# TREINAMENTO DE CLASSIFICADORES DE APRENDIZAGEM DE MÁQUINA PARA O DIAGNÓSTICO DO GLAUCOMA COM O USO DE DADOS DE PERIMETRIA AUTOMATIZADA PADRÃO (SAP)

**Bruno F. Hashimoto[1], Leonardo S. Shigueoka[2], Vital P. Costa[2], Edson S. Gomi[1]**

[1] Universidade de São Paulo (USP), São Paulo, SP

[2]Universidade Estadual de Campinas (UNICAMP), Campinas, SP

brunohashimoto@usp.br, leonardo.seidi@yahoo.com.br

vp.costa@uol.com.br, gomi@usp.br

***Abstract.*** *Glaucoma is an optical neuropathy that causes damage to the optic nerve and consequent loss of the visual field. This research project investigates the performance of Machine Learning Classifiers in the diagnosis of glaucoma using data from the SAP exam (*Standard Automated Perimetry*). Due to the* dataset *with real patient data being small, a synthetic* dataset *was generated to increase the amount of data available for the training of the classifiers. The classifiers* Random Forest*,* Gradient Boosting *and DNN (*Deep Neural Network*) were tested. The results obtained show that these classifiers present acceptable performance for the diagnosis of glaucoma, all classifiers presented AUC above 83,9% and accuracy above 72,0%.*

***Resumo.*** *O glaucoma é uma doença ocular que provoca danos ao nervo óptico e consequente perda do campo visual. Este projeto de pesquisa tem como objetivo investigar o desempenho de Classificadores de Aprendizagem de Máquina no diagnóstico do glaucoma usando dados do exame SAP (*Standard Automated Perimetry*). Por conta do* dataset *com dados de pacientes reais ser pequeno, foi gerado um* dataset *sintético para aumentar a quantidade de dados disponíveis. Foram testados os classificadores* Random Forest*,* Gradient Boosting *e DNN (*Deep Neural Network*). Os resultados obtidos mostram que esses classificadores apresentam desempenho aceitável para o diagnóstico do glaucoma, todos com AUC acima de 83,9% e acurácia acima de 72,0%.*

## 1. Introdução

O glaucoma é uma doença ocular que provoca danos ao nervo óptico e consequente perda do campo visual [Azuara-Blanco et al. 2001]. O tratamento do glaucoma exige um acompanhamento prolongado do paciente. Médicos oftalmologistas utilizam exames como SAP (*Standard Automated Perimetry*) e OCT (*Optical Coherence Tomography*) para fazer o diagnóstico do glaucoma. Entretanto, os resultados desses exames nem sempre são conclusivos, especialmente no diagnóstico de casos precoces. O uso de técnicas de Aprendizagem de Máquina pode ajudar os médicos a realizar diagnósticos mais precisos. Especificamente, o uso de classificadores baseados em *Deep Learning* tem se mostrado

uma alternativa promissora para melhorar o desempenho no diagnóstico do glaucoma [Asaoka et al. 2016].

O objetivo desse trabalho é investigar se classificadores de aprendizagem de máquina tem desempenho aceitável no diagnóstico de glaucoma usando dados de SAP, com foco nos classificadores *Random Forest*, *Gradient Boosting* e *Deep Neural Network*. Para tornar esses classificadores eficazes, na etapa de treinamento prepara-se uma base de dados que englobe a maior variedade de possíveis casos. No entanto, custos e tempo dificultam a coleta de dados e acabam limitando a quantidade de dados nesses estudos, podendo gerar classificadores com problemas de *overfitting*. O *overfitting* ocorre quando um modelo apresenta bom desempenho em um certo conjunto de dados, mas perde a capacidade de generalização quando apresentado a dados nunca antes vistos. Por conta disso, um objetivo adicional é a geração de pacientes sintéticos a fim de se aumentar o conjunto de dados, de forma a gerar amostras populacionais artificiais mais representativas e melhorar o desempenho dos classificadores [Shorten and Khoshgoftaar 2019].

A Seção 2 apresenta os conceitos do exame SAP e caracteriza os dados reais utilizados neste trabalho. A Seção 3 apresenta os classificadores de aprendizado de máquina utilizados nessa pesquisa. A Seção 4 aborda o método e os procedimentos para obtenção dos resultados. A Seção 5 apresenta os resultados obtidos, a Seção 6, as conclusões e a Seção 7, os agradecimentos.

## 2. Diagnósticos de Glaucoma

A Perimetria Automatizada Padrão (SAP) [Liu et al. 2003] é um exame funcional que avalia a sensibilidade no campo visual do paciente. No exame são apresentados estímulos luminosos ao paciente que, por sua vez, aciona um botão para indicar se o sinal luminoso foi visto por ele. Uma das estratégias empregadas na apresentação dos sinais luminosos nesse exame é a Central 24-2 (C24-2), na qual são testados 54 pontos, sendo 2 deles correspondentes ao ponto cego, que é uma área da retina desprovida de fotorreceptores. Um exemplo de exame SAP é mostrado na Figura 1.

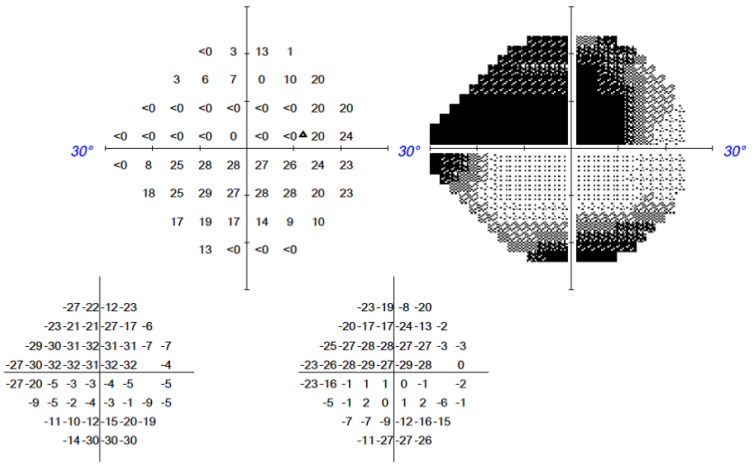


**Figura 1. Exemplo de exame da Perimetria Automatizada Padrão (SAP).**

Alguns dos resultados desse exame são: os valores dos Limiares de Sensibilidade (LS), o *Mean Deviation* (MD) e o *Pattern standard deviation* (PSD). Os LS são os valores das intensidades dos estímulos medidos diretamente dos 54 pontos pelo exame. Como dois deles fazem parte do ponto cego, eles são descartados. O MD é a média ponderada das diferenças entre os LS do paciente e da população normal. O PSD é o desvio-padrão da média dos dados de *Pattern Deviation*, que é a comparação dos limiares de sensibilidade esperados para o próprio paciente. Valores altos de PSD indicam defeitos localizados no campo visual. Outro exame funcional é o *Frequency Doubling Perimetry* (FDT), o qual consiste em estímulos com alternação entre faixas pretas e brancas. Ocorre uma ilusão de duplicação das faixas por conta disso, e é determinado o menor limiar, que permite a percepção dessa duplicação.

O conjunto de dados utilizado neste trabalho foi obtido num projeto de pesquisa realizado no Hospital das Clínicas da UNICAMP [Shigueoka et al. 2018] e que teve o aval do Comitê de Ética dessa Universidade. Os dados são compostos por 227 casos, dos quais 115 são pacientes com diagnóstico de glaucoma e 112 pacientes saudáveis. Os critérios de seleção de pacientes foram os utilizados no estudo publicado em [Shigueoka et al. 2018]. Os pacientes normais tinham visão maior que 20/40, pressão intraocular normal, ângulo aberto, refração não maior que 5 dioptrias esféricas e não maior que 3 dioptrias cilíndricas, 2 exames de FDT normais e disco óptico normal ao exame de fundoscopia. Os pacientes com glaucoma tinham 2 pressões intraoculares maior que 20, alteração no exame do FDT (dentro do critério utilizado) e dano glaucomatoso no nervo óptico. Os pacientes que tinham doenças de retina, catarata ou glaucoma avançado foram excluídos, pois o intuito dos classificadores desse projeto é conseguir fazer o diagnóstico em casos mais difusos da doença, nos quais os exames tradicionais não dão um diagnóstico mais assertivo. Os dados utilizados desses pacientes no treinamento dos classificadores foram o LS, o MD e o PSD oriundos dos exames SAP. A criação do conjunto de dados sintéticos se deu através da adição de ruído gaussiano com média 0 e variância de 0.75 nos valores de LS e cada caso gerou outros 9 casos sintéticos. O conjunto de dados foi organizado aleatoriamente para se obter uma melhor distribuição dos dados.

## 3. Classificadores de Aprendizagem de Máquina

Nos experimentos realizados, utilizamos os seguintes classificadores: Random Forest, Gradient Boosting e Rede Neural.

### 3.1. Random Forest

*Random Forest* (RF) é um método de *Ensemble Learning* utilizado para problemas de classificação. Ele é constituído pela combinação de várias árvores de decisão, cada uma contribui com um único voto e a floresta de árvores como um todo irá escolher a classificação com maior quantidade de votos [Breiman 2001]. Cada árvore é treinada com o mesmo conjunto de dados de treinamento, mas escolhendo variáveis diferentes para montar os seus nós. Há dois principais parâmetros para serem escolhidos na RF: a quantidade de árvores e o número máximo de características na construção de uma dada árvore.

### 3.2. Gradient Boosting

*Gradient Boosting* (GB) [Friedman 2001] é um outro método de *Ensemble Learning* utilizado para classificação. Faz uso uma combinação de árvores de decisão, no entanto, essas

árvores são adicionadas ao modelo sequencialmente e são treinadas para corrigir os erros de predição de árvores previamente alocadas (*Boosting*). Os modelos são treinados com o intuito de otimizar a função *Loss*, reamostrando e variando os pesos ao adicionar novas árvores.

### 3.3. Neural Network

*Neural Networks* (NN) ou Redes Neurais são compostas por unidades de processamento não-lineares (neurônios) em suas camadas de entrada, intermediárias e de saída totalmente conectadas. São bastante utilizadas no reconhecimento de padrões. O treinamento de uma rede neural usa aprendizagem supervisionada com o uso de *backpropagation error*. Na primeira etapa o vetor entrada propaga pela rede, camada por camada, até a saída, utilizando-se de pesos inicialmente fixos e gerados aleatoriamente, *bias* e função de ativação. A segunda etapa ocorre da saída em direção à entrada, o erro quadrado médio entre a resposta obtida e desejada é minimizado, através do reajuste dos pesos sinápticos por cada iteração, até que a rede obtenha um erro aceitável [Chollet 2017].

*Autoencoders* (AE) são redes neurais auto-supervisionadas que reduzem a dimensão da entrada numa primeira etapa (*encoder*) e a reconstituem na saída (*decoder*). A saída de um *encoder* é utilizada para o treinamento de outra AE e, em seguida, conectadas para formar as *Stacked Autoencoders* (SAE). As SAE podem ser utilizadas como forma de pré-treinamento ao inicializar os pesos das camadas de uma Rede Neural para classificação ao se substituir os *decoders* por uma camada classificadora [Vincent et al. 2010]. Dessa forma, o pré-treinamento auxilia na generalização pois garante que as informações contidas nos pesos venham da modelagem dos dados, e as informações dos *targets* de classificação são utilizados para ajustar levemente os pesos obtidos previamente no pré-treinamento [Hinton and Salakhutdinov 2006].

## 4. Metodologia e Procedimentos

A avaliação dos classificadores foi feita com o uso da validação cruzada (*K-Fold cross-validation*) (K-FCV). O conjunto de dados é dividido em k sub-conjuntos e a partir daí, k–1 conjuntos são utilizados para o treinamento dos classificadores e apenas o conjunto restante é utilizado para a validação do modelo. Os sub-conjuntos são usados para treinar iterativamente k modelos, sempre utilizando um conjunto diferente para validação e os outros k–1 como treinamento, até que todos os dados sejam usados na validação. Neste trabalho foram escolhidos k = 10 *folds*. Foram realizados 3 experimentos para cada classificador: o experimento E1 foi realizado no uso apenas dos LS reais, apresentando um vetor de entrada com 52 variáveis (features). O experimento E2 fez uso tanto dos LS como também dos valores de MD e PSD,apresentando 54 variáveis de entrada. No experimento E3, utilizou-se dos dados sintéticos conjuntamente com os dados reais, totalizando 2270 casos, 10 vezes maior que a quantidade de dados reais, com a mesma configuração de vetor de entrada como no E2.

No RF e no GB, o número de árvores foi fixado em 10000 unidades. Outros parâmetros foram mantidos como *default* utilizando-se da ferramenta *Scikit-learn* para *Python*. Implementou-se duas redes neurais *Multilayer Perceptron* (MLP) classificadoras utilizando *Keras* como *backend*. As duas redes tinham a mesma estrutura: 1 camada de entrada, 2 escondidas, uma com 10 e outra com 3 neurônios, e 1 camada de saída. Foi

utilizado o algoritmo de otimização Adam, o algoritmo de *backpropagation error* com função *Loss Mean Squared Error* (MSE), uso de EarlyStopping com 1000 épocas iniciais e mini-batches de tamanho 8 no treinamento das redes. A primeira rede neural (NN) foi implementada com os pesos iniciais aleatórios, enquanto a segunda rede neural (PTNN) teve pré-treinamento de seus pesos com o uso do Stacked Autoencoder.

Os resultados de predição dos modelos sobre todos os dados e suas respectivas classificações verdadeiras foram utilizados para calcular as figuras de mérito: Acurácia (Acc) e a área sob a curva ROC (AUC). Essas métricas são utilizadas para avaliar o desempenho de classificadores binários. A AUC permite resumir a precisão geral do diagnóstico do teste. Uma AUC de valor 0,5 sugere que o desempenho do classificador é o mesmo de um sorteio aleatório; 0,7 a 0,8 é considerado aceitável; 0,8 a 0,9 é considerado bom e maior que 0,9 é considerado excelente [Mandrekar 2010]. Por exemplo, uma AUC de valor 0,8000 sugere uma chance de 80,00% do oftalmologista conseguir distinguir corretamente um paciente normal de um paciente com glaucoma com base na ordem das avaliações dos dados do exame.

## 5. Resultados

Na Tabela 1 são apresentados os resultados da Acurácia (Acc) e da área sob a curva ROC (AUC) obtidos sobre todo o conjunto de dado quando usado para validação através do método K-FCV, dos experimentos E1, E2 e E3.

**Tabela 1. Resultados obtidos dos experimentos**

| **Classificadores** | **E1 (LS)** | | **E2 (LS, MD e PSD)** | | **E3 (10x)** | |
|---|---|---|---|---|---|---|
| | Acc | AUC | Acc | AUC | Acc | AUC |
| Gradient Boosting | 78,41% | 84,94% | 82,38% | 88,90% | 82,73% | 87,53% |
| Random Forest | 82,38% | 87,26% | 82,82% | 89,18% | **84,32%** | **89,49%** |
| Neural Network | 77,97% | 84,39% | 78,85% | 85,86% | 79,96% | 89,45% |
| Pre-Trained Neural Network | 78,41% | 86,58% | 80,62% | 87,14% | 72,03% | 83,98% |

## 6. Conclusão

Os resultados obtidos através de experimentos com *K-fold Cross-validation* em todos os Classificadores de Aprendizagem de Máquina mostraram que é possível criar um classificador de Aprendizagem de Máquina capaz de diagnosticar o glaucoma a partir de dados do exame SAP, mostrando um bom desempenho em relação aos valores obtidos de suas respectivas AUCs. Os resultados experimentais indicam que o RF apresenta o melhor desempenho comparado com os demais classificadores. Esta conclusão faz sentido considerando-se que a quantidade de atributos do exame SAP utilizados nos modelos dos classificadores é pequena e isso leva a um melhor desempenho dos classificadores mais simples. Outra conclusão é que inserir mais atributos do exame SAP como os resultados de MD e PSD aos resultados de LS elevou o desempenho de todos os classificadores. No entanto, adicionar dados sintéticos não apresentou melhoria significativa dos resultados

obtidos, inclusive, houve até redução do desempenho em alguns classificadores, mais evidenciado pela PTNN, o que indica que o método utilizado para gerar os dados sintéticos não foi adequado. Nas pesquisas futuras serão testados outros métodos para gerar dados sintéticos mais representativos em relação aos dados reais, como por exemplo, Redes Adversárias Generativas (GANs) e outros classificadores como Redes Neurais Convolucionais (CNN), uma vez que podemos ver o exame de SAP da Figura 1 também como uma imagem de tons de cinza.

## 7. Agradecimentos

Este trabalho foi realizado com o apoio do Itaú Unibanco S.A., por meio do Programa de Bolsas Itaú (PBI), vinculado ao Centro de Ciência de Dados da Escola Politécnica da USP.

## Referências

Asaoka, R., Murata, H., Iwase, A., and Araie, M. (2016). Detecting preperimetric glaucoma with standard automated perimetry using a deep learning classifier. *Ophthalmology*, 123(9):1974 – 1980.

Azuara-Blanco, A., Costa, V., and Wilson, R. (2001). *Handbook of Glaucoma*. Taylor & Francis.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Chollet, F. (2017). *Deep Learning with Python*. Manning Publications Co., USA, 1st edition.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507.

Liu, J. H. K., Zhang, X., Kripke, D. F., and Weinreb, R. N. (2003). Twenty-four-Hour Intraocular Pressure Pattern Associated with Early Glaucomatous Changes. *Investigative Ophthalmology and Visual Science*, 44(4):1586–1590.

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, 5(9):1315–1316.

Shigueoka, L. S., Vasconcellos, J. P. C. d., Schimiti, R. B., Reis, A. S. C., Oliveira, G. O. d., Gomi, E. S., Vianna, J. A. R., Lisboa, R. D. d. R., Medeiros, F. A., and Costa, V. P. (2018). Automated algorithms combining structure and function outperform general ophthalmologists in diagnosing glaucoma. *PloS one*, 13(12):e0207784.

Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A., and Bottou, L. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12):3371–3408.