

# Classificação de Mamografias Usando MLP e PHoW sobre a base de dados MIAS

André Luiz Tragancin Filho<sup>1</sup>, Gilmário Barbosa dos Santos<sup>2</sup>

<sup>1</sup>Aluno do Curso de Bacharelado em Ciência da Computação da Universidade do Estado de Santa Catarina (UDESC) – Joinville – SC – Brasil.

<sup>2</sup> Professor do Departamento de Ciência da Computação, Universidade do Estado de Santa Catarina (UDESC) – Joinville – SC – Brasil

andre.filho@edu.udesc.br, gilmario.santos@udesc.br

**Abstract.** *Breast cancer is a very lethal and highly occurring form of carcinoma. It is important to detect breast cancer in early stages. CAD (Computer Aided Diagnosis) systems are developed for improving the medical performance specially in diagnostic process. This paper describes the application of a Multilayer Perceptron (MLP) neural network applied for classification of mammographies from MIAS database processed without patches and described by features Pyramid Histogram Of visual Words (PHoW), a method based on SIFT which considers the global and local distribution of information (features) of the image. The experiments were conducted on cases normality, presence of microcalcification and breast density. The potential of the proposal for aid the clinical routine is characterized. The experimental procedures are described, and results are presented and discussed. The results are very good, nevertheless it is important to emphasize that the experiments performed were based only on the MIAS image database, with a PHoW vocabulary restricted to the characteristics of the images in this dataset.*

**Resumo.** *O câncer de mama é um tipo letal e recorrente de carcinoma, sendo por isso importante detectar nos estágios iniciais. Dessa forma, têm se destacado os sistemas do tipo Computer Aided Diagnosis (CAD) construídos a partir de modelos de para apoio à decisão. Este trabalho descreve a aplicação de uma rede neural Multilayer Perceptron (MLP) aplicada à classificação de mamografias do banco de dados MIAS processado sem patches e descrita por características Pyramid Histogram Of visual Words (PHoW), método baseado no Scale Invariant Feature Transform (SIFT) que considera a distribuição global e local de informações (características) da imagem. Os experimentos foram conduzidos em casos de normalidade, presença de microcalcificação e densidade mamária. O potencial da proposta de auxílio à rotina clínica é caracterizado. Os procedimentos experimentais são descritos e os resultados são apresentados e discutidos. Os resultados são muito bons, porém é importante enfatizar que os experimentos se basearam apenas no banco de imagens MIAS, com um vocabulário restrito às características das imagens dessa base de dados.*

## 1. Introdução

Segundo a Organização Mundial de Saúde (OMS) [WHO 2020], o câncer de mama é o tipo mais frequente e o maior causador de mortes relacionadas a câncer na população feminina. Trata-se de uma enfermidade que não se limita a fronteiras socioeconômicas, visto que as taxas de ocorrência de câncer de mama estão crescendo em praticamente todo o mundo [WHO 2020].

É de conhecimento geral que a prevenção do câncer está baseada nas precauções quanto aos fatores de risco e na aplicação de fatores protetivos. Basicamente significa que qualquer fator que aumente as chances de desenvolvimento de câncer representa um fator de risco e deve ser evitado e, ao mesmo tempo, deve-se cultivar hábitos saudáveis, como a boa alimentação e a prática de atividade física [NCI 2020]. Porém, embora alguns fatores de risco sejam plenamente evitáveis, tais como o contato com pesticidas ou o hábito de fumar, outros não o são. É o caso da herança de mutações nos genes BRCA1 e BRCA2, os quais são sérios fatores de risco para o desenvolvimento de câncer de mama [Narod e Salmena 2011]. Sendo herdados, não há como evitá-los.

Se a prevenção por si só não protege efetivamente contra a ocorrência de câncer de mama, a alternativa é tentar detectar qualquer sinal desta doença o mais cedo possível e com isso maximizar as chances de sucesso no seu tratamento. De fato, a detecção precoce é motivadora de políticas públicas para o exame da população feminina, o que ocorre em diversos países inclusive no Brasil, onde o INCA (Instituto Nacional do câncer – José Alencar Gomes da Silva) recomenda o rastreamento bienal do câncer de mama em mulheres na faixa etária dos cinquenta aos sessenta e nove anos de idade [INCA 2020]. O exame de rastreamento utilizado nessas campanhas é basicamente a radiografia da mama, ou mamografia, a qual é visualmente inspecionada pelo especialista médico que determinará um laudo técnico do caso em análise.

*Computer Aided Diagnosis* (CAD) é uma designação geral para *softwares* desenvolvidos para aperfeiçoamento do processo de diagnóstico e da sua acurácia. Esse tipo de *software* é particularmente importante para a prática médica no rastreamento do câncer de mama. Conforme foi mencionado, em geral o rastreamento é feito pela análise das mamografias das pacientes decorrente de inspeção visual, uma estratégia suscetível à fadiga e outros fatores humanos que podem levar ao erro no diagnóstico/prognóstico de um caso.

Muita discussão tem sido feita em torno do uso de sistemas CAD e seu papel no cotidiano de clínicas e hospitais. Por exemplo, Azavedo et al. (2012) comparam a acurácia da análise de exames realizados por um radiologista com o apoio de um sistema CAD em relação à análise dos mesmos casos por diferentes radiologistas, em separado. Independentemente do resultado de estudos desse tipo, é comum a percepção de que o avanço tecnológico (*hardware* e *software*) tende a produzir sistemas CAD que deem cada vez mais suporte ao trabalho do radiologista, compensando eventuais deficiências sensoriais (passageiras ou não) por ele apresentadas. Por exemplo, o sistema poderia destacar regiões em uma mamografia contendo alterações sutis e importantes, porém indetectáveis ao olho humano, poderia realizar uma pré-classificação de mamografias em um processo de triagem para análise posterior por um especialista ou um outro sistema computacional complementar, ou poderia estar focada no desenvolvimento de uma plataforma que incorporasse conceitos avançados como Explainable AI (XAI), agregando ao sistema a importante característica da auditabilidade.

Este trabalho se concentrou em uma proposta de auxílio no cotidiano da prática médica. São relatados experimentos de classificação de mamografias utilizando recurso computacional relativamente modesto (*notebook*), por outro lado, bastante acessível. Os experimentos descritos dizem respeito à aplicação de uma rede neural multicamadas, a qual será designada como MLP, para a classificação de imagens de mamografias. Trata-se de um estudo preliminar visando uma ferramenta útil à triagem clínica de casos.

A imagem mamográfica é composta de *pixels* em tons de cinza, variando de intensidade do preto ao branco. Por se tratar de uma captura de imagem baseada em radiação penetrante (Rx), quanto mais denso é o tecido sujeito a essa radiação, mais claro se apresentará o seu *pixel* correspondente na imagem. A Fig. 1 exibe exemplos de mamografias contendo diferentes densidades de tecidos mamários.

É importante enfatizar que as microcalcificações são sinais muito importantes para a detecção precoce do câncer de mama, assemelham-se a grãos brancos e microscópicos sobre a mamografia [Chan et al. 1990] [Betal et al. 1997] [Chan et al. 1987] [Suckling et al. 1994], conforme pode ser visto na Fig. 1-D. Portanto, quanto mais denso for o tecido mamário, menos contraste haverá entre as possíveis regiões de *pixels* de microcalcificações e *pixels* do tecido mamário de fundo, dificultando a detecção das microcalcificações [Yaghjyan et al. 2017] [Titus-Ernstoff 2017]. A densidade da mama interfere na acurácia da inspeção visual, sendo valioso para o especialista médico conhecer a classe da mamografia quanto a sua densidade [Yaghjyan et al. 2017] [Titus-Ernstoff 2017].

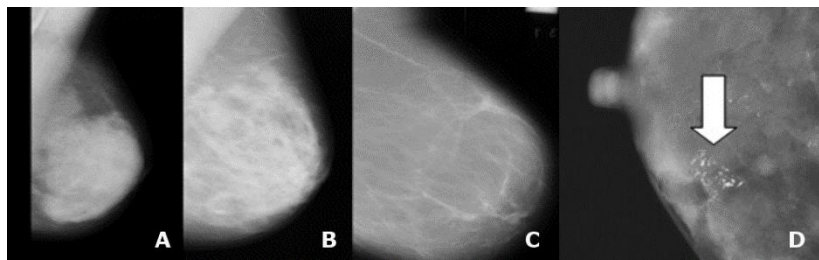
As próximas seções se distribuem da seguinte forma: na seção 2 se descreve o problema tratado. A seção 3 é dedicada a solução proposta especialmente no que tange aos materiais e métodos aplicados nos experimentos, os resultados obtidos são exibidos e analisados na seção 4. As conclusões ocorrem na seção 5.

## 2. Descrição do Problema

A inspeção visual de imagens de mamografias pode ser afetada por diversos fatores (acuidade visual comprometida por cansaço, por exemplo) com impactos sobre o diagnóstico baseado em sistemas CAD. É importante realizar estudos que apliquem técnicas de Aprendizado de Máquina (*Machine Learning*) no campo da análise de mamografias digitais visando avaliar a eficiência desses métodos em relação aos fatores que impactam a análise dessas imagens.

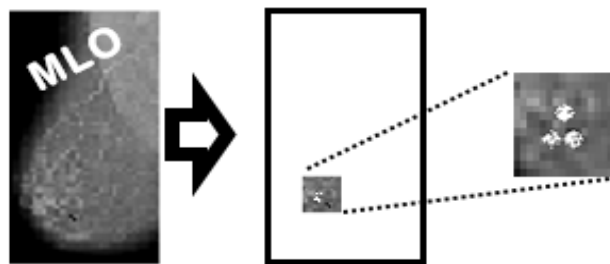
Com a finalidade de contribuir nesse campo de estudo e aprimorar o desenvolvimento de sistemas CAD e tendo como base no anteriormente exposto, foram realizados três tipos de experimentos de classificação visando aplicações na prática de triagem clínica:

1. Classificação binária de acordo com a normalidade do caso: casos normais *versus* casos anormais;
2. Classificação em múltiplas classes baseada na densidade do tecido mamário: baixa densidade *versus* média densidade *versus* alta densidade;
3. Classificação em múltiplas classes baseada na presença de microcalcificações: casos normais *versus* casos que apresentam microcalcificações *versus* outros. A classe “outros” representa os casos anormais que não contém microcalcificações.



**Figura 1. (A) média densidade, (B) alta densidade, (C) baixa densidade (C). À direita, (D) grupo de microcalcificações.**

Uma prática usual em trabalhos em CAD que utilizam mamografias consiste em extrair *patches*, ou seja pequenas regiões de interesse sobre a imagem, é o que ocorre nos trabalhos de Ibrahim et al. (2016), Mehdy et al. (2017), Bozek et al. (2009) e Diamant et al. (2013). A Fig. 2 exibe o esquema dessa operação. A extração dessas regiões pode ser feita por procedimento computacional ou até mesmo por meio de processo manual, conforme ocorre no trabalho de Ibrahim et al. (2016), o que consome muito tempo e é de pouca valia para o uso prático. Posteriormente à etapa de extração de um ou mais *patches* ocorre o treinamento e teste do modelo para classificação dos casos mamográficos.



**Figura 2. Evidenciação e de um patch (região alvo de interesse).**

A proposta deste trabalho não utiliza qualquer determinação de regiões de interesse (*patches*), conseqüentemente, o processo de aprendizagem/teste do modelo leva em conta a imagem como um todo. Essa restrição indica a adoção de um modelo baseado em Aprendizado de Máquina, a escolha recaiu sobre o modelo das Redes Neurais Artificiais em multicamadas (MLP). Neste estudo, o modelo de classificação (MLP) foi treinado e testado utilizando o banco de imagens MIAS [Suckling et al. 1994] – *Mammographic Image Analysis Society*.

### 3. Solução Proposta

Conforme descrito na seção introdutória deste artigo, a proposta desse trabalho consiste na realização de três experimentos de classificação de imagens de mamografias. Os experimentos foram executados em uma máquina virtual Oracle Virtualbox (10 GB RAM) rodando SO Ubuntu 16.04 LTS em um *notebook* Vaio, processador Intel, 16 GB RAM. A seguir são descritos os materiais e métodos aplicados na implementação desses experimentos.

### 3.1. Modelo para a Classificação

O modelo MLP adotado é inspirado nos tecidos neuronais reais. O tecido neuronal natural é composto por uma densa rede de neurônios individualmente capazes de receber estímulos de entrada (via dendritos), que determinam o estado de ativação do neurônio, e enviar estímulos de saída (via terminais do axônio) dependentes do estado assumido pelo núcleo do neurônio. O aprendizado decorre da capacidade de cada neurônio capturar parte de um conhecimento complexo (Fig. 3).

Na Fig. 3, o neurônio artificial emula a capacidade de ativação da sua contraparte natural por meio de um sinal de saída na forma de uma função de ativação ( $f$ ), a qual pode assumir diferentes perfis desde linear até a *sigmoidal*. De fato, a função de ativação representa a fração neuronal do aprendizado instanciado na rede neural. A MLP emula a rede neural real utilizando uma rede de neurônios artificiais em multicamadas, conforme a Fig. 4.

A Eq. 1 descreve a função de ativação ( $f$ ) de entradas ( $w_i, x_i$ ), onde  $X = \{x_1, x_2, \dots, x_n\}$  é o vetor de estímulos de entrada,  $W = \{w_1, w_2, \dots, w_n\}$  o vetor de pesos representando a respectiva ponderação de cada estímulo de entrada e  $b$  é um vetor de viés utilizado para ajustes dos estímulos de entrada.

$$y = f(X.W + b) \quad (1)$$

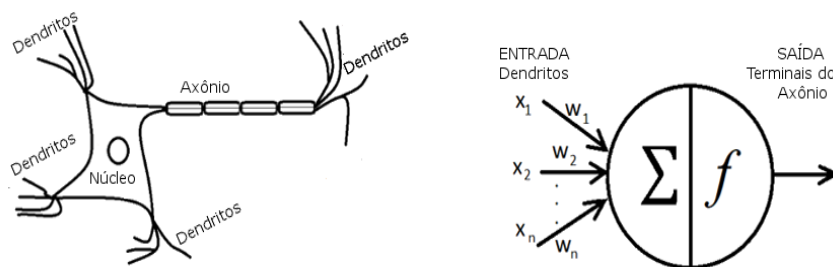


Figura 3. Um neurônio real (esq) e seu equivalente artificial (dir).

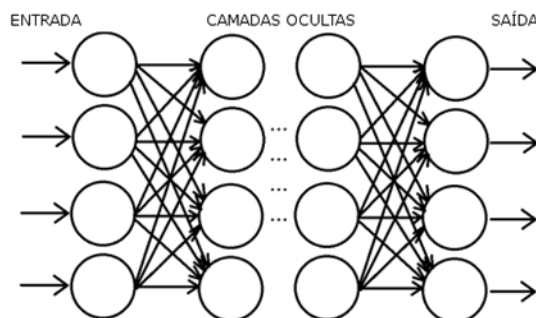


Figura 4. A rede na forma de uma rede multicamadas de neurônios artificiais.

O ajuste dos chamados hiperparâmetros (número de camadas e neurônios, tipo de função de ativação, etc.) é fator importante no processo de aprendizado da MLP. Aqui são tratados alguns aspectos fundamentais, maiores detalhes podem ser vistos nos trabalhos de Buduma e Locascio (2017) e Géron (2017).

A rede utilizada necessita de treinamento supervisionado para assimilar conhecimento, em um processo que busca ajustar o modelo ao seu conjunto de dados de treinamento. Este treinamento não deve ser tal que o modelo se ajuste tanto aos dados de treinamento que não consiga classificar corretamente os dados de testes utilizados para avaliar a acurácia da classificação. É preciso que o modelo seja capaz de generalizar, apresentado uma alta acurácia tanto na classificação de dados de treinamento, quanto na classificação de dados de validação e de teste. De fato, o treinamento pode gerar certas “anomalias de aprendizado” conhecidas como superajuste ou subajuste.

O superajuste (*overfitting*) ocorre quando a MLP não generaliza. Géron (2017) informa que a solução nesse caso é simplificar o modelo, por exemplo, reduzindo o número de atributos dos dados ou utilizando mais dados no conjunto de treinamento. O oposto do superajuste é o subajuste (*underfitting*), o qual acontece quando a rede é muito simples para apreender a complexidade inerente aos dados. Nesse caso, Géron (2017) indica como solução a reparametrização do modelo visando aumentar a sua complexidade. Por exemplo, aumentando o número de camadas e neurônios.

Como neste trabalho se utilizou o processamento das imagens como um todo sem a extração prévia de *patches*, tornou-se necessário um modelo capaz de aprender as sutis diferenças entre mamografias consideradas em seu todo. Por ser composto de uma larga rede de neurônios, individualmente capazes de capturar parte do complexo conhecimento que diferencia os casos mamográficos, o modelo MLP é uma escolha natural nesse contexto.

Poderia se aventar a utilização de uma rede convolucional, porém, não está disponível para este trabalho a capacidade computacional exigida por este modelo.

### 3.2. Banco de Imagens

As imagens utilizadas neste trabalho foram obtidas na base de dados MIAS [Suckling et al. 1994], a qual apresenta 322 mamografias em tons de cinza de dimensões 1024 por 1024 *pixels*. Trata-se de uma base referenciada desde a década dos anos noventa, portanto é confiável e estável. A documentação dos casos é realizada em um simples, porém é bem informativo, arquivo de texto. A Tabela 1 exibe um exemplo da informação contida no MIAS, maiores detalhes podem ser obtidos na referência Suckling et al. (1994).

**Tabela 1. Fragmento do MIAS**

Caso	Densidade do tecido	Anormalidade
mdb066	D (média densidade)	NORM (sem anormalidade)
mdb069	F (baixa densidade)	CIRC (lesão circular)
mdb070	F (baixa densidade)	NORM (sem anormalidade)
mdb209	G (alta densidade)	CALC (contém microcalcificações)

### 3.3. Pyramid of Histograms of Bag of Words (PHoW)

A caracterização dos dados utilizados neste trabalho é realizada por meio do descritor PHoW, o qual foi concebido para aprimorar o método *Bag of visual Words* (BoW) [Lazebnik et al. 2006] [Bosch et al. 2007] [Gao et al. 2015]. O BoW representa uma

imagem como uma coleção de características (*features*) globais da imagem, sem destacar aspectos mais locais contidos em amostras regionais das características extraídas. Por ser composto de uma pirâmide de amostras, o descritor PHoW possibilita uma representação mais detalhada, em relação ao BoW. A Fig. 5 apresenta uma interpretação simplificada da construção da representação BoW e da PHoW.

Os polígonos apresentados na Fig. 5 são chamados de “palavras visuais”, essas palavras são centros de *clusters* de características determinadas por um extrator denso, tal como o operador *Scale-Invariant Feature Transform* (SIFT), aplicado sobre a imagem. No lado esquerdo na Fig. 5: o descritor BoW é uma informação global da imagem e corresponde ao histograma de ocorrência dos polígonos. Por sua vez, ao lado direito da mesma figura, o descritor PHoW é apresentado como uma concatenação de três histogramas regionais determinados por quatro subdivisões da imagem. Essas regiões permitem uma coleção de características por região, portanto PHoW representa mais detalhes da imagem do que BoW [Lazebnik et al. 2006] [Bosch et al. 2007] [Gao et al. 2015], tendo sido o descritor adotado neste trabalho.

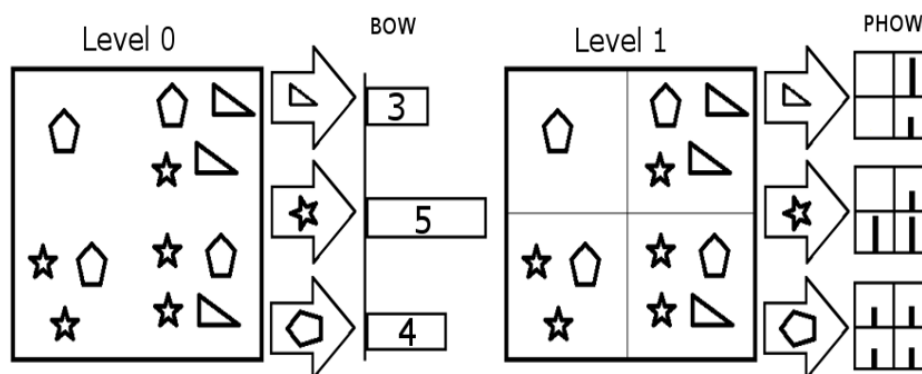


Figura 5. Esquerda: BoW como um único histograma. Direita: PHoW como uma concatenação de três histogramas regionais. Desenho adaptado [Lazebnik et al. 2006].

### 3.4 Experimentos

Foram implementados três tipos de experimentos envolvendo um total de oito classes de casos mamográficos:

- Experimento-1: classificação de casos com base na normalidade (presença ou não de alguma anomalia) utilizando rótulos *Normal* e *Not\_Normal*.
- Experimento-2: classificação pela densidade do tecido mamário utilizando rótulos *Glandular* (alta densidade), *Dense* (média densidade) e *Fatty* (baixa densidade).
- Experimento-3: classificação baseada na presença de microcalcificação utilizando os rótulos *Normal*, *Calc* e *Outros*. Onde *Calc* se refere a casos apresentado microcalcificações e *Outros* se refere aos casos anormais, porém, sem microcalcificações.

A biblioteca *PHoW-Code* está disponível na internet [VGG 2020] e foi utilizada para a extração de características. Programas em linguagem Python foram implementados para a preparação de dados e outros procedimentos, conforme é descrito abaixo.

Para o modelo MLP foi utilizado o pacote *MLPClassifier* do módulo Python *sklearn.neural\_network* [MLPClassifier 2020], o qual implementa uma rede neural multicamadas e métodos para treinamento/teste, otimização de hiperparâmetros, bem como geradores de gráficos e outros recursos. Também foram utilizados recursos da biblioteca Python-OpenCV, bem como Python-Numpy. A Tabela 2 exibe a sequência dos procedimentos aplicados, cujas descrições são realizadas a seguir.

**Tabela 2. Sequência dos Procedimentos Aplicados**

Pré-processamento		Implementação do modelo MLP	
A) Preparação de dados	B) Processamento de Imagens	C) Extração de características	D) Treinamento e testes da MLP

### 1) Preparação de Dados

Foi aplicada uma operação de aumento de dados por meio de rotações visando ampliar a disponibilidade de casos. Para tanto, cada imagem foi rotacionada em diferentes ângulos bem como sofreu operação de espelhamento vertical (*flip*).

Considerando duas imagens  $I_i$  e  $I_j = T(I_i)$ , onde  $T$  corresponde a uma operação de rotação ou espelhamento,  $I_i$  e  $I_j$  apresentam os mesmos histogramas globais, porém diferentes descritores PHoW, pois  $T$  altera a localização dos *pixels*. Ao mesmo tempo,  $I_i$  e  $I_j$  pertencem a uma mesma classe, pois  $T$  não altera a atribuição de classe nem o histograma da imagem. Dessa forma,  $T$  amplia a base de dados e ambas as imagens podem ser utilizadas para treinamento e testes do modelo. Cada uma das oito classes utilizadas nos experimentos foi preenchida com 600 casos.

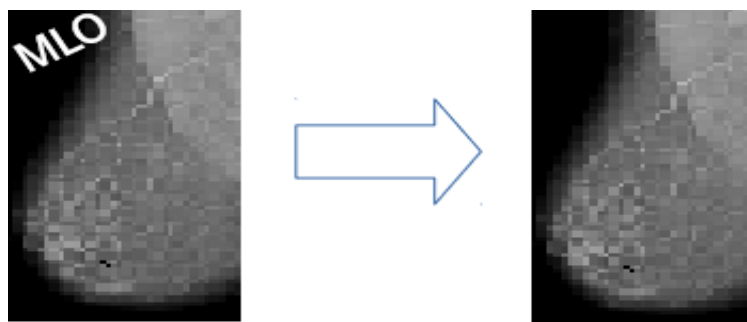
Três experimentos foram conduzidos utilizando a mesma base de dados contendo casos originalmente listados em uma única tabela. Foi necessário implementar as relações de casos de acordo com o tipo de experimento e os rótulos de classe envolvidos. Cada relação representou uma “visão” da base de dados a partir dos respectivos rótulos de classe, permitindo acessar a base de dados selecionando os casos necessários a cada experimento.

### 2) Processamentos de Imagens

É necessário segmentar a região da mama eliminando marcações. Ao lado esquerdo na Fig. 6 é possível verificar a existência de uma imagem contendo o marcador “MLO”, o qual é inserido pelo operador do equipamento radiológico para identificar a direção do eixo de Rx incidente sobre a mama, no caso, o “MLO” indica o eixo médio-lateral oblíquo. Esse tipo de marcação é considerado um artefato e deve ser eliminada, conforme se verifica na imagem ao lado direito na Fig. 6.

Para a segmentação desejada, é aplicada uma “binarização” seguida por uma separação da região da mama. Para esta segmentação considera-se que a mama corresponde à maior região em *pixels*.





**Figura 6.** Esquerda: imagem apresentando marcação “MLO” utilizada para indicar a direção do Rx. Direita: imagem pós-segmentação da mama.

### 3) Extração de Características

Para cada imagem mamográfica segmentada no passo anterior, foi extraído o seu vetor de PHoW, o qual é associado ao respectivo rótulo de classe (conhecido previamente por meio da documentação do MIAS) formando uma linha da tabela, conforme exibido no Quando 1. A casuística de mamografias de um experimento é então caracterizada pela sequência de vetores (PHoW+Rótulo) constituindo uma tabela de dados, essa base inicial é então submetida a uma redução de atributos por meio da transformada PCA (*Principal Component Analysis*), obtendo-se uma base mais concisa. Esse procedimento foi repetido para cada experimento formando as três bases de dados necessárias. Por exemplo, para o experimento n. 1 a transformada PCA reduziu de 6300 atributos para 1200 atributos e para os experimentos n. 2 e n. 3 a redução foi de 6300 atributos para 1800 atributos.

É importante otimizar o uso de memória com a redução de número de atributos, especialmente no caso em questão, quando se utiliza um recurso computacional relativamente limitado (*notebook*).

←	PHoW				→	Rótulo
10.1	...	...	2.2	8.2		0

**Quadro 1.** Uma Linha da Base de Dados: PHoW e Rótulo.

### 4) Treinamento e teste da rede neural

#### 4.1) Método aplicado para o Experimento-1 e o Experimento-2:

##### 4.1.1) Carga da base de características (PHoWs+rótulos);

4.1.2) Divisão da base em conjunto-treino e conjunto-teste. Foram utilizados 67% de dados para treinamento e 33% para testes.

4.1.3) Escolha do Classificador: foi realizado a partir das considerações já discutida em seção anterior, foi escolhido o modelo MLP implementado no módulo Python Sklearn;

4.1.4) Treinamento e especificação dos parâmetros por cross-validation: para cada iteração de treinamento, o conjunto-treino inicialmente determinado é dividido em duas partes: treino (80%) e validação (20%). Esses novos conjuntos são utilizados no processo de cross-validation, no qual várias iterações de treinamento são realizadas para diferentes combinações de casos entre treino/validação. Além de ser utilizado treinamento a cross-

validation também é aplicada no processo de otimização de parâmetros, o qual é realizado via funcionalidade `gridSearchCV(...)` da API e é descrito a seguir.

#### 4.1.5) Determinação dos hiperparâmetros:

Os principais hiperparâmetros são:

- Número de camadas escondidas e de neurônios por camada: 3 camadas de 60 neurônios cada. Determinadas empiricamente;
- O solver, o qual otimiza os pesos ( $W$  na Eq. 1), foi determinado de acordo com a documentação do módulo Python-Sklearn [MLPClassifier 2020] para bases de dados relativamente pequenas. Foi determinado o solver `lbfgs`;
- Taxa de aprendizado: seguindo a documentação do Python, foi determinada como adaptativa;
- Tipo de função de ativação: por meio de grid-search, entre as três alternativas – `tanh`, `logistic-sigmoid` e `relu` a melhor performance foi atribuída a `tanh` (tangente hiperbólica).

4.1.6) Análise da curva de aprendizado e, se necessário, ajuste do modelo retornando ao passo 4 (ou até mesmo ao passo 3, se necessário);

4.1.7) Teste final do modelo usando o conjunto-teste.

## 4. Resultados

Para cada experimento são apresentados o gráfico da curva de aprendizado, a matriz de confusão e a acurácia média em termos das métricas *Precision* e *Recall* conforme a Eq. 1 e a Eq. 2. Para essas equações, *TP*, *FP* e *FN* correspondem aos totais de casos classificados como verdadeiro-positivos (*True Positive*), falso-positivos (*False Positive*) e falso-negativos (*False Negative*).

$$\text{Precision} = TP / (TP + FP) \quad (1)$$

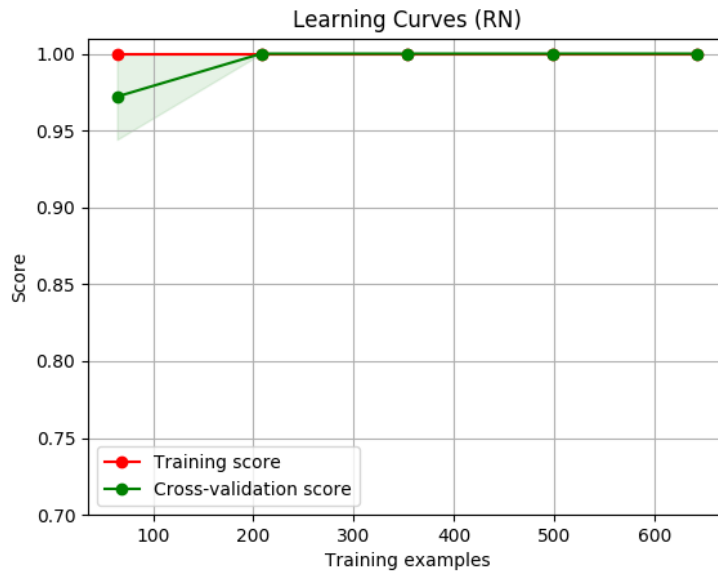
$$\text{Recall} = TP / (TP + FN) \quad (2)$$

A matriz de confusão também é apresentada e resulta dos testes finais do modelo utilizando o conjunto-teste. *Precision* (Eq. 1) e *Recall* (Eq. 2) são representações concisas do conteúdo da matriz de confusão. *Precision* indica quanto das predições positivas para certa classe foram corretamente detectadas pelo classificador (baseia-se nas colunas da matriz de confusão). Em complemento, *Recall* informa quanto dos casos verdadeiramente positivos para certa classe foram corretamente classificados/detectados pelo classificador (baseia-se nas linhas da matriz de confusão).

As curvas de aprendizado são obtidas durante o treinamento da MLP. Para cada iteração de *cross-validation*, o conjunto-treino inicial é subdividido em dois: *treino* e *validação*. Ao final de cada iteração de treinamento são calculadas as pontuações *training-score* e *validation-score* sobre a taxa de acertos obtidos pelo uso dos respectivos conjuntos *treino* e *validação*. Essas pontuações determinam as curvas *cross-validation-score* e *training-score* que constituem os respectivos gráficos de aprendizado (Fig. 7, Fig. 8 e Fig. 9).

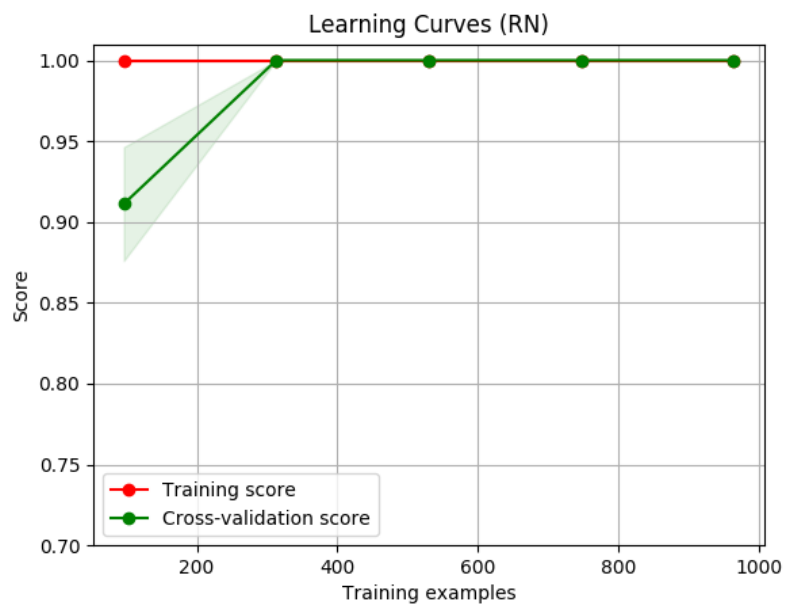
Um gráfico de aprendizado caracteriza o ajuste (*fitting*) do modelo MLP ao treinamento e possibilita avaliar a ocorrência de *superajuste* ou *subajuste*, bem como permite a tomada de decisões para corrigir o modelo e/ou o conjunto de dados.

*A. Experimento 1 – Normal vs Not\_Normal*



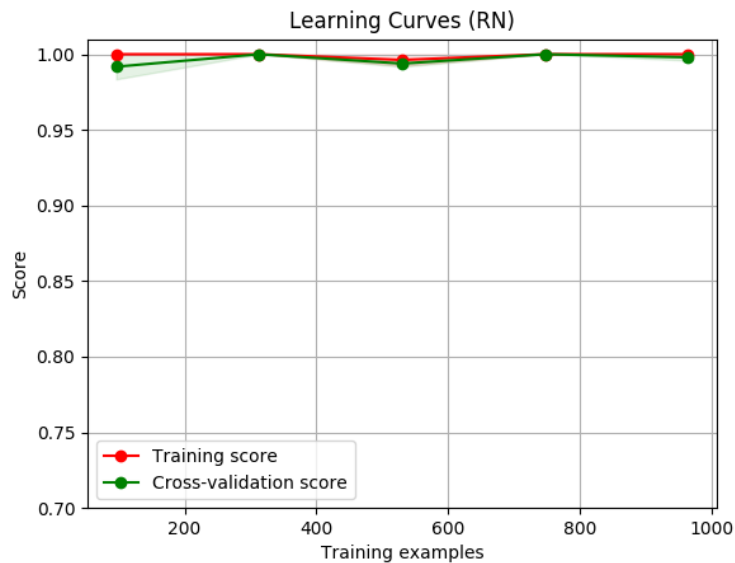
**Figura 7. Curva de aprendizado durante treinamento - Experimento 1.**

*B. Experimento 2 – Dense vs Fatty vs Glandular*



**Figura 8. Curva de aprendizado durante treinamento - Experimento 2.**

### C. Experimento 3 – Normal vs Calc vs Outros



**Figura 9. Curva de aprendizado durante treinamento - Experimento 3.**

### D. Análise

A seguir são levantados alguns pontos importantes para as análises dos resultados. Cada instância do gráfico de treinamento é composta pela curva *training-score* e pela curva *validation-score*, se a curva *training-score* e a curva *validation-score* exibem um perfil de valores baixos, significa que a MLP não capturou as sutilezas dos casos de treinamento, estando um estado de subajuste (*underfitting*) [Scikit-learn 2020].

Se a curva *training-score* é alta enquanto a curva *validation-score* é baixa, então a MLP está em estado de superajuste (*overfitting*) e não consegue generalizar para os dados de validação e teste. Caso ambas curvas *training-score* e *validation-score* converjam em valores altos, significa que a MLP aprendeu com o treinamento e é capaz de generalizar seu conhecimento para os dados de validação e teste. Esta é a situação que se deseja.

A matriz de confusão é construída com base nas predições da MLP (treinada) submetida aos casos no conjunto-teste. Para a análise da matriz de confusão se observa sua diagonal principal. Quanto maiores forem os valores da diagonal principal em relação aos outros valores da matriz de confusão, mais acurado é o modelo.

Quanto às métricas *Precision* e *Recall*, quanto mais próximos de zero estiverem FN e FP, mais próximos da unidade estarão as métricas *Precision* e *Recall* e melhor será a acurácia da MLP.

No caso do Experimento-1, em relação às duas curvas que compreendem o gráfico na Fig. 7, a curva *training-score* apresenta valores altos durante todo o treinamento ao passo que curva *cross-validation* apresenta-se inicialmente mais baixa, porém as duas curvas se aproximam à medida que o treinamento avança. As duas curvas se igualam, na máxima acurácia, a partir do 200º exemplo de treinamento, continuando juntas e altas até o final do treinamento.

Para esse experimento, 1/3 da base de 1200 casos foi separada para a fase de teste do modelo treinado, ou seja, 396 casos foram dedicados aos testes finais do modelo.

A matriz de confusão na Tabela 3 apresenta uma classificação sem erros, o que é corroborado pelas métricas *Recall* e *Precision* apresentadas na Tabela 4.

**Tabela 3.**  
**Matriz de Confusão - Teste do Modelo - Experimento 1**

	Predições	
	Normal	Not_Normal
Verdade	Normal	201
	Not_Normal	0

**Tabela 4.**  
**Acurácia - Teste do Modelo - Experimento 1**

Classe/Média	Precision	Recall
Normal	1.0	1.0
Not_Normal	1.0	1.0
Média	1.0	1.0

O perfil apresentado pelo gráfico de treinamento demonstra uma evolução de aprendizado no qual os testes com os casos do conjunto *validação* confirmam a capacidade da MLP em generalizar, classificando corretamente tanto o conjunto *treino* quanto o conjunto *validação*.

Ao mesmo tempo, a matriz de confusão se apresenta como uma matriz-diagonal, um resultado de classificação perfeita corroborado pelos valores (máximos) obtidos pelas métricas *Precision* e *Recall*. Tais resultados demonstram que o modelo MLP representou internamente um conhecimento capaz de diferenciar mamografias do banco MIAS, no que se refere à normalidade clínica dos casos, categorizando-as nas classes: *Normal* ou *Not\_Normal*.

Para o Experimento-2, o gráfico apresentado pela Fig. 8 é similar àquele na Fig. 7, apesar da curva *validation-score* se apresentar inicialmente mais distante da curva *training-score*, ambas convergem para um mesmo perfil na abscissa 260 aproximadamente, mantendo-se em um perfil de altos valores até o final do treinamento.

Para esse experimento, 1/3 da base de 1800 casos foi separada para a fase de teste do modelo treinado, ou seja, 594 casos foram dedicados aos testes finais do modelo. Os resultados são exibidos na Tabela 5, representando a matriz de confusão. As métricas *Precision* e *Recall* apresentam valores na Tabela 6.

A matriz de confusão apresenta-se na forma de uma matriz diagonal, o que significa que a fase de teste apresentou um resultado bastante acurado. O mesmo pode ser dito dos valores obtidos para as métricas *Precision* e *Recall*. A alta acurácia apresentada é similar àquela obtida para o experimento anterior. Tais resultados demonstram que o modelo MLP apreendeu um nível de conhecimento capaz de diferenciar mamografias do

banco MIAS, no que se refere à densidade do tecido mamário, categorizando-as nas classes: *Dense*, *Fatty* ou *Glandular*.

**Tabela 5.**  
**Matriz de Confusão - Teste do Modelo - Experimento 2**

	Predições			
		Dense	Fatty	Glandular
Verdade	Dense	203	0	0
	Fatty	0	184	0
	Glandular	0	0	207

**Tabela 6.**  
**Acurácia - Teste do Modelo - Experimento 2**

Classe/Média	Precision	Recall
Dense	1.0	1.0
Fatty	1.0	1.0
Glandular	1.0	1.0
Média	1.0	1.0

Em relação ao Experimento-3, o gráfico na Fig. 9 apresenta duas curvas de valores próximos do máximo (1,0) durante todo o treinamento, com a curva *cross-validation* praticamente unida à curva *training-score* já no início do processo. No decorrer do treinamento, as duas curvas praticamente se equivalem, apresentando algumas pequenas variações, mas sempre se mantendo próximas do valor máximo (1,0).

A quantidade de dados selecionados para teste do modelo é a mesma utilizada para o experimento anterior: 1/3 da base de 1800 casos equivalente a 594 casos para os testes finais do modelo. Os resultados são exibidos na matriz de confusão presente na Tabela 7, enquanto as métricas *Precision* e *Recall* constam na Tabela 8.

A matriz de confusão é diagonal (Tabela 7) e as métricas *Precision* e *Recall* apresentam valores ótimos (Tabela 8).

**Tabela 7.**  
**Matriz de Confusão - Teste do Modelo - Experimento 3**

	Predições			
		Normal	CALC	Outros
Verdade	Normal	203	0	0
	CALC	0	184	0
	Outros	0	0	207

**Tabela 8.**  
**Acurácia - Teste do Modelo - Experimento 3**

Classe/Média	Precision	Recall
<b>Normal</b>	1.0	1.0
<b>CALC</b>	1.0	1.0
<b>Outros (casos anormais sem microcalcificações)</b>	1.0	1.0
<b>Média</b>	1.0	1.0

Conforme se observa pelos resultados obtidos, foi bem-sucedida a utilização da versão retreinada da MLP do Experimento-2. Os resultados do Experimento-3 alcançaram o nível ótimo. Tais resultados demonstram que o modelo MLP apreendeu o conhecimento necessário à discriminação de mamografias do banco MIAS no que se refere ao critério da presença de microcalcificação, referido por meio dos rótulos de classes: *Normal*, *Calc* e *Outros*.

É importante destacar que o método PHoW depende fortemente do vocabulário obtido na etapa precedente, na qual se determinam as *BoW features*. Os experimentos realizados se basearam no vocabulário obtido a partir do banco de imagens MIAS, podendo apresentar variações se realizados em outros bancos de imagens.

## 5. Conclusões

Este artigo descreveu a aplicação do modelo MLP utilizando o módulo Python-Sklearn para a classificação de mamografias do banco de imagens MIAS processadas no seu todo (sem o uso de *patches*) e descritas por meio do modelo PHoW. Foram realizados experimentos de classificação de casos baseados: na presença de anormalidade, na densidade de tecido mamário e na presença de microcalcificações. Esses casos são considerados importantes pela literatura.

Considerando o escopo deste estudo e de acordo com as análises dos resultados obtidos, conclui-se que:

- A. *O modelo e seu treinamento*: o processo aplicado para o treinamento do modelo foi bem sucedido. O modelo adotado, implementado pelo módulo *Python SKlearn-MLP*, foi capaz de apreender o nível de conhecimento capaz de diferenciar mamografias do banco no que se refere às classes experimentadas;
- B. *Descritor*: o uso do descritor PHoW foi efetivo para representar a complexidade das características presentes nas imagens das casuísticas tratadas (mamografias MIAS);
- C. *Patches*: foi demonstrada a viabilidade da classificação desejada utilizando a imagem mamográfica no seu todo, sem o uso de *patches*;
- D. *Proposta*: foi demonstrado o potencial da proposta para a classificação de mamografias nas classes consideradas importantes para rotinas clínicas, em apoio a ações de triagem de casos.

É importante destacar que o desempenho do método PHow depende fortemente do vocabulário obtido na etapa precedente, na qual se determinam as BoW features. Os experimentos realizados neste trabalho se basearam apenas no banco de imagens MIAS, com um vocabulário BoW mais restrito às características das imagens desse banco de dados.

Como não há uma padronização rigorosa para as configurações dos sistemas de radiografia no processo de aquisição de imagens, os bancos de imagens podem apresentar características (de imagens) diferentes entre si. Por exemplo, a obtenção da mamografia digital pode ocorrer diretamente pelo uso de um equipamento de radiografia digital ou pode advir de um processo de “escaneamento” digital de uma radiografia convencional da mama, com o risco de apresentarem-se artefatos na imagem.

Seria interessante que esse potencial diferenciação fosse capturada pelo vocabulário na fase de extração dos descritores BoW. Como trabalho futuro se recomenda a utilização de imagens de diferentes bases de mamografias testando a capacidade do método em adquirir um vocabulário mais generalista/abrangente e atingir um ótimo desempenho.

Ainda sobre as recomendações para trabalhos futuros, a alta disponibilidade de dispositivos móveis associada a vastidão territorial brasileira, motiva o desenvolvimento de sistemas CAD embarcados na forma de aplicativos de *smartphones*. Esse tipo de sistema poderia servir como base de conhecimento para localidades remotas que não disponham de médicos especialistas, nesse caso um profissional generalista poderia triar casos com mais segurança. A proposta apresentada poderia servir como base para esse tipo de experimentação.

## Agradecimentos

Os autores gostariam de agradecer à Universidade do estado de Santa Catarina (UDESC) e a Fundação de Amparo à Pesquisa e Inovação do Estado de Santa Catarina (FAPESC) pelo suporte a este trabalho.

## Referências

- CANCER. Documento disponível no URL: <https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/>. Último acesso em 17/09/2020.
- NCI- Breast Cancer Prevention (PDQ®)–Patient Version. Documento disponível no URL: <https://www.cancer.gov/types/breast/patient/breast-prevention-pdq> Último acesso em 17/09/2020.
- S.A. Narod, L. Salmena. BRCA1 and BRCA2 Mutations and Breast Cancer. *Discovery Medicine*; ISSN: 1539-6509; *Discov Med* 12(66):445-453, November 2011. Documento disponível no URL: <http://www.discoverymedicine.com/Steven-A-Narod/2011/11/25/brca1-and-brca2-mutations-and-breast-cancer/> Último acesso em 17/09/2020.
- INCA – Instituto Nacional de Câncer. Detecção Precoce do Câncer de Mama. Documento acessível no URL: <https://www.inca.gov.br/controlado-cancer-de-mama/acoes-de-controlado/deteccao-precoce> Último acesso em 17/09/2020.



- E. Azavedo, S. Zackrisson, I. Mejåre, A. M. Heibert. Is single reading with computer-aided detection (CAD) as good as double reading in mammography screening? A systematic review. *BMC Medical Imaging*. 12:22, 2012.
- L. Yaghjyan et al. Tissue-based associations of mammographic breast density with breast stem cell markers. *Breast Cancer Research*. 19:10, 2017.
- L. Titus-Ernstoff et al. Breast cancer risk factors in relation to breast density (United States). *Cancer Causes Control*. 17:1281–1290, 2006.
- H. P. Chan et al. Improvement in radiologist’s detection of clustered microcalcifications on mammograms – the potential of computer aided-diagnosis. *Invest. Radiol.*, 25, 1990.
- D. Betal, N. Roberts, G.H. Whitehouse. Segmentation and numerical analysis of microcalcifications on mammograms using mathematical morphology. *The British Journal of Radiology*, 70(837), 1997.
- H. P. Chan et al. Image feature analysis and computer-aided diagnosis in digital radiography. i. automated detection of microcalcifications in mammography. *Medical Physics*, 14(4), 1987.
- J. Suckling et al. The Mammographic Image Analysis Society Digital Mammogram Database. *Excerpta Medica. International Congress Series* 1069 pp375-378, 1994. Documento e base de dados disponíveis no URL: <http://www.mammoimage.org/databases/>. Último acesso em 17/09/2020.
- A. M. Ibrahim, B. Baharudin, A. M. Said, P. N. Hashimah. Classification of Breast Tumor in Mammogram Images Using Unsupervised Feature Learning. *American Journal of Applied Sciences*, 13 (5): 552.561, 2016.
- M.M. Mehdy, P.Y. Ng, E.F. Shair, N.I. Md Saleh and C. Gomes. Survey: Artificial Neural Networks in Image Processing for Early Detection of Breast Cancer. *Hindawi Computational and Mathematical Methods in Medicine*. Volume 2017.
- J. Bozek, M. Mustra, K. Delac, M. Grgic. A Survey of Image Processing Algorithms in Digital Mammography. In: Grgic M., Delac K., Ghanbari M. (eds) *Recent Advances in Multimedia Signal Processing and Communications. Studies in Computational Intelligence*, vol 231, 2009. Springer, Berlin, Heidelberg.
- I. Diamant, J. Goldbergerb, H. Greenspana. Visual Words Based Approach for Tissue Classification in Mammograms. *Medical Imaging, Proc. of SPIE*, vol. 8670, 2013.
- N. Buduma and N. Locascio. *Fundamentals of deep learning: designing next-generation machine intelligence algorithms*. O’Reilly, December 2017.
- A. Géron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow, Concepts, Tools, and Techniques to Build Intelligent Systems*. O’Reilly, December 2017.
- S. Lazebnik, C. Schmid, J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2169–2178). 2006.
- A. Bosch, A. Zisserman, and X. Munoz. Image classification using random forests and ferns. In *ICCV*, 2007.

H. Gao, W. Chen and L. Dou. Image classification based on support vector machine and the fusion of complementary features. Journal CoRR, 2015. Document available at Universal Resource Locator: <http://arxiv.org/abs/1511.01706>

PHOW-code VGG. Biblioteca disponível pelo Library available at Uniform Resource Locator: <http://www.robots.ox.ac.uk/~vgg/research/caltech/code/phow.zip> Último acesso em 17/09/2020.

Python-Sklearn MLPClassifier. Documento, tutorial e exemplos disponíveis no URL: [http://scikit-learn.org/stable/modules/neural\\_networks\\_supervised.html#multi-layer-perceptron](http://scikit-learn.org/stable/modules/neural_networks_supervised.html#multi-layer-perceptron) Último acesso em 17/09/2020.

Python-Sklearn. Validation curves: plotting scores to evaluate models. Documento disponível no URL: [http://scikit-learn.org/stable/modules/learning\\_curve.html](http://scikit-learn.org/stable/modules/learning_curve.html) Último acesso em 17/09/2020.