analysis. *Parasites Vectors* **2015**, *8*, 557. [CrossRef] [PubMed]
2. Méndez-Ortíz, F.A.; Sandoval-Castro, C.A.; Vargas-Magaña, J.J.; Sarmiento-Franco, L.; Torres-Acosta, J.F.J.; Ventura-Cordero, J. Impact of gastrointestinal parasitism on dry matter intake and live weight gain of lambs: A meta-analysis to estimate the metabolic cost of gastrointestinal nematodes. *Vet. Parasitol.* **2019**, *265*, 1–6. [CrossRef] [PubMed]
3. Torres-Acosta, J.F.; Hoste, H.; Sandoval-Castro, C.A.; Torres-Fajardo, R.A.; Ventura-Cordero1, J.; González-Pech, P.G.; Mancilla-Montelongo, M.G.; Ojeda-Robertos, N.F.; Martínez-Ortíz-de-Montellano, C. The "Art of War" against gastrointestinal nematodes in sheep and goat herds of the tropics. *Rev. Acad. Ciênc. Anim.* **2019**, *17*, 39–46.
4. Chagas, A.C.S.; Tupy, O.; Santos, I.B.; Esteves, S.N. Economic impact of gastrointestinal nematodes in Morada Nova sheep in Brazil. *Braz. J. Vet. Parasitol.* **2022**, *31*, e008722. [CrossRef]
5. Adduci, I.; Sajovitz, F.; Hinney, B.; Lichtmannsperger, K.; Joachim, A.; Wittek, T.; Yan, S. Haemonchosis in Sheep and Goats, Control Strategies and Development of Vaccines against Haemonchus contortus. *Animals* **2022**, *12*, 2339. [CrossRef]
6. Szewc, M.; Waal, T.; Zintl, A. Biological methods for the control of gastrointestinal nematodes. *Vet. J.* **2021**, *268*, 105602. [CrossRef]
7. Charlier, J.; Rinaldi, L.; Musella, V.; Ploeger, H.W.; Chartier, C.; Rose Vineer, H.; Hinney, B.; Von Samson-Himmelstjerna, G.; Băcescu, B.; Mickiewicz, M.; et al. Initial assessment of the economic burden of major parasitic helminth infections to the ruminant livestock industry in Europe. *Prev. Vet. Med.* **2020**, *182*, 300–304. [CrossRef]
8. Alba-Hurtado, F.; Muñoz-Guzmán, M.A. Immune Responses Associated with Resistance to Haemonchosis in Sheep. *BioMed Res. Int.* **2013**, *2013*, 162158. [CrossRef]
9. Palomo-Couoh, J.G.; Aguilar-Caballero, A.J.; Torres-Acosta, J.F.; Magaña-Monforte, J.G. Evaluation of different models to segregate Pelibuey and Katahdin ewes into resistant or susceptible to gastrointestinal nematodes. *Trop. Anim. Health Prod.* **2016**, *48*, 1517–1524. [CrossRef]
10. Amarante, A.F.; Bricarello, P.A.; Rocha, R.A.; Gennari, S.M. Resistance of Santa Ines, Suffolk and Ile de France sheep to naturally acquired gastrointestinal nematode infections. *Vet. Parasitol.* **2004**, *120*, 91–106. [CrossRef]

11. Woolaston, R.R.; Baker, R.L. Prospects of breeding small ruminants for resistance to internal parasites. *Int. J. Parasitol.* **1996**, *26*, 845–855. [CrossRef] [PubMed]
12. Albers, G.A.A.; Gray, G.D.; Piper, L.R.; Barker, J.S.F.; Jambre, L.F.L.; Barger, I.A. The genetics of resistance and resilience to Haemonchus contortus infection in young Merino sheep. *Int. J. Parasitol.* **1987**, *17*, 1355–1363. [CrossRef] [PubMed]
13. Bishop, S.C. A consideration of resistance and tolerance for ruminant nematode infections. *Front. Genet.* **2012**, *3*, 62–68. [CrossRef] [PubMed]
14. Pollot, G.E.; Karlsson, L.J.E.; Eady, S.; Greeff, J.C. Genetic parameters for indicators of host resistance to parasites from weaning to hogget age in Merino sheep. *J. Anim. Sci.* **2004**, *82*, 2852–2864. [CrossRef]
15. McManus, C.; Paim, T.P.; Melo, C.B.; Brasil, B.S.A.F.; Paiva, S.R. Selection methods for resistance to and tolerance of helminths in livestock. *Parasite* **2014**, *21*, 56. [CrossRef]
16. Marques, C.A.T.; Saraiva, L.A.; Torreao, J.N.C.; Silva, T.P.D.; Bezerra, L.R.; Edvan, R.L.; Marcos, J.A.; Romilda, R.N. The use of targeted selective treatments on controlling gastrointestinal nematodes in different sheep categories under grazing system. *Pesq. Vet. Bras.* **2018**, *38*, 470–476. [CrossRef]
17. Goldstein, B.A.; Navar, A.M.; Carter, R.E. Moving beyond regression techniques in cardiovascular risk prediction: Applying machine learning to address analytic challenges. *Eur. Heart J.* **2017**, *38*, 1805–1814. [CrossRef]
18. Mullainathan, S.; Spiess, J. Machine learning: An applied econometric approach. *J. Econ. Perspect.* **2017**, *31*, 87–106. [CrossRef]
19. Andrade, V.V.; Bernardes, P.A.; Vicentini, R.R.; Oliveira, A.P.; Veroneze, R.; Ujita, A.; Negrão, J.A.; El Faro, L. Estrus prediction models for dairy gyr heifers. *Animals* **2021**, *11*, 3103. [CrossRef]
20. Bollig, N.; Clarke, L.; Elsmo, E.; Craven, M. Machine learning for syndromic surveillance using veterinary necropsy reports. *PLoS ONE* **2020**, *15*, e0228105. [CrossRef]
21. Moawed, S.A.; Osman, M. The Robustness of Binary Logistic Regression and Linear Discriminant Analysis for the Classification and Differentiation between Dairy Cows and Buffaloes. *J. Stat. Appl. Probab.* **2017**, *7*, 304–310. [CrossRef]
22. Gordon, H.M.; Whitlock, H.V. A new technique for counting nematode eggs in sheep feces. *J. Sci. Ind. Res.* **1939**, *12*, 50–52.
23. Roberts, F.H.S.; O'Sullivan, J.P. Methods for egg counts and larval cultures for Strongyles infesting the gastrointestinal tract cattle. *Aust. J. Agric. Res.* **1950**, *1*, 99–102. [CrossRef]
24. Oliveira, E.J.; Savegnago, R.P.; Freitas, L.A.; Freitas, A.P.; Maia, S.R.; Simili, F.F.; EL Faro, L.; Costa, R.L.D.; Santana, M.L.J.; Paz, C.C.P. Estimates of genetic parameters and cluster analysis for worm resistance and resilience in Santa Ines meat sheep. *Pesq. Agropec. Bras.* **2018**, *53*, 1338–1345. [CrossRef]
25. Schalm, O.W.; Jain, N.C.; Carrol, E.J. *Veterinary Haematology*, 3rd ed.; Lea and Febiger Publication: Philadelphia, PA, USA, 1975; pp. 197–199.
26. Thompson, J.M.; Meyer, H.H. *Body Condition Scoring of Sheep*; Oregon State University, Extension Service: Corvallis, OR, USA, 1994.
27. Oliveira, E.J.; Savegnago, R.P.; Freitas, A.P.; Freitas, L.A.; Paz, A.C.A.R.; El Faro, L.; Simili, F.F.; Vercesi Filho, A.E.; Costa, R.L.D.; Paz, C.C.P. Genetic parameters for body weight and morphometric traits in Santa Ines sheep using Bayesian inference. *Small Rum. Res.* **2021**, *201*, 106446. [CrossRef]
28. Van Wyk, J.A.; Bath, G.F. The FAMACHA system for managing haemonchosis in sheep and goats by clinically identifying individual animals for treatment. *Vet. Res.* **2002**, *33*, 509–529. [CrossRef]
29. Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 18–22.
30. Venables, W.N.; Ripley, B.D. *Modern Applied Statistics with S*, 4th ed.; Springer: New York, NY, USA, 2002; pp. 271–300. [CrossRef]
31. Fan, J.; Upadhye, S.; Worster, A. Understanding receiver operating characteristic (ROC) curves. *Can. J. Emerg. Med.* **2006**, *8*, 19–20. [CrossRef]
32. Hosmer, D.W.; Lemeshow, S. *Applied Logistic Regression*, 2nd ed.; John Wiley and Sons: New York, NY, USA, 2000; pp. 160–164. [CrossRef]
33. Dorea, J.R.R.; Rosa, G.J.M.; Weld, K.; Armentano, L. Mining data from milk infrared spectroscopy to improve feed intake predictions in lactating dairy cows. *Int. J. Dairy Sci.* **2018**, *101*, 5878–5889. [CrossRef]
34. Shetty, N.; Løvendahl, P.; Lund, M.; Buitenhuis, A. Prediction and validation of residual feed intake and dry matter intake in Danish lactating dairy cows using mid-infrared spectroscopy of milk. *J. Dairy Sci.* **2017**, *100*, 253–264. [CrossRef]
35. Cornelius, M.P.; Jacobson, C.; Besier, R.B. Body condition score as a selection tool for targeted selective treatment-based nematode control strategies in Merino ewes. *Vet. Parasitol.* **2014**, *206*, 173–181. [CrossRef] [PubMed]
36. Soto-Barrientos, N.; Chan-Pérez, J.I.; España-España, E.; Novelo-Chi, L.; Palma-Ávila, I.; Ceballos-Mendoza, A.; Sarabia-Hernández, J.A.; Santos-Ricalde, R.H.; Cámara-Sarmiento, R.; Torres-Acosta, J.F. Comparing body condition score and FAMACHA© to identify hair-sheep ewes with high faecal egg counts of gastrointestinal nematodes in farms under hot tropical conditions. *Small Rumin. Res.* **2018**, *167*, 92–99. [CrossRef]
37. Lambertz, C.; Poulopoulou, I.; Wuthijaree, K.; Gauly, M. Anthelmintic resistance in gastrointestinal nematodes in sheep raised under mountain farming conditions in Northern Italy. *Vet. Rec. Open* **2019**, *6*, e000332. [CrossRef] [PubMed]
38. Burke, J.M.; Miller, J.E. Sustainable approaches to parasite control in ruminant livestock. *Vet. Clin. N. Am. Food Anim. Pract.* **2020**, *36*, 89–107. [CrossRef]

39. Hoglund, J.; Baltrusis, P.; Enweji, N.; Gustafsson, K. Signs of multiple anthelmintic resistance in sheep gastrointestinal nematodes in Sweden. *Vet. Parasitol. Reg. Stud. Rep.* **2022**, *36*, 100789. [CrossRef]
40. Dos Santos, I.B.; Anholeto, L.A.; de Sousa, G.A.; da Silva Nucci, A.; Gainza, Y.A.; Figueiredo, A.; Dos Santos, L.A.L.; Minho, A.P.; Barioni-Junior, W.; Esteves, S.N.; et al. Investigating the benefits of targeted selective treatment according to average daily weight gain against gastrointestinal nematodes in Morada Nova lambs. *Parasitol. Res.* **2022**, *121*, 2433–2444. [CrossRef]

Supplementary Materials

Table S1. Number of records for each farm and each class of the parasitic infection (resistant, resilient and susceptible).

Class of the parasitic infection	Farm					
	A	B	C	D	E	F
Resistant	397	301	322	57	684	844
Resilient	243	128	145	42	110	271
Susceptible	28	5	19	7	4	47

CHAPTER 3. IMAGE ANALYSIS TO AUTOMATICALLY CLASSIFY ANEMIA BASED ON FAMACHA© SCORE IN SHEEP INFECTED WITH *HAEMONCHUS CONTORTUS* USING OCULAR CONJUNCTIVA IMAGES

Abstract: *Haemonchus contortus* is the most pathogenic blood-feeding parasitic in sheep, responsible to cause anemia and consequently changes in the color of the ocular conjunctiva, from the deep red of healthy sheep to shades of pink to practically white of non-healthy sheep. In this context, the Famacha© method has been created for detecting sheep unable to cope with the infection by *H. contortus*, through visual assessment of ocular conjunctiva coloration. Thus, the objectives of this study were (1) to extract ocular conjunctiva image features to automatically classify Famacha© score and compare two classification models (multinomial logistic regression - MLR and random forest - RF), and (2) to evaluate the applicability of the best classification model on three sheep farms. The dataset consisted of 1,156 ocular conjunctiva images from 422 animals. RF model was used to segment the images, i.e. to select the pixels that belong to the ocular conjunctiva. After segmentation, the quantiles (1, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 99%) of color intensity in each image channel (red, blue, and green) were determined and used as explanatory variables in the classification models, and the Famacha© scores 1 (non-anemic) to 5 (severely anemic) were the target classes to be predicted (1= 162 images, 2= 255 images, 3= 443 images, 4= 266 images, 5= 30 images). For objective 1, the performance metrics (precision and sensitivity) were obtained using MLR and RF models considering data from all farms randomly split. For objective 2, a leave-one-farm-out cross-validation technique was used to assess prediction quality across three farms (A= 726 images, B= 205 images, and C= 225 images). The RF provided the best performances in predicting anemic animals, as indicated by the high values of sensitivity for Famacha© score 3 (80.9%), 4 (46.2%), and 5 (60%) compared to the MLR model. The precision of the RF was 72.7% for Famacha© score 1 and 62.5% for Famacha© score 2. These results indicate that is possible to successfully predict Famacha© score, especially for scores 2 to 4, in sheep via image analysis and RF model using ocular conjunctiva images collected in farm conditions. As expected, model validation excluding entire farms in cross-validation presented a lower prediction quality. Nonetheless, this setup is closer to reality because the developed models are supposed to be used across farms, including new ones, and with different environments and management conditions.

Keywords: gastrointestinal nematodes, machine learning, *Ovis aries*, random forest

1. INTRODUCTION

Gastrointestinal nematode infection represents a major threat to the health, welfare, and productivity of sheep populations worldwide (Mcrae et al., 2015). *Haemonchus contortus* is one of the most common and economically significant blood-feeding parasites of small ruminants, responsible for causing anemia and consequently changes in the color of the ocular conjunctiva, from the deep red of healthy sheep to shades of pink to practically white of non-healthy sheep (Singh and Swarnkar, 2012; El-Ashram et al., 2017). The most commonly used methods to control this gastrointestinal parasite are based on the use of anthelmintics (Szewc et al., 2021). However, parasite resistance to anthelmintics has escalated to such an extent that new strategies are urgently required to address this problem (Kuiseu et al., 2021).

Strategies that are not based on the chemical treatment of entire herds or expensive laboratory tests (blood or fecal) to identify parasitized animals have been receiving special attention. In this context, the Famacha© method has been widely used for detecting sheep unable to cope with the infection by *H. contortus*, through visual assessment of ocular conjunctiva coloration (Oliveira et al., 2018; Freitas et al., 2022; Freitas et al., 2023). It is the easiest and cheapest method to identify parasitized animals with *H. contortus* as Famacha©

only requires trained professionals. From a productive standpoint, this is a very important ability because it enables real-time diagnosis of parasitic disease.

On the other hand, this method is error-prone, since it requires a subjective evaluation from the professionals performing the analysis. This can lead to misdiagnoses, resulting in parasitized animals not receiving treatment, or non-parasitized animals being unnecessarily treated. Moreover, visual evaluation by trained professionals is labor-intensive and difficult to scale up in large production systems, highlighting the importance of developing high-throughput measurements of economically important traits, such as diagnostic by image analyses.

Due to the evolution of technology in image capture and the growing demand for more accurate and quickly accessible information, several computer vision techniques have been developed recently (Szeliski, 2010). Digital image processing techniques are characterized by their attempts to reproduce, using statistical and mathematical methods, the recognition and interpretation mechanisms of the human vision system. For this recognition, it is necessary to extract information from the images, and to facilitate the extraction of this information, computer vision employs a set of machine learning techniques (Lee et al., 2015), including random forest and artificial neural networks. In this context, image analysis can be of great help for more accurate identification of parasitized animals using images of the ocular conjunctiva, eliminating human subjectivity in assessment. Thus, the objectives of this study were (1) to extract ocular conjunctiva image features to automatically classify anemia based on Famacha© scores and compare two classification models (multinomial logistic regression – MLR and random forest – RF), and (2) to evaluate the applicability of the best classification model on three sheep farms.

2. MATERIAL AND METHODS

2.1 Animals and data acquisition

The dataset consisted of 422 sheep and 1,156 ocular conjunctiva images and records of Famacha© score collected in three farms located in São Paulo State, Brazil: Farm A (Jardinopolis, latitude: 21°01′04″ S; longitude: 47°45′50″ W), Farm B (Cravinhos, latitude: 21°20′25″ S; longitude: 47°43′46″ W) and Farm C (Nova Odessa, latitude: 22°46′39″ S, longitude: 47°17′45″ W) (Table 1). The experimental procedures were conducted following the recommendations of the Institutional Animal Care and Use Committee of the Animal Science Institute, Nova Odessa, São Paulo, Brazil (protocol code CEUA Nº. 267-18, 3rd October 2018).

The Famacha© score was measured using Famacha© card (Figure 1) to compare different color shades of the ocular conjunctiva with a standard scale from 1 to 5: 1 (non-anemic, n= 162 images), 2 (non-anemic, n= 255 images), 3 (mildly anemic, n= 443 images), 4 (anemic, n= 266 images), and 5 (severely anemic, n= 30 images), which correspond to red, pink-red, pink, pink-white and white, respectively (Van Wyk and Bath, 2002).

A pool of feces samples was collected directly from the rectal ampoule and separated for the preparation of larvae culture to establish the genera of nematodes prevalent in the herd (Roberts and O'Sullivan, 1950). The results showed that the main genus of gastrointestinal nematode found during the study was *Haemonchus* (63%) followed by *Trichostrongylus* (24%), *Cooperia* (7%), and *Oesophagostomum* (6%) (Oliveira et al., 2018; Freitas et al., 2023).

The images from the ocular conjunctiva were acquired using an Apple smartphone camera (model iPhone 8, digital 12 megapixels). The data acquisition was performed over six months, from January to July 2019, and Famacha© score was measured by the same team of trained professionals.

2.2 Trainable segmentation using local features and random forest model

Local features refer to a pattern or distinct structure found in an image. They are usually associated with an image patch that differs from its immediate surroundings by texture, color, or intensity (Crommelinck et al., 2016). A pixel-based segmentation was computed using local features based on color intensity and textures at different scales, extracted using the feature.multiscale_basic_features function from the scikit-image library implemented in Python. The automated segmentation method was developed and trained based on an image of the animal's right face with exposure of the ocular conjunctiva inside the barn and collected during the day with natural lighting (Figure 2A). In the training image, we manually created different boxes of identification for each component of the image (Figure 2B). Four regions were segmented: ocular conjunctiva, hands or gloves, hair, and eye, and the ocular conjunctival region was selected and captured through a mask (Figure 2C). The pixels contained in the mask were used to train a random forest classifier using the scikit-learn library in Python. Once the classifier was trained using the user-selected feature data, the trained classifier was applied to other images (Figure 2E).

From the pixels predicted as ocular conjunctiva using the trained pixel classifier for each new image, we included two conditions to select the ocular conjunctiva region with more accuracy. The first condition was the size of the ocular conjunctiva, with a threshold between 500 and 50,000 pixels, and the second condition was that the correct region should be the one closest to the center of the image, ensuring that no other region would be identified as the ocular conjunctiva. After segmentation, 768 color intensities were extracted from the image channels (red =256, blue =256, and green =256).

2.3 Segmentation evaluation with Intersection over Union

After the segmentation process, the segmentation evaluation was performed using the

Intersection over Union (IoU) metric, which measures the ratio of intersection area over the union area of the ground truth region and the segmented region. The identification of ground truth region was performed using the Draw ROIs tool, available for Matlab program (MathWorks Inc.). We calculate the IoU metric for 65 random images between the prediction and ground truth (Figure 3). The IoU metric can be described as:

$$\text{IoU} = \frac{|Reference \cap Prediction|}{|Reference \cup Prediction|}$$

where reference = ground truth (real area of the ocular conjunctiva), and prediction = segmented results.

2.4 Canonical correlation analysis

To reduce the dimension of the color intensity traits extracted from the ocular conjunctiva region (256 color intensities for each channel, blue, green, and red) during the segmentation process, we determined the quantiles (1, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 99%) of color intensity in each channel (red, blue, and green) and we performed canonical correlation analysis (CCA) to find the linear relationship between the intensity of color extracted in each image and quantiles of the intensity of the color in each channel, aiming to determine if the selected quantiles could be used as an indicator to represent the original color intensity of each image channel.

2.5 Classification models

Multinomial logistic regression (MLR) and random forest (RF) models were used to classify the animals according to Famacha© score. The hyperparameter values were set to the default values available in the library used for the MLR. For RF, we performed tuning analyses to find the best hyperparameters for the model. For the tuning of the RF model, 80% of the total dataset was used in a 5-fold cross-validation strategy with 10 random replicates. The best architectures for the RF model had 1 mtry (number of features to consider at each split point) and 500 trees with an accuracy of 0.51. After finding the best hyperparameters for RF, we performed MLR and RF analyses with a full tuning set (80% of the dataset) to train the models, whereas 20% excluded from the dataset from the tuning analysis was used to test the model's predictive ability.

The classification methods were fitted using the information of quantiles for blue, green, and red channels (1, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 99%) as explanatory variables, and the Famacha© scores (1 to 5) as the target classes to be predicted. The metrics used for comparing the models' performance were precision and sensitivity. The classification metrics were calculated as follows:

$$Precision\ = \frac{TP}{(TP + FP)}\ x\ 100\%$$

$$\text{Sensitivity}\ = \frac{TP}{(TP + FN)}\ x\ 100\%$$

in which TP = number of true positives, FP = number of false positives, and FN = number of false negatives.

Sensitivity was used to select the best model for classifying anemic animals that presented Famacha© scores from 3 to 5. Sensitivity is important when it is concerned with identifying positive outcomes (anemic animals) and the cost of a false positive is low (animals classified as anemic when they are not), as long as the model identifies as many actual positives as possible. Precision was used to select the best models to classify non-anemic animals, with Famacha© scores from 1 or 2. Precision looks at the ratio of true positives to predicted positives. This metric is most often used when there is a high cost for having false positives (in this case, classification of non-anemic animals when they are not indeed).

The analyses were performed through the R software (R Core Team, 2018), using the mlbench, caret, and e1071 packages for tuning analyses, randomForest package for RF analyses

(Liaw and Wiener, 2002), and nnet package for MLR analyses (Venables and Ripley, 2002).

2.6 Extrapolating classification for farms

We performed classification analyses using the leave-one-farm-out cross-validation technique for three farms (Farm A = 726 images, Farm B = 205 images, and Farm C = 225 images). Before performing the classification analyses, the tuning analyses of the RF model were performed using images from two farms in a 5-fold cross-validation strategy with 10 random replicates. Namely, the RF model was tuned three times, excluding a different farm at each round of tuning. The best architectures for the RF model were 1 mtry and 1,500 trees with an accuracy of 0.49 for Farm A, 7 mtry and 1,500 trees with an accuracy of 0.57 for Farm B, and 6 mtry and 1,000 trees with an accuracy of 0.62 for Farm C.

After finding the best hyperparameters for RF, we performed the analysis with a full tuning set (2 farms) to train the models, whereas one farm excluded from the tuning dataset was used to test the model's predictive ability. The RF was fitted using the information of quantiles for blue, green, and red channels (1, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 99%) as explanatory variables, and the Famacha© scores (1 to 5) as the target classes to be predicted. These analyses were performed to explore the classification performance of the models in predicting Famacha© scores for a specific farm that was not used to build the model, simulating the use of our predictive system in new, previously unseen farms. The metrics used for comparing the models' performance were precision and sensitivity, as described previously.

3. RESULTS AND DISCUSSION

3.1 Segmentation evaluation

The proposed segmentation method with local features and RF model presented an average IoU of 0.70 ± 0.08. Figure 3 shows the segmentation results along with the corresponding ground truth of the ocular conjunctiva area of five animals. It is observed in the results of the average of IoU (0.70± 0.08) and in Figure 3, where the segmentation results of ocular conjunctiva areas achieved by the proposed segmentation model are close to the ground truth. It is demonstrated that the proposed local features combined with the RF method have good localization accuracy and reliable segmentation ability. Additionally, a benefit of using our proposed method, as opposed to convolution neural networks (CNNs), is the potential for efficient model training with just a single image, offering a more streamlined approach.

3.2 Canonical correlation analysis

After segmentation, we extracted 768 values of intensity of the color in each image (256 intensity of color for each channel, blue, green, and red) and determined the quantiles of color intensity in each channel (1, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 99%), in order to reduce the dimension of the color intensity traits extracted in the ocular conjunctiva region by segmentation process.

In this context, the CCA was performed to assess whether the selected quantiles could be used as an indicator of the intensities of color extracted from the original image. The number of canonical correlations extracted from the analysis was equal to the number of variables in the smaller set, resulting in 33 canonical correlations (Table 2). Thus, the canonical correlation ranged from 0.70 for the 99% quantile of the red channel to 0.94 for the 1% quantile of the blue channel (Table 2). The CCA suggested that the 11 quantiles evaluated of color intensity in each image channel measured are highly correlated with the 768 values extracted from the original image. Therefore, 11 quantiles were selected for each channel as explanatory variables, totaling 33 explanatory variables.

3.3 Classification models

Among the metrics used to compare the classification performance across models, the most suitable model combines higher sensitivity for anemic animals (Famacha© scores 3, 4, and 5) and the highest precision for non-anemic animals (Famacha© scores 1 and 2). This specific criterion was chosen to evaluate the prediction ability of the models, because it considers the balance among productivity, cost reduction, and treating the anemic animals (parasitized animals) in the herd, making the production system more sustainable. It was found that among the two models tested, the RF model presented the best classification performances for all Famacha© scores.

The RF model provided the best performances in predicting anemic animals, as indicated by the high values of sensitivity for Famacha© scores 3 (80.9%), 4 (46.2%), and 5 (60%) compared to the MLR model (Table 3). The poorest performance for the anemic animals' class was achieved with MLR for the Famacha© score 5, which did not identify any animal (sensitivity = 0) (Table 3). The precision of the RF was 72.7% for Famacha© score 1 and 62.5% for Famacha© score 2 presenting better results than the MLR model (Table 3).

The differences between the classification performances of the models can be explained by the ability of each algorithm to learn the complex relationship between the input variables (quantiles) and the Famacha© score. Machine learning methods, different from regression methods, can derive a model from available data without previous knowledge of the relationship between variables (McQueen et al., 1995; Kotsiantis et al., 2007). In addition, RF offers a non-linear approach, robust and able to capture non-linear relationships between input and target variables, as opposed to MLR which is a generalized linear model (Goldstein et al., 2017; Mullainathan and Spiess, 2017). The higher predictive performances found using RF might indicate that there is some non-linear component in the relationship between ocular conjunctiva color and Famacha© score.

3.4 Classification performance of leave-one-farm-out cross-validation technique

According to the results, the RF was the most suitable model for classifying all Famacha© scores (Table 3), as indicated by the metrics used to evaluate the performance of the classification models. We further investigated the performance of RF model across farms by alternating the data from one specific farm as the testing set and the remaining farms as the training set. The prediction quality of RF for Famacha© score using this leave-one-farm-out cross-validation technique presented better sensitivity for Famacha© score 3 in farms A (66.4%), and C (77.8%), and for score 4 in farm B (44.4%) (Table 3). While the precision for Famacha© score 1 (non-anemic animals) showed the best results in farm C (50%) (Table 3). The identification of anemic animals with Famacha© score of 5 was not possible on any farm (Table 3).

The strategy that omits records for the same animal from the training data set has been carried out in some areas of animal production and it is employed for validating the model's predictive ability in a setup that is closer to reality (Shetty et al., 2017; Dorea et al., 2018; Freitas et al., 2023). The RF model performed using leave-one-farm-out cross-validation technique did not present competitive predictive ability compared with RF performed with data randomly split. Inflation in the classification performance may have occurred when sheep were split randomly, and animal images of the same farm were kept in both the training and testing sets.

Dorea et al. (2018) investigated this topic hypothesizing that even excluding an animal from the training data set, conditions related to external factors such as weather, diet, season, management, and others would inflate the prediction quality. As expected, model validation excluding entire farms in cross-validation presented a lower prediction quality. Nonetheless, this setup is closer to reality because the developed models are supposed to be used across farms, including new ones, with different environments and management conditions. However, the difference in the unbalanced number of images and Famacha© scores from each farm may have contributed to the lower quality of the results (Table 1).

In summary, it is feasible to predict the Famacha© score, especially for scores 2 to 4, from ocular conjunctiva images in sheep, providing key information for management decisions on the farm eliminating human subjectivity in assessment. Nonetheless, future studies will be required to improve Famacha© score classification performance metrics by using a larger dataset and images obtained from other breeds. Furthermore, in order that the Famacha© score estimation method based on image analysis can be easily applied to production of sheep farming, automation of the method would be considered in future studies. In addition, it should be noted that our dataset was imbalanced (Table 1), the number of observations categorizing extreme Famacha© scores (1 and 5) was smaller than the other Famacha© scores (2, 3, and 4). Therefore, a broader assessment using a balanced dataset will be required to increase the reliability of Famacha© score classification models.

Overall, the use of image analysis to automatically classify Famacha© score in sheep farming is still incipient. There are no reports in the literature on the use of image analysis combined with RF model for classifying Famacha© score using ocular conjunctiva images in sheep. This reinforces the importance of the present work, which will certainly open the way for further studies and future applicability at the field level.

4. CONCLUSION

The results indicate that it is possible to successfully predict Famacha© scores, especially for scores 2 to 4, in sheep via image analysis and random forest classification model using ocular conjunctiva images collected in farm conditions. As expected, model validation excluding entire farms presented a lower prediction quality. Nonetheless, this setup is closer to reality because the developed models are supposed to be used across farms, including new ones, with different environments and management conditions.

5. REFERENCES

Crommelinck, S., R. Bennett, M. Gerke, F. Nex, M. Y. Yang, and G. Vosselman. 2016. Review of Automatic Feature Extraction from High-Resolution Optical Sensor Data for UAV-Based Cadastral Mapping. Remote Sens. 8:689. doi:10.3390/rs8080689

Dorea, J. R. R., G. J. M. Rosa, and K. Weld, L. Armentano. 2018. Mining data from milk infrared spectroscopy to improve feed intake predictions in lactating dairy cows. Int. J. Dairy Sci. 101: 5878-5889. doi: 10.3168/jds.2017-13997.

El-Ashram, S., I. Al Nasr, R. Mehmood, M. Hu, L. He, and X. Suo. 2017. *Haemonchus Contortus* and Ovine Host: A Retrospective Review. Int. J. Adv. Res. 5:972–999. doi: 10.21474/IJAR01/3597

Freitas, L. A., R. P. Savegnago, L. S. Menegatto, R. D. Bem, N. B. Stafuzza, A. C. A. R. Paz, B. V. Pires, R. L. D. Costa, and C. C. P. Paz. 2021. Cluster analysis to explore additive-genetic patterns for the identification of sheep resistant, resilient and susceptible to gastrointestinal nematodes. Vet. Parasitol. 301:109640. doi: 10.1016/j.vetpar.2021.109640

Freitas, L. A., R. P. Savegnago, A. A. C. Alves, R. L. D. Costa, D. P. M. Munari, N. B. Stafuzza, G. J. M. Rosa, and C. C. P. Paz. 2023. Classification performance of machine learning methods for identifying resistance, resilience, and susceptibility to gastrointestinal nematode infections in sheep. Animals 13:374. doi: 10.3390/ani13030374

Goldstein, B. A., A. M. Navar, and R. E Carter. 2017. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur. Heart J. 38:1805-1814. doi:10.1093/eurheartj/ehw302.

Kotsiantis, S. B. 2007. Supervised machine learning: A review of classification techniques. Informatica 31:249-268.

Kuiseu, J., F. T. A. Zinsou, P. A. Olounlade, G. G. Alowanou, A. D. Adenile, C. C. Dansou, S. Hounzangbe-Adote, O. J. Babayemi, and P. A. Edorh. 2021. Prevalence, effects and alternative control methods of *Haemonchus contortus* in small ruminants: A review. J. Vet. Med. Anim. Health 13:84-97. doi: 10.5897/JVMAH2020.0868.

Lee, W. P., M. A. Osman, A. Z. Talib, K. Yahya, J. C. Burie, J. M. Ogier, and J. Mennesson. 2015. "Recognition of fish based on generalized color fourier descriptor". Science and Information Conference (SAI), 680–686. doi: 10.1109/SAI.2015.7237215.

Liaw, A., and M. Wiener. 2002. Classification and Regression by randomForest. R News 2:18-22.

McQueen, R. J., S. R. Garner, C. G. Nevill-Manning, and I. H. Witten. 1995. Applying machine learning to agricultural data. Comput. Electron. Agric. 12:275-293. doi: 10.1016/0168-1699(95)98601-9

McRae, K. M., M. J. Stear, B. Good, and O. M. Keane. 2015. The host immune response to gastrointestinal nematode infection in sheep. Parasite Immunol. 37:605-613. doi: 10.1111/pim.12290.

Mullainathan, S., and J. Spiess. 2017. Machine learning: an applied econometric approach. J. Econ. Perspect. 31:87-106. doi: 10.1257/jep.31.2.87.

Oliveira, E. J., R. P. Savegnago, L. A. Freitas, A. P. Freitas, S. R. Maia, F. F. Simili, L. EL Faro, R. L. D Costa, M. L. J. Santana, and C. C. P Paz. 2018. Estimates of genetic parameters and cluster analysis for worm resistance and resilience in Santa Ines meat sheep. Pesq. Agropec. Bras. 53:1338–1345. doi: 10.1590/S0100-204X2018001200006

Roberts, F. H. S., and J. P O'Sullivan. 1950. Methods for egg counts and larval cultures for Strongyles infesting the gastrointestinal tract cattle. Aust. J. Agric. Res. 1:99-102. doi: 10.1071/AR9500099

Shetty, N., P. Løvendahl, M. Lund, and A. Buitenhuis. 2017. Prediction and validation of residual feed intake and dry matter intake in Danish lactating dairy cows using mid-infrared spectroscopy of milk. J. Dairy Sci. 100:253–264. doi: 10.3168/jds.2016-11609.

Singh, D., and C. P. Swarnkar. 2012. Evaluation of targeted selective treatment strategy in sheep farm of Rajasthan. Indian J. Anim. Res. 82:679-686.

Szeliski, R. 2010. Computer Vision: Algorithms and Applications. 1st ed. Springer London, London. doi: 10.1007/978-1-84882-935-0

Szewc, M., T. Waal, and A. Zintl. 2021. Biological methods for the control of gastrointestinal nematodes. Vet. J. 268:105602. doi: 10.1016/j.tvjl.2020.105602

Van Wyk, J. A., and G. F Bath. 2002. The FAMACHA system for managing haemonchosis in sheep and goats by clinically identifying individual animals for treatment. Vet. Res. 33:509-529. doi: 10.1051/vetres:2002036.

Venables, W. N., and B. D. Ripley. 2002. Modern Applied Statistics with S. Fourth. Springer, New York.

Tables

Table 1. Number of images obtained according to farm and Famacha© score.

Famacha© Score	Farm			
	A	B	C	Total
1	58	1	103	162
2	188	24	43	255
3	277	121	45	443
4	183	54	29	266
5	20	5	5	30
Total	726	205	225	1,156

Table 2. Canonical correlation for the quantiles of color intensity in each image channel (red, blue, and green) and 768 color intensities extracted from the original image (256 intensity of color for each channel, blue, green and red).

	Canonical Correlation										
Quantiles	1%	10%	20%	30%	40%	50%	60%	70%	80%	90%	99%
Blue	0.94	0.91	0.90	0.90	0.90	0.88	0.88	0.87	0.87	0.86	0.86
Green	0.85	0.84	0.84	0.83	0.83	0.82	0.82	0.82	0.81	0.81	0.80
Red	0.79	0.78	0.77	0.77	0.76	0.76	0.74	0.74	0.72	0.72	0.70

Table 3. Precision and sensitivity for the classification of Famacha© score (1 to 5in commercial sheep according to different methods (multinomial logistic regression - MLR, and random forest - RF) trained with information of quantiles for blue, green, and red image channel (1, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 99%) as explanatory variables.

Famacha© Score	Metrics (%)	Farms randomly split[1]		Leave one-farm-out[2]		
		Model[3]		RF model		
		MLR	RF	Farm A	Farm B	Farm C
1	Sensitivity	51.61	51.61	55.17	0	1.94
	Precision	55.17	72.73	15.69	0	50.00
2	Sensitivity	43.64	54.54	5.85	41.67	53.49
	Precision	51.06	62.50	36.67	33.33	25.27
3	Sensitivity	69.66	80.90	66.43	42.98	77.78
	Precision	55.86	60.00	39.15	62.65	29.66
4	Sensitivity	42.31	46.15	5.46	44.44	17.24
	Precision	51.16	61.54	45.45	30.00	41.67
5	Sensitivity	0	60.00	0	0	0
	Precision	0	100.00	0	0	0

[1] Training set: 80% of the dataset; Testing: 20% of the dataset.

[2] Cross-validation technique: performance of RF across farms by alternating the data from one specific farm as the testing set and the remaining farms as the training set.

[3] MLR, Multinomial logistic regression; RF, random forest.

Figures

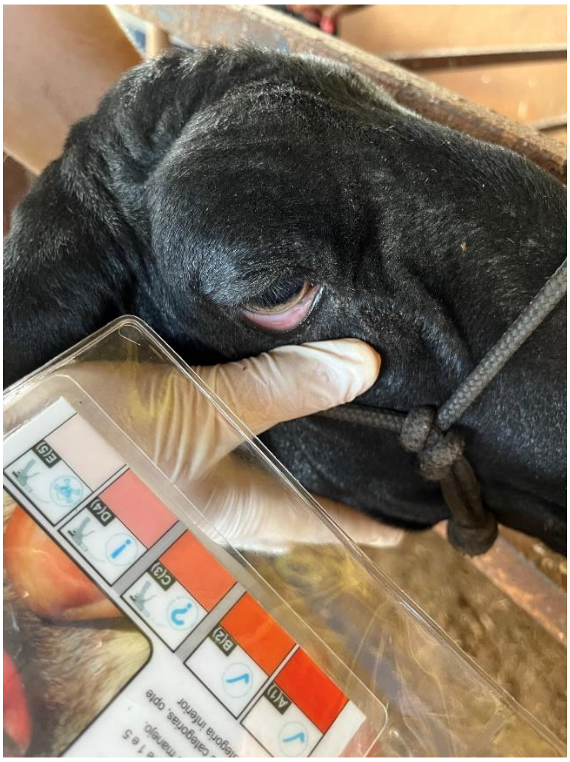

Figure 1. Application of the Famacha© method.

Source: Personal Archive.

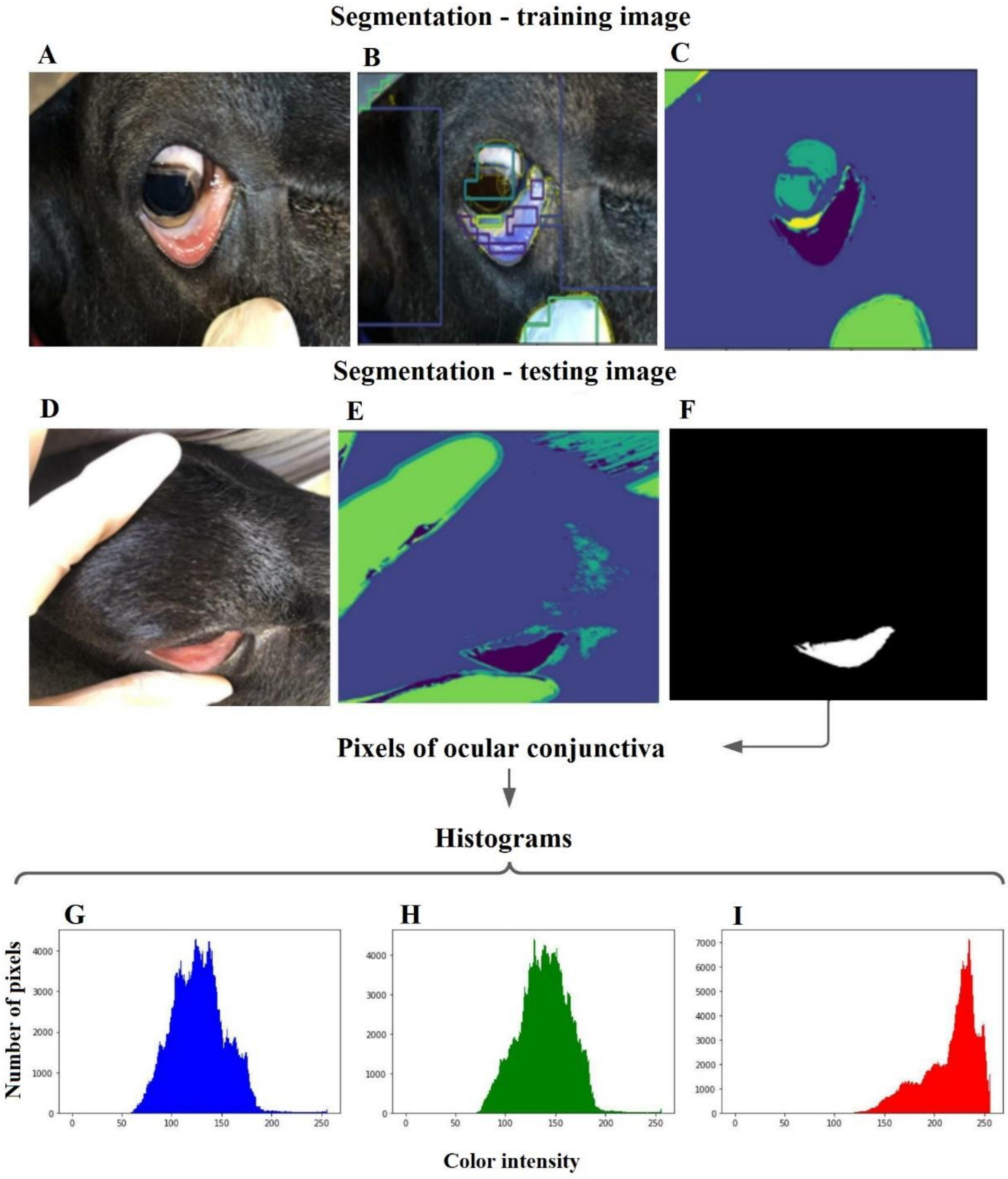

Figure 2. Diagram of the segmentation steps comprising image analyses. According to the steps performed in image used for training, **A** is the original image, **B** is the image with boxes created to identify different regions (ocular conjunctiva, hands or gloves, hair, and eye), and **C** is the image after segmentation by random forest model. In segmentation steps for testing images, **D** is the original image, **E** is image after segmentation with the random forest model, and **F** is the ocular conjunctiva region selected after segmentation and condition of size (threshold between

500 to 50,000 pixels) and localization of the region of interest (region closest to the center of the image). **G**, **H**, and **I** are histograms of the intensities of blue (n= 256), green (n= 256), and red (n= 256), respectively, using only the pixels contained in the detected ocular conjunctiva region.

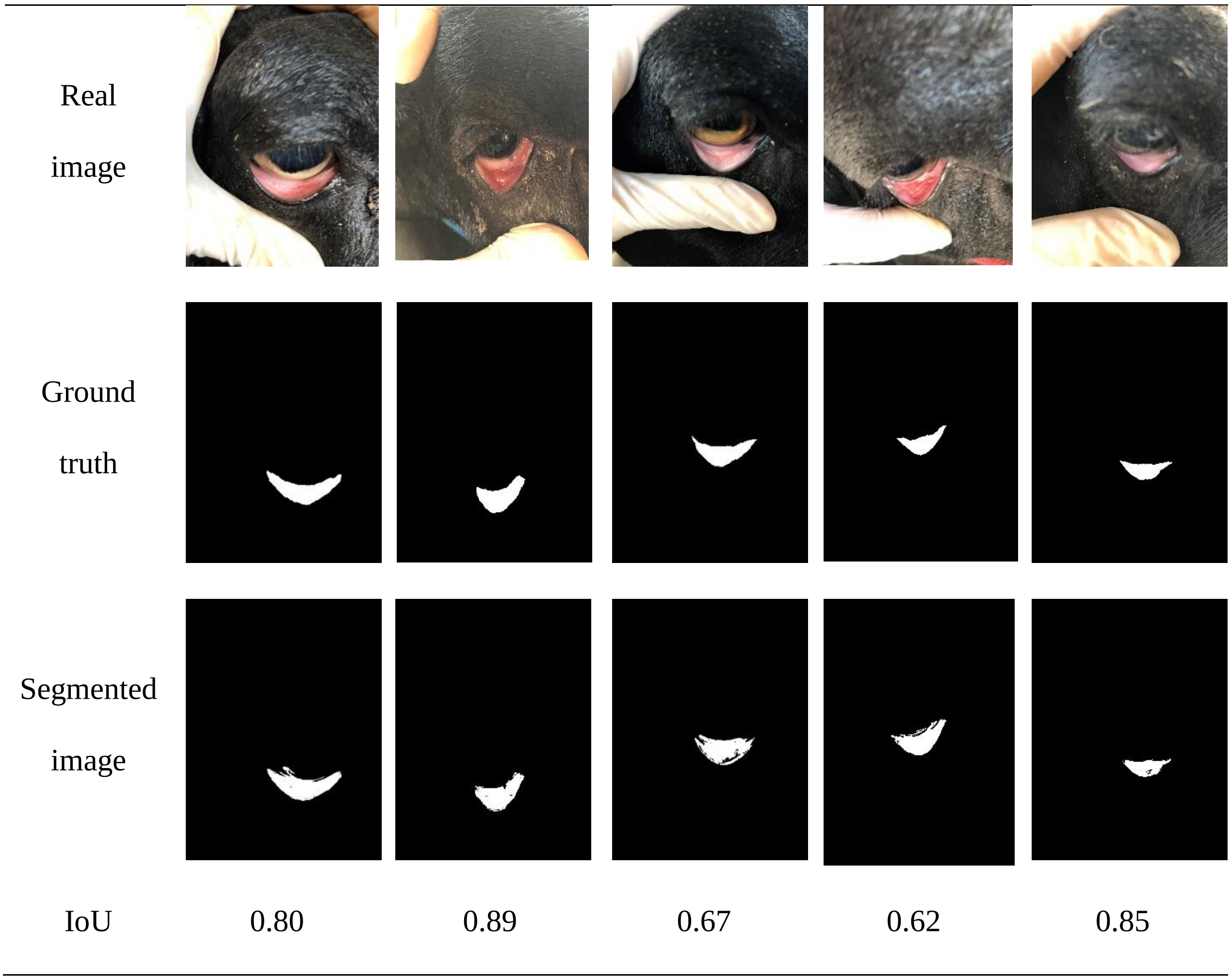

Figure 3. From top to bottom are real images, ground truth, segmentation results, and IoU metric of the proposed segmentation method with local features and random forest model.

CHAPTER 4. CLUSTER ANALYSIS TO EXPLORE ADDITIVE-GENETIC PATTERNS FOR THE IDENTIFICATION OF SHEEP RESISTANT, RESILIENT AND SUSCEPTIBLE TO GASTROINTESTINAL NEMATODES

Manuscript published in the journal of Veterinary Parasitology (Online) as part of the results obtained in this Ph.D. study. DOI: https://doi.org/10.1016/j.vetpar.2021.109640

Received 26 September 2021; Received in revised form 20 December 2021; Accepted 21 December 2021; Available online 23 December 2021.

Following is the manuscript accepted version according to the Journal of Veterinary Parasitology sharing policies for subscription articles after the embargo period (12 months).

Cluster analysis to explore additive-genetic patterns for the identification of sheep resistant, resilient and susceptible to gastrointestinal nematodes

Luara Afonso de Freitas[a*], Rodrigo Pelicioni Savegnago[b], Leonardo Sartori Menegatto[a], Ricardo Dutra do Bem[c], Nedenia Bonvino Stafuzza[c], Ana Carolina Almeida Rollo de Paz[c], Bianca Vilela Pires[a], Ricardo Lopes Dias da Costa[d], Claudia Cristina Paro de Paz[a,c]

[a] University of Sao Paulo, Department of Genetics, Ribeirão Preto, São Paulo, Brazil, 14049-900. luaraa.freitas@gmail.com, leonardomenegatto@gmail.com, bianca2510@outlook.com

[b] Michigan State University, Department of Animal Science, East Lansing, USA. 48864. rodrigopsa@yahoo.com

[c] Instituto de Zootecnia, Beef Cattle Research Center, Animal Science Institute, Sertãozinho, São Paulo, Brazil, 14174-000. ricardodutradobem@hotmail.com, nedeniabs@gmail.com, ac-paz@outlook.com, claudiacristinaparopaz@gmail.com

[d] São Paulo Agency of Agribusiness and Technology, Animal Science Institute, Nova Odessa, Brazil, 13380-011. rldcosta@sp.gov.br

[*]Corresponding author's: luaraa.freitas@gmail.com
Department of Genetics, Ribeirão Preto Medical School,
University of São Paulo, Ribeirão Preto, São Paulo,
Brazil, 14049-900.

Abstract: Infection caused by gastrointestinal nematodes is an important issue for animal health and production. Controlling worm infections improves the sustainability of the sheep industry. Genetic selection of animals that are resistant to gastrointestinal nematodes is another strategy to render sheep production more sustainable by decreasing the use of anthelmintics. The aims of this study were (1) to explore the additive-genetic pattern of EBVs for Famacha© (FAM), packed-cell volume (PVC), and fecal egg counts (FEC) of Santa Ines sheep, (2) to propose a classification of animals that are resistant, resilient and susceptible to gastrointestinal nematodes based on their additive-genetic patterns, and (3) to identify the most suitable animals for selection based on their genetic pattern. A dataset of 2,241 records from 747 animals was used to predict the breeding values for indicator traits of resistance to gastrointestinal nematodes with THRGIBBS1F90 and to carry out cluster analyses was used R software. Three clusters of animals were found in the population using hierarchical cluster analysis of the breeding values for FAM, PCV and FEC. Each cluster was characterized by different additive-genetic patterns identified by k-means non-hierarchical cluster analysis. Among a total of 747 animals, 196 were classified as resistant, 288 as resilient, and 263 as susceptible. Cluster analysis is a valuable tool for data screening that permits to evaluate only selection candidates based on their additive-genetic pattern for gastrointestinal nematode resistance. EBVs for FEC were decisive to divide the population into resilient, resistant and susceptible animals. It is also important to include the EBVs for PCV and FAM to adequately distinguish resistant from resilient animals. Finally, the resistant cluster consisted of the most desirable animals to be used as selection candidates in order to genetically improve resistance to infection with gastrointestinal nematodes. This cluster contained animals with the most appropriate additive-genetic pattern to achieve the breeding goal, with positive breeding values for PCV and negative breeding values for FAM and FEC.

Keywords: estimated breeding values, fecal egg count, *Haemonchus contortus*, multivariate analysis, *Ovis aries*

Abbreviations: estimated breeding values (EBVs), packed cell volume (PCV), Famacha© method (FAM), fecal egg count (FEC).

1. INTRODUCTION

Haemonchus contortus is a common gastrointestinal nematode that affects sheep in tropical and subtropical areas (Starling et al., 2019). This nematode causes anemia, decreasing productivity and increasing sheep mortality and thus leading to economic losses (Tehrani et al., 2012; Atlija et al., 2016; Starling et al., 2019).

Sheep infected by gastrointestinal nematodes can be classified as resistant, resilient and susceptible according to their immune response to parasitic infections (Woolaston and Baker, 1996). The immune system of resistant animals controls the parasite infection by preventing the establishment of infective larvae and/or eliminating those already implanted in the gastrointestinal tract (Albers et al., 1987). Resilient animals exhibit normal productivity and remain healthy despite high levels of infection (Bishop, 2012, Albers et al., 1987). Susceptible animals do not have resilience or resistance and have high parasite egg counts in feces and anemia, resulting in low meat production and health problems (Woolaston and Baker, 1996). The detection of resilient, resistant and susceptible animals is therefore important to minimize economic losses.

Anthelmintics are the most common treatment to control gastrointestinal infections caused by nematodes (Alba-Hurtado; Muñoz-Guzmán, 2012; Calvete et al., 2014; Atlija et al., 2016). However, their intensive use has rendered the parasites resistant to the drugs and has increased anthelmintic residues in the environment (Beynon, 2012, Shalaby, 2013).

Atlija et al. (2016) found that genetic selection of sheep for resistance decreased the use of anthelmintics. In addition, some sheep breeds, like Santa Ines, are naturally more resistant to parasitic infections than others (Rocha et al., 2011; McManus et al., 2014). Oliveira et al (2018) reported that traits related to parasite resistance (FAM and FEC) were negatively correlated with body weight and condition score while PCV had positive correlation with body weight and condition score. So, those traits can be taking into account to improve the herd genetically to parasite resistance and for productivity. However, selection must be done in a balanced because some traits related to parasite resistance can affect productivity traits positively and others negatively. Multivariate cluster analysis is an exploratory tool to divide samples into groups based on a set of traits. It groups similar individuals based on a set of traits, minimizing the heterogeneity of animals within groups and maximizing heterogeneity between groups (Hair et al., 2009). Thus, grouping animals based on their estimated breeding values (EBVs) may help detect the most suitable candidates for selection based on the additive-genetic patterns for traits associated with resistance to gastrointestinal nematodes.

The aims of this study were (1) to explore the additive-genetic pattern of EBVs for Famacha© (FAM), packed-cell volume (PVC), and fecal egg counts (FEC) of Santa Ines sheep, (2) to propose a classification of animals that are resistant, resilient and susceptible to gastrointestinal nematodes based on their additive-genetic patterns, and (3) to identify the most suitable animals for selection based on their genetic pattern.

2. MATERIAL AND METHODS

2.1 Description of the data

Experimental procedures were conducted following the recommendations of the Institutional Animal Care and Use Committee of the Beef Cattle Research Center, Animal Science Institute, Sertãozinho, São Paulo, Brazil (CEUA No. 267-18).

Pedigree had 4,780 animals and dataset had 2,241 phenotypes from 747 Santa Ines sheep (666 females and 81 males) monthly recorded during two periods: from July 2013 to October 2014 and from August 2018 to December 2020. The animals were on average 2.5 years old and 84% of the animals with known birth date records were adults and lambs. The animals came from five farms in São Paulo State: Cravinhos (latitude: 21°20′25″ S; longitude: 47°43′46″ W), Jardinopolis (latitude: 21°01′04″ S; longitude: 47°45′50″ W), Nova Odessa (latitude: 22°46′39″ S, longitude: 47°17′45″ W), Pontal (latitude: 21°01′21″ S; longitude: 48°02′14″ W), and Serrana (latitude: 21°12′41″ S; longitude: 47°35′44″ W). According to Köppen's climate classification, these cities have a humid subtropical climate (Cravinhos, Nova Odessa, and Serrana) and a tropical climate with a dry season (Jardinopolis and Pontal) (Alvares et al., 2014).

Faecal samples were collected directly from the animals using a rectal ampoule. FEC was determined using the modified McMaster technique described by Gordon and Whitlock (1939) and modified by Ueno and Gonçalves (1998). FEC was $\log_{10}$ transformed (FEC + 1) to approximate it to a normal distribution.

Blood samples were collected into vacutainer tubes containing $EDTAK_3$ by puncture of the jugular vein. PCV was measured by the microhematocrit centrifugation technique (Schalm et al., 1975).

FAM was measured using the Famacha© card (Van Wyk; Bath, 2002) to compare different color shades of the ocular conjunctiva with a standard scale from 1 to 5: 1 (non-anemic), 2 (non-anemic), 3 (mildly anemic), 4 (anemic), and 5 (severely anemic), which correspond to red, pink-red, pink, pink-white and white, respectively (Vatta et al., 2001).

2.2 Estimated breeding values

The covariance components for FAM, PCV and FEC were estimated by Bayesian inference in a single-trait animal model using the THRGIBBS1F90 program (Misztal et al., 2008). The model included the fixed effects of contemporary group (farm, year and season of record), sex, and additive genetic, permanent environment and residual random effects.

A total of 100,000 samples were generated after a stochastic simulation process, with 10,000 burn-in samples and a thinning interval of 25 samples in order to obtain 3,600 effective samples for the estimation of (co)variance components and genetic parameters. Convergence was monitored using the Geweke criterion (p>0.05) and by visual inspection of trace plots of the variance components. The stationary trend around the mean of the trace plots indicated convergence of the parameters (Fig. 1).

INSERT FIGURE 1

The posterior estimates were obtained with POSTGIBBSF90 (Misztal, 2008). The complete model can be written in matrix notation as:

$$y = X\beta + Za + Wpe + \ \varepsilon$$

where *y* is the vector of trait values in all individuals; β, a, and *pe* are the solution vectors for the fixed effects, additive genetic and permanent environmental random effects, respectively; X, Z, and W are the incidence matrices for the fixed effects, additive genetic and permanent environmental effects, respectively, and ε is the vector for the residual random effect.

The assumptions for the random effects and (co)variance components are:

$$\beta \propto constant$$

$$y|\beta, a, pe, G_a, P, R \sim NMV[X\beta + Za + Wpe + \ \varepsilon]$$

$$a|G_a \sim NMV[0, (G_a \otimes A)]$$

$$\mathrm{pe}|P \sim NMV[0, (P \otimes I_n)]$$

$$\varepsilon|R \sim NMV[0, (R \otimes I_n)]$$

$$G_a|S_g, \upsilon_g \sim IW\left[S_g\, \upsilon_g\, , \upsilon_g\right]$$

$$P|S_{pe}, \upsilon_{pe} \sim IW\left[S_{pe}\, \upsilon_{pe}\, , \upsilon_{pe}\right]$$

$$R|S_r, \upsilon_r \sim IW[S_r\, \upsilon_r\, , \upsilon_r],$$

where A, G_a, P, R, and I_n are the relationship matrices based on the pedigree, direct genetic covariance, permanent environment, residual and identity matrices, respectively; $\otimes$ is the Kronecker product, and S_g and υ_g; S_{pe} and υ_{pe}; S_r and υ_r are the a priori values and degrees of freedom for additive genetic, permanent environmental and residual (co)variances, respectively.

A threshold model was used for FAM, which assumes that the underlying (liability) scale has a normal continuous distribution (Gianola and Sorensen, 2002):

$$U \mid \theta \sim N(W\theta, I\sigma_e^2)$$

where U is the underlying scale vector of order r; $\theta' = (\beta', a', pe')$ is the vector of location parameters of order s, with β defined as fixed effect; a and pe are the solution vectors for additive genetic and permanent environmental random effects, respectively; W is the known incidence matrix of order $r_x s$; I is the identity matrix of order $r_x r$, and σ_e^2 is the residual variance.

Considering that the variable in the underlying distribution is not observable, $\sigma_e^2 = 1$ is adopted which allows identification of the variable in the likelihood function (Gianola and Sorensen, 2002). This assumption is a standard procedure in the analysis of categorical data using threshold models. Categorical traits are determined by non-observable continuous variables on an underlying scale, in which initial threshold values were fixed for t (t_{min}, t_1, ... , t_{j-1}, t_{max}), $t_1 < t_2 ... < t_{j-1}$, where j is the number of categories (thresholds). Thus, the observable data depend on the underlying variable, which is limited by two unobservable thresholds (Gianola and Foulley, 1983). To understand the joint probability density function of the parameters, given the hyperparameters, it is necessary to know the observed values y distribution, which is conditioned by the continuous and threshold observations. FAM was divided into five categories (1 to 5). Thus, the categories of FAM for each animal i (y_i) were defined by U_i on the underlying scale:

$$y_i = \begin{cases} 1\ if\ t_0 < U_i \leq t_1 \\ 2\ if\ t_1 < U_i \leq t_2 \\ 3\ if\ t_2 < U_i \leq t_3 \\ 4\ if\ t_3 < U_i \leq t_4 \\ 5\ if\ t_4 < U_i \leq t_5 \end{cases} i = 1, \ldots, \mathrm{n}$$

where *n* is the number of observations for each category.

After the t_0 to t_5 thresholds were specified for FAM, one of the thresholds (t_1 to t_5) was adjusted to an arbitrary constant. In this case, it was assumed that $t_1 = 0$, with the vector of estimable thresholds being defined as:

$$t = \begin{bmatrix} t_2 \\ t_3 \\ t_4 \end{bmatrix}.$$

Since the observations are conditionally independent, the likelihood function is defined by the product of contributions of each record. Thus, the conditional probability that y_i falls into category *j* (j = 1 to 5), given vectors $\beta, a, pe,$ and t, can be written as:

$$Pr(y_i = j \mid \beta, a, pe, t) = Pr\left(t_{j-1} < U_j < t_j | \beta, a, pe, t\right)$$

$$= \Phi\left(t_j - X'_i\beta - Z'_i a - W'_i pe\right) - \Phi\left(t_{j-1} - X'_i\beta - Z'_i a - W'_i pe\right)$$

$$= Pr(y_i \mid \beta, a, pe, t),$$

where y_i is the response of animal *i* assuming value 1, 2, 3, 4, or 5 if the observation belongs to the first, second, third, fourth, or fifth category, respectively; *t* is the threshold value for which an arbitrary value will be fixed since it is not estimable; U_i is the value for the underlying variable for animal *I,* and Φ is the cumulative distribution function of a standard normal variable (i.e., indicator of a function with value 1 if the expression is true and 0 if it is false).

2.3 Cluster analysis

Cluster analyses were done using the predicted breeding values (EBVs) for FAM, PCV and FEC in order to identify animals that are resistant, resilient, and susceptible to gastrointestinal nematodes based on the additive-genetic pattern of each trait. Hierarchical cluster analyses were performed with the R program (Development Core Team, 2018) with the Ward algorithm (1963). The dissimilarity measure between animals was the Euclidean distance. EBVs for all traits were standardized using the standard normal distribution. The NbClust package (Charrad et al., 2014) was used to determine the optimum number of clusters in the population. Non-hierarchical clustering with k-means method was done to visualize the additive-genetic pattern of the clusters from hierarchical cluster analysis.

3. RESULTS

According to 26 indices from NbClust package (Charrad et al., 2014), three clusters divided the population based on the EBVs for FAM, PCV and FEC (Fig. 2). The additive-genetic patterns of the EBVs for the traits were different in each cluster (Fig. 3). Resistant, resilient and susceptible animals were classified according to FEC values in each cluster. This trait is the main indicator of gastrointestinal nematode resistance (Pollot et al., 2004; Bishop, 2012; McManus et al., 2014). Among a total of 747 animals, 196 were classified as resistant (26.24%), 288 as resilient (38.55%), and 263 as susceptible (35.21%).

INSERT FIGURES 2 AND 3

The additive-genetic pattern of the resistant cluster revealed animals with positive EBVs for PCV and negative EBVs for FAM and FEC (Fig. 3). Resilient animals had positive EBVs for PCV and FEC and negative EBVs for FAM (Fig. 3). The susceptible cluster included animals with positive EBVs for FAM and FEC and negative EBVs for PCV (Fig. 3).

Animals in the resilient cluster had mean FEC and PCV of 1,719.44 eggs/g and 32%, respectively, and 77.87% of the animals were classified as Famacha© 1 and 2. Animals in the resistant cluster had mean FEC and PCV of 243.37 eggs/g and 31%, respectively, and 68.37% of the animals were classified as Famacha© 1 and 2. The mean FEC and PCV in the susceptible cluster were 2,057.41 eggs/g and 28%, respectively, and 40.3% of the animals were classified as Famacha© 3, 4 and 5 (Table 1 and Fig. 3).

INSERT TABLE 1

4. DISCUSSION

Non-hierarchical cluster analysis using three traits (EBVs for FAM, FEC and PCV) divided the population into three clusters. It can be used as a pre-screening tool to identify desirable animals for selection based on their additive-genetic pattern. Resistant animals can be used as selection candidates to genetically improve the resistance to gastrointestinal nematode infection because of their positive EBVs for PCV and negative EBVs for FAM and FEC (Fig. 3).

In this study, the EBVs and phenotypic values for FEC differed between the resistant and resilient groups despite their similar FAM classification (Fig. 3 and Table 1). This can be explained because resilient animals tolerate parasitic infections without production losses. Thus, these animals may not have anemia even if they are infected with parasites. According to Albers et al. (1987) and Marques et al. (2018), resilient animals can withstand the effects of infection even in the presence of high FEC. On the other hand, resistant animals resist parasitic infection by eliminating the parasite.

It is extremely important to distinguish resilient and resistant animals because the former may have a high parasitic load and can contaminate the environment, causing environmental impacts (Bishop, 2012). Romjali et al. (1996) and Gauly and Ehrhardt (2001) suggested the use of FEC to identify and select animals resistant to gastrointestinal nematodes. If selection were done based only on FAM and PCV, without considering FEC, resilient animals would have been selected together with resistant animals, compromising the breeding goals.

According to Albers et al. (1987), selection of genetically resistant sheep would considerably increase productivity and genetic gains in resistance to gastrointestinal nematode infection. However, selection of resilient animals would slow down the rates of genetic progress and is not desirable because these animals have high parasitic loads even in the absence of anemia (Albers et al., 1987). Furthermore, high loads of gastrointestinal nematodes can affect the animal's shape and reproduction rate, decreasing genetic progress and increasing the generation interval (Woolaston and Baker, 1996).

Using EBVs for FEC, FAM, and PCV together is appropriate to distinguish resistant and resilient animals. Stear et al. (1995) and Marques et al. (2018) suggested that the use of more than one trait increases the efficiency in identifying resistant, resilient and susceptible animals.

5. CONCLUSION

Cluster analysis is a valuable tool for data screening that permits to evaluate only selection candidates based on their additive-genetic pattern for gastrointestinal nematode resistance. EBVs for FEC were decisive to divide the population into resistant, resilient and susceptible animals. It is also important to include the EBVs for PCV and FAM to adequately distinguish resistant from resilient animals. Finally, the resistant cluster consisted of the most desirable animals to be used as selection candidates in order to genetically improve resistance to infection with gastrointestinal nematodes. This cluster contained animals with the most appropriate additive-genetic pattern to achieve the breeding goal, with positive breeding values for PCV and negative breeding values for FAM and FEC.

ACKNOWLEDGMENTS

This work was supported by FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) [grant numbers 2012/15982-0 and 2016/14522-7]. L. A. Freitas received a fellowship from FAPESP (grant number 2018/01540-2). C.C.P. Paz was the recipient of a productivity research fellowship from CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

REFERENCES

Alba-Hurtado, F., Muñoz-Guzmán, M.A., 2012. Immune responses associated with resistance to Haemonchosis in sheep. Biomed Res. Int. 2013, 1-11.

Albers, G.A.A., Gray, G.D., Piper, L.R., Barker, J.S.F., Le Jambre, L.F., Barger, I.A., 1987. The genetics of resistance and resilience to *Haemonchus contortus* infection in young Merino sheep. Int. J. Parasitol. 17, 1355-1363.

Alvares, C.A., Stape, J.L., Sentelhas, P.C.S., Gonçalves, J.L.M., Sparovek, G., 2014. Köppen's climate classification map for Brazil. Meteorol. Z. 22, 711–728.

Atlija, M., Prada, J.M., Gutiérrez-Gil, B., Rojo-Vázquez, F.A., Stear, M.J., Arranz, J.J., Martínez-Valladares, M., 2016. Implementation of an extended ZINB model in the study of low levels of natural gastrointestinal nematode infections in adult sheep. Vet. Res. 12, 97.

Beynon, S.A., 2012. Potential environmental consequences of administration of anthelmintics to sheep. Vet. Parasit.189, 113-124.

Bishop, S.C., 2012. A consideration of resistance and tolerance for ruminant nematode infections. Front. Genet. 3, 62-68.

Calvete, C., Ferrer L.M., Lacasta, D., Calavia, R., Ramos, J.J., Ruiz-de-Arkaute, M., Uriarte, J., 2014. Variability of the egg hatch assay to survey benzimidazole resistance in nematodes of small ruminants under field conditions. Vet. Parasit. 203, 102-113.

Charrad, M., Ghazzali, N., Boiteau, V., Niknafs A., 2014. NbClust: An R package for determining the relevant number of clusters in a data set. J. Stat. Softw. 61, 1-36.

Gauly, M., Erhardt, G., 2001. Genetic resistance to gastrointestinal nematode parasites in Rhon sheep following natural infection. Vet. Parasitol. 102, 253 –259.

Gianola, D., Foulley, J. L. 1983. Sire evaluation for ordered categorical data with a threshold model. Genet. Sel. Evol. 15, 201-224.

Gianola, D., Sorensen, D. (Ed.1), 2002. Likelihood, Bayesian, and MCMC methods in quantitative genetics. Springer-Verlag, New York, 740p.

Gordon, H.M., Whitlock, H.V., 1939. A new technique for counting nematode eggs in sheep faeces. J. Sci. Ind. Res. 12, 50-52.

Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E. (Ed. 7), 2009. Multivariate Data Analysis. Prentice Hall, New Jersey, 816p.

Marques, C.A.T., Saraiva, L.A., Torreão, J. N. C, Silva, T.P.D, Bezerra, L.R., Edvan, R.L., Marcos, J.A., Romilda, R.N., 2018. The use of targeted selective treatments on controlling gastrointestinal nematodes in different sheep categories under grazing system. Pesq. Vet. Bras. 38, 470-476.

McManus, C., Paim, T.P., Melo, C.B., Brasil, B.S.A.F, Paiva, S.R. 2014. Selection methods for resistance to and tolerance of helminths in livestock. Parasite. 21, 56.

Misztal, I. 2008. Reliable computing in estimation of variance components. J. Anim. Breed. Genet. 125, 363-370.

Oliveira, E.J., Savegnago, R.P., Freitas, L.A., Freitas, A.P., Maia, S.R., Simili, F.F., EL Faro, L., Costa, R.L.D., Santana, M.L.J., Paz, C.C.P. 2018. Estimates of genetic parameters and cluster analysis for worm resistance and resilience in Santa Inês meat sheep. Pesq. Agropec. Bras. 53, 1338-1345.

Pollot, G.E., Karlsson, L.J.E., Eady, S., Greeff, J.C., 2004. Genetic parameters for indicators of host resistance to parasites from weaning to hogget age in Merino sheep. J. Anim. Sci. 82, 2852-2864.

Rocha, R.A., Bricarello, P.A., Silva, M.B., Houdijk, J.G.M., Almeida, F.A., Cardia, D.F.F., Amarante, A.F.T., 2011. Influence of protein supplementation during late pregnancy and lactation on the resistance of Santa Ines and Ile de France ewes to *Haemonchus contortus*. Vet. Parasitol. 181, 229–238.

Romjali, E., Pandey, V.S., Batubara, A., Gatenby, R.M., Verhulst, A., 1996. Comparison of resistance of four genotypes of rams to experimental infection with *Haemonchus contortus*. Vet. Parasitol. 65, 127-137.

Schalm, O.W., Jain, N.C., Carrol, E.J. (Ed. 3), 1975. Veterinary Haematology. Lea and Febiger, Philadelphia, pp. 197-199.

Shalaby, H.A. 2013. Anthelmintics resistance; How to overcome it? Iran J. Parasitol. 8, 18-32.

Stear, M.J. Bishop, S.C., Duncan, J.L., McKellar, Q.A., Murray, M., 1995. The repeatability of faecal egg counts, peripheral eosinophil counts, and plasma pepsinogen concentrations during deliberate infections with *Ostertagia circumcincta*. Int. J. Parasitol. 25, 375-380.

Starling, R.Z.C, Almeida, F.A., Viana, M.V.G., Castilhos, A.M., Amarante, A.F.T., 2019. Losses Caused by Gastrointestinal Nematode Infections in Dorper Lambs Under Two Nutritional Status. Rev. Bras. Parasitol. Vet. 28, 652-660.

Tehrani, A., Javanbakht J., Jani M., Sasani F., Solati A., Rajabian M., Khadivar F., Akbari H., Mohammadian M., 2012. Histopathological study of *Haemonchus contortus* in Herrik sheep abomasum. J. Bacteriol. Parasitol. 3, 2-5.

Ueno, H., Gonçalves, P.C. (Ed. 4), 1998. Manual para diagnóstico das helmintoses de ruminantes. Press Color, Salvador, 143p.

Van Wyk, J.A., Bath G.F., 2002. The FAMACHA© system for managing haemonchosis in sheep and goats by clinically identifying individual animals for treatment. Vet. Res. 33, 509-529.

Vatta, A.F., Letty B.A., van der Linde M.J., van Wijk E.F., Hansen J.W., Krecek R.C., 2001.

Testing for clinical anaemia caused by *Haemonchus* spp. in goats farmed under resource-poor conditions in South Africa using an eye colour chart developed for sheep. Vet. Parasitol. 99, 1-14.

Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58, 236–244.

Woolaston, R.R., Baker, R.L., 1996. Prospects of breeding small ruminants for resistance to internal parasites. Int. J. Parasitol. 26, 845-855.

Tables

Table 1. Number of animals (N) and means and standard deviations of PCV and FEC phenotype values by cluster in Santa Ines sheep.

		Traits		
Clusters	N	PCV (%)	FEC (eggs/g)	FEC (log)[1]
Resilient	288	32±0.04	1,719.00±3,376.71	2.67±0.93
Resistant	196	31±0.04	243.37±1,112.96	0.69±1.16
Susceptible	263	28±0.04	2,057.41±4,792.64	2.70±0.96

PCV: packed cell volume; FEC: fecal egg count.

[a] $\log_{10}$ (FEC + 1) transformation.

Figure Captions

Figure 1. Trace-plots of (co)variance components for Famacha© method (FAM), packed cell volume (PCV) and fecal egg count (FEC) for additive-genetic variance (A), permanent effect variance (B), and residual variance (C).

Figure 2. Dendrogram obtained by hierarchical cluster analysis using Euclidian distance based on estimated breeding values for Famacha© method (FAM), packed cell volume (PCV), and fecal egg count (FEC). Rectangles delimit the individuals in each cluster.

Figure 3. Additive-genetic pattern obtained by k-means non-hierarchical cluster analysis using standardized estimated breeding values (EBVs) for Famacha© method (FAM), packed cell volume (PCV), and fecal egg count (FEC).

Figures

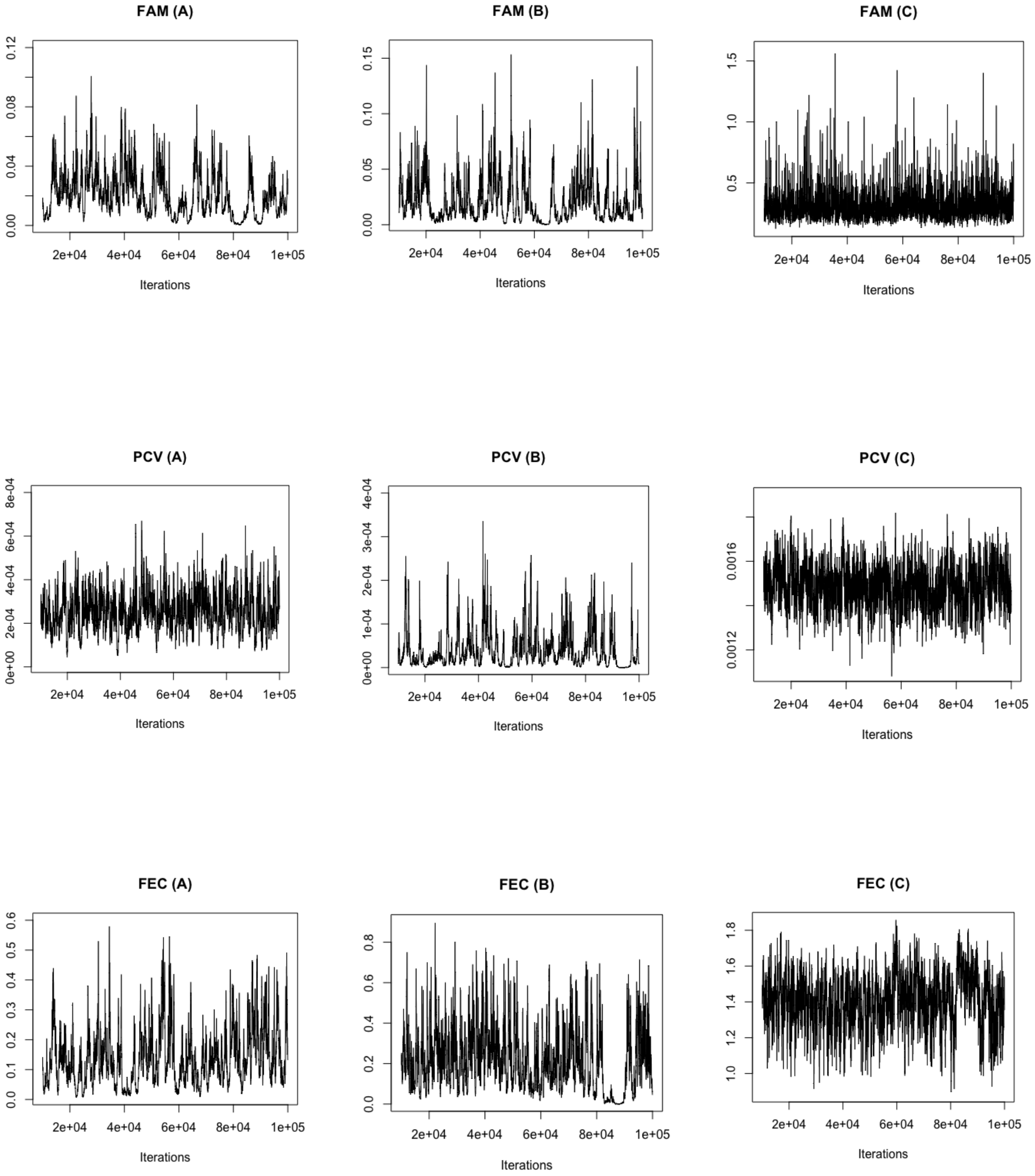

Figure 1

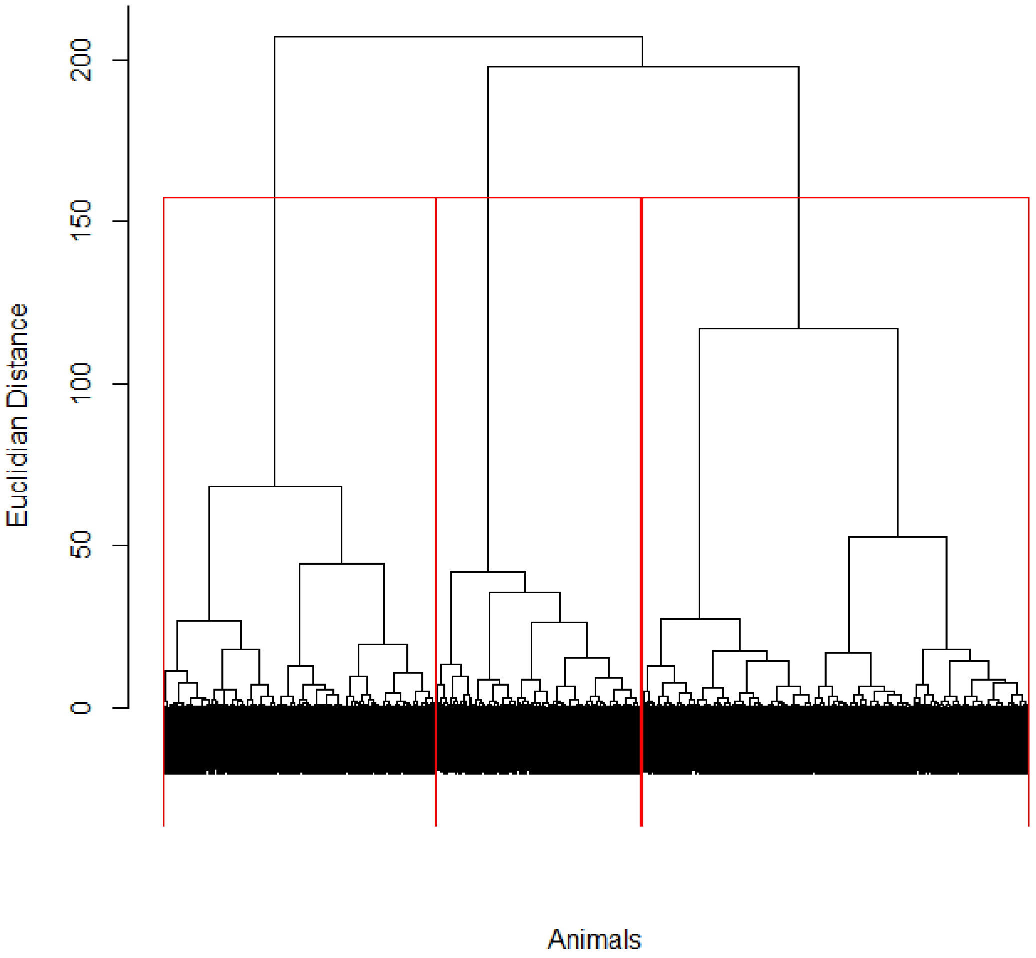

Figure 2

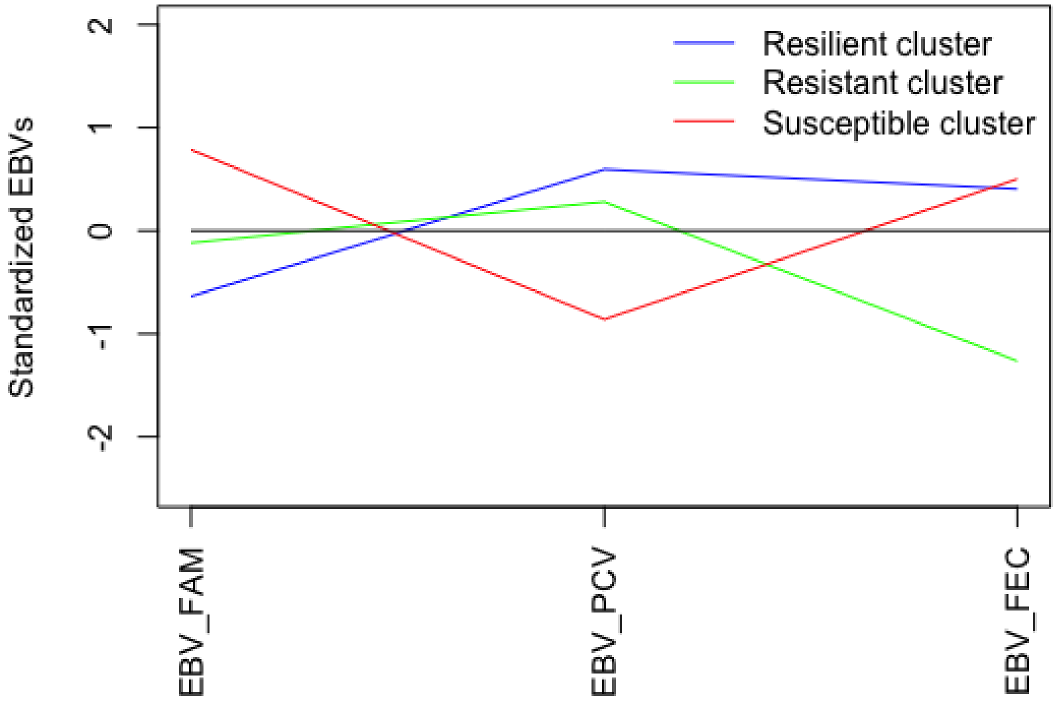
Resilient cluster
Resistant cluster
Susceptible cluster
Standardized EBVs
2
1
0
-1
-2
EBV_FAM
EBV_PCV
EBV_FEC

Figure 3

CHAPTER 5. GENOME-ENABLED PREDICTION OF INDICATOR TRAITS OF RESISTANCE TO GASTROINTESTINAL NEMATODES IN SHEEP USING PARAMETRIC MODELS AND ARTIFICIAL NEURAL NETWORKS

Abstract: This study aimed to assess the predictive ability of parametric models and artificial neural network method for genomic prediction of the following indicator traits of resistance to gastrointestinal nematodes in Santa Inês sheep: packed cell volume (PCV), fecal egg count (FEC), and Famacha© method (FAM). After quality control, the number of genotyped animals was 551 (PCV), 548 (FEC), and 565 (FAM), and 41,676 SNPs. The average prediction accuracy (ACC) calculated by Pearson correlation between observed and predicted values and mean squared errors (MSE) were obtained using genomic best unbiased linear predictor (GBLUP), BayesA, BayesB, Bayesian least absolute shrinkage and selection operator (BLASSO), and Bayesian regularized artificial neural network (three and four hidden neurons, BRANN_3 and BRANN_4, respectively) in a 5-fold cross-validation technique. The average ACC varied from moderate to high according to the trait and models, ranging between 0.418 and 0.546 (PCV), between 0.646 and 0.793 (FEC), and between 0.414 and 0.519 (FAM). Parametric models presented nearly the same ACC and MSE for the studied traits and provided better accuracies than BRANN. The GBLUP, BayesA, BayesB and BLASSO models provided better accuracies than the BRANN_3 method, increasing by around 23.16, 23.16, 23.44 and 22.88% (PCV), and 18.54, 18.54, 18.64, 18.54% (FEC), respectively. Likewise, there was an increase of ACC of 20.23% (FAM) comparing all parametric models and BRANN_4. In conclusion, parametric models are suitable for genome-enabled prediction of indicator traits of resistance to gastrointestinal nematodes in sheep. Due to the small differences in accuracy found between them, the use of the GBLUP model is recommended due to its lower computational costs.

Keywords: bayesian alphabet, genomic selection, machine learning, *Ovis aries*, predictive

ability

1. INTRODUCTION

Infection by gastrointestinal nematodes (GIN) is one of the main sanitary issues in small ruminants (Szewc et al., 2021) and has a great impact on animal health and productivity, negatively affecting the quantity and quality of production (Mavrot et al., 2015, Castagna et al., 2021). The adoption of anthelmintic is widely used to reduce in-farm GIN infections. However, the indiscriminate use of these compounds has resulted in serious problems of parasite resistance to anthelmintics (Kuiseu et al., 2021; Szewc et al., 2021). In addition, there is a growing demand from consumers to produce chemical-free food and increasing concern about animal welfare (Aboshady et al., 2020).

One of the alternatives for controlling gastrointestinal parasite infections is the selection of genetically resistant animals, a promising avenue to increase the productivity of animals in the presence of infectious diseases (Mulder and Rashidi, 2017; Aboshady et al., 2020). The search for genetically resistant animals, which tolerate the effects of helminthic infections, combined with efficient pasture management, has become a sustainable alternative for the prophylaxis of parasites, resulting in increased production (Greeff & Karlsson, 2020).

The identification of animals infected by GIN can be performed by fecal egg count (FEC), packed cell volume (PCV), and Famacha© method (FAM), being these measures highly correlated with the resistance to gastrointestinal nematode (Van Wyk & Bath, 2002; Selvam, 2021; 2022; Ferreira et al., 2021).

The genetic selection of animals for economically important traits is traditionally based on the phenotypic records and the pedigree information, which have been combined to compute best linear unbiased predictions (BLUP) of the genetic merit, as described by Henderson (1975). Additionally, the inclusion of modern molecular information in genetic evaluations has contributed to increasing the accuracy of breeding values prediction and reducing the generation interval (Meuwissen et al., 2001; Meuwissen et al. 2013).

Genomic selection (Meuwissen et al., 2001) has been successfully used in some animal breeding programs, and in others, its practice has been investigated (Knol et al., 2016; Brito et al., 2017; Mrode et al., 2018; Alves et al., 2020). The availability of high-density single nucleotide polymorphism (SNP) panels broadened the horizons for the study of complex traits controlled by several genes with small effects and enabled improvements in marker-assisted selection in both animal and plant breeding programs (Xu et al., 2020). Meuwissen et al. (2001) introduced the concept of predicting the genetic merit of selection candidates based on their genotypes for SNPs. This methodology is known as genomic prediction and can be used in genomic selection, in which the selection of animals is based on breeding values estimated through information from SNPs, accelerating genetic gains in breeding programs (Voss-Fels et al., 2019).

Several models for applying genomic selection have been proposed (Heslot et al. 2012). The differences between statistical modeling in genomic studies are related to the type of relationship assumed between target traits and explanatory covariates (linear or nonlinear), the type of genomic information (full marker matrix, subset of genotypes, genomic relationship matrix) and strategies used to address the problem of estimating genomic breeding values from models that have a large number of parameters with a small number of observations (González-Recio et al., 2014).

Parametric approaches have been shown to be effective for whole genome-enabled prediction. Nonetheless, these methods assume an assumption about the genetic architecture of the trait that does not always hold in practice. In recent years there has been a growing interest in using Machine Learning (ML) methods (semi and non-parametric methods) for genome-enabled prediction (Ehret et al. 2015; Naderi et al. 2016; Alves et al., 2020), mainly due to the theoretical flexibility offered by such models and the ability to intuitively explore linear and non-linear relationships between a response variable (the phenotype) and several other predictor variables (the genotypes) (Alves et al., 2020).

In addition, the increasing availability of mixed content information combined with a boost in computational processing power lead to the development and application of ML approaches in different areas (Nayeri et al., 2019), such as in genomic selection. According to Xu et al. (2020), the performance of genomic selection is associated with numerous factors, one of which is the methodology for estimating genomic breeding values. Therefore, the choice of the statistical method to be applied must be considered before the implementation of genomic selection in breeding programs. Thus, this study aimed to assess the predictive ability of parametric models and artificial neural networks for genomic prediction of indicator traits of resistance to gastrointestinal nematodes in Santa Inês sheep.

2. MATERIAL AND METHODS

2.1 Phenotypic data

The experimental procedures were conducted following the recommendations of the Institutional Animal Care and Use Committee of the Animal Science Institute, Nova Odessa, São Paulo, Brazil (protocol code CEUA Nº. 267-18, 3rd October 2018).

The dataset comprised 5,108 phenotypic records of 1,549 Santa Inês sheep for packed cell volume (PCV), 4,228 phenotypic records of 1,459 Santa Inês sheep for fecal egg count (FEC), and 5,522 phenotypic records of 1,701 Santa Inês sheep for Famacha© method (FAM), recorded in two periods: from 2013 to 2014, and from 2018 to 2020. The animals came from six farms located in the Southeast and South regions of Brazil: Cravinhos (latitude: 21°20′25″ S; longitude: 47°43′46″ W), Jardinopolis (latitude: 21°01′04″ S; longitude: 47°45′50″ W), Nova Odessa (latitude: 22°46′39″ S, longitude: 47°17′45″ W), Pontal (latitude: 21°01′21″ S;

longitude: 48°02′14″ W), Serrana (latitude: 21°12′41″ S; longitude: 47°35′44″ W), and Ventania (latitude: 24°14'45" S; longitude: 50°14'34" W).

Blood samples were collected by puncture of the jugular vein and disposed into vacutainer tubes containing $EDTAK_3$. PCV was measured by the microhematocrit centrifugation technique (Schalm et al., 1975). Fecal samples were collected directly from the rectal ampoule of the animals. FEC was determined using the modified McMaster technique described by Gordon and Whitlock (1939). FEC was $\log_{10}$ transformed (FEC + 1) to approximate it to a normal distribution. The Famacha© diagnosis was made using the Famacha© card, which compares the different shades of the ocular conjunctiva on a five-point scale corresponding to the colors robust red (non-anemic - 1), red-pink (non-anemic - 2), pink (mildly anemic - 3), pink-white (anemic - 4), and white (severely anemic - 5), respectively (Van Wyk & Bath, 2002).

2.2 Estimated breeding values

Before performing the genomic prediction analyses, the variance components for PCV, FEC, and FAM were estimated by Bayesian inference in a single-trait animal model using the THRGIBBS1F90 program (Misztal et al., 2008). The model included the fixed effects of contemporary group (farm, year, and season of record), sex, birth season, and the linear effect of age class (1: 0-150 days of age, 2: 151-550 days of age; 3: > 550 days of age) as a covariate for all traits, and additive genetic, permanent environment and residual random effects. The seasons of record collection and birth were defined as dry winter (May to October) and rainy summer (November to April). The number of animals included in the pedigree was 4,823.

A total of 500,000 Gibbs samples were generated after a stochastic simulation process, with the first 100,000 samples used as burn-in and a thinning interval of 25 samples, resulting in 16,000 effective samples for the estimation of variance components and genetic parameters.

Convergence was monitored using the Geweke criterion ($p>0.05$) and by visual inspection of trace plots of the variance components. The posterior estimates were obtained with POSTGIBBSF90 (Misztal, 2008). The complete model can be written in matrix notation as:

$$y = X\beta + Za + Wpe + \varepsilon$$

where y is the vector of trait values in all individuals; β, a, and pe are the solution vectors for the systematic effects, additive genetic and permanent environmental random effects, respectively; X, Z, and W are the incidence matrices for the systematic effects, additive genetic and permanent environmental effects, respectively, and ε is the vector for the residual random effect.

A threshold model was used for FAM, which assumes that the underlying (liability) scale has a normal continuous distribution (Gianola & Sorensen, 2002):

$$U \mid \theta \sim N(W\theta, I\sigma_e^2)$$

where U is the underlying scale vector of order r; $\theta' = (\beta', a', pe')$ is the vector of location parameters of order s, with β defined as fixed effect; a and pe are the solution vectors for additive genetic and permanent environmental random effects, respectively; W is the known incidence matrix of order $r_x s$; I is the identity matrix of order $r_x r$, and σ_e^2 is the residual variance.

The estimated breeding values (EBV) obtained with single-trait animal models were then used as response variables in genomic prediction analyses for PCV, FEC and FAM. The variance components and heritability are presented in Table 1.

2.3 Genotypic data

Genotyping was performed using the Ovine SNP50 Genotyping BeadChip (Illumina, Inc), presenting 54,516 SNPs with an average interval of 50.9 kb, which were validated in more than 75 breeds of the *Ovis aries* species, including the Santa Inês breed.

Genotype quality control (QC) was performed by an iterative process using R software

(R Development Core Team 2018). Non-autosomal and duplicated SNPs were removed from the genotypic dataset, SNP with a call rate lower than 0.90, minor allele frequency (MAF) lower than 0.05, and with a p-value less than 10^{-5} for the Hardy-Weinberg Equilibrium (HWE) test were removed from the dataset. Animals with a call rate lower than 0.90 were excluded. After QC, the number of genotyped animals was 551 for PCV, 548 for FEC and 565 for FAM. The number of SNPs passing the QC criteria was 41,676. The final dataset structure and summary statistics for the response variables used in the genome-based analyses are presented in Table 2.

2.4 Genomic prediction models

2.4.1 Parametric prediction models

In this study, the genomic best linear unbiased predictor (GBLUP), BayesA, BayesB, and Bayesian least absolute shrinkage and selection operator (BLASSO) models were employed using the BGLR (Bayesian generalized linear regression) package, available for R (Pérez & De Los Campos 2014). These models are well-documented methodologies for genomic selection, presenting conceptual differences concerning the prior assumptions assigned to marker effects. The GBLUP model can be described as follows:

$$\boldsymbol{y}^{*} = \boldsymbol{1_n}\mu + \boldsymbol{Zg} + \boldsymbol{e}$$

in which $\boldsymbol{y}^{*}$is the vector of response variables (EBV); $\boldsymbol{1_n}$ is an n × 1 vector of 1 s; μ represents the overall mean; $\mathbf{Z}$ is the incidence matrix relating the animals to the direct genomic breeding values (DGV); $\mathbf{g}$ is the vector of DGV, assumed to follow a normal distribution N(0, $\mathbf{G}\sigma_g^2$), where $\mathbf{G}$ is the marker-based genomic relationship matrix and σ_g^2 stands for the markers' genetic variance; and e is the vector of random residual effects, assumed to follow a normal distribution $N(0, \mathbf{I}\sigma_e^2)$, where $\mathbf{I}$ is an identity matrix and σ_e^2 is the residual variance. The **G**-matrix was calculated according to VanRaden (2008):

$$\boldsymbol{G} = \frac{(\boldsymbol{M}-\boldsymbol{P})(\boldsymbol{M}-\boldsymbol{P})^T}{2\sum_{j=1}^{m} p_j(1-p_j)}$$

in which **M** represents a matrix containing individual animal marker genotypes (coded as 0, 1, and 2 for the genotypes AA, AB and BB respectively) of dimension $n \times s$ (n = number of animals and s = number of SNPs) and **P** is a matrix with elements of column j coded as $2p_j$, where p_j is the estimated frequency of the allele B for the jth SNP.

BayesA was described by Meuwisse et al., (2001), and considers that the variance can be different for the segments of the genome. For this method, the prior distribution for the SNPs effects has a scaled-t density. BayesB also considers that the genetic variance explained by each locus can be different (Meuwissen et al., 2001). In summary, BayesB uses a two-component mixture prior, with a point of mass at zero and nonnull effects following a scaled-t distribution with prior probability π.

As presented by Meuwissen et al. (2001), Bayesian regression can be used when there are more predictors than observations (p>>n) as is the case with genomic data. Therefore, the estimates $\hat{\beta}$ are not unique, as the matrix $X'X$ does not have a complete rank. An interesting approach to the situation described above is the use of the LASSO regression method (Tibshirani, 1996), which solves the problem using the penalty function:

$$\hat{\beta}_{LASSO} = \underset{\beta}{\operatorname{argmin}} \left\| y - X\beta \right\|_2 + \lambda \left\| \beta \right\|_1 = \sum_{i=1}^{n} \left(y_i - x_i'\beta \right)^2 + \lambda \sum_{j=1}^{p} \left| \beta_j \right|$$

where $\underset{\beta}{\operatorname{argmin}}$ is the minimum argument that minimizes the model sum of squares, $\left\| y - X\beta \right\|_2 = \sum_{i=1}^{n} \left(y_i - x_i'\beta \right)^2$ is the residual sum of squares, x_i' is the i-th row of X, $\lambda \geq 0$ is the model regularization parameter, and $\left\| \beta \right\|_1 = \sum_{j=1}^{p} \left| \beta_j \right|$ is the sum of the absolute values of the marker effects β. In this model, the penalty is applied to the sum of the model parameters.

LASSO has its Bayesian version called Bayesian least absolute shrinkage and selection operator (BLASSO) (Park; Casella, 2008). Tibshirani (1996) reported that LASSO estimates can be interpreted as the mode of the β posterior distribution of models with Gaussian distribution for likelihood, $p\left(y \mid \beta, \sigma_e^2\right) = \Pi_{i=1}^{n} N\left(y_i \mid \sum_{j=1}^{p} x'_{ij}\beta_j, \sigma_e^2\right)$ and the prior distribution of marker effects as the product of p-independent double-exponential distributions, with zero mean, such that $p(\beta \mid \lambda) = \Pi_{j=1}^{p}(\lambda/2)\exp\left(-\lambda\left|\beta_j\right|\right)$, being $\hat{\beta}_{LASSO} = \underset{\beta}{\mathrm{argmax}}\ p\left(\beta \mid y, \sigma^2, \tau\right)$

The parametric models were implemented using a Gibbs sampling algorithm with 50,000 iterations, considering a burn-in period of 20,000 and a thinning interval of 10 iterations.

2.4.2 Bayesian regularized artificial neural networks

Artificial neural networks (ANN) are connectionist models that use sophisticated nonlinear techniques, capable of modeling complex functions. ANN is inspired by the human nervous system, and the learning acquired from the dataset is stored in synaptic connections found between artificial neurons. Linking with the linear models, to facilitate the understanding of ANN, the synapses would be the weights of linear regression models, the neurons of the input layer would be the predictor variables and the neurons of the output layer would be the response variable. In the present study, a single hidden layer feed-forward ANN was used, formed by an input layer that receives the explanatory variables, one hidden layer containing the neurons (processing units), and an output layer (Fig. 1).

Each input neuron *j* (*j*=1, 2, ..., s input neurons) connects to each neuron k of the adjacent middle layer via synaptic weights (w_{jk}, *k* = 1, 2, …, *n* hidden layer neuron), and a model intercept neuron (b_k) connected to hidden neurons and output layer neurons (Fig. 1). The response generated in each neuron *k* of the hidden layer is given by y_k = g(a_k), where a_k is the synaptic function, which is the linear combination between the normalized variables *x* weighted by the synaptic weights w_{jk} and added to the intercept b_k. The synaptic function a_k is given by:

$$a_k = b_k + \sum_{j=1}^{n} w_{jk} x_j$$

The value generated by a_k is transformed in the hidden layer through the activation function (or transfer function) g(a_k). A widely used nonlinear activation function is the hyperbolic tangent, where the response amplitude is in the interval (-1,1). This function is ideal for predicting genomic values since they can assume positive and negative values. Here, the hyperbolic tangent activation function was adopted:

$$g(a_k) = \frac{e^{a_k} - e^{-a_k}}{e^{a_k} + e^{-a_k}}$$

where *e* is the neperian number.

The response activated in the hidden layer neuron is propagated to the output layer neuron as:

$$a_l = \sum_{k=1}^{m} w_k g_k \left(b_k + \sum_{j=1}^{n} w_{kj} x_j \right) + b_l$$

where a_l is the input value in the output layer neuron and b_l is the output layer neuron intercept. After that, a_l is activated by the activation function g(a_l), which is a linear activation function.

Different numbers of neurons in the hidden layer (three, BRANN_3; and four, BRANN_4) were tested to assess the ANN architecture with the best predictive ability, and 1,000 epochs were adopted in the analyses. The analyses of neural networks with Bayesian regularization were performed using the BRNN package (Pérez-Rodríguez et al., 2013) of the R program (R Development Core Team, 2018).

2.5 Cross-validation and comparison criteria

The models' predictive ability was evaluated using a 5-fold cross-validation technique, in which the full dataset was randomly divided into five folds with approximately equal sizes. The genotypes and phenotypes from four folds were used as training data for fitting the models, whereas the fold held out from the training set was used to test the model's predictive ability. This process was performed five times, using different folds for training in each round and omitting a fold for validation of the models. The first adopted criterion for comparing the predictive ability among models was prediction accuracy (ACC). The ACC was computed as the linear correlation between observed and estimated values, *r*(EBV, GEBV). The second prediction ability criterion was the mean squared error (MSE), computed as:

$$\frac{1}{N_{val}}\sum_{1}^{N_{val}}(y_{val}-\hat{y}_{val})^2$$

in which N_{val} represents the number of animals in the *k*th cross-validation testing fold (k = 1, 2, ..., 5), $\hat{y}_{val}$ is the vector of estimated values (GEBV) and y_{val} is the vector of observed values (EBV). The predictive ability metrics (ACC and MSE) are presented as the averages of the values computed for each cross-validation testing fold.

3. RESULTS AND DISCUSSION

3.1 Predictive ability of models

The average ACC varied from moderate to high according to the trait and models considered, ranging between 0.418 and 0.546 for PCV, between 0.646 and 0.793 for FEC, and between 0.414 and 0.519 for FAM. The MSE presented higher variations between parametric models and the BRANN method, with differences ranging from 0.130 to 0.325 for FEC, and from 0.026 to 0.058 for FAM, and smaller differences for PCV (0.0002 to 0.0003) (Table 3).

We expected to obtain higher prediction accuracies for traits with higher heritability.

However, prediction accuracies for PCV ($h^2 = 0.30 \pm 0.04$) were lower than those obtained for FEC ($h^2 = 0.21 \pm 0.03$) (Table 3). The same was observed by Alves et al. (2020) for reproductive traits in Nellore cattle reporting higher prediction accuracy using different models for early pregnancy ($h^2 = 0.30 \pm 0.01$) compared with scrotal circumference ($h^2 = 0.43 \pm 0.01$). Some factors might influence genomic evaluation accuracy other than the accuracy of SNP effect estimation that depends on heritability magnitudes and the size (N) of the reference population, such as the structure of the genome and the genetic architecture of the trait, besides the type of response variable, the relationship between training and validation populations and training population design (Lorenz & Smith 2015; Boichard et al., 2016; Fernandes Junior et al., 2016; Naderi et al., 2016; Mehrban et al., 2017; Zhu et al., 2021).

Additionally, the accuracy of genomic evaluation may be also affected by the statistical method, marker density, and MAF (Atefi et al., 2016; Zhang et al., 2019, Zhu et al., 2021). The investigation of factors that influence the accuracies of genomic evaluation is necessary, especially when adopting new parameters, for example, by increasing the number of markers, the computational time also increases considerably, especially for Bayesian alphabet methods, due to the increase in the number of unknown parameters that need to be estimated. Furthermore, for simple traits that are controlled by one or several genes with large effects, higher marker density can cause a lower prediction accuracy, whereas, for traits with complex genetic architecture, the prediction accuracy can be improved by increasing the marker density (Zhang et al., 2019).

Considering the different models, small differences were observed in terms of predictive ability between the parametric genome-enabled prediction models for all traits (Table 3). This finding agrees with the literature, which has reported small differences in accuracies between GBLUP and Bayesian regression models (Mehrban et al., 2017; Alves et al., 2020; Lopes et al., 2020; Kjetså et al., 2022). Some characteristics belonging to the data may limit the observation of greater differences in prediction accuracy between the models, such as the size of the linkage disequilibrium, the effective population size, insufficient sample size to infer a large number of coefficients (n<p) and the complex nature of the traits studied (a large number of small-effect QTLs) (Howard et al., 2014; Mehrban et al., 2017; Alves et al., 2020).

The BRANN method showed major differences in predictive abilities, providing the poorest predictive skill for the studied traits (Table 3). The GBLUP, BayesA, BayesB, and BLASSO models provided better accuracies than the BRANN_3 method, increasing the ACC by around 23.16, 23.16, 23.44 and 22.88% for PCV, and 18.54, 18.54, 18.64, 18.54% for FEC, respectively. Likewise, there was an increase of ACC of 20.23% for FAM, comparing all parametric models and BRANN_4. The BRANN_3 model showed lower ACC than the BRANN_4 model in 2.34 and 0.31% for PCV and FEC, respectively, and higher ACC than the BRANN_4 model in 0.96% for FAM. The lowest MSE values were also observed for the parametric models for all studied traits. Accordingly, the parametric models are considered less biased than the BRANN method. Considering our results, the BRANN method did not present competitive predictive ability compared with the parametric approaches, presenting lower prediction accuracies and higher MSE than parametric models for all traits evaluated.

Previous studies demonstrated that the performance of ML methods can vary between different species and traits (Yan & Wang, 2022). In addition, studies suggested that ML methods can produce competitive predictions to classical animal breeding models like parametric models (e.g., GBLUP) (Naderi et al., 2016; Nayeri et al., 2019). Nonetheless, the simulation study reported that ML methods tend to present worse behavior when the traits are affected by many loci with small effects (Naderi et al. 2016), which seems to be the case in the present study. Other factors such as the size of the training set, the magnitude of heritability, the size of the linkage disequilibrium in the population, and the process of fitting parameters can also impact the performance of ML models (Montesinos-López et al., 2021).

The results found by the BRANN method can be explained by the simple network architecture adopted in this study, consisting of a single hidden layer with three or four neurons. Meher et al. (2022), evaluating the performance of different methods for genomic prediction, also found that artificial neural networks did not perform best for any trait. However, the authors identified strategies, such as feature selection and seeded starting weights, that increased the ANN's performance to near the level of other algorithms (Bayesian and BLUP alphabets). Possibly, one could achieve further improvement in predictive performance with different strategies, as mentioned by Meher et al. (2022), in addition, to other regularization techniques, such as dropout and early stop, and other optimization algorithms. Alves et al. (2020) also mentioned other approaches to improvement of ANN's performance, such as increasing the number of hidden layers and neurons and using other activation functions.

Overall, the empirical results from this study indicate that parametric models are suitable for genome-enabled prediction of indicator traits of resistance to gastrointestinal nematodes in Santa Inês sheep. In this regard, the application of BRANN for predicting indicator traits of resistance to gastrointestinal nematodes in this sheep population should be further evaluated and elucidated in future studies, especially different architectures and deep learning techniques, in addition, other ML methods, such as support vector regression and random forest.

3.2 Computational time

Regarding the mean computational time, the GBLUP (11.12 mins) was the most efficient method evaluated (Table 4). In contrast to the methods of the Bayesian alphabet that have a greater number of unknown parameters, the GBLUP has only n+1 unknown parameters, due to its architecture being based on (semi)definite (G) symmetric positive kernels to find the similarity between the animals (Alves et al., 2020). This finding agrees with Meher et al. (2022), where the GBLUP is computationally efficient and meets the requirements predicting especially for traits with complex genetic architecture.

Considering the different methods, GBLUP demands considerably less time than BayesA (380.12 mins), BayesB (425.04 mins), BLASSO (395.10 mins) BRANN_3 (510.06 mins), and BRANN_4 (1,301.85 mins) methods. The less time-efficient methods were BRANN_3 and BRANN_4 (Table 4). The high computational burden in ANNs methods is due to increases in the number of hidden neurons for the ANNs and, for this reason, BRANN_4 presented 60.82% less computational efficiency than BRANN_3. The machine used for implementing the models for PCV, FEC and FAM had an Intel Xeon Bronze 3104 with CPU of 6 cores and 1.70 GHz frequency, and 120GB total RAM memory + 15x HPE 8GB 1Rx8 PC4-2666V-R.

4. CONCLUSION

The results indicate that parametric models (GBLUP, BayesA, BayesB and BLASSO) are suitable methods for genome-enabled prediction of indicator traits of resistance to gastrointestinal nematodes in Santa Inês sheep. Due to the small difference in accuracy found between them, the use of the GBLUP model is recommended due to its lower computational cost and the possibility of incorporating non-genotyped animals into the analysis using single-step procedures.

5. REFERENCES

Aboshady H.M., Stear M.J., Johansson A., Jonas E. & Bambou, J.C. (2020) Immunoglobulins as Biomarkers for Gastrointestinal Nematodes Resistance in Small Ruminants: A systematic review. *Scientific Reports* **10**, 7765, doi: 10.1038/s41598-020-64775-x

Alves A.A.C., Espigolan R., Bresolin T., Costa R.M., Fernandes Júnior G.A., Ventura R.V., Carvalheiro R. & Albuquerque L.G. (2020) Genome-enabled prediction of reproductive traits in Nellore cattle using parametric models and machine learning methods. *Animal Genetics* **52**, 32-46. doi: 10.1111/age.13021

Atefi A., Shadparvar A.A. & Hossein-Zadeh N.G. (2016) Comparison of whole genome prediction accuracy across generations using parametric and semi parametric methods. Acta Scientiarum *Animal Sciences* **38**, 447-453, doi: 10.4025/actascianimsci.v38i4.32023

Boichard D., Ducrocq V., Croiseau P. & Fritz S. (2016) Genomic selection in domestic animals: Principles, applications and perspectives. *Comptes Rendus Biologies* **339**, 274-277, doi: 10.1016/j.crvi.2016.04.007

Brito L.F., Clarke S.M., McEwan J.C., Miller S.P., Pickering N.K., Bain W.E., Dodds K.G., Sargolzaei, M. & Schenkel F.S. (2017) Prediction of genomic breeding values for growth, carcass and meat quality traits in a multi-breed sheep population using a HD SNP chip. *BMC Genetics* **18**, 7, doi: 10.1186/s12863-017-0476-8

Castagna F., Piras C., Palma E., Musolino V., Lupia C., Bosco A., Rinaldi L., Cringoli G., Musella V., Britti D. (2021) Green Veterinary Pharmacology Applied to Parasite Control: Evaluation of Punica granatum, Artemisia campestris, Salix caprea Aqueous Macerates against Gastrointestinal Nematodes of Sheep. *Veterinary sciences* **8**, 237. doi: 10.3390/vetsci8100237

De Los Campos G., Gianola D., Rosa G.J.M., Weigel K.A. & Crossa J. (2010) Semiparametric genomic enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genetics Research* **92**, 295-308, doi: 10.1017/S0016672310000285

Ehret A., Hochstuhl D., Gianola D. & Thaller G. (2015) Application of neural networks with back-propagation to genome-enabled prediction of complex traits in Holstein-Friesian and German Fleckvieh cattle. *Genetics Selection Evolution* **47**, 1–9, doi: 10.1186/s12711-015-0097-5

Ferreira G.F., Ciappesoni G., Castells D., Amarilho-Silveira F., Navajas E.A., Giorello D., Banchero G. & De Barbieri I. (2021) Feed conversion efficiency in sheep genetically selected for resistance to gastrointestinal nematodes. *Animal Production Science* **61**, 754-760, doi: 10.1071/AN20121

Fernandes Junior G.A., Rosa G.J.M., Valente B.D., Carvalheiro, R., Baldi, F., Garcia, D.A., Gordo D.G.M., Espigolan R., Takada L., Tonussi R.L., de Andrade W.B.F., Magalhães A.F.B., Chardulo L.A.L., Tonhati H. & Albuquerque L.G. (2016) Genomic prediction of breeding values for carcass traits in Nellore cattle. *Genetics Selection Evolution* **48**, 1–8, doi: 10.1186/s12711-016-0188-y

Gianola D. & Sorensen D. (2002) Likelihood, Bayesian, and MCMC methods in quantitative genetics (ed.1), 740p. Springer-Verlag, New York.

Gianola D, De Los Campos G (2008) Inferring genetic values for quantitative traits non-arametrically. *Genetics Research* **90**, 525-540, doi: 10.1017/S0016672308009890

González-Recio O., Rosa G.J.M. & Gianola D. (2014) Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. *Livestock Science* **166**, 217-231, doi: 10.1016/j.livsci.2014.05.036.

Gordon H.M. & Whitlock H.V. (1939) A new technique for counting nematode eggs in sheep feces. *Journal of Scientific & Industrial Research* **12**, 50-52.

Greeff J. & Karlsson L. (2020) Production benefits of breeding for worm resistance in Merino sheep. *Animal Production Science* **60**, 1643-1653, doi: 10.1071/AN19368.

Habier D., Tetens J., Seefried F-R., Lichtner P. & Thaller G. (2010) The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genetics Selection Evolution* **42**, 1-5.

Henderson, C.R. (1975) Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31**, 423-447.

Heslot N, Yang H-P, Sorrells ME, Jannink J-L (2012) Genomic selection in plant breeding. A comparison of models. *Crop Science* **52**, 146, doi: 10.2135/crops ci201 1.06.0297

Howard R., Carriquiry A.L. & Beavis W.D. (2014) Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. *G3 (Bethesda)* **4**, 1027-46, doi: 10.1534/g3.114.010298.

Kjetså M.V., Gjuvsland A.B., Nordbø Ø., Grindflek E. & Meuwissen T. (2022) Accuracy of genomic prediction of maternal traits in pigs using Bayesian variable selection methods. *Journal of Animal Breeding and Genetics* **139**, 654– 665, doi: 10.1111/jbg.12729

Knol E.F., Nielsen B. & Knap P.W. (2016) Genomic selection in commercial pig breeding. *Animal Frontiers* **6**, 15–22, doi: 10.2527/af.2016-0003

Kuiseu J., Zinsou F.T.A., Olounlade P.A., Alowanou G.G., Adenile A.D., Dansou C.C., Hounzangbe-Adote S., Babayemi O.J. & Edorh P.A. (2021) Prevalence, effects and alternative control methods of *Haemonchus contortus* in small ruminants: A review. *Journal of Veterinary Medicine and Animal Health* **13**, 84-97, doi: 10.5897/JVMAH2020.0868.

Lopes F.B., Magnabosco C.U., Passafaro T.L., Brunes L.C., Costa M.F.O., Eifert E.C., Narciso M.G., Rosa G.J.M., Lobo R.B. & Baldi F. (2020) Improving genomic prediction accuracy for meat tenderness in Nellore cattle using artificial neural networks. *Journal of Animal Breeding and Genetics* **00**, 1-11, doi: 10.1111/jbg.12468

Lorenz A.J. & Smith K.P. (2015) Adding genetically distant individuals to training populations reduces genomic prediction accuracy in barley. *Crop Science* **55**, 2657–67, doi: 10.2135/cropsci2014.12.0827

Mavrot F., Hertzberg H. & Torgerson P. (2015) Effect of gastro-intestinal nematode infection on sheep performance: a systematic review and meta-analysis. *Parasites Vectors* **8**, 557, doi: 10.1186/s13071-015-1164-z.

Meher P.K., Rustgi S. & Kumar A. (2022) Performance of Bayesian and BLUP alphabets for genomic prediction: analysis, comparison and results. *Heredity* **128**, 519–530. doi: 10.1038/s41437-022-00539-9

Mehrban H., Lee D.H., Moradi M.H., Ilcho C., Naserkheil M. & Ibanez-Escriche N. (2017) Predictive performance of genomic selection methods for carcass traits in Hanwoo beef cattle: impacts of the genetic architecture. *Genetics Selection Evolution* **49**, 1–13, doi: 10.1186/s12711-016-0283-0

Meuwissen T.H.E., Hayes B.J. & Goddard M.E. (2001) Prediction of total genetic value using genome wide dense marker maps. *Genetics* **157**, 1819-1829.

Meuwissen T.H.E., Hayes B.J. & Goddard M.E. (2013) Accelerating improvement of livestock with genomic selection. *Annual Review of Animal Biosciences* **1**, 221–37. doi: 10.1146/annurev-animal-031412-103705

Misztal I. (2008) Reliable computing in estimation of variance components. *Journal of Animal Breeding and Genetics* **125**, 363-370.

Montesinos-López O.A., Montesinos-López A., Pérez-Rodríguez P., Barrón-López J.A., Martini J.W.R., Fajardo-Flores S.B., Gaytan-Lugo L.S., Santana-Mancilla P.C. & Crossa, J. (2021) A review of deep learning applications for genomic selection. *BMC Genomics* **22**, 19, doi: 10.1186/s12864-020-07319-x

Mrode R., Tarekegn G.M., Mwacharo J.M. & Djikeng, A. (2018) Invited review: Genomic selection for small ruminants in developed countries: how applicable for the rest of the world? *Animal* **12**, 1333-1340, doi: 10.1017/S1751731117003688.

Mulder H.A. & Rashidi, H. (2017) Selection on resilience improves disease resistance and tolerance to infections. *Journal of Animal Science* **95**, 3346-3358, doi: 10.2527/jas.2017.1479.

Naderi S., Yin T. & Konig S. (2016) Random Forest estimation of genomic breeding values for disease susceptibility over different disease incidences and genomic architectures in simulated cow calibration groups. *Journal of Dairy Science* **99**, 7261–73, doi: 10.3168/jds.201610887

Nayeri S., Sargolzaei M. & Tulpan, D. (2019) A review of traditional and machine learning methods applied to animal breeding. *Animal Health Research Reviews* **20**, 31-46, doi: 10.1017/S1466252319000148.

Park T. & Casella G. (2008) The bayesian Lasso. *Journal of American Statistical Association* **103**, 681-686.

Pérez P. & De Los Campos G. (2014) Genome-wide regression & prediction with the BGLR statistical package. *Genetics* **198**, 483–495, doi: 10.1534/genetics.114.164442

Pérez-Rodríguez P., Gianola D., Weigel K.A.; Rosa G.J.M.; Crossa J. 2013 Technical Note: An R package for fitting Bayesian regularized neural networks with applications in animal breeding. *Journal of Animal Science* **91**, 3522-3531, doi: 10.2527/jas.2012-6162

Schalm O.W., Jain N.C. & Carrol E.J. (1975) Veterinary Haematology (ed. 3), pp. 197-199. Lea and Febiger Publication, Philadelphia.

Selvam R. (2021) Estimation of genetic parameters of traits for evaluation of resistance to gastrointestinal nematode infection in Vembur sheep. *The Pharma Innovation Journal* **10**, 686-688.

Selvam R. (2022) Genetic Parameter Analysis of Indicator Traits of Gastrointestinal Nematode Infection in Sheep Breeds. *International Journal of Bio-resource and Stress* **13**, 17-21.

Szewc M., Waal T. & Zintl A. (2021) Biological methods for the control of gastrointestinal nematodes. *Veterinary Journal* **268**, 105602, doi: 10.1016/j.tvjl.2020.105602

Tibshirani R. (1996) Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B (Methodological)* **58**, 267-288.

Van Wyk J.A. & Bath G.F. (2002) The FAMACHA system for managing haemonchosis in sheep and goats by clinically identifying individual animals for treatment. *Veterinary Research* **33**, 509-29, doi: 10.1051/vetres:2002036.

VanRaden P.M. (2008) Efficient methods to compute genomic predictions. *Journal of Dairy Science* **91**, 4414–4423, doi: 10.3168/jds.2007-0980.

Voss-Fels K.P., Cooper M. & Hayes B.J. (2019) Accelerating crop genetic gains with genomic selection. *Theoretical and Applied Genetics* **132**, 669–686, doi: 10.1007/s00122-018-3270-8

Xu Y., Liu X., Fu J., Wang H., Wang J., Huang C., Prasanna B.M., Olsen M.S., Wang G. & Zhang A. (2020) Enhancing Genetic Gain through Genomic Selection: From Livestock to Plants. *Plant Communications* **1**, 100005, doi: 10.1016/j.xplc.2019.100005

Yan J. & Wang X. (2022) Machine learning bridges omics sciences and plant breeding. *Trends in Plant Science* Article 101890, doi: 10.1016/j.tplants.2022.08.018

Zhang H., Yin L., Wang M., Yuan X. & Liu X. (2019) Factors Affecting the Accuracy of Genomic Selection for Agricultural Economic Traits in Maize, Cattle, and Pig Populations. *Frontiers in Genetics* **10**, 189, doi: 10.3389/fgene.2019.00189

Zhu S., Guo T., Yuan C., Liu J., Li J., Han M., Zhao H., Wu Y., Sun W., Wang X., Wang T., Liu J., Tiambo C.K., Yue Y. & Yang B. (2021) Evaluation of Bayesian alphabet and GBLUP based on different marker density for genomic prediction in Alpine Merino sheep. *G3 Genes|Genomes|Genetics* **11**, jkab206, doi: 10.1093/g3journal/jkab206

Tables

Table 1 Additive (σ_a^2), permanent environment (σ_{pe}^2), and residual (σ_e^2) variance components and heritability (h^2) estimates for indicator traits of resistance to gastrointestinal nematodes in Santa Inês sheep.

Trait[1]	$\sigma_a^2 (SE)$	$\sigma_{pe}^2 (SE)$	$\sigma_e^2 (SE)$	$h^2\ (SE)$
PCV (%)	0.0006 (9.04 x 10^{-5})	0.0003 (6.71 x 10^{-5})	0.0011 (6.71 x 10^{-5})	0.30 (0.04)
FEC (eggs/g)	0.37 (0.05)	0.09 (0.04)	1.26 (0.03)	0.21 (0.03)
FAM (score)	0.12 (0.04)	0.05 (0.03)	0.38 (0.12)	0.22 (0.05)

[1] PCV, packed cell volume; FEC, fecal egg count; FAM, Famacha© method.

Table 2 Descriptive statistics for the estimated breeding values (EBV) used in the genomic analyses.

Trait[1]	Genotyped animals			EBV		
	Males	Females	Total	Mean (SD)	Minimum	Maximum
PCV	102	449	551	0.003 (0.02)	-0.07	0.05
FEC	102	446	548	-0.17 (0.58)	-1.57	1.01
FAM	103	462	565	-0.06 (0.19)	-0.61	0.56

[1] PCV, packed cell volume; FEC, fecal egg count; FAM, Famacha© method.

Table 3 Prediction accuracy (ACC), mean squared error (MSE) and standard deviations (SD) for Packed cell volume (PCV), Fecal egg count (FEC) and Famacha© method (FAM) were obtained with different genome-enabled linear prediction models and artificial neural networks in a 5-fold cross-validation technique.

Trait	**Model**[1]	**ACC (SD)**	**MSE (SD)**
PCV (%)	GBLUP	0.544 (0.06)	0.0002 (3.72×10^{-5})
	BayesA	0.544 (0.06)	0.0002 (3.72×10^{-5})
	BayesB	0.546 (0.06)	0.0002 (3.78×10^{-5})
	BLASSO	0.542 (0.06)	0.0002 (3.62×10^{-5})
	BRANN_3	0.418 (0.06)	0.0003 (3.44×10^{-5})
	BRANN_4	0.428 (0.06)	0.0003 (3.44×10^{-5})
FEC (eggs/g)	GBLUP	0.793 (0.03)	0.130 (0.004)
	BayesA	0.793 (0.03)	0.130 (0.004)
	BayesB	0.794 (0.03)	0.130 (0.004)
	BLASSO	0.793 (0.03)	0.130 (0.004)
	BRANN_3	0.646 (0.05)	0.325 (0.04)
	BRANN_4	0.648 (0.05)	0.322 (0.05)
FAM (score)	GBLUP	0.519 (0.05)	0.026 (0.0008)
	BayesA	0.519 (0.04)	0.026 (0.0008)
	BayesB	0.519 (0.04)	0.026 (0.0008)
	BLASSO	0.519 (0.04)	0.026 (0.0007)
	BRANN_3	0.418 (0.11)	0.056 (0.0131)
	BRANN_4	0.414 (0.12)	0.058 (0.0152)

[1] GBLUP, Genomic best linear unbiased prediction; BLASSO, Bayesian least absolute shrinkage and selection operator; BRANN_3 and BRANN_4, Bayesian regularized artificial

neural networks using 3 and 4 hidden neurons, respectively.

Table 4 Average computation time (in minutes) to complete all iterations of the 5-fold cross-validation technique according to the different genome-enabled prediction methods.

Trait[1]	Method[2]					
	GBLUP	BayesA	BayesB	BLASSO	BRANN_3	BRANN_4
PCV	10.78	375.60	428.40	391.20	410.27	1,450.49
FEC	11.19	381.35	421.72	396.91	446.51	1,144.32
FAM	11.40	383.40	425.01	397.20	673.40	1,310.73
Average	11.12	380.12	425.04	395.10	510.06	1,301.85

[1] PCV, packed cell volume (%); FEC, fecal egg count (eggs/g); FAM, Famacha© method (score). [2] GBLUP, Genomic best linear unbiased prediction; BLASSO, Bayesian LASSO; BRANN_3 and BRANN_4, Bayesian regularized artificial neural networks using 3 and 4 hidden neurons, respectively.

Figure

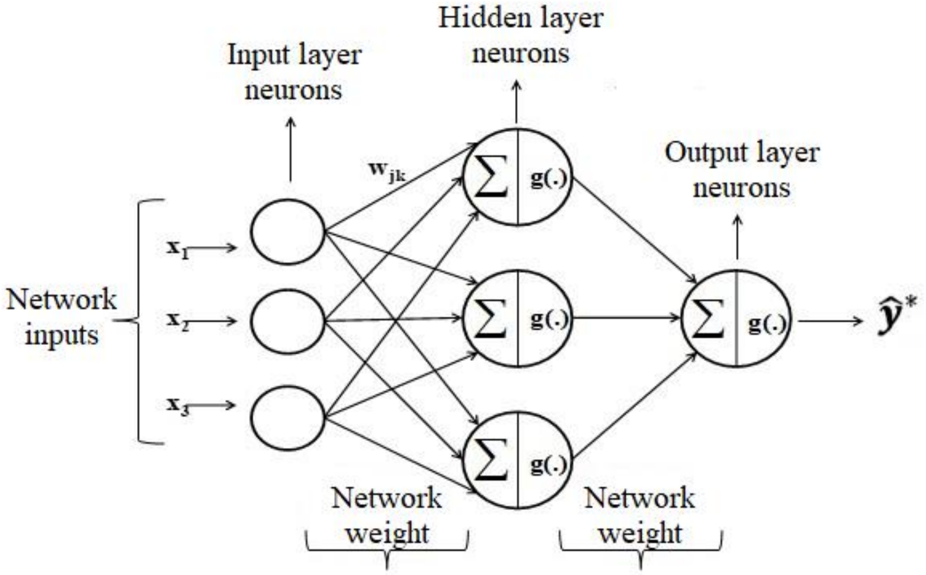

Figure 1 Topology of a single hidden layer feed-forward artificial neural network with three neurons in the input layer, three in the intermediate layer and one in the output layer. The input information is x_1, x_2 and x_3. The synaptic weights between neurons in adjacent layers are represented by w_{jk}, where j are the neurons in the input layer and k are the neurons in the intermediate layer; $\sum$ is the weighted sum given by $\sum_{j=1}^{n} x_k w_{jk}$, g(.) is the activation function and $\hat{\boldsymbol{y}}^*$ is the predicted value.

FINAL CONSIDERATIONS

The results of machine learning models to classify resistant, resilient, and susceptible animals suggest that the use of readily available records and easily-measured variables such as body weight, body condition score, farm, sex, age class, record month, and Famacha© score may contribute to the identification of susceptible animals and support management decisions at the farm level, potentially reducing economic losses due to parasitic infection with higher prevalence of *H. contortus* parasites. The best performances for classifying susceptible and resistant animals were achieved by multinomial logistic regression and linear discriminant analysis. The animals identified as resistant can also be incorporated as selection candidates into breeding programs for the genetic improvement of sheep populations.

The image analysis results indicate that it is possible to successfully predict Famacha© scores, particularly scores 2 to 4, in sheep through image analysis and the use of a random forest classification model using ocular conjunctiva images collected in farm conditions. Model validation excluding entire farms in cross-validation resulted in lower prediction quality, but it is closer to reality as the developed models are intended to be used across farms, including new ones, with different environments and management conditions. In the future, automation of the method for Famacha© score estimation based on image analysis may be considered to facilitate its application in sheep farming production.

Cluster analysis was found to be a valuable tool for data screening that allows the evaluation of only selection candidates based on their additive-genetic pattern for gastrointestinal nematode resistance. Estimated breeding values for fecal egg count were decisive in dividing the population into resistant, resilient, and susceptible animals, and it is important to include estimated breeding values for packed cell volume and Famacha© score to distinguish resistant from resilient animals effectively. The resistant cluster consisted of the most desirable animals to be used as selection candidates for genetic improvement of resistance to infection with gastrointestinal nematodes.

The results of genome-enabled prediction analysis indicate that parametric models (GBLUP, BayesA, BayesB, and BLASSO) are suitable methods for genome-enabled prediction of indicator traits of resistance to gastrointestinal nematodes in

Santa Inês sheep. The GBLUP model was recommended due to its lower computational cost and the possibility of incorporating non-genotyped animals into the analysis using single-step procedures, despite the small difference in accuracy found between the models.

Finally, the evaluation of the use of machine learning models for classifying resistance, resilience, and susceptibility to gastrointestinal nematodes and image analysis for classifying anemia using Famacha© scores in Santa Ines sheep is still in its early stages. There are also few reports in the literature on the use of genomic prediction of indicator traits of resistance to gastrointestinal nematodes, emphasizing the importance of the present work, which opens the way for further studies and future applicability at the field level.

In conclusion, the results of this study suggest that the use of machine learning models, image analysis, and genome-enabled prediction techniques can effectively contribute to the identification and classification of susceptible, resilient, and resistant animals in Santa Ines sheep populations with respect to gastrointestinal nematode infections. The use of readily available records and easy-to-measure variables in machine learning models, the ability of image analysis to predict Famacha© scores, and the suitability of parametric models for genome-enabled prediction of resistance indicator traits are all promising results. The resistant animals identified through these techniques can be considered for selection in breeding programs to improve the genetic resistance to gastrointestinal nematodes infections in sheep populations. Although the evaluation of these techniques is still incipient, this work highlights their potential and opens the way for further studies and future applicability in the field of livestock genomics.