Rede neural artificial para predição de brucelose bovina a partir de dados desbalanceados

Caio Donizetti Queiroz Alves
Departamento de Automática
Universidade Federal de Lavras
MG, Brasil
caio.alves2@estudante.ufla.br

Danton Diego Ferreira
Departamento de Automática
Universidade Federal de Lavras
MG, Brasil
danton@ufla.br

Danielle Abreu Fortunato
Departamento de Automática
Universidade Federal de Lavras
MG, Brasil
danielle.fortunato@estudante.ufla.br

Christiane Maria Barcellos Magalhães da Rocha
Departamento de Medicina Veterinária
Universidade Federal de Lavras
MG, Brasil
rochac@ufla.br

***Abstract*—The expressiveness of Brazilian livestock is unquestionable. According to data from the United States Department of Agriculture (USDA), in 2021 Brazil was the world's largest exporter of beef. Bovine brucellosis is one of the most worrying diseases for the sector. In Brazil, bovine brucellosis causes annual losses of around 448 million dollars. Several factors threaten the establishment of actions of the current animal defense programs in Brazil, the main ones being: lack of distinct guidelines for the diagnosis of brucellosis cases, infected animals remain asymptomatic when infected, extensive Brazilian territory and large number of herds. The use of Artificial Neural Networks (ANNs) can be very useful in health and epidemiological surveillance services, helping to screen properties with different risks for the disease. The objective of this work is the development of ANN with class balancing techniques and selection of variables via genetic algorithm, for the classification and segregation of bovine herds, regarding seroprevalence for brucellosis. Five ANNs were designed combining different approaches of class balancing technique and variable selection, in order to compare which approach would perform better results. The results showed that ANN combined with the variable selection technique, via Genetic Algorithms, and class balancing, is a promising approach.**

***Index Terms*—Artificial Neural Networks, brucellosis, class balancing, feature selection**

I. Introdução

O setor agropecuário é de fundamental importância econômica para o Brasil. Segundo dados da *United States Department of Agriculture* (USDA), em 2022 o Brasil foi o segundo maior produtor de carne bovina, sexto maior produtor de leite de vaca e o maior exportador de carne bovina, do *ranking* mundial. Ainda de acordo com dados da USDA, o Brasil detém o segundo maior rebanho bovino do planeta, ficando atrás apenas da Índia [1].

Diante da expressividade da pecuária Brasileira, a brucelose bovina é uma das doenças mais preocupantes para o setor agropecuário. A Food and Agriculture Organization (FAO) e a World Health Organization (WHO) consideram a brucelose como uma das zoonoses mais prevalentes, principalmente em países em desenvolvimento e de baixa renda [2]. Causada por Cocobacilos Gram-negativos, geralmente *Brucella abortus* e raramente por *Brucella melitensis* e *Brucella suis*, a brucelose bovina é uma das zoonoses economicamente mais importantes do mundo [3]. Abortos, perdas de bezerros recém-nascidos resultantes de abortos e natimortos, redução na produção de leite, impedimento na exportação e comércio de animais, são alguns exemplos de encargos econômicos que podem decorrer da brucelose bovina. Além do mais, há casos que faz-se necessário sacrificar ou realizar o abate sanitário de animais infectados [2]. No Brasil foram estimadas perdas anuais, atribuíveis à brucelose, em 448 milhões de dólares [4].

Para que programas de prevenção e erradicação da brucelose sejam eficazes devem ser realizados diagnósticos acurados e precisos da doença. Apesar da Organização Mundial da Saúde (OMS) relatar em sua ficha informativa que cerca de milhões de casos de brucelose são contabilizados todos os anos, estima-se que a taxa real de incidência ainda é 10 a 25 vezes maior do que o número declarado de casos. A falta de diretrizes distintas para o diagnóstico de casos de brucelose é uma das razões importantes por trás dessa condição [2]. O fato da maioria dos animais infectados permanecerem assintomáticos quando infectados e de não ser esperado lesões inflamatórias intensas, nem lesões patognomônicas causadas por Brucella spp., corroboram com a estatística [5]. Estas características somadas a extensa área territorial Brasileira, grande efetivo de rebanhos, e outros fatores ameaçam o estabelecimento de ações dos programas de defesa sanitária vigentes no Brasil. À vista disso, a eficiência e a estratégia dos programas de sanidade animal podem ser otimizadas através de métodos que auxiliem na triagem de propriedades com riscos diferenciados para a doença. Neste sentido, a utilização de inteligência com-

Os autores gostariam de agradecer à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) e a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo suporte financeiro.

putacional para classificação tem contribuído no diagnóstico de doenças infecciosas [6].

As Redes Neurais Artificiais (RNAs) são ferramentas de inteligência computacional utilizadas amplamente nas áreas da psicologia, robótica, biologia, ciência da computação, engenharia e em diversas outras áreas [7]. Por possuírem grande capacidade de previsão de processos complexos e não lineares aliados à facilidade e flexibilidade de implementação, nos últimos houve crescente uso de RNAs para fornecer modelos matemáticos de previsão e classificação de processos biológicos [8]–[10].

A sensibilidade, especificidade e/ou acurácia da classificação feita por RNAs correm risco de serem comprometidas caso as categorias de classe não estejam representadas suficientemente. É importante que essa representação aconteça de forma equilibrada para as classes envolvidas no problema. Em outras palavras, supondo um problema de classificação de duas classes, A e B, sendo estas de complexidade similar, é importante que o número de eventos (amostras) da classe A seja equivalente ao da classe B. Quando este equilíbrio não está disponível na base de dados, nestes casos podem ser adotadas técnicas de balanceamento das classes para o conjunto de dados de treinamento [11].

Outro ponto crítico que pode comprometer o resultado da classificação feita por RNAs é o número de variáveis do conjunto de dados. Grande número de variáveis não implica na melhoria de desempenho da RNA, provavelmente a redundância de informações reduzirá o desempenho da RNA [12]. Neste ponto, diversas abordagens podem ser empregadas visando reduzir a dimensionalidade dos dados. Os métodos estatísticos como Discriminante Linear de Fisher (FDR – *Fisher Discriminat Ratio*) e Análise de Componentes Principais (PCA - *Principal Component Analysis*) são bastante usados como ferramenta para pré-seleção de variáveis [13]. Os Algoritmos Genéticos (AGs) também podem ser aplicados como ferramenta para determinar dependências de informações e diminuir o número de variáveis em um conjunto de dados. O Algoritmo Genético escolherá um subconjunto de variáveis com a mesma discernibilidade do conjunto de variáveis iniciais, resultando em melhores desempenhos de classificação. Algoritmo Genéticos são programas de computador empregados para solucionar problemas complexos e que evoluem de maneira semelhante à seleção natural [14], [15].

A classificação de rebanhos potencialmente positivos ou negativos para brucelose permite aperfeiçoamento das estratégias de diagnóstico e controle da doença. Deste modo, o uso de tecnologias que proporcionem esta classificação de forma automática é capaz de dar suporte aos órgãos de defesa sanitária, facilitando a triagem de propriedades e proporcionando melhor alocação de recursos humanos e financeiros. Tendo em vista esses benefícios este artigo tem como finalidade avaliar modelos de RNA com técnicas de balanceamento de classes e seleção de variáveis, para a classificação e segregação dos rebanhos bovinos quanto à soroprevalência para brucelose.

II. Fundamentação Teórica

A. Brucelose

A brucelose é causada por bactérias do gênero Brucella spp. e provoca uma doença grave, crônica e debilitante em humanos. Mesmo diante de sua gravidade, esta zoonose ainda é muito negligenciada. A transmissão da brucelose ao ser humano pode ocorrer através do contato direto com secreções de animais infectados, inoculação acidental durante programas de vacinação, ou por transmissão aérea para o vacinador, veterinários, laboratoristas, trabalhadores de matadouros e trabalhadores de campo [16].

A bactéria Brucella spp. pode viver por um longo tempo no solo, água, pastagem e esterco. Pode se espalhar facilmente através de uma excreção profusa de organismos na placenta, fluidos fetais e secreções vaginais dos animais. A excreção de Brucella no meio ambiente é um fator extremamente preocupante para a saúde pública das famílias rurais, uma vez que não existem vacinas que possam prevenir a brucelose em humanos atualmente [16].

Além dos grandes danos para a saúde humana, a brucelose causa alto prejuízo econômico para o setor pecuário, seja na cadeia do corte ou do leite. As perdas econômicas podem decorrer de custos diretos (redução da produção de carne e leite, aumento da mortalidade, aumento da infertilidade, entre outros) ou de custos indiretos (vacinação, descarte, abate sanitário, entre outros). De acordo com [17], estudo realizado no Brasil em 2013 indicaram que a brucelose bovina acarretava uma perda de US$ 2,10 por animal. Observou-se também que cada aumento ou diminuição de 1% na prevalência da doença, aumenta ou decresce, respectivamente, o ônus econômico em US$ 0,37 por animal.

As estratégias eficazes de controle da brucelose bovina incluem vigilância, prevenção da transmissão e controle do reservatório da infecção por diferentes métodos, incluindo o descarte. Porém, as dificuldades em realizar o diagnóstico clínico, devido a sintomas inespecíficos, em localizar os animais infectados com brucelose, em conter ou regular o movimento e compra de animais sem teste para brucelose, e falta de educação e interesse dos agricultores em relação à doença, atrapalham o controle e erradicação [2], [4].

B. Redes Neurais Artificiais

Uma Rede Neural Artificial é uma rede de neurônios artificial desenvolvida para modelar como o cérebro humano reconhece novos objetos a partir de um aprendizado supervisionado anterior. Cada neurônio equivale a uma regressão logística. O conceito de neurônio artificial foi introduzido pela primeira vez por McCulloch e Pitts em 1943. Logo depois, o primeiro modelo perceptron de RNA foi desenvolvido [18], [19].

Atualmente existem diversas estruturas de RNAs, a mais utilizada em aplicações de classificação é a perceptron de múltiplas camadas (MLP - *Multilayer Perceptron*). Geralmente uma RNA MLP contém uma camada de entrada, camadas ocultas, e uma camada de saída. As camadas de entrada e saída

contêm um número fixo de neurônios, que são habitualmente definidos com base na estrutura do problema e no conjunto de dados. Na RNA MLP, cada neurônio é uma unidade de algoritmo computacional simples que recebe múltiplas entradas. Após processar as entradas, o neurônio produz uma saída que será usada como entrada dos neurônios na camada seguinte, caso a sua camada não seja a camada de saída. Posteriormente ao calculo da saída, uma função de ativação pode ser usada para calcular a saída final do neurônio [20]. Com base na saída dos neurônios da camada de saída é criado a lógica de classificação.

C. Balanceamento de classes

Domínios desequilibrados levantam desafios significativos ao construir modelos para realizar predição. Comumente a classe com menor quantidade de dados costuma ser a de maior interesse do ponto de vista do aprendizado. A escassa representação dos casos mais importantes leva a modelos que tendem a ser mais focados nos exemplos da classe com maior número de registros, negligenciando os eventos raros [21].

Diversas abordagens foram desenvolvidas na última década para lidar com a classificação em situações em que as classes estão desbalanceadas. As técnicas a nível de dados são as mais utilizadas para balanceamento de classe. Estas técnicas reequilibram o espaço amostral para um conjunto de dados desequilibrado com a finalidade de aliviar o efeito da distribuição de classes distorcida durante o treinamento da RNA. Os métodos de reamostragem independem do classificador selecionado o que os tornam mais versáteis. As técnicas de reamostragem se dividem basicamente em três grupos segundo [22], sendo eles:

- ***Oversampling:*** elimina os danos da distribuição distorcida criando novas amostras para as classes minoritárias.
- ***Undersampling:*** elimina os danos da distribuição assimétrica, descartando as amostras da classe majoritária.
- **Métodos híbridos:** combinam métodos de *oversampling* e métodos de *undersampling*.

As estratégias usadas com maior frequência para criar ferramentas de *oversampling*, *undersampling* ou híbridos são métodos baseados em *cluster* (por exemplo: K-means), métodos baseados em distância (por exemplo: KNN [23]) e métodos evolutivos (por exemplo: Algoritmo Genético) [22].

A depender da abordagem de *oversampling* adotada pode-se provocar *overfitting* durante a fase de treinamento do classificador, uma vez que, o classificador é exposto às mesmas informações. O SMOTE *(Synthetic Minority Oversampling Technique)* é uma abordagem alternativa de *oversampling* que gera novos dados de uma maneira que visa minimizar esse problema. Os dados sintéticos gerados com a técnica SMOTE são criados ao longo do segmento de linha que liga exemplos de classe minoritária. Apesar de mitigar a ocorrência *overfitting*, o SMOTE tem a desvantagem de criar exemplos sintéticos da classe minoritária, desconsiderando os exemplos da classe majoritária, podendo então criar dados ruidosos e limítrofes. Diante desta problemática várias modificações do SMOTE foram propostas. O Borderline-SMOTE [24], SMOTE-ENN [25], SMOTE-RSB [26], MSMOTE [27], ADASYN [28], Safe-Level-SMOTE [29] e diversos outros, são exemplos dessas variações do SMOTE.

D. ADASYN

A questão fundamental do algoritmo ADASYN *(Adaptive Synthetic Sampling)* proposto por [28] é compensar as distribuições desbalanceadas alterando, de forma adaptativa, os pesos de diferentes exemplos minoritários. Para isso, o ADASYN utiliza uma distribuição de densidade como critério para decidir automaticamente o número de amostras sintéticas que precisam ser geradas para cada exemplo minoritário. Como resultado, os dados sintéticos adicionais são gerados para exemplos da classes minoritária que são mais difíceis de aprender [30].

De acordo com [28], o ADASYN não apenas reduz o viés de aprendizado introduzido pela distribuição de dados desequilibrada, como também pode mudar o limite de decisão, de forma adaptativa, para se concentrar nos exemplos da classe minoritária mais difíceis de aprender. Como resultado, o ADASYN melhora o aprendizado da distribuição de dados de duas formas: (1) reduzindo o viés causado pelo desequilíbrio de classe e (2) alterando o limite de decisão de classificação na direção de exemplos mais difíceis de aprender.

O ADASYN atinge seus objetivos primeiramente calculando-se a quantidade total de exemplos da classe minoritária a serem gerados de acordo com a equação (1):

$$G = (c_n - c_p) \cdot \beta, \tag{1}$$

na qual, c_n corresponde a quantidade de exemplos da classe majoritária, c_p a quantidade de exemplos da classe minoritária e $\beta \in [0, 1]$ especifica o nível de balanceamento após a criação dos exemplos sintéticos. $\beta = 1$ implica em um conjunto de dados totalmente balanceado. Então, para cada exemplo $p_i \in P$, em que P é o conjunto minoritário, encontra-se os K vizinhos mais próximos, com base na distância euclidiana no espaço v dimensional, e calcula-se a proporção r_i de exemplos da classe majoritária vizinhos de p_i conforme a equação (2):

$$r_i = \frac{\Delta_i}{K}, \tag{2}$$

sendo, Δ_i o número de exemplos nos K vizinhos mais próximos de p_i que pertencem à classe majoritária, portanto $r_i \in [0, 1]$. Calculado r_i, deve-se normalizá-lo de acordo com a equação (3), de modo que $\hat{r_i}$ é uma distribuição de densidade ($\sum_i \hat{r_i} = 1$).

$$\hat{r_i} = \frac{r_i}{\sum_{n=1}^{c_p} r_i}, \tag{3}$$

Por fim, é calculado o número de exemplos de dados sintéticos que precisam ser gerados (q_i) para cada exemplo minoritário p_i, para isso utiliza-se a equação (4):

$$q_i = \hat{r_i} \cdot G, \tag{4}$$

Para gerar os exemplos sintéticos para cada exemplo p_i deve-se repetir os seguintes passo para cada p_i:

1) Escolha aleatoriamente um exemplo $p_{zi} \in P$ dentre os K vizinhos mais próximos de p_i;
2) Gere o exemplo sintético s_i utilizando a equação (4):

$$s_i = p_i + (p_{zi} - p_i) \cdot \lambda, \tag{5}$$

onde, $(p_z i - p_i)$ é o vetor diferença no espaços de v dimensões, e λ corresponde a um número aleatório na qual $\lambda \in [0, 1]$.
3) Repita os passos 1 e 2 até que a quantidade de exemplos sintéticos gerados para p_i seja igual a q_i.

E. Algoritmo Genético

Algoritmo Genético é um método de busca heurística que pode ser utilizado para buscar uma solução ótima em espaços que são excessivamente expansivos. O Algoritmo Genético gera soluções para problemas de otimização com base na mecânica da genética natural e evolução biológica [31], [32]. As principais etapas de um Algoritmo Genético são basicamente:

1) **Criar população:** o primeiro passo é criar, de forma aleatória, os indivíduos na população. Cada indivíduo (cromossomo) na população representa uma solução candidata para o problema;
2) **Avaliar aptidão:** após criar a população é calculada a aptidão de cada indivíduo. A aptidão refere-se a qualidade do indivíduo;
3) **Seleção:** na etapa de seleção são escolhidos os indivíduos que irão se recombinar para a próxima geração;
4) **Cruzamento:** o operador de cruzamento recombina as características dos indivíduos escolhidos para gerar novos descendentes para a nova população;
5) **Mutação:** no cruzamento geralmente são produzidos descendentes muito semelhantes aos pais. Isso promove uma nova geração com baixa diversidade, o que pode acarretar convergência prematura do Algoritmo Genético para uma solução que não seja a ótima. O operador de mutação contorna esse problema alterando aleatoriamente o valor de algumas características nos descendentes;
6) **Atualizar população:** a população é atualizada com os novos descendentes gerados.

As etapas: avaliar aptidão, seleção, cruzamento, mutação e atualizar população, são repetidas por um número fixo de gerações ou até que uma condição de término seja satisfeita [32]. As etapas de seleção, cruzamento e mutação são ditas como operadores genéticos, visto que, nestes passos são simulados os eventos de genética natural e evolução biológica.

III. Metodologia

Para o desenvolvimento da RNA foram utilizados os dados do inquérito de brucelose do Estado de Minas Gerais. O banco de dados é composto pelas variáveis coletadas por meio dos questionários e pelos resultados dos testes de diagnóstico para a doença. A base de dados conta com dados de rebanhos de 2.185 propriedades, dentre eles, 2103 com diagnósticos negativos e 82 com diagnósticos positivos para brucelose. Os casos e controles de brucelose (rebanhos positivos e negativos) foram consideradas como variáveis dependentes, e as coletadas por meio das entrevistas, relacionadas às características e ao manejo geral e sanitário da propriedade, como variáveis independentes. Realizou-se análises para a validação do banco de dados. Variáveis que não apresentaram variações ou apresentaram muitos dados perdidos foram descartadas. As variáveis qualitativas foram codificadas e atribuiu-se a elas valores numéricos padronizados de acordo com as respostas dos questionários. As variáveis quantitativas foram normalizadas (padronizadas em uma dada escala numérica). Após pré-processamento o banco de dados passou a contar com 49 variáveis.

A. Oversampling

A base de dados possui um desiquilíbrio muito grande, visto que, a classe de rebanhos positivos é composta por apenas 82 propriedades, enquanto que a de rebanhos negativos conta com 2103 propriedades. Este desequilíbrio prejudica o treinamento da RNA. Para minimizar este problema foram aplicadas duas técnicas de *oversampling*: o algoritmo de gerações de dados sintéticos ADASYN e a técnica de replicação de dados [33].

Das 82 propriedades positivas foram selecionadas aleatoriamente 60 para que, em conjunto com as 2103 propriedades negativas, fossem expostas ao ADASYN. O intuito foi gerar aproximadamente 140 propriedades positivas sintéticas. Para que isto fosse possível configurou-se o ADASYN com valor de $\beta = 0,10$. β é um parâmetro usado para especificar o nível de equilíbrio após a geração dos dados sintéticos, onde, $\beta \in [0, 1]$, e $\beta = 1$ denota que um conjunto de dados totalmente balanceado será criado após o processo de generalização. Optou-se por utilizar o valor de β consideravelmente baixo visto que o aumento desse valor provocou redução acentuada no desempenho da RNA. Definiu-se também o valor de $K = 6$. K corresponde a quantidade de vizinhos mais próximos, com base na distância euclidiana no espaço v dimensional, que o algorítimo considerou para aplicar os cálculos necessários. No caso deste trabalho definiu-se $v = 49$, dado que o número de variáveis da base de dados original é igual a 49.

A técnica de replicação de dados constitui-se da replicação dos dados de 60 propriedades positivas selecionadas aleatoriamente.

As 22 propriedades positivas que não participaram das técnicas de *oversampling* foram destinadas para compor o conjunto de dados da validação da RNA.

B. Undersampling

Mesmo aplicando a técnica de *oversampling* ADASYN e de replicação de dados para amenizar o desequilíbrio das classes e melhorar o aprendizado da RNA, a base de dados ainda contava com um desequilíbrio considerável entre as classes. O número de propriedades positivas destinadas ao treinamento subiu para 200 utilizando o ADASYN e para 120 utilizando replicação, enquanto o número de propriedades negativas é 2103. Este desequilíbrio ainda poderia prejudicar o treinamento da RNA. Com o objetivo de equilibrar os

conjuntos que seriam destinados ao treinamento adotou-se também técnica de *undersampling* para classe de rebanhos negativos. Para garantir que o conjunto da classe de rebanhos negativos fosse composto por propriedades que incluíssem informações da diversidade da classe, ou seja, que represente bem a classe, as propriedades negativas foram divididas em 4 agrupamentos. A divisão foi obtida usando o algoritmo k-means.

O algoritmo k-means é um dos algoritmos mais usados para agrupamento. É um algoritmo simples e eficiente que na maioria dos problemas converge num número reduzido de iterações [34] [35]. A forma como é mais utilizado atualmente foi proposta por [36]. O K-means encontra a melhor divisão de um conjunto de dados em Q grupos (para este trabalho $Q = 4$), de maneira que a distância total entre os dados de um grupo e o seu respectivo centro, somada por todos os grupos, seja minimizada.

C. Seleção de variáveis com Algoritmo Genético

Objetivando diminuir o número de variáveis em um conjunto de dados, garantindo a mesma discernibilidade do conjunto de variáveis iniciais, para resultar em melhores desempenhos da RNA aplicou-se a seleção de variáveis através de Algoritmo Genético. Os parâmetros utilizados na implementação do Algorítimo Genético foram avaliados experimentalmente e estão dispostos na Tabela I.

TABLE I
PARÂMETROS UTILIZADOS NA EXECUÇÃO DO AG.

Parâmetro	Valor
Tamanho da população (p)	100
Número de gerações	200
Probabilidade de mutação	0,15
Probabilidade de cruzamento	0,85
Tipo de mutação	Uniforme
Tipo de cruzamento	Uniforme
Seleção	Torneio
Elitismo	4 indivíduos

A representação cromossomial foi feita de forma binária, sendo que cada gene g do cromossomo C indica o uso (1) ou não (0) da variável j na RNA. Desta forma o número de gene de cada cromossomo é igual a v.

$$C_i = [g_1, ..., g_{j=v}], \tag{6}$$

onde $i \in [1, p]$ e $j \in [1, v]$.

A qualidade de cada indivíduo, aferida na fase de avaliação da aptidão, diz respeito a acurácia da RNA utilizando o respectivo conjunto de variáveis indicado por C. A acurácia da RNA foi quantificada de acordo com (7):

$$Acuracia_i = \frac{VN + VP}{VN + VP + FN + FP}, \tag{7}$$

onde VP, FP, VN e FN são o número de propriedades positivas corretamente classificadas, negativas incorretamente classificadas, negativas corretamente classificadas e positivas incorretamente classificadas, respectivamente.

A seleção de variáveis com Algorítimo Genético foi executada posteriormente a aplicação das técnicas de *oversampling* na base de dados. Sendo assim, foram feitas duas seleções de variáveis com Algorítimo Genético, uma para cada base de dados estabelecida com as respectivas técnicas de *oversampling*.

D. Projeto da RNA

Neste trabalho foram utilizadas RNAs multicamadas MLP (*Multi Layer Perceptron*). As RNAs contavam com uma camada de entrada, uma camada intermediária e uma camada de saída. A camada intermediária foi composta por 30 neurônios. Adotou-se um único neurônio na camada de saída, tal que valores maiores do que zero indicam propriedades positivas para a Brucelose e valores menores do que zero indicam propriedades negativas. A função de ativação adotada foi a tangente hiperbólica. Os parâmetros da RNA foram definidos através de avaliação experimental.

O algoritmo *Scaled conjugate gradiente backpropagation algorithm*, proposto por [37], foi adotado para o treinamento da RNA. Foi escolhido por ser mais rápido no treinamento e por apresentar bons resultados para a base de dados utilizada.

A metodologia de treinamento e validação das RNAs foi realizada 100 vezes e, portanto, para cada abordagem 100 RNAs foram obtidas. Os resultados de desempenho no projeto (treino e teste) e na validação foram apresentados na forma de média ± desvio padrão. A cada execução, os dados foram sorteados e divididos em dois conjuntos, projeto (treino e teste) e validação. A composição dos conjuntos de projeto e validação foi feita de acordo com o método de *oversampling* empregado. Para os casos que foi empregado o algoritmo ADASYN, o conjunto de projeto contou com 200 propriedades positivas (140 sintéticas e 60 selecionadas aleatoriamente da base de dados) e 200 propriedades negativas. As 200 propriedades negativas foram concebidas através da seleção aleatória de 50 propriedades de cada um dos 4 grupos gerados pela técnica de *undersampling*.

Nos casos que utilizou-se a técnica de replicação dos dados, o conjunto de projeto contou com 120 propriedades positivas (60 propriedades que foram replicadas) e 120 propriedades negativas. As 120 propriedades negativas foram obtidas através da seleção aleatória de 30 propriedades de cada um dos 4 grupos gerados pela técnica de *undersampling*. A Figura 1 esquematiza as etapas seguidas para seleção das propriedades que compuseram os modelos das RNAs nas abordagens em que utilizou-se técnicas de *oversampling*.

Nas abordagens em que não utilizou-se técnicas de *oversampling*, o conjunto de projeto contou com 64 propriedades positivas e 64 propriedades negativas. As 64 propriedades negativas foram obtidas através da seleção aleatória de 16 propriedades de cada um dos 4 grupos gerados pela técnica de *undersampling*.

Dos conjuntos de projeto utilizado para treinamento das RNAs, 85% das propriedades foram usadas para treino e 15% para teste. O restante dos dados foram usados para validação. O critério de parada utilizado para o treinamento foi o *early stopping*, em que um número máximo de 200 épocas foi

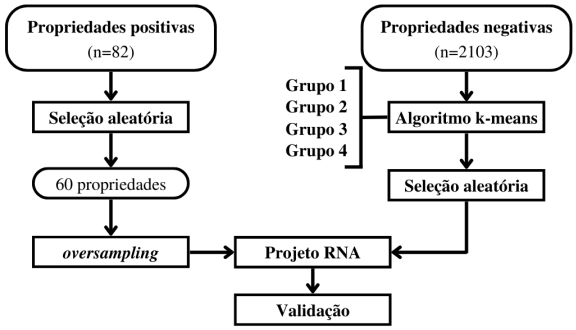

Fig. 1. Esquema de seleção de propriedades que compuseram os modelos das RNAs.

utilizado e 15% dos eventos de projeto foram utilizados para o teste [38]. Com isso, o *overfitting* (treinamento em excesso) foi evitado e garantiu-se a generalização da RNA.

As RNAs foram projetadas e validadas usando o software MatLab, que apresenta uma biblioteca bem completa para treinamento e testes de RNA.

E. Comparação entre os modelos de RNAs

Foram projetadas cinco abordagens a fim de se comparar o desempenho, sendo elas:

1) **RNA com técnica ADASYN e seleção de variáveis via AG;**
2) **RNA com técnica de replicação de dados e seleção de variáveis via AG;**
3) **RNA com técnica de replicação de dados e sem seleção de variáveis;**
4) **RNA com técnica ADASYN e sem seleção de variáveis; e**
5) **RNA sem técnica de *oversampling* e sem seleção de variáveis.**

Com o objetivo de se comparar as diferentes abordagens, o Teste Tukey com 95% de probabilidade foi utilizado para a comparação entre os valores de sensibilidade e especificidade das RNAs. Para o cálculo dos valores de sensibilidade e especificidade, as seguintes equações foram utilizadas:

$$Sensibilidade = \frac{VP}{VP + FN}, \tag{8}$$

$$Especificidade = \frac{VN}{VN + FP}, \tag{9}$$

IV. Resultados e Discussão

O Algoritmo Genético aliado ao balanceamento de classe com o ADASYN conseguiu reduz o número de variáveis de 49 para 26. Durante a execução da seleção de variáveis a acurácia da RNA saltou de 61,67% para 77,50%, considerando a acurácia com todas as variáveis e a acurácia do melhor indivíduo da última geração do Algoritmo Genético, respectivamente. Utilizando a replicação de dados para balanceamento de classe o Algorítimo Genético conseguiu reduzir de 49 para 12 variáveis. E a acurácia saltou de 74,50% para 83,25%.

A Tabela II e Tabela III apresentam a média de desempenho das diferentes abordagens no projeto (treino e teste) e na validação, respectivamente.

TABLE II
Resultados de desempenho das RNAs no conjunto de treino e teste para as diferentes abordagens.

Abordagens	Desempenho	Média treino e teste ± DP (%)
1	Sensibilidade	79,03 ± 11,66
	Especificidade	72,09 ± 12,88
2	Sensibilidade	81,52 ± 20,08
	Especificidade	81,78 ± 13,62
3	Sensibilidade	83,92 ± 17,00
	Especificidade	82,66 ± 14,49
4	Sensibilidade	94,93 ± 4,93
	Especificidade	91,34 ± 8,37
5	Sensibilidade	73,98 ± 18,12
	Especificidade	73,23 ± 18,70

TABLE III
Resultados de desempenho das RNAs no conjunto de validação para as diferentes abordagens.

Abordagens	Desempenho	Média validação ± DP (%)
1	Sensibilidade	52,31 ± 11,92
	Especificidade	63,74 ± 10,72
2	Sensibilidade	54,82 ± 15,01
	Especificidade	61,16 ± 10,30
3	Sensibilidade	50,27 ± 14,00
	Especificidade	61,10 ± 9,75
4	Sensibilidade	38,09 ± 11,62
	Especificidade	69,54 ± 6,87
5	Sensibilidade	57,79 ± 19,69
	Especificidade	50,67 ± 15,63

Os resultados de desempenho para os dados do treino e teste, quanto para os dados de validação apresentaram maior variação na sensibilidade, quando comparados a especificidade, conforme demonstrado pelos desvios padrões. Os desvios padrões maiores obtidos para sensibilidade podem ser justificados pelo reduzido número de dados da classe de propriedades positiva, comparado com a classe de propriedades negativas. Desvios padrões altos podem indicar maior dispersão da classe, ou seja, são geralmente encontrados para classes mais heterogêneas. Esta discrepância entre os desvios padrões não é observada para a abordagem em que foi aplicado a técnica ADASYN e a seleção de variáveis com Algoritmo Genético. Isto submete que o fundamento do ADSYN, de concentrar os dados sintéticos criados em locais de difícil aprendizado, provocou redução na dispersão ou heterogeneidade da classe de propriedades positivas. Toda via, este fato não colaborou para melhoras substanciais nos indicadores de desempenho.

Os gráficos da Figura 2 e Figura 3 demonstram, respectivamente, a sensibilidade e especificidade das abordagens com seus desvios-padrão. Através dos gráficos é possível fazer a comparação dos valores médios (Intervalo de Confiança 95%) entre as diferentes abordagens. As diferentes letras presentes nos gráficos para cada abordagem representam, no caso de

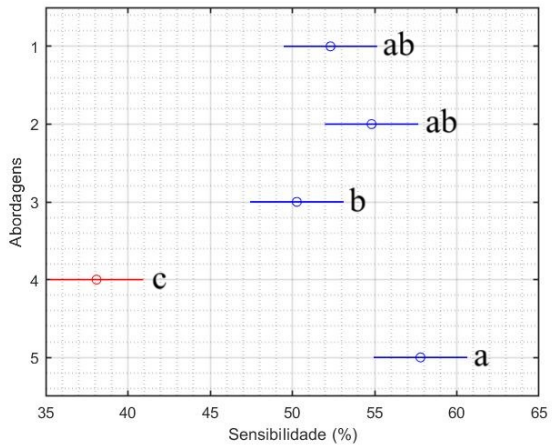

Fig. 2. Gráfico de comparação da sensibilidade (IC 95%).

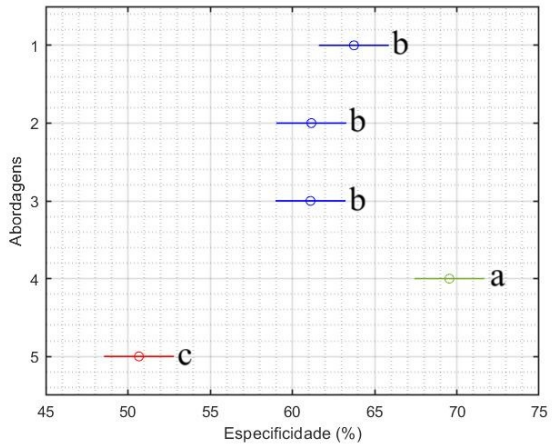

Fig. 3. Gráfico de comparação da especificidade (IC 95%).

letras diferentes, diferenças significativas de média entre as abordagens.

A aplicação de seleção de variáveis com Algoritmo Genético em conjunto com o ADASYN acarretou melhoras significativas para sensibilidade, porém, reduziu o resultado de especificidade, quando comparado a aplicação somente do ADASYN sem seleção de variáveis. A aplicação de seleção de variáveis com Algoritmo Genético não acarretou melhoras significativas para sensibilidade e especificidade dos demais modelos em que foram aplicadas técnicas de *oversampling*. Visto que, a abordagem com balanceamento de classe via replicação de dados sem seleção de variáveis apresentou resultados estatisticamente iguais aos resultados obtidos com seleção de variáveis. Por outro lado, o Algoritmo Genético permitiu reduzir consideravelmente o número de variáveis sem comprometer o desempenho das abordagens. Entende-se então que o Algoritmo Genético conseguiu eliminar da base dados variáveis que eram redundantes e que não explicavam bem a base de dados para cada respectiva abordagem. Este fato é benéfico visto que o menor número de variáveis reduz a complexidade computacional do modelo.

Como mostrado na Figura 3, nota-se que, para todas as abordagens em que técnicas de *oversampling* foram aplicadas, os resultados de especificidade foram superiores (IC 95%). O emprego de técnicas de *oversampling* permitiu explorar em maior quantidade os dados da classe de propriedades negativas. Isso colaborou para que as RNAs das abordagens com *oversampling* fossem treinadas de forma a permitir melhor generalização do modelo para classe negativa, resultando então, na melhoria da especificidade. A melhoria na especificidade pode contribuir para os órgãos de defesa sanitária definirem suas estratégias. Visto que, a especificidade diz respeito a capacidade da RNA classificar como negativa as propriedades que realmente possuam rebanho negativo para brucelose. Dessa maneira, a melhoria na especificidade evita a mobilização desnecessária de recursos dos órgãos de defesa sanitária.

V. Conclusão

Este trabalho teve por objetivo avaliação de modelos de RNA com técnicas de balanceamento de classes e seleção de variáveis, para a classificação e segregação dos rebanhos bovinos quanto à soroprevalência para brucelose. Para isto, foram projetadas cinco RNAs combinando diferentes abordagens de técnica de balanceamento de classe e seleção de variáveis. Foram avaliadas duas técnicas de *oversampling*, sendo elas: ADASYN e replicação de dados. A seleção de variáveis foi feita utilizando Algoritmo Genético. As abordagens com balanceamento de classe obtiveram resultados de especificidade superiores a abordagem em que essas técnicas não foram utilizadas. Os resultados mostraram que RNA aliada a técnica de balanceamento de classe e seleção de variáveis são promissoras para auxiliar no aperfeiçoamento das estratégias de diagnóstico, controle e erradicação da Brucelose bovina. A classificação de forma automática é capaz de dar suporte aos órgãos de defesa sanitária, facilitando a triagem de propriedades e proporcionando melhor alocação de recursos humanos e financeiros.

Em trabalhos futuros pretende-se:

- Analisar o desempenho de algoritmos de *one-class classifcation* frente ao desafio provocado pelo desbalanceamento das classes da base de dados.
- Analisar a seleção de variáveis pautando-se em modelos de interpretação de algoritmos de aprendizagem de máquina.

References

[1] "Production, supply, and distribution database," USDA Foreign Agricultural Service, 2022.

[2] S. Khurana, A. Sehrawat, R. Tiwari, M. Prasad, B. Gulati, M. Shabbir, R. Chhabra, K. Karthik, S. Patel, M. Pathak, M. Yatoo, V. Gupta, K. Dhama, R. Sah, and W. Chaicumpa, "Bovine brucellosis – a comprehensive review," *Veterinary Quarterly*, vol. 41, pp. 61–88, 2021.

[3] K. A. C. Kothalawala, K. Makita, H. Kothalawala, A. M. Jiffry, S. Kubota, and H. Kono, "Association of farmers' socio-economics with bovine brucellosis epidemiology in the dry zone of sri lanka," *Preventive veterinary medicine*, vol. 147, pp. 117–123, 2017.

[4] L. Cárdenas, L. Awada, P. Tizzani, P. Cáceres, and J. Casal, "Characterization and evolution of countries affected by bovine brucellosis (1996–2014)," *Transboundary and emerging diseases*, vol. 66, no. 3, pp. 1280–1290, 2019.

[5] M. R. S. Souza, P. M. Soares Filho, M. A. Hodon, P. G. de Souza, and C. H. O. Silva, "Evaluation of diagnostic tests' sensitivity, specificity and predictive values in bovine carcasses showing brucellosis suggestive lesions, condemned by brazilian federal meat inspection service in the amazon region of brazil," *Preventive veterinary medicine*, vol. 200, pp. 105 567–105 567, 2022.

[6] M. Wang, Z. Wei, M. Jia, L. Chen, and H. Ji, "Deep learning model for multi-classification of infectious diseases from unstructured electronic medical records," *BMC medical informatics and decision making*, vol. 22, no. 1, pp. 41–41, 2022.

[7] T. Varrecchia, S. F. Castiglia, A. Ranavolo, C. Conte, A. Tatarelli, G. Coppola, C. Di Lorenzo, F. Draicchio, F. Pierelli, and M. Serrao, "An artificial neural network approach to detect presence and severity of parkinson's disease via gait parameters," *PloS one*, vol. 16, no. 2, 2021.

[8] K. Benfodil, M. A. Benbouras, S. Ansel, A. Mohamed-Cherif, and K. Ait-Oudhia, "Prediction of trypanosoma evansi infection in dromedaries using artificial neural network (ann)," *Veterinary parasitology*, vol. 306, pp. 109 716–109 716, 2022.

[9] V. Biourge, S. Delmotte, A. Feugier, R. Bradley, M. McAllister, and J. Elliott, "An artificial neural network-based model to predict chronic kidney disease in aged cats," *Journal of veterinary internal medicine*, vol. 34, no. 5, pp. 1920–1931, 2020.

[10] E. A. Bauer and W. Jagusiak, "The use of multilayer perceptron artificial neural networks to detect dairy cows at risk of ketosis," *Animals (Basel)*, vol. 12, no. 3, p. 332, 2022.

[11] S. Sreejith, H. Khanna Nehemiah, and A. Kannan, "Clinical data classification using an enhanced smote and chaotic evolutionary feature selection," *Computers in biology and medicine*, vol. 126, pp. 103 991–103 991, 2020.

[12] D. Paul, R. Su, M. Romain, V. Sébastien, V. Pierre, and G. Isabelle, "Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier," *Computerized medical imaging and graphics*, vol. 60, pp. 42–49, 2016.

[13] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New Jersey: Wiley, 2012.

[14] E. Aličković and A. Subasi, "Breast cancer diagnosis using ga feature selection and rotation forest," *Neural computing applications*, vol. 28, no. 4, pp. 753–763, 2015.

[15] W. Brand, A. Wells, S. Smith, S. Denholm, E. Wall, and M. Coffey, "Predicting pregnancy status from mid-infrared spectroscopy in dairy cow milk using deep learning," *Journal of dairy science*, vol. 104, no. 4, pp. 4980–4990, 2021.

[16] O. L. Herrán Ramirez, H. Azevedo Santos, I. L. Jaramillo Delgado, and I. da Costa Angelo, "Seroepidemiology of bovine brucellosis in colombia's preeminent dairy region, and its potential public health impact," *Brazilian journal of microbiology*, vol. 51, no. 4, pp. 2133–2143, 2020.

[17] R. P. Deka, U. Magnusson, D. Grace, and J. Lindahl, "Bovine brucellosis: prevalence, risk factors, economic cost and control options with particular reference to india- a review," *Infection ecology epidemiology*, vol. 8, no. 1, 2018.

[18] J. Wang, P. Jia, D. F. Cuaos, M. Xu, X. Wang, W. Guo, B. A. Portnov, Y. Bao, Y. Chang, G. Song, N. Chen, and A. Stein, "A remote sensing data based artificial neural network approach for predicting climate-sensitive infectious disease outbreaks: A case study of human brucellosis," *Remote sensing (Basel, Switzerland)*, vol. 9, no. 10, p. 1018, 2017.

[19] B. J. Nartowt, G. R. Hart, D. A. Roffman, X. Llor, I. Ali, W. Muhammad, Y. Liang, and J. Deng, "Scoring colorectal cancer risk with an artificial neural network based on self-reportable personal health data," *PloS one*, vol. 14, no. 8, pp. e0 221 421–e0 221 421, 2019.

[20] S. Ahmadian, S. M. J. Jalali, S. Raziani, and A. Chalechale, "An efficient cardiovascular disease detection model based on multilayer perceptron and moth-flame optimization," *Expert systems*, vol. 39, no. 4, p. n/a, 2022.

[21] P. Branco, L. Torgo, and R. Ribeiro, "A survey of predictive modeling on imbalanced domains," *ACM computing surveys*, vol. 49, no. 2, pp. 1–50, 2016.

[22] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert systems with applications*, vol. 73, pp. 220–239, 2017.

[23] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.

[24] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-smote: A new over-sampling method in imbalanced data sets learning," in *Advances in Intelligent Computing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 878–887.

[25] G. Batista, R. Prati, and M. Monard, "A study of the behavior of several methods for balancing machine learning training data," *SIGKDD explorations*, vol. 6, no. 1, pp. 20–29, 2004.

[26] E. Ramentol, Y. Caballero, R. Bello, and F. Herrera, "Smote-rsb: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using smote and rough sets theory," *Knowledge and information systems*, vol. 33, no. 2, pp. 245–265, 2012.

[27] S. Hu, Y. Liang, L. Ma, and Y. He, "Msmote: Improving classification performance when training data is imbalanced," in *2009 Second International Workshop on Computer Science and Engineering*, vol. 2. IEEE, 2009, pp. 13–17.

[28] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, June 2008, pp. 1322–1328.

[29] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem," in *Advances in Knowledge Discovery and Data Mining*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 475–482.

[30] D. Datta, P. K. Mallick, A. V. N. Reddy, M. A. Mohammed, M. M. Jaber, A. S. Alghawli, and M. A. A. Al-qaness, "A hybrid classification of imbalanced hyperspectral images using adasyn and enhanced deep subsampled multi-grained cascaded forest," *Remote sensing (Basel, Switzerland)*, vol. 14, no. 19, p. 4853, 2022.

[31] N. Maleki, Y. Zeinali, and S. T. A. Niaki, "A k-nn method for lung cancer prognosis with the use of a genetic algorithm for feature selection," *Expert systems with applications*, vol. 164, p. 113981, 2021.

[32] Z. Ceylan and A. Atalan, "Estimation of healthcare expenditure per capita of turkey using artificial intelligence techniques with genetic algorithm-based feature selection," *Journal of forecasting*, vol. 40, no. 2, pp. 279–290, 2021.

[33] C. L. d. Castro and A. P. Braga, "Aprendizado supervisionado com conjuntos de dados desbalanceados," *Sba: Controle Automação Sociedade Brasileira de Automatica*, vol. 22, no. Sba Controle Automação, 2011 22(5), 2011.

[34] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[35] A. G. Evsukoff, *Inteligência computacional: Fundamentos e aplicações [recurso eletrônico]*. Rio de Janeiro: E-papers, 2020.

[36] J. MacQueen, "Classification and analysis of multivariate observations," in *5th Berkeley Symp. Math. Statist. Probability*. University of California Los Angeles LA USA, 1967, pp. 281–297.

[37] M. F. Moller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Networks*, vol. 6, no. 4, pp. 525–533, 1993.

[38] L. Prechelt, "Automatic early stopping using cross validation: quantifying the criteria," *Neural Networks*, vol. 11, no. 4, pp. 761–767, 1998.