XXVII Congresso Nacional de Estudantes de Engenharia Mecânica
08 a 12 de fevereiro de 2021, Curitiba, PR, Brasil

PERCEPTRON DE MÚLTIPLAS CAMADAS PARA AVALIAÇÃO DO IMPACTO DA POLUIÇÃO DO AR NA SAÚDE POPULACIONAL

João Luiz Miranda Meyer, joao_lmm@hotmail.com
Thiago Antonini Alves, antonini@utfpr.edu.br
Hugo Valadares Siqueira, hugosiqueira@utfpr.edu.br
Yara de Souza Tadano, yaratadano@utfpr.edu.br

Universidade Tecnológica Federal do Paraná (UTFPR), *Câmpus* Ponta Grossa, Av. Doutor Washington Subtil Chueire, 330, Jardim Carvalho, Ponta Grossa/PR, 84.017-220

Resumo.*O objetivo foi realizar uma Revisão Bibliográfica Sistemática (RBS) relacionada à poluição atmosférica, saúde humana e redes neurais artificiais. Em seguida, a rede neural Perceptron de Múltiplas Camadas - MLP (do inglês,* Multi-Layer Perceptron*) foi utilizada para prever o número de internações por doenças respiratórias causadas pelo* MP_{10} *na cidade de São Paulo. A RBS foi realizada em 6 bases de dados, utilizando as palavras-chave "*air pollutants*", "*neural network*" e "*health*", sendo os anos de publicações restritos entre 2009 e janeiro de 2020. A MLP foi implementada utilizando a linguagem de programação Python versão 3.7 com o auxílio da biblioteca* Scikit-learn*. Foi possível concluir que apenas uma pequena parcela dos artigos encontrados foi selecionada para a RBS, mostrando ser um tema ainda pouco explorado no meio acadêmico. As previsões encontradas pela MLP apresentam valores próximos aos observados. Além disso, a MLP se mostrou uma ferramenta eficaz para esse tipo de previsão, podendo ser explorado em trabalhos futuros.*

Palavras chave*: Inteligência Artificial. Saúde. Poluentes Atmosféricos.*

Abstract.*The objective was to carry out a Systemic Bibliographic Review (SBR) related to atmospheric, human health and artificial neural networks. Then, the MLP (Multi-Layer Perceptron) neural network was used to predict the number of hospitalizations for respiratory diseases caused by* PM_{10} *in the city of São Paulo. The RBS was carried out in 6 databases, using the keywords "air pollutants", "neural network", and "health", with the years of publications being restricted from 2009 to January 2020. MLP was implemented using the language Python programming version 3.7 with the aid of the Scikit-learn library. It was possible to conclude that only a small amount of the articles was considered in the RBS, showing that it is a topic that is still little explored on academia. The results of MLP are close to observed data. In addition, MLP has proven to be an effective tool for this type of forecast, and can be explored in future work.*

Keywords:*Artificial Intelligence. Health. Atmospheric Pollutants.*

1. INTRODUÇÃO

O alto número de poluentes lançados diariamente na atmosfera por veículos que utilizam combustíveis fósseis e/ou indústrias que não possuem um controle efetivo de seus resíduos atmosféricos pode impactar diretamente na saúde da população. De acordo com Arbex *et al*. (2012), as principais doenças causadas por poluentes atmosféricos são câncer de pulmão, asma e doença pulmonar obstrutiva crônica. Isso é intensificado em grandes centros urbanos, onde há uma maior concentração dessas fontes emissoras, sendo esse problema agravado para a população local.

Vários podem ser os poluentes, segundo o Ministério do Meio Ambiente (MME, 2020), os mais comuns são CO, CO_2, CH_4, SO_2, Hidrocarbonetos (HC), NO e NO_2. Além desses, há também os poluentes que são constituídos por material sólido ou líquido de pequenas dimensões, que acabam ficando suspensos no ar e entrando no aparelho respiratório, sendo denominados de material particulado (MP). Suas concentrações servem como indicadores para a qualidade do ar; quanto maior a concentração, pior para a saúde humana. O foco deste artigo ficará relacionado ao impacto causado especificamente pelo MP_{10}, ou seja, partículas com diâmetro aerodinâmico menor que 10 μm.

Um indicativo para relacionar poluição atmosférica e saúde humana pode ser através do número de internação hospitalares relacionadas às doenças respiratórias (DAPPER, 2016). É possível prever o número de internações com o uso de Redes Neurais Artificiais (RNA), que utilizam dados como parâmetros e identificam suas relações, obtendo-se

previsões, ou então, classificações. O tipo de RNA a ser usado dependerá do problema, nesse trabalho será utilizada a rede Perceptron de Múltiplas Camadas (MLP - do inglês *Multi-Layer Perceptron*).

Segundo Manzan (2016), a MLP é uma RNA do tipo *feedfoward*, ou seja, as informações dos dados de entrada são processadas em apenas um fluxo. A propagação das informações é feita pelo produto de cada entrada com seus respectivos pesos, sendo somados e servindo como valor de entrada para as funções de ativação dos neurônios nas camadas escondidas, ocorrendo o mesmo processo entre a camada escondida e a de saída. Após o cálculo do valor de saída, é calculada a função erro e feita a atualização dos pesos, a fim de diminuir o erro na próxima iteração de treinamento da rede, começando pelos pesos da camada de saída e indo até os pesos da camada de entrada. Esse processo de atualização dos pesos é chamado de *backpropagation*. Na Figura 1 é possível ver uma imagem ilustrativa do funcionamento de uma MLP.

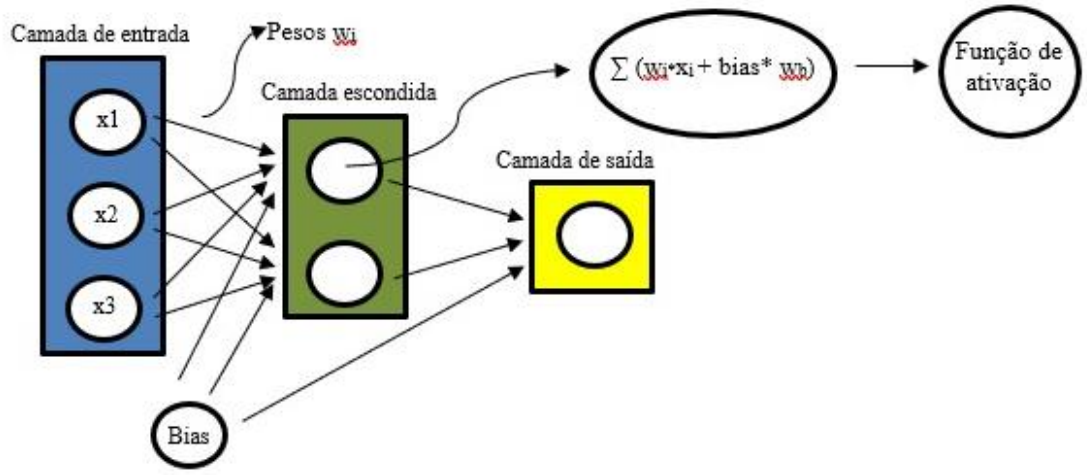

Figura 1. Esquema de funcionamento de uma MLP.

Nesse contexto, o objetivo desse trabalho foi realizar uma Revisão Bibliográfica Sistemática (RBS) referente ao tema e utilizar a MLP para prever o número de internações por doenças respiratórias causadas pelo MP_{10} na cidade de São Paulo.

2. MATERIAIS & MÉTODOS

2.1. Revisão Bibliográfica Sistemática (RBS)

A consulta de artigos, para a parte da RBS, foi executada utilizando 6 bases de dados (*IEEE, PubMed, Science Direct, ACM digital Library, Springer Link* e *Scopus*). As palavras-chave utilizadas foram "*air pollutants*", "*neural network*" e "*health*", com o conectivo "*AND"* sendo usado entre elas. Os anos de publicação dos artigos foram limitados ao período entre 2009 e janeiro de 2020. Para os resultados encontrados em cada banco de dados, foi necessário realizar uma pesquisa manual dos artigos, para identificar quais realmente estavam relacionados à poluição atmosférica e seus danos à saúde com o uso de redes neurais artificiais.

Durante a pesquisa manual, foram selecionados artigos que remetessem ao tema pelo seu título, sendo realizada a leitura dos seus resumos, para que fosse possível ter certeza sobre o tema abordado em cada artigo. Para obter um panorama das publicações no tema, informações específicas foram observadas: RNA usada, principais variáveis e *outputs*.

2.2. Rede Neural Artificial (RNA)

Após o conhecimento e aprofundamento no tema, uma das RNA mais conhecidas, segundo Manzan (2016), e usada no tema abordado (MLP), foi implementada utilizando a linguagem de programação *Python* versão 3.7. Além disso, foi usada a biblioteca *Scikit-learn*, a qual possibilitou a implementação da rede. Foi utilizada a ferramenta MLP *Regressor*, responsável pela criação propriamente da rede. Para tanto, os parâmetros utilizados foram:

a) Quantidade de camadas escondidas: 1
b) Quantidade de neurônios na camada escondida: 5
c) Número máximo de iterações: 200
d) Taxa de aprendizagem: 0,001

e) Função de ativação: logística
f) Valor do erro permitido para interromper o treinamento: 10^{-4}
g) Otimizador dos pesos: *Broyden–Fletcher–Goldfarb–Shanno algorithm* (BFGS).

O otimizador de pesos *Broyden–Fletcher–Goldfarb–Shanno algorithm* (BFGS) foi escolhido devido à melhor convergência dos pesos quando comparado à outros. O BFGS consegue convergir melhor nesse caso, por usar uma quantidade de dados relativamente pequena. De acordo com Cintra (2018), a quantidade de neurônios para a camada escondida e a quantidade de camadas escondidas pode variar dependendo do problema. A taxa de aprendizagem sendo um valor pequeno permite que sejam alcançados os mínimos globais em função do erro. Os dados foram divididos em: 70% treinamento, 15% validação e 15% teste. Para avaliar a qualidade dos resultados, foram utilizados os valores de Erro Quadrático Médio (MSE - do inglês *Mean Square Error*) e Erro Médio Absoluto (MAE – do inglês *Mean Absolute Error*), apresentados nas Equações (1) e (2):

$$MSE = \frac{1}{N_{td}} \sum_{t=1}^{N} (rt - yt)^2 \ , \tag{1}$$

$$MAE = \frac{1}{N_{td}} \sum_{t=1}^{N} |rt - yt| \ , \tag{2}$$

sendo que, *rt* é o valor real observado, *yt* o *output* da rede e N_{td} o número total de dados utilizados na fase de teste.

O tipo de validação utilizada para esse trabalho foi *k-fold*. Os dados de treinamento e validação foram divididos em 5, variando para que fosse possível analisar qual possuía um melhor treinamento da rede. Os dados de poluição atmosférica e meteorológicos foram obtidos pela Companhia Ambiental do Estado de São Paulo (CETESB, 2010), contendo um total de 1.068 dados, que datam de 1º de janeiro de 2014 até 31 de dezembro de 2016. São Paulo é, de acordo com Kachba *et al*. (2020), o maior centro financeiro e comercial da América do Sul, possuindo, segundo IBGE (2019), 12.176.866 habitantes. A cidade possui mais de 8 milhões de veículos circulando diariamente e, conforme a Organização Mundial da Saúde (OMS, 2017), o dobro da concentração de poluentes para os seus padrões. Além disso, de acordo com Araujo *et al*. (2020), a cidade de São Paulo possui uma área total de 1.521,11 km^2, sendo 968,32 km^2 de área urbana.

As variáveis utilizadas foram: concentração de MP_{10}, temperatura ambiente média, umidade relativa do ar, dia da semana e se o dia é feriado ou não. De acordo com Tadano *et al.* (2016), mesmo não sendo variáveis climáticas, saber o dia da semana e de feriados influencia no número de internações, por isso foram consideradas. Os dados de internação foram obtidos pelo Sistema Nacional de Saúde (DATASUS, 2020). Dias que possuíam dados faltantes foram excluídos para evitar problemas na hora de serem usados. Foi necessário realizar uma normalização dos dados para que não ocorresse interferência devido à diferença entre os valores de uma variável e outra, sendo normalizados entre +1 e –1. Para a normalização, foi feito o seguinte procedimento com os dados:

$$v' = \frac{v - min}{max - min} * (novo_{max} - novo_{min}) + novo_{min}, \tag{3}$$

sendo v' o valor normalizado, v o valor que se deseja normalizar, min o valor mínimo da variável em que está realizando a normalização, max o valor máximo da variável que está realizando a normalização, novo_max o valor máximo para o intervalo da normalização (neste caso, será 1) e o novo_min, o valor mínimo para o intervalo da normalização (neste caso, será -1).

Após o treinamento da rede e obtenção dos *outputs*, foi necessário desnormalizar os resultados, para que fossem testados com os valores reais e obtidos o MSE e MAE. A desnormalização dos resultados obtidos foi feita da seguinte maneira:

$$y' = \frac{y - min}{max - min} * (novo\_max - novo\_min) + novo\_min \tag{4}$$

sendo y' o valor obtido pela rede desnormalizado, y o valor obtido pela rede normalizado, min o valor mínimo do intervalo normalizado (-1), max o valor máximo do intervalo normalizado (1), novo_max o valor máximo de internações da fase de teste, e novo_min o valor mínimo de internações da fase de teste.

3. RESULTADOS & DISCUSSÃO

3.1. Revisão Bibliográfica Sistemática (RBS)

Durante a pesquisa dos artigos da RBS foram encontrados, somando todas as seis bases de dados, um total de 2.984 artigos, porém, no final foram selecionados apenas 21 artigos que de alguma forma relacionavam os três temas (poluição atmosférica, saúde humana e RNA). Foi possível observar, por meio da RBS, que as variáveis temperatura, umidade relativa do ar, concentração de poluentes atmosféricos e direção/velocidade do vento são as mais usadas pelos pesquisadores. É importante salientar que não existe uma relação direta quanto maior o número de variáveis, melhor o resultado. Isso acaba gerando a necessidade de seleção de variáveis, baseado no conhecimento físico do problema, gerando um melhor desempenho da rede. Além disso, a RBS ainda mostrou uma predominância no uso da MLP como a RNA mais usada pelos pesquisadores, contabilizando 8 artigos dentre os 21 selecionados, e um foco na previsão de internações por doenças respiratórias, 9 artigos, e na identificação do índice de qualidade do ar para a saúde humana, 6 artigos.

3.2. Rede Neural Perceptron de Múltiplas Camadas (MLP)

Como estudo de caso para aplicação da MLP, foi considerada a cidade de São Paulo, por ser a maior cidade brasileira e ter uma alta concentração de poluentes atmosféricos, de acordo com Kachba *et al*. (2020). A Tabela 1 mostra os resultados de MSE e MAE obtidos das 5 diferentes divisões dos dados de treinamento e validação.

Tabela 1. MSE e MAE para cada teste

Teste	MSE	MAE
#1	2.607	36,71
#2	2.674	37,67
#3	2.730	38,68
#4	2.717	38,37
#5	2.590	36,62

Os menores valores, tanto de MSE quanto de MAE, são do Teste #5, sendo esse o melhor resultado. A Figura 2 mostra os valores reais (em tracejado) junto com os 5 testes realizados pela MLP. Quanto mais sobrepostos estiverem os gráficos dos valores calculados com o real, melhor o resultado. A Figura 2 compara os valores reais com os valores ajustados pelo Teste #5, que foi o que obteve melhor desempenho. Como é possível observar, os valores obtidos pela MLP (linha azul) possuem uma tendência de seguir os valores reais, porém ficam distantes nos picos.

Cabe salientar que existem incontáveis fatores responsáveis por agravar o número de internações por doenças respiratórias, além da poluição do ar e variáveis meteorológicas, portanto, é esperado que as curvas (observada e ajustada) tenham discrepâncias.

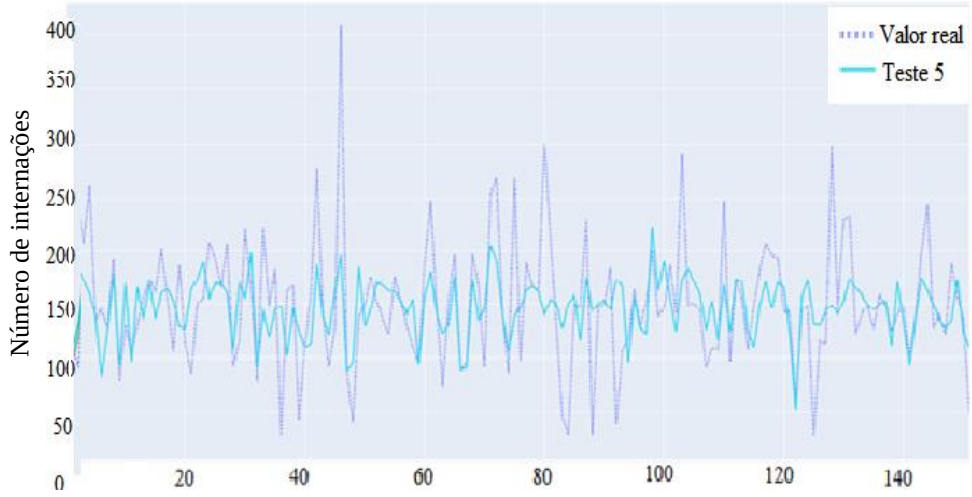

Figura 2. Número de internações hospitalares por doenças respiratórias observadas *versus* estimadas pelo Teste #5

4. CONCLUSÕES

A Revisão Bibliográfica Sistemática (RBS) mostrou ao total 2.984 resultados, sendo selecionados apenas 21 artigos, o que indica uma área de conhecimento ainda pouco explorada no ambiente acadêmico. Houve uma predominância no uso da MLP como rede neural, sendo usada principalmente na tentativa de previsão de internações por doenças respiratórias e no índice de qualidade do ar para a saúde humana. Além disso, é importante notar que as variáveis mais usadas pelos pesquisadores foram temperatura, umidade relativa do ar e concentração de poluentes atmosféricos, sendo as mesmas utilizadas nesse trabalho. Os resultados obtidos pela MLP mostraram uma tendência de aproximação com os dados observados, sendo assim, a utilização da MLP para esse tipo de previsão é válida. Para trabalhos futuros existe a possibilidade do uso de mais variáveis de entrada e o uso de outros tipos de redes neurais artificiais.

5. AGRADECIMENTOS

Agradecimentos são externados ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) pelo apoio financeiro fornecido ao primeiro autor em 2020 através do Programa de Bolsas de Iniciação em Desenvolvimento Tecnológico e Inovação (PIBITI) da Universidade Tecnológica Federal do Paraná (UTFPR).

6. REFERÊNCIAS

Araujo, L.N., Belotti, J.T., Antonini Alves, T., Tadano, Y.S., Siqueira, H., 2020. “Ensemble method based on Artificial Neural Networks to estimate air pollution health risks”. *Environmental Modelling & Software*, Vol. 123, #104567.

Arbex, M.A., Santos, U.P., Martins, C., Saldiva, P.H.N., Pereira, A.A., Braga, A.L.F., 2012. “A poluição do ar e o sistema respiratório”. *Jornal Brasileiro Pneumologia*, Vol. 38, p. 643-655.

Cintra, R., 2018. “Introdução a neurocomputação”. Instituto Nacional de Pesquisas Espaciais.28 de janeiro de 2021. < http://www.inpe.br/elac2018/arquivos/ELAC2018_MC3_apostila.pdf>.

Companhia Ambiental do Estado de São Paulo (CETESB), 2010. 22 de agosto de 2020.<https://cetesb.sp.gov.br/ar/publicacoes-relatorios/>.

Dapper, S. N., Spohr, C., Zanini, R. R., 2016. “Poluição do ar como fator de risco para a saúde: uma revisão sistemática no estado de São Paulo”. *Instituto de Estudos Avançados da Universidade de São Paulo,* Vol 30.

Departamento de Informática do Sistema Único de Saúde (DATASUS), 2020. 22 de agosto de 2020. <http://www2.datasus.gov.br/DATASUS/%20index.php?area=02>.

Instituto Brasileiro de Geografia e Estatística (IBGE), 2019. 22 de agosto de 2020. <https://www.ibge.gov.br/estatisticas-novoportal/sociais/populacao/9103-estimativas-de-populacao.html?¼&t¼resultados>.

Kachba, Y.R., Chiroli, D.M.G., Belotti, J.T., Antonini Alves, T., Tadano, Y.S., Siqueira, H., 2020. “Artificial neural networks to estimate the influence of vehicular emission variables on morbidity and mortality in the largest metropolis in South America”. *Sustainability*, Vol. 12, #2621.

Manzan, J.R. G., 2016. “Análise de desempenho de redes neurais artificiais do tipo multilayer perceptron por meio do distanciamento dos pontos do espaço de saída”. Universidade Federal de Uberlândia.

Ministério do Meio Ambiente (MME), 2020. “Poluentes Atmosféricos”. 22 deagosto de 2020. <https://www.mma.gov.br/cidades-sustentaveis/qualidade-do-ar/poluentes-atmosf%C3%A9ricos.html>.

Organização Mundial da Saúde (OMS), 2017. *Evolution of WHO Air Quality Guidelines: Past, Present and Future*. WHO Regional Office for Europe.

Tadano, Y.S., Siqueira, H.V., Antonini Alves, T., 2016. “Unorganized machines to predict hospital admissions for respiratory diseases”. In *Proceedings of the 2016 IEEE Latin American Conference - LA-CII.*Cartagena, Colombia.

7. RESPONSABILIDADE PELAS INFORMAÇÕES

Os autores são os únicos responsáveis pelas informações incluídas nesse trabalho.