Análise ao *dataset – Skin MNIST*

Depois de concluída a análise do *dataset* anterior, *Breast Histopathology*, seguiu-se a análise de outro problema, relacionando o âmbito do trabalho.

O *Skin Cancer MNIST: HAM10000*, é representado por um elevado conjunto de imagens dermatológicas, espelhando diferentes lesões associadas ao cancro da pele[1].

Estas imagens foram obtidas recorrendo a um processo denominado por dermatoscopia. Os dermatologistas recorrem ao dermatoscópio, de modo a conseguir avaliar, com maior rigor e segurança, as lesões do seu paciente. Permitindo-lhe assim identificar a gravidade da lesão sofrida.

Os dermatoscópios permitem aos profissionais de saúde, a visualização de lesões na pele, em escalas muito amplificados, possibilitando a análise de uma lesão, com um aumento de/até 400 vezes, o seu tamanho.

Para além disso, estes aparelhos permitem a digitalização das lesões, para um computador. Ajudando assim, a estabelecer um mecanismo de análise mais segura e gradual (verificação do avanço ou recuo das lesões).

As imagens relativas a este estudo, foram obtidas ao longo de um extenso período de tempo, nomeadamente 20 anos. Contando com o apoio de: o *Department of Dermatology at the Medical University of Vienna*, e ainda com o apoio do Médico/Professor *Cliff Rosendahl*.

O *Skin Cancer MNIST: HAM10000*, apesar de ser um *dataset* recente: disponibilizado em 2018, é já considerado o *standard,* no que toca à análise de lesões associadas ao Cancro da Pele.

São diversas as razões, que ajudam a destacar a sua importância, sendo essencialmente: (1) elevado nº de imagens disponíveis: 10015, (2) elevado nº de classes consideradas: 7 e (3) percentagens de amostras para as classes menos balanceadas, ajustada em conformidade com as possibilidades[1].

São 7 os tipos de lesões identificadas, ao longo do *dataset*. Sendo estas: *akiec*, *bcc*, *bkl, df*, *nv*, *mel* e *vasc*. Estas "*labels"* são apenas identificadores do tipo de lesão associada, porque cada classe pode conter várias patologias associadas à lesão associada[1]. Seguidamente, num contexto resumido, é contextualizada cada uma das classes descritas atrás.

A classe ***Akiec,*** agrega dois tipos de cancro não invasivo: *Actinic Keratoses* e ainda Intraepithelial Carcinoma*. Actinic Keratoses* também conhecidos por *Solar Keratoses*, descrevem lesões causadas pela excessiva exposição solar, sendo que a probabilidade de evolução para cancro invasivo é baixa. *Intraepithelial Carcinoma*, também conhecido por *Bowen's disease*, representa uma lesão semelhante à *Actinic Keratoses*, com a exceção da causa associada, isto é, a lesão é causada por um vírus e não pela exposição solar. Ambas as patologias foram inseridas pelos autores na mesma classe, devido ao facto de ambas serem consideradas e citando [1]: "variantes de squamous cell carcinoma".

BCC é a abreviação de *Basal cell carcinoma,* e representa o tipo mais comum de cancro de pele. As principais causas associadas a esta lesão, são sobretudo: a exposição solar excessiva, pessoas com pele mais branca tem maior risco ou exposições a radioterapias. Normalmente, e quando detetada a tempo, não resulta em morte, ainda assim é necessário ter em consideração, que o tumor cresce e pode-se tornar invasivo.

A classe ***BKL***, *Benign Keratosis* é segundo os autores uma classe genérica, sendo a mesma composta por três sub-grupos, sendo estes: *seborrheic keratoses*, *solar lentigo* e ainda *lichen-planus like keratoses*. Estes três sub-grupos formaram uma única classe, porque e citando os autores: "apresentam semelhanças biológicas e são normalmente avaliadas considerando o mesmo termo histopatológico". Uma curiosidade destas lesões, é que o seu aspeto varia de acordo com o local da lesão[1].

DF, *Dermatofibroma*, representam lesões benignas que podem aparecer na pele, sendo desconhecidas as causas associadas ao seu aparecimento. Na maioria das situações, não existe a necessidade de recorrer a qualquer tipo de cirugia, quer para o seu tratamento, quer para a sua eliminação.

NV, *Melanocytic nevi*, representam neoplasmos benignos, e que podem resultar de uma forma adquirida, como por exemplo: exposição à luz solar, ou ainda de uma forma congénita, isto é, genética. Estas lesões aparecem normalmente, nos primeiros 20 anos de vida dos pacientes. Na maioria dos casos, a probabilidade de evoluir para Melanoma, é reduzida.

Mel, *Melanoma*, representam uma lesão na pele maligna, e que se desenvolve a partir de *melanocytes*. Existem vários tipos de melanomas, como: Superficial, Nodular, Lentigo maligna, ou Acral lentiginoso. Existem várias as causas associadas a esta lesão, mas a que revela uma maior incidência é a excessiva exposição solar. Os índices de cura são muito elevados, quando não existem metástases associadas.

Finalmente, ***Vasc,*** *Vascular skin lesions*, representam lesões que originam problemas nos vasos linfáticos, podendo incluir tumores benignos ou malignos ou malformações. Algumas das variantes consideradas nestes *dataset*: *cherry angiomas*, *angiokeratomas, pyogenic granulomas* e *Hemorrhage*. Algumas destas lesões são facilmente detetadas, contudo outras necessitam de um maior cuidado, na sua análise.

Seguidamente, irá ser demonstrada a análise exploratória, que fora realizada ao *dataset*, com o objetivo de ajudar o autor, a perceber as "peças chave" do problema. Para tal, foi criado um *Jupyter notebook*, para auxiliar na análise e gráficos, que surgiram subsequentemente da análise.

Análise Exploratória

Antes de avançar para a análise dos dados, é necessário explicar um conjunto de passos iniciais, que foram estabelecidos, e que são vitais, durante todo o processo de análise dos dados. Sendo estes: Armazenamento das imagens e Leitura/Adaptação do *Dataframe,* que agrega a informação relativa a cada uma das amostras do problema.

Sendo assim, primeiramente foram transferidos todos os ficheiros associados ao problema em questão, os mesmos encontram-se disponibilizados, na comunidade *Kaggle* [2].

As imagens encontram-se divididas, ao longo de duas pastas. As imagens estão em formato *.jpg* e apresentam um tamanho de 600 pixéis por 450 pixéis, respetivamente largura e altura.

Para além das imagens, foram disponibilizados ainda cinco ficheiros *.csv*. O primeiro ficheiro *HAM10000_metadata*, contêm um conjunto de informações relevantes sobre cada uma das amostras do problema, tais como: idade do paciente, local da lesão, sexo do paciente ou classe associada. Um maior destaque a este documento, irá ser demonstrada através da análise exploratória realizada. Já os restantes ficheiros, descrevem para cada amostra os valores

referentes aos pixéis em cada posição da imagem, considerando um tamanho pré-estabelecido, sendo considerado um tamanho de 8*8 e 28*8. A representação dos pixéis, foi demonstrada num formato *RGB* e ainda *greyscale*.

Os ficheiros e imagens foram devidamente incluídos na raíz do projeto, de modo a facilitar, o processo de leitura dos dados. De salientar ainda, que ao longo do projeto e também da análise exploratória, não foi considerada a utilização de *path's* relativos, de modo a aumentar a flexibilidade do código. Esta preocupação revela-se essencial, porque o código necessita de ser trabalho em diferentes máquinas e diferentes sistemas operativos, evitando assim a necessidade de "alterar configurações de caminhos".

A Figura 1 ilustra a estrutura que agrega os vários ficheiros alusivos ao *dataset* em estudo, e que foram descritos atrás. A Figura 2 enumera várias "variáveis de configuração" estabelecidas no roteamento e definição dos caminhos inerentes aos dados e ficheiros, a analisar.

```
ROOT_DIR = os.getcwd()
print(ROOT_DIR)
INPUT_DIR = os.path.join(ROOT_DIR, config.INPUT_FOLDER)
print(INPUT_DIR)
PATIENTS_INFO = os.path.join(INPUT_DIR, config.INFO_PATIENTS)
print(PATIENTS_INFO)
```

```
C:\Users\gusta\Desktop\Projetos_Python\skin_mnist
C:\Users\gusta\Desktop\Projetos_Python\skin_mnist\input
C:\Users\gusta\Desktop\Projetos_Python\skin_mnist\input\HAM10000_metadata.csv
```

Figura 1 - Estrutura que agrega os vários ficheiros do dataset

```
INPUT_FOLDER = 'input'
IMAGES_ACESS = 'images/*.jpg'
INFO_PATIENTS = 'HAM10000_metadata.csv'
PIXEL_28_RGB_CSV = 'HMNIST_28_28_RGB.csv'
```

Figura 2 - Variáveis de configuração (route)

A raíz do projeto é designada pela pasta *skin_mnist*, observável através da Figura 1, sendo este o ponto "mais distante", da estrutura em árvore do projeto. No interior do projeto, para além dos *scripts* de código, e de outros diretórios, existe uma pasta *input*. É esta pasta apelidada de *Input*, que contém as imagens e os ficheiros *.csv* descritos anteriormente. Como, podemos visualizar através da Figura 1, a o ficheiro *HAM10000_metadata.csv*, está no interior da pasta *Input*.

A Figura 2, descreve os nomes dos ficheiros e diretórias, associadas ao problema em análise. Esta abordagem, permite a eventual alteração da estrutura e/ou caminhos destas pastas e ficheiros com facilidade.

Para terminar, esta análise basta apenas identificar o local onde as imagens são armazenadas. A Figura 3, ilustra o caminho único onde são armazenadas todas as imagens alusivas ao *dataset*.

```
IMAGES_REGEX = os.path.join(INPUT_DIR, config.IMAGES_ACESS)
images_paths = config_func.getImages(IMAGES_REGEX)
images_paths[0]
```

```
'C:\\Users\\gusta\\Desktop\\Projetos_Python\\skin_mnist\\input\\images\\ISIC_0024306.jpg'
```

Figura 3 - Exemplo de um caminho, associado às várias imagens do dataset

Tal como tinha sido referido anteriormente, a pasta *Input* armazena os ficheiros *.csv*, mas também as amostras do problema. Recorrendo, à visualização da Figura 3 é possível verificar que a pasta *Input*, agrega no seu interior uma outra pasta: *images*. Esta pasta, é responsável por armazenar unicamente, todas as imagens do problema. Ou seja, a utilização da pasta *images*, permite isolar as imagens, dos ficheiros *.csv*. Na Figura 3, é ainda possível visualizar o *output* referente a uma imagem: ISIC_0024306*.jpg*. A extensão da imagem, é *.jpg*, tal como já fora exposto anteriormente. Já o indicador ISIC_0024306, representa um identificador associado a cada uma das imagens (este aspecto, volta a ser referido, mais adiante).

Após esta breve contextualização sobre os ficheiros, e também à forma como foram armazenados, segue-se uma análise concreta dos dados.

O primeiro passo aplicado consistiu na leitura do ficheiro *HAM10000_metadata.csv*. Este ficheiro reúne informações específicas, sobre cada imagem presente no *dataset*.

A Figura 4 ilustra as 5 primeiras linhas deste ficheiro.

	lesion_id	image_id	dx	dx_type	age	sex	localization
0	HAM_0000118	ISIC_0027419	bkl	histo	80.0	male	scalp
1	HAM_0000118	ISIC_0025030	bkl	histo	80.0	male	scalp
2	HAM_0002730	ISIC_0026769	bkl	histo	80.0	male	scalp
3	HAM_0002730	ISIC_0025661	bkl	histo	80.0	male	scalp
4	HAM_0001466	ISIC_0031633	bkl	histo	75.0	male	ear

Figura 4 - Exemplo do conteúdo do ficheiro - HAM10000_metadata.csv

Este ficheiro foi convertido para um objeto *DataFrame* (biblioteca *Pandas*), de modo a facilitar a leitura, escrita, manipulação e análise dos dados.

O *DataFrame* reúne ao todo sete colunas, sendo estas: *lesion_id, image_id, dx, dx_type, age, sex* e *localization*. A coluna *lesion_id* simboliza o id da lesão associada à cada imagem. Ou seja, podem existir várias imagens associadas à mesma lesão, como podemos constatar, através das 2ª imagens da Figura 4. O campo *image_id* é um identificador único associado a cada imagem. A coluna *dx* representa o diagnóstico associado à lesão do paciente (se existirem várias imagens, as mesmas apresentam o mesmo valor neste campo), isto é, o *target* associado a cada uma das imagens. O campo *dx_type* reflete o tipo de diagnóstico, que foi considerado na avaliação da lesão, existindo 4 possibilidades: *Histopathology* (*histo*), *Reflectance confocal microscopy*, *Consensus of at least three expert dermatologists from a single image* (*consensos*), *Lesion did not change during digital dermatoscopic follow up over two years with at least three images* (*follow-up*). Os dois campos seguintes descrevem o sexo e a idade do paciente. Finalmente, a coluna *localization* indica o local do corpo, onde se localiza a lesão, por exemplo: *ear, scalp, foot, neck*, entre outros.

Mas, antes de proceder à análise dos dados, é necessário adicionar uma nova coluna ao *DataFrame,* isto é, é necessário identificar para cada uma das imagens (via *image_id*), qual a imagem que lhe está associada. Sendo que, a nova coluna irá conter o *path* associado, a essa imagem. Como já fora evidenciado atrás, cada imagem reúne um identificador único, *image_id*, sendo possível através desse identificador, verificar qual a imagem associada a cada *image_id*, visto que o nome da imagem, é descrito através desse mesmo identificador. As Figuras 3 e 4 ilustram a presença desse identificador único associado a cada imagem ISIC_*****.

Dessa forma, foi adicionada uma nova coluna no *DataFrame,* "apelidada" de *path*. O resultado final do *DataFrame* está representado através da Figura 5.

	lesion_id	image_id	dx	dx_type	age	sex	localization	path
0	HAM_0000118	ISIC_0027419	bkl	histo	80.0	male	scalp	C:\Users\gusta\Desktop\Projetos_Python\skin_mn...
1	HAM_0000118	ISIC_0025030	bkl	histo	80.0	male	scalp	C:\Users\gusta\Desktop\Projetos_Python\skin_mn...
2	HAM_0002730	ISIC_0026769	bkl	histo	80.0	male	scalp	C:\Users\gusta\Desktop\Projetos_Python\skin_mn...
3	HAM_0002730	ISIC_0025661	bkl	histo	80.0	male	scalp	C:\Users\gusta\Desktop\Projetos_Python\skin_mn...
4	HAM_0001466	ISIC_0031633	bkl	histo	75.0	male	ear	C:\Users\gusta\Desktop\Projetos_Python\skin_mn...

Figura 5 - DataFrame - após adição de nova coluna: path

Desta forma, o objeto *DataFrame,* já se encontra devidamente formulado, para que se possa dar seguimento à realização de outras tarefas, nomeadamente: análise exploratória (e que vai ser abordada de seguida), e ainda na preparação dos dados do problema (este objeto *DataFrame*, é um "ponto de partida", para a obtenção dos vários conjuntos de dados a utilizar na criação, treino e previsão dos modelos.

Inicialmente, foi efetuada uma breve inspeção ao *DataFrame*, como medida de precaução, e de modo a verificar se tudo estava em conformidade. Sendo assim, existiu a preocupação de: (1) verificar se o nº total de imagens coincidia com o valor indicado em [1], isto é, 10015 imagens no total, (2) confirmação dos atributos do ficheiro *metadata,* com o objeto *DataFrame*, (3) verificação das classes do problema e (4) confirmação de que o *path* de cada imagem está coerente com o seu *id*.

Relativamente ao nº total de imagens presentes no *DataFrame* foi possível concluir que o nº total de linhas no *DataFrame*, é igual ao nº total de imagens declaradas em [1]. Ou seja, o *DataFrame* é composto por 10015 imagens. A Figura 6 enumera o nº total de linhas presentes no objeto.

```
data.shape[0]
```

```
10015
```

Figura 6 - Nº de linhas total do DataFrame

Já, o nº total de colunas também se encontra em conformidade com o ficheiro *metadata*, isto é, o objeto *DataFrame* reúne as 7 colunas presentes no ficheiro: *lesion_id, image_id, dx, dx_type, age, sex, localization*, contando ainda com a nova coluna adicionada *path*. A Figura 7 demonstra as colunas do objeto.

```
data.columns
```

```
Index(['lesion_id', 'image_id', 'dx', 'dx_type', 'age', 'sex', 'localization',
       'path'],
      dtype='object')
```

Figura 7 - Colunas do DataFrame

As classes do problema, tal como expectável estão em conformidade, com o ficheiro *metadata*. Sendo ao todo 7 as classes do problema. A Figura 8 ilustra as "abreviações", já abordadas, e que referem as classes do problema.

```
classes = data.dx.unique()
classes
```

```
array(['bkl', 'nv', 'df', 'mel', 'vasc', 'bcc', 'akiec'], dtype=object)
```

Figura 8 - Labels relativas às classes do problema

Finalmente, foi necessário verificar se o conteúdo inerente à nova coluna adicionada: *path*, estava em concordância com o *image_id*. Isto é, é necessário garantir, que o caminho de cada imagem é o correto, referenciando assim o *image_id*, correto da imagem. A Figura 9 demonstra a relação entre ambas as colunas, *image_id* e *path*.

	image_id	path
0	ISIC_0027419	C:\Users\gusta\Desktop\Projetos_Python\skin_mnist\input\images\ISIC_0027419.jpg
1	ISIC_0025030	C:\Users\gusta\Desktop\Projetos_Python\skin_mnist\input\images\ISIC_0025030.jpg
2	ISIC_0026769	C:\Users\gusta\Desktop\Projetos_Python\skin_mnist\input\images\ISIC_0026769.jpg
3	ISIC_0025661	C:\Users\gusta\Desktop\Projetos_Python\skin_mnist\input\images\ISIC_0025661.jpg
4	ISIC_0031633	C:\Users\gusta\Desktop\Projetos_Python\skin_mnist\input\images\ISIC_0031633.jpg

Figura 9 - Relação entre as colunas Image_ID e Path

Como podemos observar através da Figura 9, o identificador presente no *path* das imagens (*ISIC_****.jpg*) é coerente com o seu *id*, garantindo assim a concordância, entre o *id* de uma imagem, e o seu caminho.

Posto isto, segue-se a análise concreta aos dados do problema. Esta análise irá ser apoiada com recurso à disponibilização de vários gráficos, que ajudam a entender melhor os dados, e a identificar padrões/particularidades nos dados.

Primeiramente, foi efetuada uma procura por valores em falta no *dataset*. A Figura 10 demonstra, um relatório genérico sobre o *dataset*.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10015 entries, 0 to 10014
Data columns (total 7 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   lesion_id     10015 non-null  object
 1   image_id      10015 non-null  object
 2   dx            10015 non-null  object
 3   dx_type       10015 non-null  object
 4   age           9958 non-null   float64
 5   sex           10015 non-null  object
 6   localization  10015 non-null  object
dtypes: float64(1), object(6)
memory usage: 547.8+ KB
```

Figura 10 - Análise de algumas propriedades do dataset

Observando a Figura 10, é possível concluir que todas as colunas, à exceção da idade, não apresentam valores em falta (o que não quer dizer, que não hajam incoerências nos dados, como vamos verificar adiante). Relativamente, ao atributo *age*, é possível constatar a presença de alguns valores em falta, mais concretamente 57 imagens, não têm associada uma idade.

Para o problema em causa, não existe a necessidade de imputar valores nesta coluna, dado que a inexistência das idades dos pacientes não invalida a utilização das imagens, visto que o objetivo

do estudo é classificar as imagens, sendo que a idade não impede a sua análise. Da mesma forma, é também descurada a eventual eliminação destas 57 amostras, visto que iria existir uma redução do nº de amostras, e não existe essa necessidade, porque o conteúdo não é afetado, pela falta destes valores.

Ainda, no que toca à análise do campo idade, foram analisadas algumas métricas, tais como: média, valor mínimo e máximo, desvio padrão, etc. A Figura 11 enumera os valores relativos às métricas analisadas.

```
count    9958.000000
mean       51.863828
std        16.968614
min         0.000000
25%        40.000000
50%        50.000000
75%        65.000000
max        85.000000
```

Figura 11 - Análise métricas - campo idade

É possível constatar que a média de idades, está na casa dos 52 anos, algo que está em conformidade com a incidência real, deste tipo de lesões. O desvio padrão é elevado, indicando-nos assim uma elevada disparidade dos dados, o que é visível através dos valores máximos e mínimos (0 anos e 85 anos). Relativamente à relação entre o 2º e o 3º quartil, é possível constatar uma distância semelhante entre a mediana e o 2º e 3º quartil, 10 e 15 respetivamente.

O gráfico ilustrado através da Figura 12, descreve a distribuição das imagens, pelo sexo do paciente.

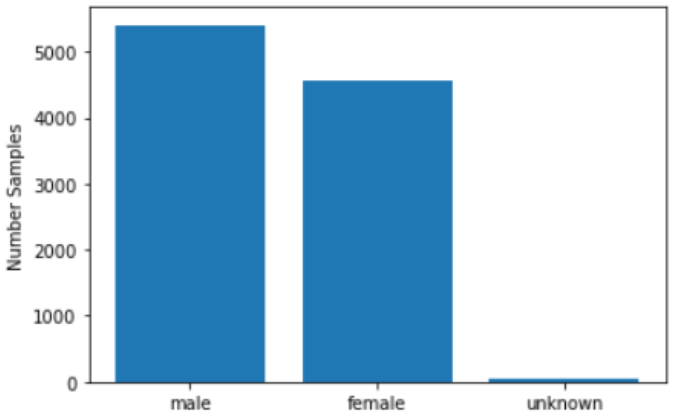

Figura 12 – Distribuição das amostras pelo sexo do paciente

Tal, como já tinha sido destacado anteriormente, não existem apenas incoerências associadas à coluna idade. O atributo sexo, também reúne um pequeno conjunto de amostras, onde a sua categoria, não é valida, sendo esta "unknown". Como podemos visualizar através do gráfico o nº de amostras categorizadas como "unknown" é baixo. O contexto abordado para a coluna idade, mantêm-se válido também para esta coluna. Dado que, esta "indefinição" do atributo sexo em algumas imagens, não interfere com a sua classificação.

Relativamente às categorias "male" e "female", é possível constatar que o nº de imagens associadas a pacientes "homens", revela-se um pouco superior face a pacientes "mulheres". Esta diferença é de cerca de 1000 amostras.

A análise prosseguiu com a exploração, da distribuição do nº de amostras, tendo em conta o local onde a lesão se situa. A Figura 13 ilustra essa mesma distribuição.

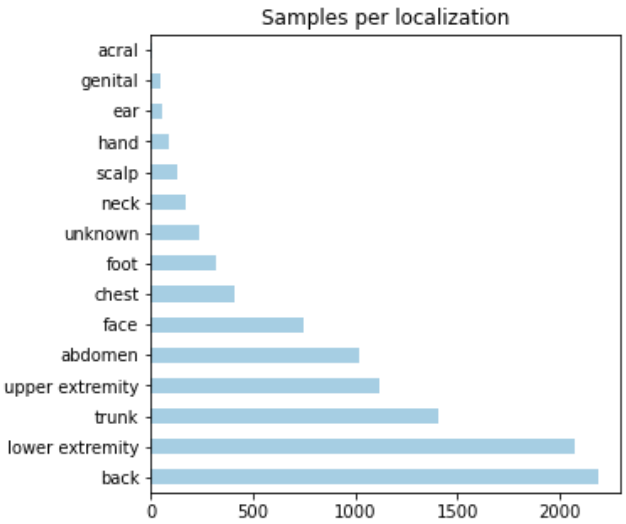

Figura 13 - Nº de imagens, por local onde a lesão se situa

Recorrendo à observação da Figura 12, constatamos a disparidade no nº de amostras, que existe entre as várias localizações das lesões.

Para além disso, constatamos novamente a presença de uma categoria, que não está em coerência com o atributo em análise, mais uma vez: "unknown". Neste caso, o nº de amostras associadas a esta categoria, é significativamente superior, aos valores identificados na coluna idade ou sexo. Existindo assim, 300 a 400 imagens, que não têm associada uma localização válida para a respetiva lesão que identificam. Ainda assim, as conclusões salientadas para as colunas analisadas anteriormente, mantêm-se válidas também para este atributo. Dado que, a inexistência de uma localização válida, não invalida a classificação da imagem.

Analisando as restantes categorias, é possível verificar que existem uma evidente disparidade entre estas, por exemplo: a categoria "acral" reúne um nº extremamente baixo de imagens. Por outro lado, existem mais de 2000 imagens, onde a lesão se situa nas costas "back". Existem 5 classes que juntas agregam quase todas as amostras do *dataset*: *back, lower extremity, trunk, upper extremity* e *abdómen*.

O próximo atributo a analisar é o *dx_type*. Tal como já fora indicado previamente, esta coluna descreve o tipo de diagnóstico, que fora considerado na avaliação de uma determinada amostra. De relembrar ainda, que os tipos de avaliação considerados foram: *Histopathology* (*histo*), *Reflectance confocal microscopy (confocal)*, *Consensus of at least three expert dermatologists from a single image* (*consensos*), *Lesion did not change during digital dermatoscopic follow up over two years with at least three images* (*follow-up*).

A Figura 14 ilustra o nº de amostras, por tipo de diagnóstico.

Da análise da Figura 14 concluímos, que existem dois diagnósticos com bem maior incidência, perante as amostras do problema, concretamente: *Histopathology* e *follow-up*. Ou seja, na maioria das situações, não foi necessário recorrer a outras vias de diagnóstico, para aferir qual o tipo de lesão de um paciente. Ainda assim, em cerca de 1000 imagens existiu a necessidade de recorrer à opinião de três ou mais médicos, para aferir com segurança, qual o tipo de lesão do paciente, demonstrando assim a dificuldade, que existiu na classificação "humana" destas

1000 imagens. O recurso ao método *Reflectance confocal microscopy*, apenas se revelou necessário num baixo nº de casos.

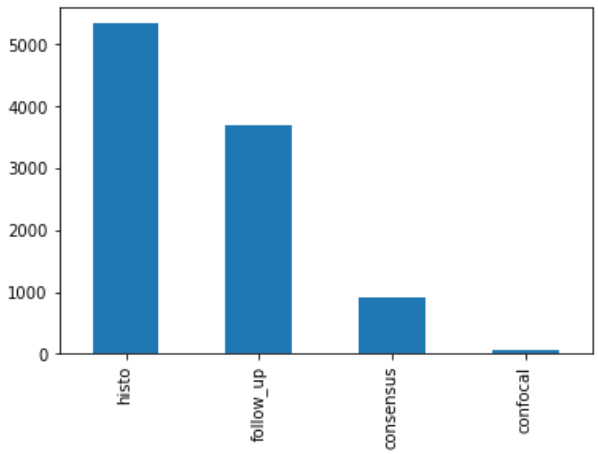

Figura 14 - Nº de amostras, por tipo de diagnóstico

Finalmente, segue-se a análise da distribuição das imagens, pelas várias classes do problema. A Figura 15 ilustra essa variação.

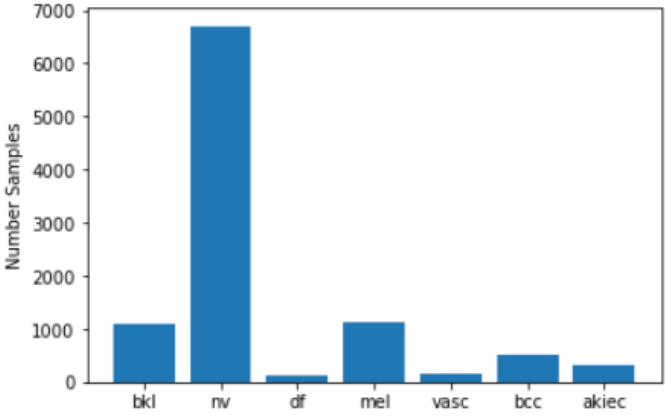

Figura 15 - Nº de imagens, por classe

Relativamente, à distribuição do nº de imagens por classe, é possível verificar que o *dataset* se revela pouco balanceado, tal como já era expectável. O artigo [1] compara a distribuição do nº de imagens por classe deste problema, com outros *dataset's* inseridos no mesmo âmbito. Dessa análise foi possível comprovar, as mesmas evidências: o não balanceamento das classes e ainda o facto das classes: *Melanocytic nevi, Melanoma* e *Benign Keratosis* serem sempre as mais representativas.

Este *dataset*, ao apresentar um maior nº de amostras, dificulta ainda mais o processo de classificação das imagens, dado que a classe *Melanocytic nevi* contêm quase 70% do nº total de imagens do *dataset*. As classes *Melanoma* e *Benign Keratosis,* agregam respetivamente 10% do nº total de imagens, já as restantes quatro classes contêm os últimos 10%.

Desta análise, fica subjacente a eventual dificuldade, que irá ser enfrentada, no processo de classificação das imagens relativas às classes menos balanceadas do problema. Dado que, o seu baixo nº de amostras dificulta a identificação de particularidades próprias destas classes.

Concluída a análise univariada do problema. Seguiu-se uma breve análise bivariada das *features* do problema, tentando apenas relacionar as classes do problema, com algumas das *features*, de modo a entender um pouco melhor o problema (não foi aplicada, uma análise severa, visto que o estudo a realizar não o necessita).

A primeira abordagem considerada, relaciona a idade dos pacientes com a lesão sofrida (classe). A Figura 16 ilustra a variação da idade dos pacientes, por cada uma das sete classes do problema.

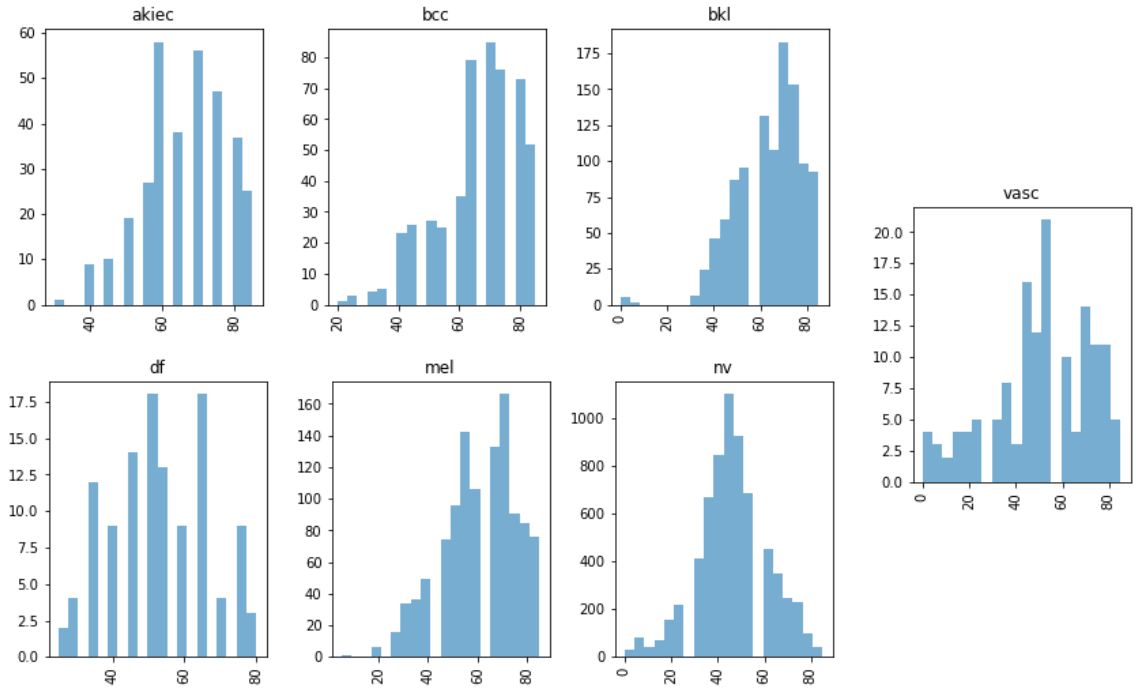

Figura 16 - Distribuição do nº de amostras por idade e por classe

Na maioria das classes é possível observar o mesmo padrão, isto é, existe uma maior incidência das lesões, em pacientes acima dos 50 anos de idade, mais concretamente nas classes: *akiec, bcc, bkl, mel e vasc*. Relativamente, às duas classes restantes: *df* e *nv*, verifica-se um maior equílibrio entre ambas as partes, ou seja, a incidência de casos revela-se semelhante para pacientes com idade inferior e superior a 50 anos. Mas, num contexto geral, independentemente do tipo de lesão sofrida, a probabilidade do paciente ter uma idade superior a 50 anos é elevada.

Seguidamente, foi efetuada uma análise à distribuição do nº de imagens por sexo, para cada uma das sete classes. A Figura 17 ilustra essa distribuição, por cada uma das classes.

Observando a Figura 17, é possível concluir que na maioria das classes, a categoria *male* é a mais representada. A única classe onde, existe uma maior incidência no sexo feminino é a *vasc*. Curiosamente, a classe com maior nº de imagens: *nv*, apresenta uma relação muito próxima, entre ambos os sexos. Demonstrando assim, que as mulheres também apresentam uma elevada incidência de apresentarem a lesão mais frequente de cancro da pele. As classes *bkl* e *nv* apresentam um nº reduzido de imagens associadas à classe "*unknown*" (este problema já foi devidamente fundamentado anteriormente).

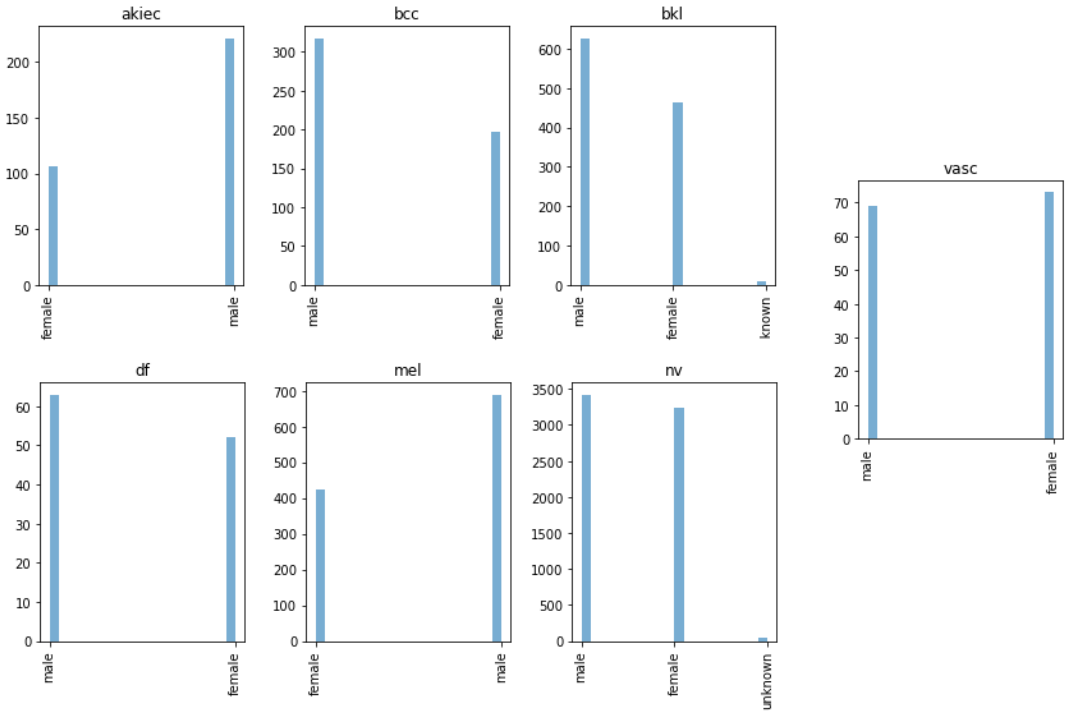

Figura 17 - Distribuição do nº de imagens por sexo, e por classe

A Figura 18, ilustra o nº de imagens, que estão associadas aos vários locais, onde se encontram as lesões, estando agrupadas pelas sete classes do problema.

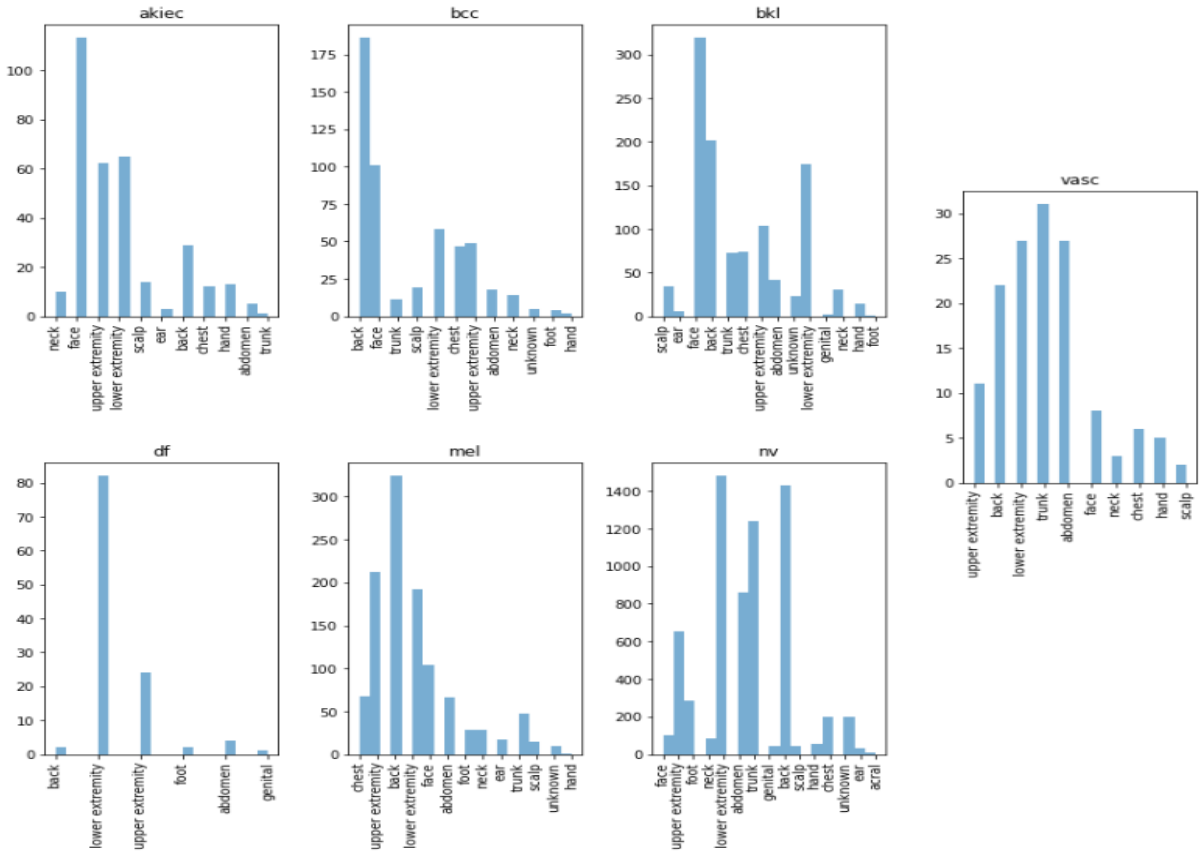

Figura 18 – Distribuição do nº de classes por localização, e por classe

Analisando, a Figura 18 é possível comprovar que os locais com maior predominância de lesões, estão dependentes do tipo de lesão sofrida pelo paciente. Isto é, observando por exemplo a categoria *back,* constatamos que a mesma é muito saliente na maioria das classes, como: *bcc, bkl, nv, mel* ou *vasc*. Contudo, nas classes *df* ou *akiec* não se revela muito dominante face às restantes categorias.

É possível ainda concluir que na maioria das classes, existem locais que estabelecem uma "área dominante", promovendo assim uma maior incidência de uma lesão se contrair num determinado local. Ou seja, exemplificando com a classe *df,* podemos constatar que as lesões são praticamente apenas inferidas, em 2 locais: *upper* e *lower extremity*. O mesmo se procede com as restantes classes, onde é notória a maior incidência em determinados locais.

Finalmente, o último passo desta análise exploratória incide na ilustração de algumas imagens referentes a cada uma das classes do problema.

Antes de avançar na execução de novas tarefas, é importante visualizar algumas imagens do *dataset*, de modo a tentar identificar algumas particularidades e diferenças, entre as várias classes do problema. Pois dessa forma, torna-se mais fácil dar resposta a uma série de questões, que são pertinentes, tais como: (1) É difícil encontrar diferenças entre as várias classes? (2) Será necessário considerar uma arquitetura mais complexa para a resolução do problema?

Esta análise ajudará futuramente, na definição dos modelos a serem criados, para a classificação de imagens.

Para tal, foi efetuada a leitura do ficheiro *HMNIST_28_28_RGB.csv*. Este ficheiro contém os valores *RGB* de cada imagem, considerando uma altura e largura igual a 28. Para além disso, contém ainda uma coluna, que especifica o *output* associado a cada imagem. **De destacar, que o estudo não vai considerar esta escala, irá considerar uma escala maior** (foi considerada esta abordagem aqui, de modo a reduzir o tempo desperdiçado).

A Figura 19 demonstra o *output* obtido da leitura do ficheiro (*DataFrame* resumido).

	pixel0000	pixel0001	pixel0002	pixel0003	pixel0004	pixel0005	pixel0006	pixel2351	label
0	192	153	193	195	155	192	197	177	2
1	25	14	30	68	48	75	123	27	2
2	192	138	153	200	145	163	201	117	2
3	38	19	30	95	59	72	143	15	2
4	158	113	139	194	144	174	215	92	2

Figura 19 - Leitura do ficheiro HMNIST_28_28_RGB.csv, e respetivo output (DataFrame)

A Figura 19, demonstra assim as cinco primeiras linhas, do objeto *DataFrame,* resultante da leitura do ficheiro indicado. O resultado obtido consiste num objeto que contém 10015 imagens (nº de linhas), e um nº total de colunas igual a 2352. As primeiras 2351 colunas correspondem aos valores dos pixéis *RGB*, expressos numa escala de 0-255. O nº 2351 resulta da multiplicação entre a altura e largura de cada imagem (28*28), sendo que é ainda necessário multiplicar este valor pelo nº total de *channels*, neste caso o nº de *channels* é igual a 3, visto que é utilizado o filtro *RGB* (28*28*3). Finalmente, a última coluna identifica o *output* de cada imagem (o tipo de lesão).

Depois de obtido o *Dataframe,* já estão definidos todos os requisitos necessários, à ilustração de algumas imagens alusivas às várias classes do problema. Sendo assim, a Figura 20 ilustra 3 imagens pertencentes a cada uma das sete classes do problema.

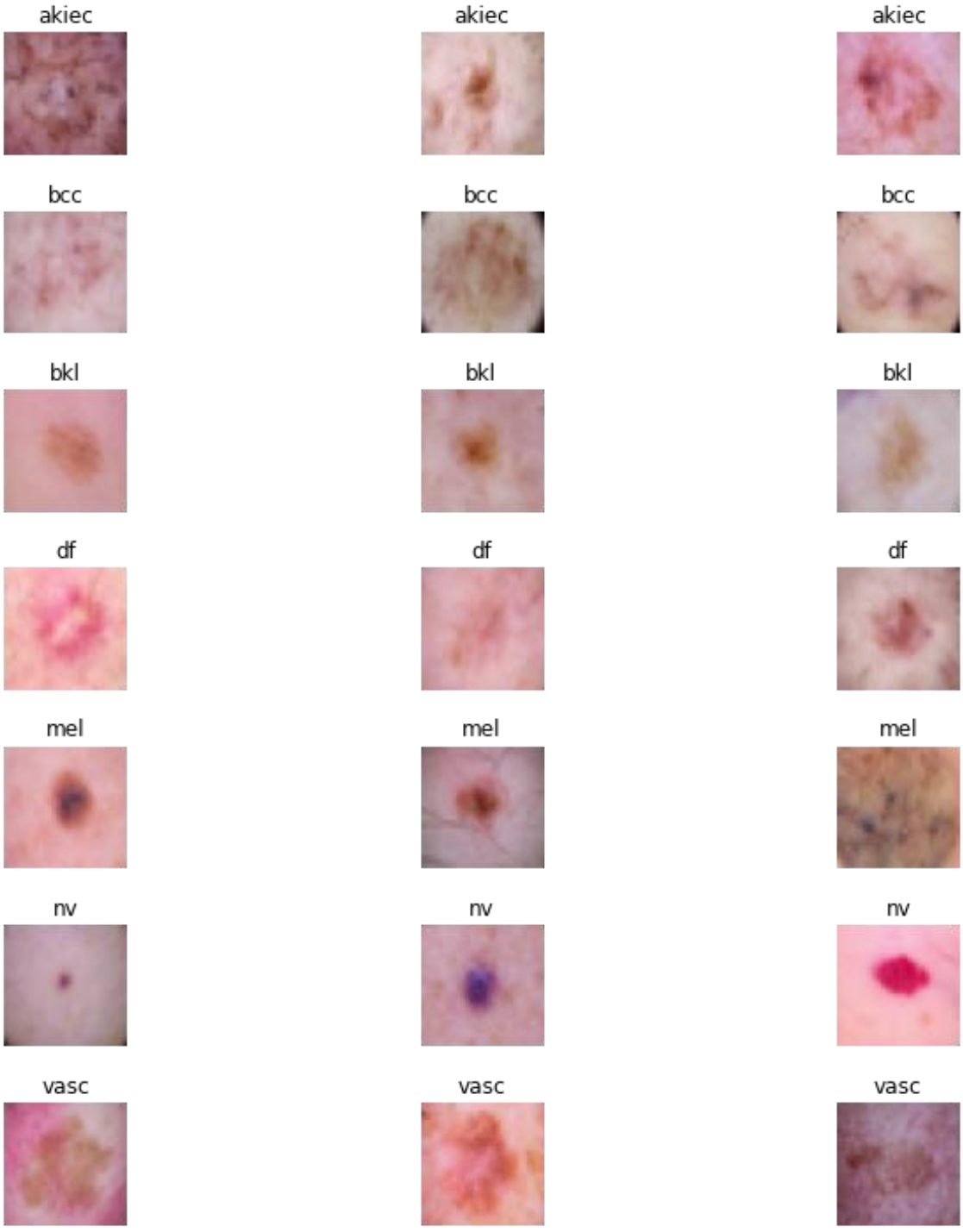

Figura 20 - Imagens alusivas, às várias classes do problema

Tal como já esperado, o problema revela-se complexo. Sendo difícil à 1ª vista identificar particularidades nas imagens, que permitam a fácil e correta classificação das imagens.

Ainda assim, é possível concluir que existem algumas evidências, que ajudam a identificar a classe associada a uma determinada imagem, como: a cor, tamanho ou forma da lesão.

Por exemplo, as lesões associadas à classe *nv* revelam uma cor e forma distintas, face aos restantes tipos de lesões. Isto é, revelam tons fora do "padrão" castanho, e apresentam uma lesão em forma circular, onde o seu tamanho é pequeno.

Já nas classes *vasc* ou *akiec*, a forma da lesão revela-se menos uniforme, não seguindo o padrão circular da classe *nv*. A cor da lesão já segue o padrão visível ao longo de todas as classes, isto é, lesão com tons castanhos.

Já as restantes classes: *mel, bcc, bkl* e *df* apresentam bastantes semelhanças entre si, sendo difícil identificar particularidades únicas, em cada uma das classes mencionadas. Mas, ainda assim é percetível, por exemplo que a classe *bkl* apresenta tons mais acastanhados do que a classe *df*, e as suas lesões apresentam tamanhos inferiores, à classe *bcc*. Ou seja, apesar de ser bastante complexa a classificação correta destas imagens (daí a dificuldade presenciada pelos profissionais de saúde, sendo necessário recorrer a métodos, como: o *follow-up*, onde é necessária a opinião de diversos médicos no diagnóstico de uma lesão), existem alguns factores que ajudam a identificar qual a classe correta.

Técnicas de Pré-Processamento aplicadas:

Concluído o estudo do *dataset*, segue-se a realização de um determinado conjunto de tarefas de processamento de dados. Estas tarefas, resumem-se essencialmente à aplicação de técnicas de obtenção, divisão, limpeza e transformação dos dados.

Esta fase deve ser analisada e definida com ponderação, pois os resultados obtidos no futuro, estarão diretamente relacionados, com as práticas implementadas previamente. Sendo assim, e tendo em conta o *dataset* em estudo, é necessário dar maior enfâse às técnicas de obtenção, divisão e transformação dos dados. As técnicas de Limpeza não se aplicam a este contexto, dado que o autor irá trabalhar diretamente com as imagens, e não com as informações (*metadata*), que foram exploradas durante a análise exploratória.

Tal como já fora referido atrás, o tamanho real das imagens é de 600*450, respetivamente largura e altura. Mas, os ficheiros *.csv* fornecidos consideram apenas, a utilização de formatos muito pequenos (28*28 e 8*8), reduzindo assim a qualidade das imagens, dado que muita informação foi comprimida/extraída. Sendo assim, foi necessário proceder à aplicação de técnicas de *resize,* das imagens reais. Os valores para a altura e largura consideradas (das imagens a redimensionar), foram de: 90 e 75 respetivamente.

Neste processo foi utilizada a biblioteca *opencv*, que permite a obtenção dos valores dos pixéis das imagens, tendo em conta o formato da imagem pretendida. De realçar ainda que foi mantido o mesmo filtro (das imagens originais), isto é, o filtro *RGB*.

Desta operação resultaram dois *arrays:* o 1º contendo os valores *RGB*, dos pixéis de cada imagem (no formato redimensionado: 90*75), e ainda um 2º *array* contendo os *targets* associados a cada imagem.

Estes dois *arrays*, agregam todos os dados do problema. Dessa forma, é necessário efetuar uma divisão das amostras totais, pelos vários conjuntos de dados, a serem utilizados posteriormente, em tarefas de treino e previsão dos modelos, ou seja, dividir o nº total de imagens disponíveis em três conjuntos de dados: treino, validação e teste.

A divisão aplicada representa o *standard*, isto é, 60%-20%-20% respetivamente treino, validação e teste. Por exemplo, considerando um nº total de imagens igual a 2000, a distribuição dos dados será de: 1200, 400, 400, respetivamente conjunto de treino, validação e teste.

Após a aplicação das técnicas de obtenção e divisão dos conjuntos de dados necessários, segue-se a aplicação de técnicas de transformação dos dados.

Este tipo de técnicas é muito importante, neste tipo de problemas, visto que a aprendizagem dos modelos está dependente dos dados. Ou seja, é necessário garantir que existe uma escala comum entre as várias *features* do problema, neste caso entre os vários pixéis. Pois, a escala destes valores varia entre 0-255, isto é, pixéis com valores mais elevados revelam uma maior "importância", comparativamente a pixéis que apresentem valores mais baixos.

Dessa forma recorreu-se à aplicação da técnica *z-score*, mais conhecida por standardização, em que a mesma tenta atribuir uma escala comum entre as várias *features* do problema. Para tal, a mesma aplica a diferença entre o valor de uma feature *X*, e o valor médio presente no conjunto de dados a considerar. O resultado desta subtração, é depois dividido pelo desvio padrão do conjunto de dados. De uma forma resumida, esta técnica redimensiona a transformação de um determinado conjunto de valores, para que a média seja 0 e o desvio padrão igual a 1. A Fórmula 1 ilustra a equação descrita.

$$X_{new} = \frac{X - \mu}{\sigma} \quad (1)$$

Esta técnica foi aplicada a todos os conjuntos de dados: treino, validação e teste. Mas, e de modo a evitar *data leakage*, antes de aplicar a Fórmula 1, é necessário calcular a média e desvio padrão do *dataset* de treino. Ou seja, de modo a evitar que os *dataset's* de validação e de teste, reúnam informações adicionais que não são desejadas, é necessário garantir que por exemplo: o *dataset* de teste mantêm unicamente informação "real", e não informação "conhecida previamente".

Sendo assim, a Fórmula 1 aplicada no conjunto de dados de validação e de teste, recorre à utilização da média e desvio padrão, dos dados de treino. Desta forma, garante-se que os modelos generalizam apenas dados que nunca antes foram vistos.

Finalmente, a última operação a aplicar consiste na transformação do *output* das imagens¸ em formato binário (*one-hot encoding*).

Análise de Resultados:

O próximo "passo" a desenvolver, baseia-se na criação de modelos convolucionais. Estes modelos irão ser utilizados, como "ferramenta" de aprendizagem dos dados que foram anteriormente "tratados".

Tal, como aconteceu no estudo do *dataset* anterior (*breast histopathology*), irão ser aplicados vários tipos de redes convolucionais, com o objetivo de comparar e aferir os benefícios, que estas promovem.

De modo a melhor a eficiência dos modelos, irá ser considerada a aplicação de várias estratégias de treino, sendo expectável que as mesmas promovam melhorias na aprendizagem e generalização dos modelos.

A análise descrita seguidamente, irá recorrer a diversas ilustrações, com o objetivo de garantir uma análise mais simplista, e visualmente agradável. Por outro lado, a análise de resultados, recorrendo a várias Figuras ajuda a entender quais os pontos a melhorar e quais as melhorias atingidas.

Por outro lado, esta análise tenta ser o mais completa possível, tentando assim promover um estudo gradual do *dataset*, bem como a aplicação de diferentes cenários, considerando os vários modelos aplicados. Ou seja, os cenários a aplicar enquadram-se na aplicação de várias condicionantes: (1) aumento da complexidade da rede, (2) aumento da profundidade da rede, (3) aplicação de estratégias de treino e (4) recursos a algoritmos de otimização, na melhoria dos resultados obtidos.

Como referi anteriormente, esta análise irá ser gradual, isto é, por exemplo: inicialmente é necessário verificar o comportamento dos modelos quando considerado um baixo nº de parâmetros (baixa complexidade), só depois é que o autor deve aumentar a complexidade da rede. Mediante os resultados obtidos, e num contexto faseado, deve-se aplicar os restantes cenários enumerados, na tentativa de melhorar os resultados obtidos, e reduzir os custos computacionais exigidos.

A aplicação de algoritmos de otimização, representa uma abordagem que "concatena" os outros três cenários identificados, tentando dessa forma identificar um modelo eficiente, robusto e com o menor custo computacional, sem a necessidade de trabalho manual, e de tentativas incrementais, de melhoria de resultado.

Finalmente, importa apenas realçar que as arquiteturas *CNN* consideradas foram: *AlexNet,* VGGNet e *U-Net*. Já, as estratégias de treino utilizadas foram: *Oversampling, Data Augmentation* e *Segmentation Mask's*.

Seguidamente, é efetuada a análise dos resultados obtidos (sendo que, a análise de cada rede *CNN*, é separada).

AlexNet:

O primeiro tipo de rede em análise, é a rede *AlexNet*. Posteriormente, é efetuado um estudo das várias condicionantes descritas atrás.

Importa apenas salientar, que algumas das representações contextualizadas seguidamente, correspondem a circunstâncias obtidas, após o estudo prévio e iterativo de outras situações

referentes ao mesmo contexto. Mas, como não é possível relatar todas as “ocorrências” criadas, então foram apenas documentadas as situações mais relevantes.

Análise à complexidade da rede:

O primeiro cenário a considerar prende-se com a aplicação de diversos cenários, relacionados com a variação da complexidade da arquitetura, mais concretamente o aumento ou a diminuição do nº total de parâmetros da rede.

É essencial efetuar este estudo, numa fase precoce da análise, visto que permite verificar o comportamento do modelo, quando se considera uma rede mais complexa. Esta análise, permite aferir qual a abordagem que permite a obtenção dos melhores resultados. Sendo que, a aplicação das estratégias de treino, estão dependentes da arquitetura considerada (complexidade, profundidade, *etc*).

Referências:

[1] https://arxiv.org/ftp/arxiv/papers/1803/1803.10417.pdf --> The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions

[2] https://www.kaggle.com/kmader/skin-cancer-mnist-ham10000