Performance Analysis Among Predictive Models of Lightning Occurrence Using Artificial Neural Networks and SMOTE

Elton Rafael Alves, Adônis Ferreira Raiol Leal, Márcio Nirlando G. Lopes and Alber da Silva Fonseca

***Abstract*—Lightning represent a potential threat to various society activities, such as damage to telecommunication systems and the distribution of electric power, as well as injury or death of humans beings. Predicting the occurrence of lightning can help in making decisions about the actions that must be taken to minimize the risks of this natural phenomenon. In this study, data from air temperature profiles, dew point temperature and historical lightning data were used to obtain two predictive models of lightning occurrence. The models were obtained by using an artificial neural network. The first model was obtained through unbalanced data and the second one with data balanced with Synthetic Minority Over-sampling Technique (SMOTE). The model performance was tested in five different classes of lightning predictions: ABSENCE, LOW, MODERATE, VERY and SEVERE, considering five prediction periods: case 1 (one hour), case 2 (two hours), case 3 (three hours), case 4 (four hours) and case 5 (five hours). It was observed that the use of the Synthetic Minority Over-sampling Technique improved accuracy in the recognition of atmospheric patterns that lead to the incidence of lightning in the five classes used in the five prediction cases.**

***Index Terms*—Artificial neural network, Forecasting lightning and SMOTE.**

I. Introdução

Raio ou descarga atmosférica é uma descarga transitória de alta corrente elétrica, cujo comprimento é medido em quilômetros. Os raios são originados em nuvens de tempestades chamadas de nuvens cumulonimbus ou Cb e estes podem ser classificados em quatro formas: raio nuvem-solo, raio nuvem-nuvem, raio nuvem-ar e raio intranuvem [1].

O Brasil é um dos países que possuem a maior incidência de raios do tipo nuvem-solo. Estima-se que, em média, cerca de 78 milhões de raios atinjam o solo brasileiro por ano [2]. Dados de Alves [2] mostram que a cada 50 mortes ocasionadas por raios no mundo, uma ocorre no Brasil. Na Amazônia Legal, local de estudo desse trabalho, mais de 115 pessoas morreram atingidas por raios entre 2009 e 2019. A região tem uma taxa de mortalidade devido a raios, maior que quatro mortes por milhão de habitantes, mais alta do que em países desenvolvidos [3].

A alta incidência de raios no Brasil, ocorre em função da grande dimensão territorial do país e sua localização próxima à linha do equador (região com clima tropical) que favorece condições climáticas de temperaturas propícias à ocorrência de raios, isto é, clima quente e úmido. O estado do Pará, no Brasil, apresenta locais com elevadas concentrações de raios (raios/km^2/ano). A Fig. 1 apresenta a densidade anual média de raios nuvem-solo sobre o estado do Pará, considerando o período base de 2013 a 2017, conforme detectados pela rede STARNET (*Sferics Timing and Ranging Network*[1]).

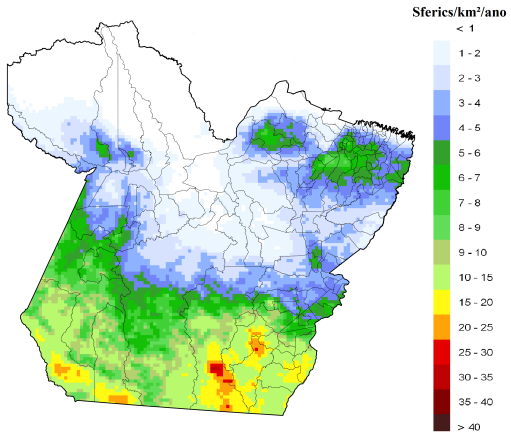

Fig. 1. Mapa de densidade de raios nuvem-solo no estado do Pará. Fonte: CENSIPAM[2]/STARNET.

Os raios são uma das principais causas de morte por desastres naturais, ocasionando cerca de 24.000 mortes e 240.000 feridos anualmente em todo o mundo [4]. Outros problemas podem ser atribuídos aos raios, como prejuízos econômicos e sociais causados em sistemas de telecomunicações, transmissão e distribuição de energia elétrica pela incidência direta ou indireta de raios, ocasionando perdas monetárias [5], [6], [7].

Uma forma de mitigação dos impactos negativos ocasionados pelos raios é o desenvolvimento de modelo de predição de sua ocorrência. Como a ocorrência de raios está fortemente

Elton Rafael Alves é professor da Faculdade de Computação e Engenharia Elétrica da Universidade Federal do Sul e Sudeste do Pará, Marabá, Pará, Brasil, e-mail: eltonalves@unifesspa.edu.br

Adônis Ferreira Raiol Leal é professor da Faculdade de Engenharia Elétrica e Biomédica da Universidade Federal do Pará, Belém, Pará, Brasil, e-mail: adonisleal@ufpa.br

Márcio Nirlando G. Lopes é Analista em Ciência & Tecnologia do Centro Gestor e Operacional do Sistema de Proteção da Amazônia, Belém, Pará, Brasil, e-mail: marcio.lopes@sipam.gov.br

Alber da Silva Fonseca é graduado em Engenharia da Computação, Marabá, Pará, Brasil, e-mail: albieri01@hotmail.com

[1]A STARNET é uma rede para detecção de raios, com cobertura na América do Sul e Caribe, coordenada pelo Laboratório T-Storm da Universidade de São Paulo. www.starnet.iag.usp.br

[2]Centro Gestor e Operacional do Sistema de Proteção da Amazônia (CENSIPAM) é um órgão federal, subordinado ao Ministério da Defesa, e que atua na proteção, inclusão social e desenvolvimento sustentável da Amazônia Legal. www.sipam.gov.br

relacionada as condições de tempo [8]. Essas condições são marcadas por eventos altamente não lineares, complexos e caóticos, o que torna esta tarefa nenhum pouco trivial. Dessa forma, a ferramenta inteligente utilizada para modelar esse fenômeno, foi uma rede neural artificial (RNA) *feedforward*, dadas suas características de aprendizado, adaptação, generalização, paralelismo massivo, robustez, armazenamento associativo de informação e processamento de informação espaço-temporal. Portanto, a RNA constitui uma boa opção para modelar sistemas meteorológicos sem o conhecimento físico do fenômeno, como é o caso das descargas atmosféricas, utilizando apenas mapeamento de dados de entrada e saída para reconhecimento de padrões.

Diversos estudos já foram desenvolvidos com o objetivo de prever raios, alguns dos quais se baseiam na utilização de índices termodinâmicos, que podem ser obtidos através de radiossondagem, que são realizadas duas vezes ao dia, uma pela manhã e outra à noite. A radiossondagem convencional consiste em um processo em que sensores (radiossonda) acoplados a um balão ascendem na atmosfera coletando dados de temperatura do ar, umidade, vento e pressão atmosférica, traçando assim, um perfil vertical da atmosfera. Neste processo, os parâmetros meteorológicos coletados permitem o cálculo de diversos indicadores termodinâmicos da atmosfera, como índice-K, índice Showalter, Energia Potencial Convectiva Disponível, também conhecida como CAPE (*Convective Available Potential Energy*), dentre outros. Todos estes índices termodinâmicos, também chamados de índices de instabilidade, podem ser utilizados como variáveis preditoras para a previsão de raios, sendo combinados com uma rede neural artificial [9], [10], [5], [11]. Abdullah et al. [12] utilizaram dados de observações atmosféricas no nível do solo, combinados com uma rede neural treinada com o algoritmo backpropagation para prever raios em Melaka, na Malásia.

Além da utilização de índices de instabilidade gerados através de dados de radiossondagens e dados de observações atmosféricas em baixa latitude, outros estudos de previsões utilizaram o campo eletrostático formado pelas tempestades de raios, como em [8].

Lu et al. [13] adotou estudos estatísticos de atividade de raios na província de Huna, na China, resumindo situações típicas de condição de tempo, como propícias para a ocorrência de raios. Enquanto Gijben at al. [4] utilizaram dados de raios nuvem-solo da rede de detecção de raios da África e parâmetros numéricos do modelo de previsão climática do modelo unificado para desenvolver um índice de ameaças de raios para a África do Sul.

Na região amazônica, Sá et al. [14] foram os pioneiros a desenvolver um estudo de previsão de raios utilizando redes neurais artificiais, tendo como variáveis preditoras índices de instabilidade obtidos por radiossondagens tradicionais, como a CAPE. Recentemente, Alves et al. [15] baseados em dados de sondagem atmosférica por satélite, mostraram a possibilidade de realização de predição de raio através de sondagens satelitais para a região amazônica. Em seu estudo, Alves et al. [15] utilizando uma rede neural do tipo perceptron de multicamadas, realizaram previsões para apenas duas classes: zero (0) para não ocorrência e um (1) para ocorrência de raios.

Visto que a incidência de raios é bastante aleatória e a produção de raios por uma Cb é um processo muito complexo. Uma nuvem Cb ordinária pode produzir desde um único raio nuvem-solo, assim como dezenas. Considerando um aglomerado de nuvens Cb, a produção de raios pode chegar a ordem de milhares de raios. Portanto, tal fenômeno pode apresentar uma grande variação no seu grau de severidade, logo uma previsão discreta quanto a severidade da tempestade elétrica se torna mais realista.

Assim, com base nos experimentos realizados em [15], este estudo propõe uma modelagem preditiva, através do treinamento de uma RNA *feedforward* utilizando o algoritmo de treinamento Levenberg-Marquardt. As vantagens inéditas apresentadas por este trabalho incluem, uma melhor discretização da previsão de ocorrência de raios nuvem-solo, diferenciando os casos com AUSÊNCIA, POUCA, MODERADA, MUITA e SEVERA ocorrência de raios. Um segundo ponto de relevância a se destacar é a utilização de dados de sondagens satelitais, o que permite a previsão de curto prazo e com maior acurácia para vários períodos do mesmo dia.

E por fim, como a série de dados de incidência de raios apresenta grande dispersão, pois pode ocorrer vários dias sem qualquer ocorrência de raios, enquanto em um outro pode haver milhares, esta característica produz uma série temporal totalmente desbalanceada para as diferentes categorias do evento, o que dificulta o treinamento das redes neurais e tende a reduzir a acurácia do previsor. Um conjunto de dados é dito desbalanceado para um problema de classificação quando o número de amostras de uma classe é superior à outra classe [16], [17]. Neste estudo, observou-se uma diferença significativa entre as classes a serem previstas, podendo ocasionar previsões tendenciosas para determinada classe majoritária. Dessa forma, foi aplicado o método *Synthetic Minority Oversampling Technique* (SMOTE) para equilibrar as classes de estudo. Por isso, uma outra contribuição importante nesta pesquisa é empregar uma técnica para balancear os dados e melhorar o desempenho do modelo previsor.

II. RNA - Redes Neurais Artificiais

Uma RNA é inspirada na forma como os neurônios biológicos funcionam no processo de propagação da informação. É utilizada nesse estudo devido sua forte capacidade no reconhecimento de padrões, implementada através de aprendizagem de dados de entrada e saída.

A RNA empregada neste estudo é vista na Fig. 2. Esta RNA *feedforward* tem como objetivo fazer uma modelagem preditiva. De acordo com [18], isto acontece quando um classificador é utilizado para identificar a qual classe um novo exemplo pertence.

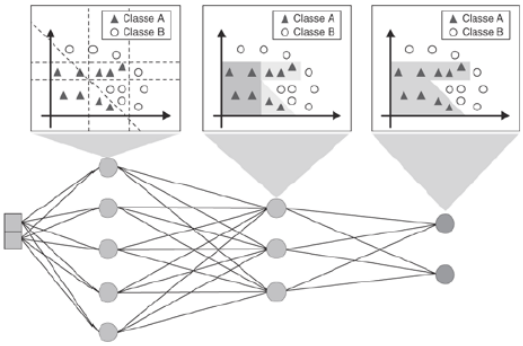

Fig. 2. RNA *feedforward* aplicada a um problema de classificação.

O treinamento da RNA foi realizado através do algoritmo Levenberg-Marquardt que é uma aproximação do método de Newton [19]. Os pesos da RNA através desse algoritmo são atualizados através da Equação 1.

$$\nabla w_{ij} = -[\nabla^2 E(w_{ij}(t)) + \eta I)]^{-1} \nabla E(w_{ij}(t)) \qquad (1)$$

Onde $\nabla^2 E(w_{ij}(t))$ é a matriz hessiana, η é a taxa de aprendizado e $\nabla E(w_{ij}(t))$ é o gradiente.

III. Materiais e Métodos

A. Descrição dos Conjuntos dos Dados e Área de Estudo

A área em estudo, localizada no nordeste do estado do Pará, foi subdividida em oito áreas, como pode ser visualizado na Fig. 3.

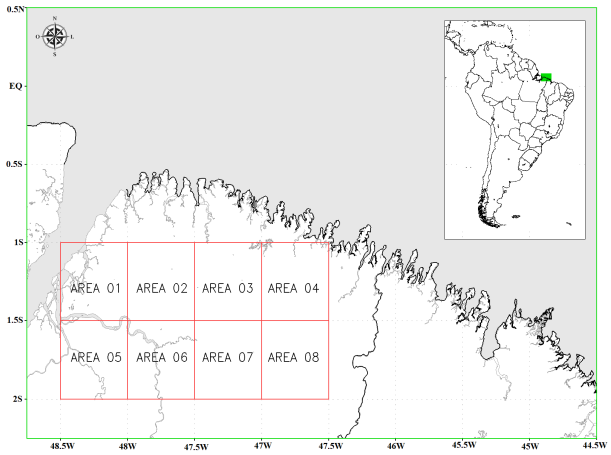

Fig. 3. Área de estudo.

A região utilizada é limitada entre as latitudes 01° S e 02° S, e entre as longitudes 46,5° W e 48,5° W. Estas áreas, além de apresentarem índices relevantes de ocorrências de raios, foram selecionadas devido a maior densidade demográfica em relação a outras cidades do Pará.

O conjunto de dados é subdividido em três grupos. Os dois primeiros grupos são variáveis atmosféricas da temperatura do ar (T) e da temperatura do ponto de orvalho (T_d). A temperatura do ar é estimada em diferentes níveis de altura pela sonda a bordo do satélite ambiental NOAA-19. O segundo atributo, é a temperatura do ponto de orvalho, que representa a temperatura que o ar deveria atingir para saturar o ambiente de vapor d'água, ou seja, para que a umidade relativa do ar alcance 100%. A cada passagem do satélite NOAA-19 foi obtida uma sondagem distinta para cada uma das oito áreas estudadas, no mesmo horário.

Como o treinamento da rede é supervisionado, foi empregado um terceiro grupo que é o atributo alvo, também denominado de classe, ou seja, a saída desejada. O atributo alvo é composto pelo número de raios detectados para o intervalo horário predefinido em um determinado dia, em cada uma das áreas anteriormente descritas. A base de dados de raios utilizada é proveniente da rede STARNET [20] e disponibilizada pelo CENSIPAM.

O intervalo das amostras que compõe o conjunto de dados, inicia-se no dia 13 de junho de 2014 e encerra-se no dia 10 de janeiro de 2017, totalizando 2.800 amostras válidas. Foram realizadas previsões de raios nas seguintes faixa horárias: 18:00 às 19:00 UTC (Caso 1), 19:01 às 20:00 UTC (Caso 2), 20:01 às 21:00 UTC (Caso 3), 21:01 às 22:00 UTC (Caso 4) e 22:01 às 23:00 UTC (Caso 5). O horário UTC significa *Universal Time Coordinated*. No caso particular do estado do Pará, o horário local corresponde a uma defasagem de três horas em relação ao horário padrão UTC. Os intervalos são não cumulativos, ou seja, a previsão realizada para o caso 1 tem antecedência menor que para o caso 2 e assim sucessivamente até o caso 5. Além disso, historicamente esses são os horários do dia com maior incidência de descargas atmosféricas na região [21].

B. Categorização dos Dados

A Fig. 4 exibe o gráfico boxplot obtido para as cinco faixas horárias de previsões empregadas, utilizando-se dados de raios coletados nas oito áreas de estudo.

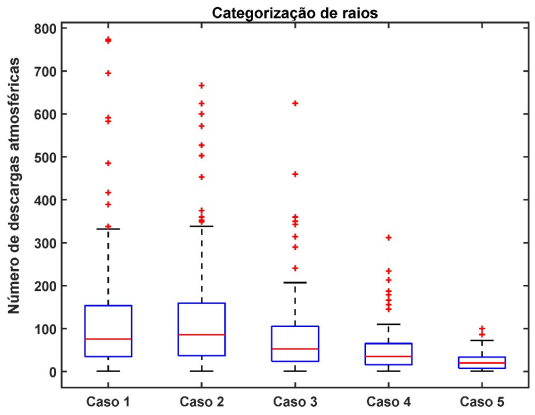

Fig. 4. Categorização dos raios.

A partir do emprego da técnica de quartis (Fig. 4), foram definidas as seguintes categorias de incidência de raios, para as áreas de estudo:

- **Categoria 1.AUSÊNCIA:** quando não há raios;
- **Categoria 2.POUCA:** compreende do valor mínimo até o 1° quartil dos dados;
- **Categoria 3.MODERADA:** compreende o primeiro valor

depois do 1° quartil até o 2° quartil dos dados;
• **Categoria 4.MUITA:** compreende o primeiro valor depois do 2° quartil até o 3° quartil dos dados;
• **Categoria 5.SEVERA:** compreende o primeiro valor depois do 3° quartil até o valor máximo discrepante dos dados;

A escolha das cinco categorias é uma abordagem inovadora e mais prática para quantificar as ocorrências de raios e melhor qualificar as previsões. Na Tabela I são apresentados os intervalos obtidos a partir das determinações dos quartis, valores mínimos e máximos discrepantes dos dados.

TABELA I
RESULTADO DA CATEGORIZAÇÃO DE RAIOS

Classificações de padrões de raios			
Classes	**Caso 1**	**Caso 2**	**Caso 3**
Severa	155 - 774	160 - 666	106 - 625
Muita	77 - 153	87 - 159	54 - 105
Moderada	36 - 76	38 - 86	25 - 53
Pouca	1 - 35	1 - 37	1 - 24
Ausência	0	0	0
Classes	**Caso 4**	**Caso 5**	
Severa	66 - 312	34 - 100	
Muita	36 - 65	21 - 33	
Moderada	17 - 35	9 - 20	
Pouca	1 - 16	1 - 8	
Ausência	0	0	

C. Pré-processamento dos Dados

1) SMOTE - Synthetic Minority Over-sampling Technique: O algoritmo SMOTE foi proposto para contrabalancear o problema de conjunto de dados desbalanceado para classificação. Ele sintetiza novas instâncias da classe minoritária, operando no espaço de recursos e não no espaço de dados. Cada amostra da classe minoritária origina uma porcentagem das amostras sintéticas. Este aumento em instâncias dos dados da classe minoritária expande as razões de decisão dos classificadores [22].

Nesse algoritmo são sintetizadas novas amostras com a finalidade de reduzir o desiquilíbrio entre as classes majoritária e minoritária no conjunto de dados. Para isso é realizada uma busca das amostras vizinhas mais próximas para cada amostra na classe minoritária. Em seguida é selecionado, aleatoriamente, instâncias entre as amostras da vizinhança mais próxima a interpolação linear. Assim, para cada uma das amostras minoritárias originais será produzido novas amostras. O próximo passo é a interpolação entre as amostras da classe minoritária original e suas amostras vizinhas, utilizando a Equação 2 de interpolação linear [23].

$$X_{novo} = X_{origin} + rand(0,1) \times (X_i - X_{origin}) \qquad (2)$$

$i = 1, 2...N$, onde X_{novo} representa a amostra da classe minoritária sintetizada; X_{origin} indica as amostras originais que são usadas para sintetizar novas amostras; $rand(0,1)$ representa um número aleatório que está entre 0 e 1 e X_i representa uma amostra que é selecionada aleatoriamente das amostras vizinhas k da amostra de classe minoritária X_{origin}.

O conjunto de dados utilizado nesse estudo foram submetidos ao SMOTE por meio do software Weka versão 3.8.3, baseando-se em [24]. Os parâmetros utilizados nesse procedimento foram o nome das classes (AUSÊNCIA, POUCA, MODERADA, MUITA E SEVERA), a porcentagem de amostras sintetizadas em relação as amostras da classe original, e o valor do k=5 vizinhos das amostras mais próximas das amostras da classe minoritária.

A Fig. 5 exibe o resultado do processo de balanceamento das classes do conjunto de treinamento, cujo total corresponde a 2.520 amostras (≃ 70%).

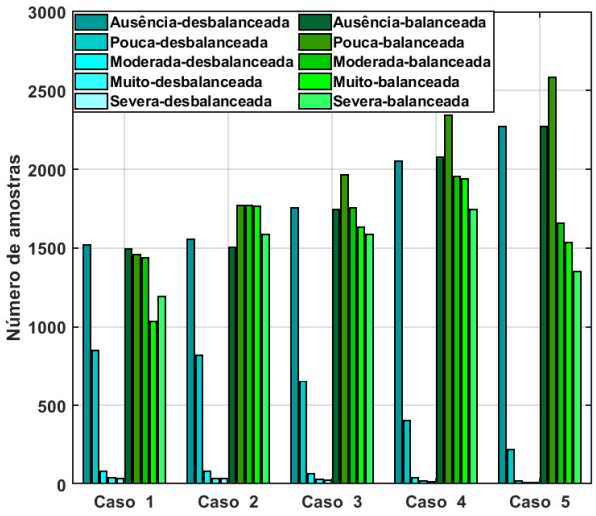

Fig. 5. Comparação das amostras desbalanceadas com as amostras balanceadas (coloração em azul - dados desbalanceados e coloração em verde - dados balanceados).

Como os dias sem ocorrência de raios ocorrem de forma majoritária (Fig. 5) em todos os casos considerados, ou seja, maior quantidade na categoria AUSÊNCIA, o número de amostras originais foi pouco alterado em relação ao número de amostras sintéticas. No entanto, o número de amostras das classes POUCA, MODERADA, MUITA e SEVERA, também considerando todos os casos, aumentaram significativamente em relação ao número de amostras originais das respectivas classes. Ressalta-se que os gráficos na coloração verde são referentes aos dados sintéticos e os gráficos na coloração azul são referentes aos dados originais.

2) Normalização dos Dados: O conjunto de dados, referente aos atributos de entrada, foi submetido ao processo de normalização, conforme a Equação 3. O objetivo foi evitar uma discrepância entre os valores de entrada.

$$valor_{normalizado} = \frac{valor_{original} - minA}{maxA - minA} \qquad (3)$$

O valor normalizado transforma os dados originais de entrada no intervalo [0, 1].

D. Arquitetura, Topologia e Estrutura da RNA

Neste estudo foi empregada uma RNA *feedforward* de três camadas (camada de entrada, intermediária e saída). O modelo desta RNA é exibido na Fig. 6.

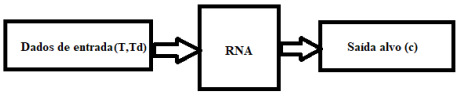

Fig. 6. Modelo da RNA.

A camada de entrada é formada por: T = dados de temperatura do ar (em °C) e T_d = dados da temperatura do ponto de orvalho (em °C). Alves et al. [15] encontraram que os seguintes níveis de pressão atmosférica: 10hPa, 15hPa, 20hPa, 25hPa, 30hPa, 50hPa, 60hPa, 70hPa, 85hPa, 100hPa, 115hPa, 850hPa, 920hPa, 950hPa e 1000hPa, possuem as maiores variabilidades para os dados da temperatura do ar, enquanto que para os dados da temperatura do ponto de orvalho, tem-se os seguintes níveis de pressão atmosférica com maior variabilidade: 150hPa, 200hPa, 250hPa, 300hPa, 350hPa, 400hPa, 430hPa, 475hPa, 500hPa, 570hPa, 620hPa, 670hPa, 700hPa, 780hPa, 850hPa, 920hPa e 1000hPa. Baseado nestes resultados, os mesmos níveis de pressão foram adotados para obtenção da temperatura do ar e temperatura do ponto de orvalho no treinamento da RNA. A saída alvo c corresponde as classes AUSÊNCIA, POUCA, MODERADA, MUITA e SEVERA, conforme definidas na subseção B.

As funções de ativação utilizadas na camada oculta e na camada de saída foram as funções sigmoide e softmax, respectivamente. Sendo que foram utilizados trinta e sete neurônios na camada oculta e cinco neurônios na camada de saída para treinar os modelos. O treinamento da RNA foi realizado através do algoritmo Levenberg-Marquardt com 70% dos dados. Adotou-se a técnica de validação cruzada k-folds para avaliar a capacidade de generalização da RNA, onde utilizou-se k=10. Na etapa de validação utilizou-se 30% dos dados. Foram adotadas as mesmas configurações da RNA para treinamento dos dados balanceados e não balanceados, considerando os cinco casos e classes de previsões.

E. Métodos de Avaliação

A avaliação de desempenho dos modelos preditivos neurais, para todos casos de estudo, consistiu na obtenção de matrizes de confusão, conforme Tabela II. Assim, os resultados foram apresentados em matrizes de confusão para avaliar a taxa de acerto dos classificadores (acurácia).

TABELA II
MATRIZ DE CONFUSÃO

		Previsto	
		1	0
Real	1	VP	FP
	0	FN	VN

Em uma matriz de confusão, amostras classificadas corretamente na classe positiva (1 - teve raio), são chamadas de verdadeiros positivos (VP), e amostras classificadas corretamente na classe negativa (0 - não teve raio), são chamadas de verdadeiros negativos (VN). As amostras da classe positiva classificadas como negativas, são chamadas de falsos negativos (FN), e as amostras da classe negativa classificada como positivas, são chamadas de falsos positivos (FP). A partir da Tabela II pode-ser obter a taxa de acerto do classificador (acurácia) dada pela Equação 4.

$$acuracia = \frac{(VP + VN)}{(VP + VN + FP + FN)} \times 100 \qquad (4)$$

As análises dos resultados também foram feitas através das curvas do gráfico ROC (Curvas de Características de Operação do Receptor). O gráfico ROC permitiu visualizar o desempenho dos classificadores [25].

IV. Resultados

Os resultados encontrados na etapa de validação demonstram que os modelos que utilizaram dados balanceados obtiveram desempenho superior aos modelos que empregaram os dados desbalanceados. Considerando, primeiramente, as matrizes de confusão, a melhoria foi mais significativa para os casos 1, 2 e 3 de previsão, cujas acurácias alcançadas para os dados desbalanceados foram, respectivamente de 62,82%, 60,36% e 69,64%, contra 80,08%, 80,49% e 83,09% para os dados balanceados com o SMOTE (Fig. 7). Ressalta-se que se utilizou o software Matlab para estas simulações.

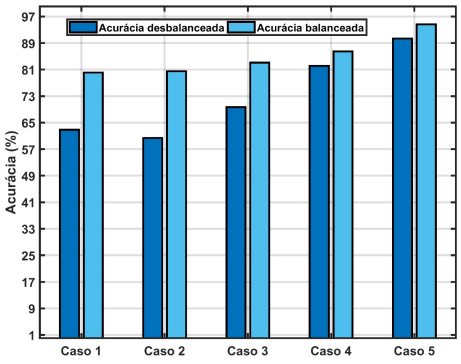

Fig. 7. Comparação entre as acurácias obtidas para os cinco casos com dados balanceados (SMOTE) e desbalanceados.

Nos casos 4 e 5, os desempenhos foram quase equivalentes, com acurácias para os dados desbalanceados de 82,14% e 90,36%, respectivamente. Enquanto com a utilização de dados balanceados, as acurácias obtidas foram de 86,49% e 94,64% para os casos 4 e 5, respectivamente.

Considerando as curvas ROC, em todos os gráficos (Fig. 8 a 12) para os dados balanceados para as cinco classes (AUSÊNCIA, POUCA, MODERADA, MUITA e SEVERA) houve aproximação da curva em relação ao eixo Y (lado esquerdo), caracterizando uma maior taxa de verdadeiros positivos e menor taxa de falsos positivos. Este comportamento diferiu significativamente para os dados desbalanceados em que a linha no ROC ficou bem próxima da diagonal secundária, caracterizando previsões com desempenho inferior. A linha diagonal de um gráfico ROC caracteriza um classificador aleatório. Dessa forma, qualquer classificador próximo dessa linha pode então ser considerado pior que o aleatório.

Para o caso 1 (Fig. 8), por exemplo, pode ser observado que a classe MUITA se aproximou do ponto (0,1) indicando uma classificação quase perfeita. Por outro lado, no caso 4 (Fig. 11), a classe AUSÊNCIA está bem próxima da linha do classificador aleatório, caracterizando um desempenho ruim do classificador. Assim, fica evidenciado que a utilização do SMOTE combinado com RNA melhorou o desempenho dos preditores, como visualizado na Fig. 7.

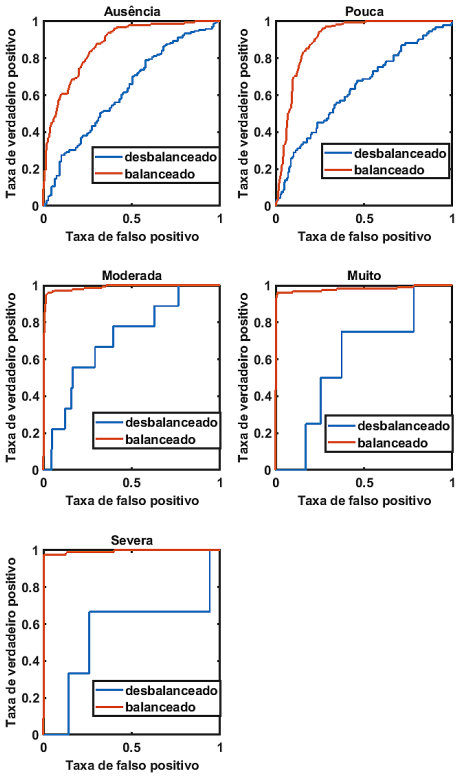

Fig. 8. Gráfico ROC para o caso 1 considerando as cinco classes.

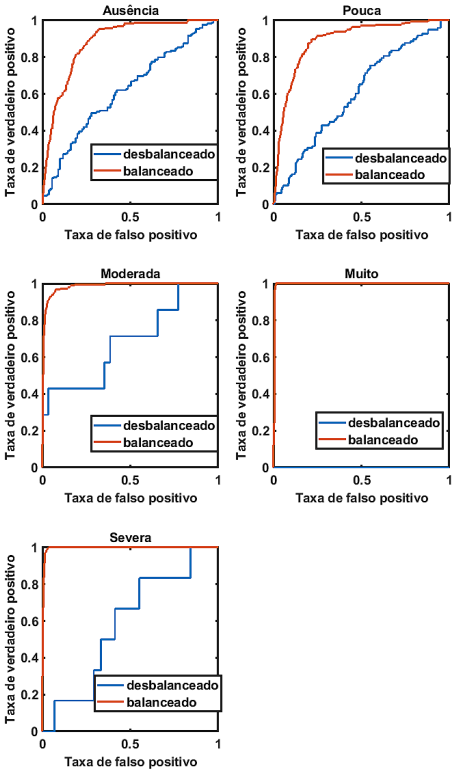

Fig. 9. Gráfico ROC para o caso 2 considerando as cinco classes.

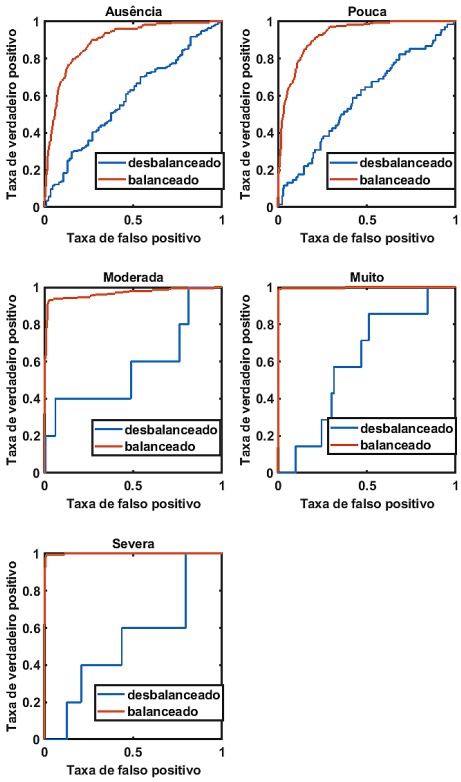

Fig. 10. Gráfico ROC para o caso 3 considerando as cinco classes.

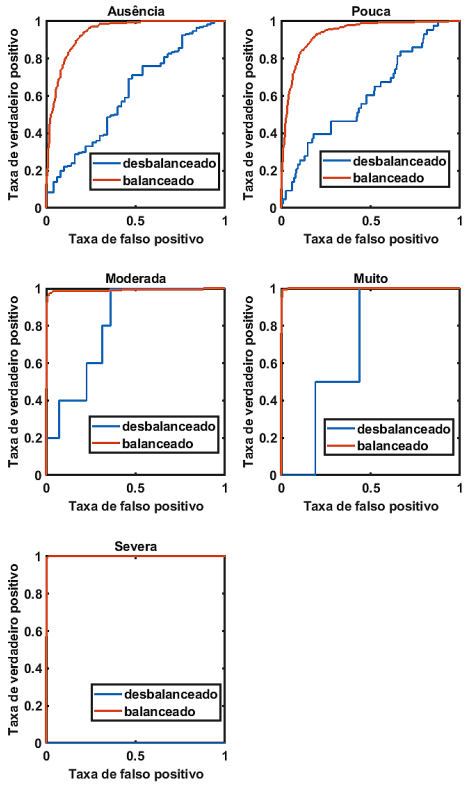

Fig. 11. Gráfico ROC para o caso 4 considerando as cinco classes.

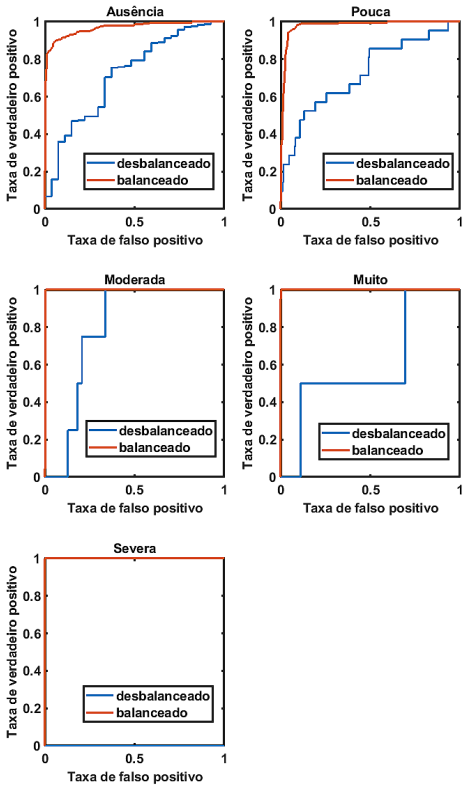

Fig. 12. Gráfico ROC para o caso 5 considerando as cinco classes.

V. Conclusões

Este trabalho realizou um estudo comparativo entre modelos preditivos que utilizam redes neurais artificiais aplicadas à previsão categorizada (AUSÊNCIA, POUCA, MODERADA, MUITA e SEVERA) de incidência de raios nuvem-solo, no nordeste paraense, para um horizonte de até cinco horas, após a passagem do satélite ambiental NOAA-19, que constitui a fonte dos dados preditores dos modelos.

O primeiro modelo foi treinado e validado considerando apenas dados brutos desbalanceados. Para o segundo modelo as classes minoritárias de ocorrência de raios foram balanceadas através do método SMOTE.

O modelo preditivo cujos dados de entrada foram equilibrados através do método SMOTE, apresentou melhor desempenho em predizer a ocorrência de raios para as cinco classes, nos cinco casos, quando comparado com o modelo preditivo obtido com dados não equilibrados. A variação do aumento percentual da acurácia para os casos 1, 2, 3, 4 e 5 foram, respectivamente: 27,47%, 33,34%, 19,31%, 5,29% e 4,73%. Dessa forma, foi evidenciado que a RNA conseguiu realizar um melhor mapeamento dos dados de entrada-saída de padrões atmosféricos balanceados.

Considerando que os modelos desenvolvidos neste trabalho podem, com elevado grau de acurácia, alertar para riscos potenciais à vida de pessoas e animais, com até cinco horas de antecedência, sobretudo produzindo informação melhor qualificada sobre o grau de severidade deste fenômeno natural que são os raios nuvem-solo; estes resultados também podem vir a corroborar sobremaneira com os tomadores de decisão de forma a permitir que estes adotem estratégias e práticas preventivas que venham minimizar os efeitos negativos da incidência de raios, assim como salvaguardar vidas.

Agradecimentos

Os autores gostariam de agradecer aos centros provedores do conjunto de dados utilizados nesta pesquisa, que foi essencial para realização deste trabalho: Laboratório T-STORM da Universidade de São Paulo-USP e Centro Gestor e Operacional do Sistema de Proteção da Amazônia-CENSIPAM.

Referencias

[1] M. A. Uman, “Natural lightning", *IEEE Transactions on Industry Applications*, vol. 30, no. 3, pp. 785-790, May/Jun, 1994.

[2] E. R. Alves, “Previsão de raios utilizando técnicas de inteligência computacional e dados de sondagem atmosférica por satélite", PhD. thesis, Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal do Pará, Belém, Pará, 2017.

[3] E. R. Ferreira, Adônis, F. R. Leal, W. L. N. Matos, G. O. Almeida, R. Shinkai and M. N. G. Lopes, “Lightning deaths and injuries in the Brazilian Amazon region in the period of 2009-2019", *in Proceedings of the International Symposium on Lightning Protection (XV SIPDA)*, São Paulo, Brazil, Sept-Oct, 2019, pp. 1-8.

[4] M. Gijben, L. L. Dyson. and M. T. Loots, “A statistical scheme to forecast the daily lightning threat over southern Africa using the Unified Model", *Atmospheric Research*, vol. 194, pp. 78-88, Sept, 2017.

[5] L. Y. Weng, J. B. Omar, Y. K. Siah, S. K. Ahmed, I. B. Z. Abidin and N. Abdullah, “Lightning forecasting using ANN-BP & radiosonde", *in Proceedings of the International Conference on Intelligent Computing and Cognitive Informatics*, Kuala Lumpur, Malaysia, Sept, 2010, pp. 1-8.

[6] G. Juntian, G. ShanQiang and F. Wanxing, “A lightning motion prediction technology based on spatial clustering method", *in Proceedings of the 7th Asia-Pacific International Conference on Lightning*, Chengdu, China, Dec, pp. 1-6, 2011.

[7] V. A. Rakov, M.A. Uman, M.I. Fernandez, C.T. Mata, K.J. Rambo, M.V. Stapleton and R.R. Sutil, "Direct lightning strikes to the lightning protective system of a residential building: Triggered-lightning experiments", *IEEE Transactions on Power Delivery*, vol. 17, no. 2, pp. 575-586, Aug, 2002.
[8] Q. Zeng, Z. Wang, F. Guo, M. Feng, M. Feng, S. Zhou and H. Wang, "The application of lightning forecasting based on surface electrostatic field observations and radar data", *Journal of Electrostatics*, vol. 71, pp. 6-13, Feb, 2013.
[9] G. S. Zepka, O. Pinto Jr. and A. C. V. Saraiva, "Lightning forecasting in southeastern Brazil using the WRF model", *Atmospheric Research*, vol. 135-136, pp. 344-362, Jan, 2014.
[10] G. S. Zepka, A. C. V. Saraiva, O. Pinto Jr and V. L. G. Gardiman, "Lightning forecasting using WRF model over EDP distribution companies areas", *in Proceedings of the International Symposium on Lightning Protection (XII SIPDA)*, Belo Horizonte, Brazil, Oct, 2013, pp. 1-8.
[11] D. Johari, T. K. A. Rahman and I. Musirin, "Artificial neural network based technique for lightning prediction", *in Proceedings of the 5th Student Conference on Research and Development*, Selangor, Malaysia, Dec, 2007, pp. 1-8.
[12] N. H. Abdullah , R. Adnan, A. M. Samad and F. A. Ruslan, "Lightning forecasting modelling using artificial neural network (ANN): Case study Sultan Abdul Aziz Shah airport or Skypark Subang", *in Proceedings of the IEEE Conference on Systems, Process and Control (ICSPC)*, Melaka, Malaysia, Dec, 2018, pp. 1-8.
[13] J. Lu, H. Zhang, L. Yang, B. Li, Z. Fang, X. Xu, "Forecast method of lightning activity based on the weather conditions", *in Proceedings of the 7th Asia-Pacific International Conference on Lightning*, Chengdu, China, Nov, 2011, pp. 1-8.
[14] J. A. S. de Sá, B. R. P. da Rocha, A. C. Almeida and J. R. Souza, "Recurrent selforganizing map for severe weather patterns recognition", *Recurrent Neural Networks and Soft Computing*, vol. 17, pp. 151-175, Oct, 2012.
[15] E. R. Alves, B. R. P. da Rocha, C. T. C. Júnior, M. N. G. Lopes and J. A. S. de Sá, "Lightning prediction using satellite atmospheric sounding data and feed-forward artificial neural network", *Journal of Intelligent & Fuzzy Systems* , vol. 33, pp. 79-92, Jun, 2017.
[16] A. A. El-Sayed, M. A. M. Mahmood, N. A. Meguid and H. A. Hefny, "Handling autism imbalanced data using synthetic minority over-sampling technique (SMOTE)", *in Proceedings of the Third World Conference on Complex Systems (WCCS)*, Marrakech, Morocco, Nov, 2015, pp. 1-8.
[17] K. U. Rani, G. N. Ramadevi and D. Lavanya, "Performance of synthetic minority oversampling technique on imbalanced breast cancer data", *in Proceedings of the 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, India, Mar, 2019, pp. 29-39.
[18] P. Tan, M. Steinbach and V. Kumar, Introduction to data mining, (First Edition), Pearson Education India, 2005.
[19] M.T. Hagan and M.B. Menhaj, "Training feedforward networks with the Marquardt algorithm", *IEEE Transactions on Neural Networks*, vol. 5, no. 6, pp. 989-993, Nov, 1994.
[20] C. A. Morales, J. R. Neves, E. M. Anselmo, K. S. Camara, W. Barreto, V. Paiva and R. L. Holle, "8 years of sferics timing and ranging network - STARNET: A lightning climatology over South America", *in International Lightning Detection Conference / International Lightning Meteorology Conference (ILDC/ILMC)*, Mar, 2014, pp. 1-8.
[21] A. C. Almeida, B. R. P. Rocha, J. R. S. Souza, J. A. S. Sá and J. A. P. Filho, "Cloud-to-ground lightning observations over the eastern Amazon Region", *Atmospheric Research*, vol. 117, pp. 86-90, Nov, 2012.
[22] M. R. Prusty, T. Jayanthi and K. Velusamy, "Weighted-SMOTE: A modification to SMOTE for event classification in sodium cooled fast reactors, *Progress in Nuclear Energy*, vol. 100, pp. 355-364, Sept, 2017.
[23] Y. Ge, D. Yue and L. Chen., "Prediction of wind turbine blades icing based on MBK-SMOTE and random forest in imbalanced data set", *in Proceedings of the TIEEE Conference on Energy Internet and Energy System Integration (EI2)*, Beijing, China, Nov, 2017, pp. 1-6.
[24] N. V. Chawla, K. W. Bowyer and L. O. Hall, "Smote: synthetic minority over-sampling technique", *Journal of articial intelligence research*, vol. 16, pp. 321-357, Jun, 2002.
[25] T. Fawcett, "An introduction to ROC analysis", *Pattern recognition letters*, vol. 27, no. 8, pp. 861-874, Jun, 2006.

Elton Rafael Alves is holds a degree in Computer Engineering from the Federal University of Pará and Doctor in Electrical Engineering (Energy System) from the Federal University of Pará. Is Adjunct Professor at the Federal University of the South and Southeast of Pará. His areas of interest are: computational intelligence, embedded systems and atmospheric discharges.

Adônis F. R. Leal received a Master and Doctor degree in electrical engineering (Power Systems) from the Federal University of Pará, Belem, Para, Brazil in 2014 and 2018 respectively. From 2016 to 2017, he worked as a visiting researcher in the Department of Electrical and Computer Engineering at the University of Florida, Gainesville, FL, USA. Since 2018 he is an Adjunct Professor at the Federal University of Para. His main interests are development of embedded systems, lightning physics, lightning detection and location systems and lightning occurrence in the Amazon.

Márcio Nirlando Gomes Lopes Graduated in Agronomy and Meteorology, with a Master's degree in Environmental Sciences and a PhD in Electrical Engineering. He is currently a Science & Technology Analyst of the Management and Operational Center for the Amazon Protection System.

Alber da Silva Fonseca Graduated in Computer Engineering from the Federal University of the South and Southeast of Pará.