Classificação de Imagens de Biópsias Renais com Glomeruloesclerose Segmentar e Focal ou com Lesões Mínimas Utilizando *Transfer Learning* em CNN

Justino Duarte Santos[1,2], Rodrigo de Melo Sousa Veras[2], Romuere Rodrigues Veloso e Silva[3], Nayze Lucena Sangreman Aldeman[4] Kelson Romulo Teixeira Aires[2], Andrea Gomes Campos Bianchi[5]

[1] Instituto Federal de Educação, Ciência e Tecnologia do Piauí (IFPI)
São Raimundo Nonato - PI - Brasil

[2]Universidade Federal do Piauí (UFPI) - Departamento de Computação
Teresina - PI - Brasil

[3]Universidade Federal do Piauí (UFPI) - Departamento de Sistemas de Informação
Picos - PI - Brasil

[4]Universidade Federal do Piauí (UFPI) - Departamento de Medicina Especializada
Teresina - PI - Brasil

[5]Universidade Federal de Ouro Preto (UFOP) - Departamento de Computação
Ouro Preto - MG - Brasil

justinoduarte@gmail.com, {rveras,romuere}@ufpi.edu.br

nayzealdeman@gmail.com, kelson@ufpi.edu.br, andrea@ufop.edu.br

Abstract. *Chronic renal diseases arise from acute or intermittent, not adequately treated pathologies such as minimal change disease (MCD) and focal segmental glomerulosclerosis (FSGS). Correct identification of these two diseases is of paramount importance because their treatments and prognoses are different. Thus, we propose a method capable of differentiating MCD and FSGS through images of pathological exams. In the proposed method, we extracted 10240 features from three pre-trained convolutional neural networks, we selected 62 from them through the mutual information algorithm, and we used the Random Forest for the classification. The method obtained an accuracy of 93.33% and Kappa of 85.47%, which is considered "Almost Perfect".*

Resumo. *Doenças renais crônicas surgem a partir de patologias agudas ou intermitentes não tratadas adequadamente como a doença de lesão mínima (DLM) e a glomeruloesclerose segmentar e focal (GESF). Identificar corretamente essas duas doenças é de suma importância, pois seus tratamentos e prognósticos são diferentes. Dessa forma, propomos um método capaz de diferenciar DLM e GESF através de imagens de exames patológicos. No método proposto, foram extraídas 10240 características de três redes neurais convolucionais pré-treinadas, foram selecionadas 62 delas através do algoritmo de informação mútua e o Random Forest foi utilizado para a classificação. O método obteve acurácia de 93,33% e Kappa de 85,47%, o que é considerado "Quase Perfeito".*

1. Introdução

No Brasil, as glomerulopatias encontram-se entre as principais causas de doença renal terminal, afetando 11% dos pacientes em diálise [Costa et al. 2017]. As glomerulopatias são doenças renais com diferentes subtipos histopatológicos. Além de crucial para o diagnóstico, a avaliação microscópica pode oferecer dados prognósticos e proporcionar orientação para o tratamento.

A síndrome nefrótica é uma das apresentações principais das doenças glomerulares e tal manifestação, quando persistente, associa-se com a progressão para doença renal crônica. Diversas anormalidades histológicas podem levar ao desenvolvimento de síndrome nefrótica, entre as quais se destacam como causa de síndrome nefrótica idiopática a doença de lesões mínimas (DLM) e a glomeruloesclerose segmentar e focal (GESF). Nas crianças, a DLM é a causa da síndrome nefrótica em 90% dos pacientes, enquanto nos adultos detectam-se em 30% dos casos doenças sistêmicas associadas (como diabetes, amiloidose ou lúpus eritematoso sistêmico, por exemplo), e os outros 70% são representados por doenças glomerulares primárias, como GESF, glomerulopatia membranosa ou mesmo DLM. Quando se considera apenas a população adulta, a GESF é a principal causa de síndrome nefrótica em vários países, inclusive no Brasil [Moura et al. 2015].

É importante a diferenciação entre essas glomerulopatias (DLM e GESF) pois do ponto de vista terapêutico, existem diferenças nos tratamentos, pelo menos no que se refere à duração da fase de ataque no caso da corticoterapia, assim como na taxa de resposta a tratamento e no prognóstico entre elas [Moura et al. 2015].

Dada a importância de se diferenciar ambas as doenças, este trabalho tem o objetivo de propor um método computacional de classificação capaz de distinguir uma imagem de biópsia renal portadora de GESF ou portadora de DLM. Para isso, foram utilizadas técnicas de *Transfer Learning* em Redes Neurais Convolucionais (CNN's) e classificadores supervisionados.

O presente trabalho está organizado da seguinte maneira: na Seção 2 são apresentados os trabalhos relacionados; na Seção 3, são descritos os materiais utilizados, as técnicas empregadas e o método proposto; as Seções 4 e 5 são destinadas, respectivamente, aos resultados obtidos e às discussões; por fim, a conclusão e trabalhos futuros são apresentados na Seção 6.

2. Trabalhos relacionados

Técnicas de *Transfer Learning* que utilizam imagens médicas como entrada são muito utilizadas em sistemas de diagnóstico auxiliado por computador (CAD), nesse contexto se encontram os trabalhos de [Vogado et al. 2018] que trabalharam com detecção de leucemia utilizando imagens de lâminas de sangue. Já [Lopez et al. 2017] lidaram com melanoma em imagens dermoscópicas. E ainda [Meng et al. 2017] que atuaram na classificação de fibrose hepática, baseada em imagens de ultrassonografia.

Outros tipos de abordagens que trabalham com a segmentação e/ou identificação de estruturas renais em imagens, em especial as relacionadas ao glomérulo, podem ser encontradas no trabalho de [Sarder et al. 2016], que se empregou na segmentação e quantificação de glomérulos, cápsula de Bowman e núcleos glomerulares. Outro exemplo pode ser visto no trabalho de [Zhao et al. 2016], eles propuseram um método automati-

zado de segmentação do glomérulo com base em micrografia de todo o tecido renal. Já [Ginley et al. 2017] utilizaram métodos não supervisionados para identificar as bordas do glomérulo, utilizando filtros de Gabor.

No trabalho de [Barros et al. 2017], também voltado para histologia renal, os autores propuseram um sistema computacional para detectar lesões glomerulares proliferativas (PGL), diferenciando-as de imagens saudáveis. As técnicas utilizadas abrangeram o processamento clássico de imagens, segmentação e métodos de reconhecimento de padrões. Por fim o KNN foi usado como classificador. A acurácia atingida foi de 88,3 $\pm$3,6%.

Durante esta pesquisa, não foram encontrados trabalhos que propuseram métodos computacionais para diferenciar FSGS e MCD. Dessa forma, essa é a principal contribuição científica deste artigo.

3. Materiais e métodos

Nesta seção dá-se ênfase ao método desenvolvido para diferenciar imagens de biópsia renal que tenham GESF ou DLM. Nas subseções a seguir serão descritas as técnicas empregadas, as métricas utilizadas para avaliar a classificação realizada, bem como a base de dados com as imagens utilizadas.

3.1. Método proposto

O método proposto é composto das seguintes etapas: aquisição da imagem e pré-processamento, extração de características, seleção das características e classificação. A Figura 1 ilustra o processo.

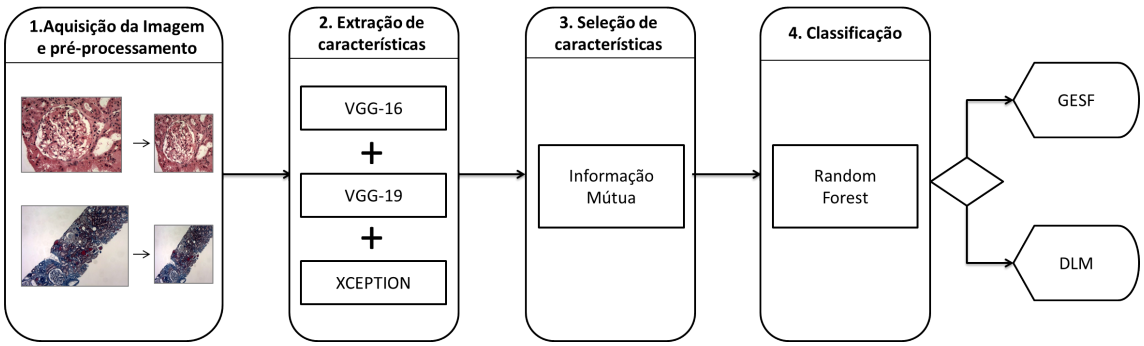

Figura 1. Etapas do método proposto. A imagem é redimensionada e passa por três CNN's, os vetores de características delas extraídos são concatenados, e têm seus atributos selecionados antes de seguir para o *Random Forest*, que classifica como GESF ou DLM

No pré-processamento, foram realizadas operações para adequar as dimensões das imagens aos tamanhos predefinidos de entrada das CNN's utilizadas. As imagens foram redimensionadas sem corte de bordas, ainda que com este procedimento se alterasse a proporção (largura/altura) original da imagem.

3.1.1. Extração de Características

As CNN's são comumente aplicadas quando se trata de aprendizagem de máquina para imagens. Uma grande vantagem de seu uso é a capacidade de detectar automaticamente as características importantes. Sua arquitetura profunda permite extrair um conjunto de características em múltiplos níveis de abstração [Tajbakhsh et al. 2016].

A arquitetura de uma CNN tipicamente inclui duas seções a primeira é uma sequência de operações de convolução seguida de operações de *pooling* e a segunda é composta por algumas camadas totalmente conectadas. A Figura 2 ilustra a arquitetura genérica de uma CNN.

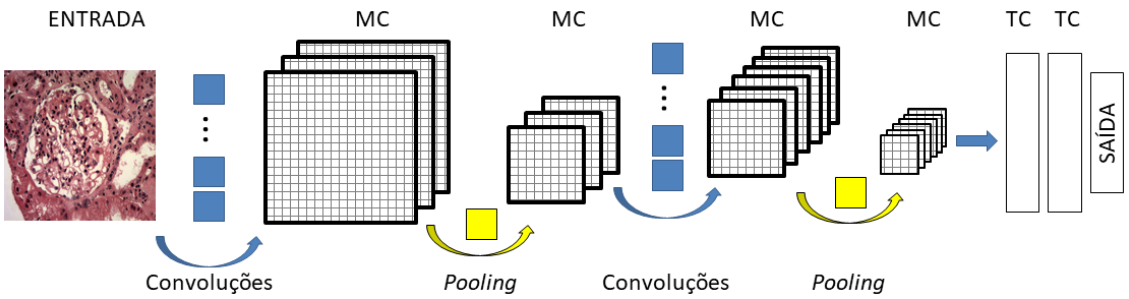

Figura 2. Arquitetura genérica de uma CNN. Os quadrados azuis e amarelos representam os filtros de convolução e *pooling* aplicados sobre a imagem. MC = mapas de características, TC = totalmente conectada.

Uma operação de convolução sobre uma imagem consiste basicamente em tomar a vizinhança um pixel, multiplicá-la por uma matriz de pesos (filtro) e computar o somatório desses produtos, repetindo esse procedimento para cada pixel, gera-se um mapa de características com as mesmas dimensões da imagem de entrada. Em cada camada de convolução de uma CNN podem ser aplicados diversos filtros, cada um gerando um mapa diferente.

A operação de *pooling* segue um mecanismo parecido com o da convolução, no entanto, ao tomar a vizinhança de um pixel, não se aplicam filtros com pesos, em vez disso tipicamente um filtro *max-pooling* é utilizado, dando como saída o maior valor dentre os vizinhos considerados. Na operação de *pooling*, geralmente, não são visitados todos os pixels da imagem e isso reduz a dimensionalidade dos mapas de características gerados, diminuindo o número de parâmetros da CNN, e consequentemente o tempo computacional para treino.

Após a sequência de convoluções e *pooling* da rede, os dados dos mapas são tomados como entrada na parte totalmente conectada, cuja arquitetura e mecanismo de funcionamento são semelhantes às de uma rede neural tradicional, sendo que na última camada já se tem os dados de saída.

Algoritmos de aprendizagem de máquina utilizam modelos estatísticos para fazer previsões. Esses modelos são construídos (na etapa de treino) com base em exemplos cuja variável a ser predita é conhecida previamente. É importante ressaltar que tradicionalmente o algoritmo é planejado para realizar sua tarefa para domínios similares nos conjuntos de treino e teste. Uma mudança de domínio necessitaria de uma reconstrução

do modelo (novo treinamento).

Transfer Learning, em contraste, permite que os domínios, tarefas e distribuições usados no treinamento e teste sejam diferentes [Pan et al. 2010]. A ideia é reutilizar o conhecimento aprendido em um campo e aplicá-lo em outro parecido, dessa forma se aproveita o processamento já realizado na etapa de treinamento de um modelo criado para uma finalidade correlata.

A aplicação de *Transfer Learning* em CNN, comumente se dá utilizando as informações geradas na saída de alguma camada da rede, por exemplo do modo feito em [Vogado et al. 2018], como forma de representar o conhecimento extraído naquele nível. Essas informações podem ser utilizadas como dados de entrada em outros métodos de aprendizagem de máquina.

A etapa de extração de características do método proposto, tem por função receber a imagem digital, e dar como saída um vetor de atributos correspondentes. Para isso, extraiu-se o vetor de saída da penúltima camada totalmente conectada (anterior à camada de classificação) de três CNN's: VGG-16, VGG-19 [Simonyan and Zisserman 2014] e XCEPTION [Chollet 2017], todas pré treinadas na base de imagens ImageNet [Russakovsky et al. 2015]. Além de dispor desses vetores separadamente, foi produzido um vetor adicional obtido pela concatenação dos três vetores anteriormente citados, este será referido como CONCAT.

3.1.2. Seleção de Características e Classificação

Cada um dos quatro vetores obtidos na etapa anterior foi submetido a uma operação de seleção de características. Esse procedimento foi realizado por meio de uma ordenação dos atributos, por ordem de importância. Foram testados dois critérios para cálculo da relevância e consequentemente da ordenação dos atributos: a estatística F da Análise de Variância (ANOVA-F) e Informação Mútua (IM).

A ANOVA-F se baseia em medidas de dispersão entre os elementos pertencentes a um grupo e na dispersão entre as médias de cada grupo. Tomando um atributo X do vetor extraído, e agrupando-os pela classe (GESF ou DLM), quanto mais distante é o X médio desses grupos e menos dispersos são os valores de X dentro dos grupos, maior é o valor da ANOVA-F e da relevância desse atributo dentro do vetor.

A informação mútua entre duas variáveis (X, Y) pode ser vista como a quantidade de informação que uma carrega da outra; ela mede quanto o conhecimento de uma destas variáveis reduz a incerteza sobre a outra. Desse modo, tomando dois atributos do vetor X1 e X2, comparando a IM de cada um deles com a variável Y (classe GESF ou DLM), será considerado mais relevante aquele que tiver o maior valor de IM.

Os vetores ordenados foram truncados iterativamente desde o tamanho 1 até o seu tamanho total e, a cada iteração, foram submetidos à classificação. Avaliamos três classificadores supervisionados: SVM [Cortes and Vapnik 1995], KNN [Aha et al. 1991] e *Random Forest* (RF) [Ho 1995].

Havendo portanto 4 vetores relativos às imagens originais oriundos das três CNN's e da concatenação, 2 seletores de atributos, e 3 classificadores, obtém-se 24 combinações

diferentes para avaliar e definir qual delas é o melhor modelo além de buscar qual o melhor número de atributos a ser tomado em cada cenário.

Após a avaliação desses cenários, tendo fixado o melhor modelo e definidos quais atributos seriam tomados para classificação, definiu-se um vasto conjunto de hiperparâmetros de inicialização do classificador selecionado e realizou-se uma busca pelos que promovessem uma melhor performance na tarefa.

3.2. Métricas de avaliação

A validação cruzada (*k-fold cross-validation*) consiste em distribuir aleatoriamente as instâncias da base de dados em *k* subconjuntos mutuamente exclusivos (*folds*) de tamanhos aproximadamente iguais. O classificador é treinado e testado *k* vezes, em cada rodada, um subconjunto diferente é tomado para teste e os *k*-1 subconjuntos restantes, para treino. Esse mecanismo garante que qualquer elemento da base de dados, em algum momento tenha servido para avaliar o classificador e em outro momento para treiná-lo.

A forma estratificada da validação cruzada segue o mesmo princípio, mas a divisão dos folds leva em consideração a classe à qual pertencem os elementos, de forma a preservar em cada *fold*, aproximadamente a mesma proporção observada na base original. Tal formato permite que, na etapa de testes, seja avaliada a capacidade de classificação para ambas as classes em todos os *folds*.

Para avaliar os classificadores selecionados, as imagens de entrada foram agrupadas em conjuntos de treino e de teste utilizando a técnica de validação cruzada (*k-fold*) estratificada com 5 grupos (*k=5*). Uma matriz de confusão foi computada para cada *fold* e a partir dela foram calculados as métricas de acurácia e Kappa. A média aritmética dos 5 valores obtidos foi considerada como forma de avaliação de cada classificador.

A matriz de confusão pode ser vista como uma relação que confronta os resultados preditos por um classificador e os resultados reais para um mesmo conjunto de testes. No caso em que temos as classes GESF e DLM, trata-se pois de um classificador binário, havendo portanto 4 valores nesta matriz: Verdadeiro Positivo (*VP*) relativo a quantidade de imagens corretamente classificadas como portadoras de DLM; Verdadeiro Negativo (*VN*) relativo ao número de classificações corretas de GESF; Falso Positivo (*FP*) relativo ao número de imagens classificadas como DLM, mas que na verdade são GESF e por fim, o Falso Negativo (*FN*) referente à quantidade de imagens classificadas erroneamente como GESF. A Tabela 1 ilustra essa configuração.

Tabela 1. Matriz de confusão.

Classificação do	Classificação computada	
Patologista	GESF	DLM
GESF	*VN*	*FP*
DLM	*FN*	*VP*

A acurácia mede a taxa de acerto geral para ambas as classes de maneira unificada, pode ser obtida pela razão entre o número de imagens corretamente classificadas e o total de imagens conforme a Equação 1:

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN}. \quad (1)$$

Outra métrica computada foi o Kappa [Cohen 1960], um valor capaz de medir o grau de concordância entre classificações nominais realizadas por dois avaliadores; no presente caso, as preditas pelo classificador e as anotadas pelo patologista. O cálculo do Kappa também pode ser obtido através dos valores da matriz de confusão descrita na Tabela 1 aplicando o disposto na Equação 2:

$$
\begin{aligned}
Kappa &= \frac{Acurácia - P_e}{1 - P_e}, \\
&\text{onde:} \\
P_e &= P_{DLM} + P_{GESF} \\
P_{DLM} &= \frac{VP + FN}{N} * \frac{VP + FP}{N} \\
P_{GESF} &= \frac{VN + FN}{N} * \frac{VN + FP}{N} \\
N &= VP + VN + FP + FN.
\end{aligned}
\tag{2}
$$

O valor máximo do Kappa é 100%, o que indica a concordância perfeita entre os avaliadores. Costuma-se utilizar os rótulos dispostos na Tabela 2, proposta por [Landis and Koch 1977] como forma de manter uma nomenclatura consistente ao descrever o grau de concordância relativo à faixa em que se enquadra o Kappa.

Tabela 2. Rótulos atribuídos às faixas correspondentes de Kappa.

Kappa (%)	Grau de concordância
<0	Sem acordo
0 ⊢ 20	Insignificante
20 ⊢ 40	Mediano
40 ⊢ 60	Moderado
60 ⊢ 80	Substancial
80 ⊢ 100	Quase perfeito

Com o objetivo de descrever melhor alguns cenários, ainda foram calculadas três métricas adicionais: a precisão que pode ser intuída como a habilidade do classificador não rotular como positivo um exemplo que é negativo, o *recall* podendo ser interpretado como a capacidade do classificador identificar todas as instâncias positivas e a métrica F1 que sintetiza um valor intermediário baseado nas duas anteriores. Essas métricas também são obtidas com base na matriz de confusão conforme o disposto na Equação 3.

$$
\begin{aligned}
Precisão &= \frac{VP}{VP + FP} \\
Recall &= \frac{VP}{VP + FN} \\
F1 &= \frac{2 * Precisão * Recall}{Precisão + Recall}
\end{aligned}
\tag{3}
$$

3.3. Base de dados

A base de dados utilizada neste trabalho é composta de 87 imagens coloridas, das quais 33 foram classificadas por um especialista como portadoras de GESF e 54 como DLM. Não constam imagens de rim saudável, pois não são realizadas biópsias renais em pacientes sem alteração de função renal.

As imagens foram obtidas com microscópios Nikon e220 e Nikon e200 adaptado com imunofluorescência, capturadas com diferentes lentes objetivas. A pigmentação aplicada sobre as lâminas foi realizada utilizando os corantes: hematoxilina-eosina, tricrômico de Masson, PAS e Prata metenamina. Tais informações não fazem parte de metadados associados a cada arquivo separadamente, não podendo portanto serem tomadas como dado adicional para classificação. As resoluções e taxa de proporção das imagens também são heterogêneas porém todas são maiores que o tamanho de entrada das CNN's utilizadas.

A Figura 3 mostra exemplos de imagens da base de dados, as dispostas na parte superior pertencem ao grupo classificado pelo especialista como portadoras de GESF e as três inferiores, ao grupo DLM.

Observa-se, em muitos casos, heterogeneidade visual entre imagens pertencentes à mesma classe bem como similitude entre imagens de classes distintas, o que certamente dificulta a tarefa de classificação.

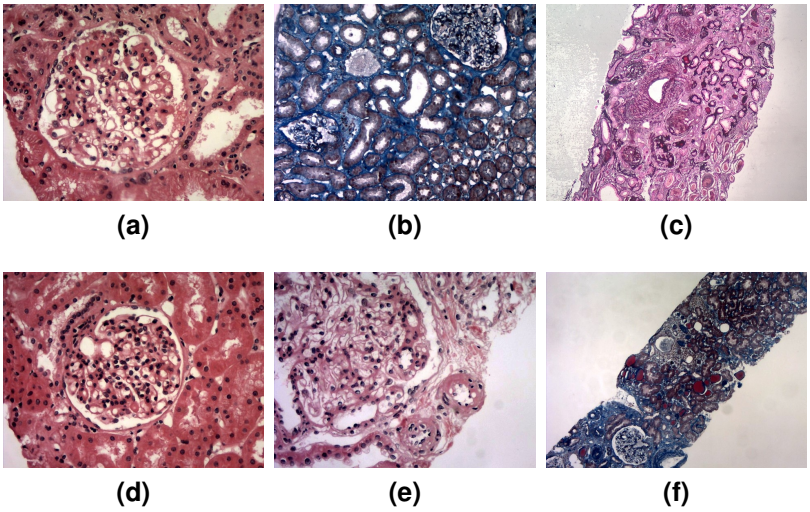

Figura 3. Exemplos de imagens da base de dados. a-c: GESF; d-f: DLM.

4. Resultados

Todos os cenários foram avaliados com base na média aritmética das métricas obtidas sobre os 5 conjuntos de testes do *k-fold*. Para cada um dos cenários analisou-se iterativamente a quantidade de atributos e a performance da classificação, como exemplo, a Figura 4 exibe a acurácia e o Kappa obtidos em função do número de atributos tomados para classificação com *Random Forest*, utilizando os atributos do vetor CONCAT, selecionados com base na Informação Mútua.

De posse desses conjuntos de valores, verificou-se quais eram os melhores tanto no que se refere ao Kappa quanto à acurácia, em cada um dos 24 cenários, para efeito

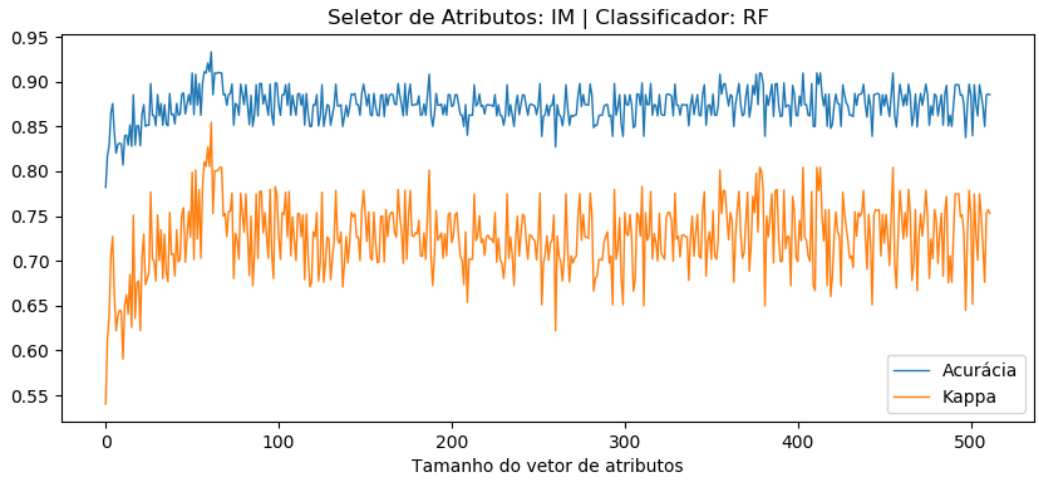

Figura 4. Iteração sobre as primeiras 512 características do vetor CONCAT, selecionadas pela IM e métricas apuradas para o classificador RF.

de comparação. A Figura 5 traz um comparativo geral de ambas as métricas, para todos os cenários testados, apresentando os maiores valores atingidos considerando todos os tamanhos de vetores tomados.

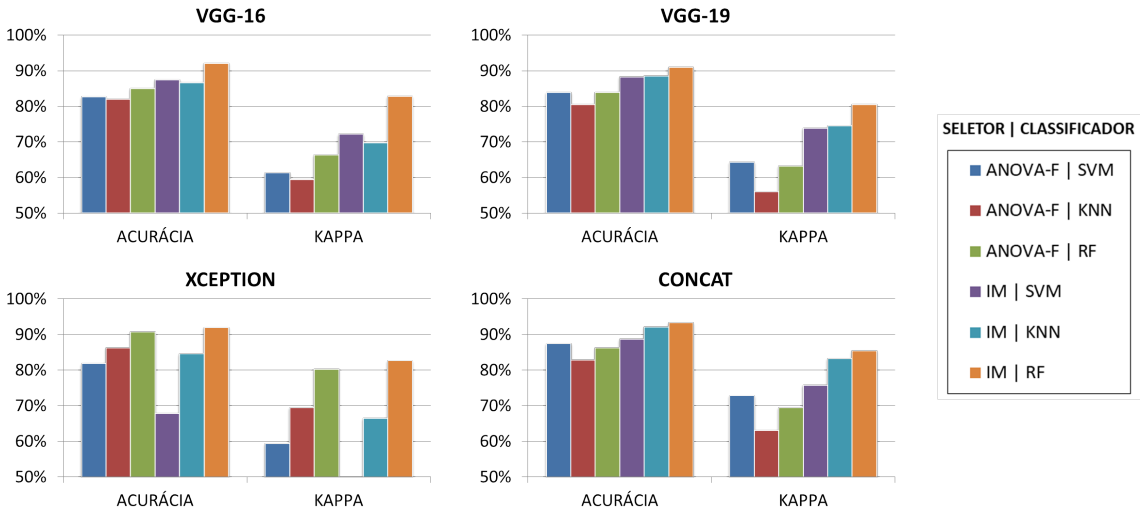

Figura 5. Melhores resultados para Kappa e Acurácia.

Observando a Figura 5, nota-se que a seleção de atributos via IM combinada ao classificador RF estão sempre melhores em todos os cenários. Utilizou-se então esse classificador e seletor, tomou-se o mesmo número de atributos que promoveram os melhores valores em cada CNN e detalhamos a performance com outras métricas. A Tabela 3 exibe valores mais detalhados, incluindo o número de atributos e métricas adicionais, para todas as CNN's consideradas.

A composição dos primeiros 62 atributos do vetor CONCAT, considerando a ordenação por IM, foi analisada para aferir a contribuição de cada uma das CNN's nos atributos selecionados. A Figura 6 mostra a porcentagem de atributos escolhidos em relação à CNN de origem, bem como sua distribuição ao longo das posições do vetor.

Tabela 3. Métricas obtidas com *Random Forest* e informação mútua, tomando o número de atributos onde o Kappa foi melhor. (Valores das métricas em termos percentuais).

VETORES	Acurácia	Kappa	Precisão	*Recall*	F1	Número de atributos
VGG-16	92,16	82,79	90,26	**98,18**	93,99	235
VGG-19	90,97	80,56	89,80	96,36	92,88	118
XCEPTION	92,02	82,63	91,77	96,36	93,82	**38**
CONCAT	**93,33**	**85,47**	**91,92**	**98,18**	**94,86**	62

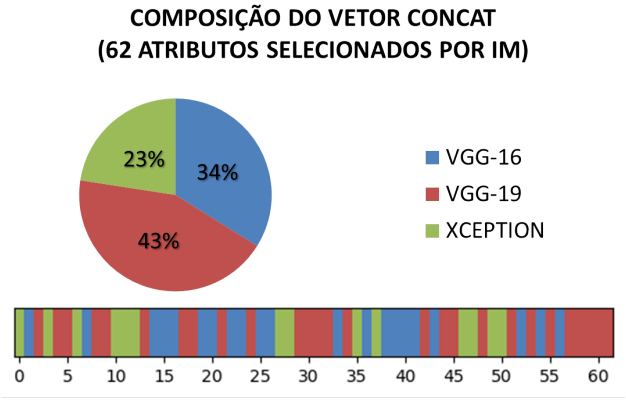

Figura 6. Origem dos 62 atributos com melhor IM do vetor CONCAT. A barra na parte inferior indica a origem dos atributos em cada posição.

5. Discussão dos resultados

Analisando a Figura 5, nota-se que utilizando a seleção de atributos via informação mútua combinada ao classificador *Random Forest*, foram alcançados os maiores valores de Kappa e acurácia entre todos os 24 cenários. No entanto também se destacam a rede XCEPTION com ANOVA-F e RF bem como o vetor CONCAT com IM e KNN, que também atingiram altos valores de Kappa.

Alguns valores próximos do Kappa e acurácia podem ser observados na Tabela 3, nas CNN's VGG-16 e XCEPTION, cuja diferença percentual ficou na ordem de poucos décimos, no entanto a segunda necessitou de menos de 17% dos atributos tomados pela primeira para prover resultados praticamente iguais.

Já o vetor CONCAT, no que se refere ao número de atributos, ficou com a segunda colocação. Porém este quesito é um critério de menor importância, uma vez que com este vetor obteve-se o maior valor em todas as métricas computadas. Tendo atingido Kappa de 85,47%, o que indica um grau de concordância quase perfeito com a classificação do patologista (Tabela 2).

6. Conclusão e trabalhos futuros

As glomerulopatias podem levar à doença renal terminal. A avaliação microscópica é crucial para o diagnóstico e pode oferecer dados prognósticos e proporcionar orientação para o tratamento. Diversas anormalidades histológicas podem levar ao desenvolvimento

de síndrome nefrótica, entre as quais se destacam como causa de síndrome nefrótica idiopática a doença de lesões mínimas (DLM) e a glomeruloesclerose segmentar e focal (GESF).

Diferenciar DLM de GESF é importante, pois existem diferenças nos esquemas medicamentosos por ocasião do diagnóstico. Diante disso, este trabalho propôs um método de classificação por imagem para diferenciar essas duas doenças. Os testes para chegar ao método final incluíram extração de características por meio de três Redes Neurais Convolucionais e sua concatenação. Em seguida essas características foram selecionadas por meio de dois critérios: informação mútua e ANOVA-F. Depois foram levados para três classificadores supervisionados: SVM, KNN e RF.

Os resultados obtidos indicaram que a utilização do vetor de atributos concatenado, seguido da seleção de atributos com base em informação mútua e classificação via *Random Forest* proveram os melhores índices de avaliação. Este modelo atingiu acurácia de 93,33% e Kappa de 85,47%, o que indica uma concordância quase perfeita com a classificação realizada pelo patologista.

Como trabalhos futuros, pretende-se avaliar outros extratores de características e métodos de seleção de atributos. Sobre a quantidade de imagens, pretende-se investigar padrões de aumento de dados indicados para o tipo de aplicação deste trabalho.

Referências

Aha, D. W., Kibler, D., and Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning*, 6(1):37–66.

Barros, G. O., Navarro, B., Duarte, A., and Dos-Santos, W. L. (2017). Pathospotter-k: A computational tool for the automatic identification of glomerular lesions in histological images of kidneys. *Scientific reports*, 7:46769.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. *arXiv preprint*, pages 1610–02357.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Costa, D. M. d. N., Valente, L. M., Gouveia, P. A. d. C., Sarinho, F. W., Fernandes, G. V., Cavalcante, M. A. G. d. M., Oliveira, C. B. L. d., Vasconcelos, C. d. A. J. d., and Sarinho, E. S. C. (2017). Comparative analysis of primary and secondary glomerulopathies in the northeast of brazil: data from the pernambuco registry of glomerulopathies-repeg. *Brazilian Journal of Nephrology*, 39(1):29–35.

Ginley, B., Tomaszewski, J. E., and Sarder, P. (2017). Automatic computational labeling of glomerular textural boundaries. In *Medical Imaging 2017: Digital Pathology*, volume 10140, page 101400G. International Society for Optics and Photonics.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282 vol.1.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Lopez, A. R., Giro-i Nieto, X., Burdick, J., and Marques, O. (2017). Skin lesion classification from dermoscopic images using deep learning techniques. In *Biomedical Engineering (BioMed), 2017 13th IASTED International Conference on*, pages 49–54. IEEE.

Meng, D., Zhang, L., Cao, G., Cao, W., Zhang, G., and Hu, B. (2017). Liver fibrosis classification based on transfer learning and fcnet for ultrasound images. *Ieee Access*, 5:5804–5810.

Moura, L. R., Franco, M. F., and Kirsztajn, G. M. (2015). Doença de lesões mínimas e glomeruloesclerose segmentar e focal em adultos: resposta a corticoide e risco de insuficiência renal. *Jornal Brasileiro de Nefrologia*, 37(4):475–480.

Pan, S. J., Yang, Q., et al. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

Sarder, P., Ginley, B., and Tomaszewski, J. E. (2016). Automated renal histopathology: digital extraction and quantification of renal pathology. In *Medical Imaging 2016: Digital Pathology*, volume 9791, page 97910F. International Society for Optics and Photonics.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., and Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312.

Vogado, L. H., Veras, R. M., Araujo, F. H., Silva, R. R., and Aires, K. R. (2018). Leukemia diagnosis in blood slides using transfer learning in cnns and svm for classification. *Engineering Applications of Artificial Intelligence*, 72:415–422.

Zhao, Y., Black, E. F., Marini, L., McHenry, K., Kenyon, N., Patil, R., Balla, A., and Bartholomew, A. (2016). Automatic glomerulus extraction in whole slide images towards computer aided diagnosis. In *e-Science (e-Science), 2016 IEEE 12th International Conference on*, pages 165–174. IEEE.