Statistical analysis for the development of a Deep Learning model for classification of images with TDP-43 pathology

Azucena Muñoz, Vasco Oliveira, and Marta Vallejo

Heriot-Watt University, Edinburgh, UK

Diagnosing Amyotrophic Lateral Sclerosis (ALS) remains a hand challenge due to its inherent heterogeneity. Notably, the occurrence of TDP-43 cytoplasmic aggregation in approximately 95% of ALS cases has emerged as a potential indicative hallmark. In order to develop deep learning models capable of distinguishing TDP-43 proteinopathic samples from their healthy counterparts, a comprehensive understanding of the sample set becomes imperative, particularly when the sample size is limited. The samples in question encompassed images obtained via an immunofluorescence procedure, employing super high-resolution microscopy coupled with meticulous processing. A feature-extracted dataset was created to collect meaningful features from every sample to approach three different classification problems (TDP-43 Pathology, TDP-43 Pathology Grades and ALS) based on the number of red and pink pixels, signifying cytoplasmic and nuclear TDP-43 presence. A series of diverse statistical approaches were undertaken. However, definitive outcomes remained elusive, although it was suggested that a classification based on the presence of TDP-43 proteinopathy was better than the one based on the presence of ALS for training the model.

The dataset was reduced by eliminating the problematic samples through curation. Analyses were repeated using t-student tests and ANOVA, and visualisation of patient inter-variability was performed using hierarchical clustering. The TDP-43 pathology classification results showed significant differences in the number of red and pink pixels, the total amount of protein and the cytoplasmic and nuclear proportions between healthy and pathological samples between groups. These findings suggested that images classified according to the presence of TDP-43 proteinopathy are more suitable for training deep learning models.

1 Introduction

1.1 Amyotrophic Lateral Sclerosis

Amyotrophic Lateral Sclerosis (ALS) is currently recognised as a multisystem neurodegenerative disease with great clinical, genetic and neuropathological heterogeneity [Masrori and Van Damme, 2020]. The estimated incidence of ALS is 1.75 to 3 per 100,000 people per year, which rises to 4-8 in the highest-risk age group (45-75 years). Prevalence has significant geographic differences as it ranges between 10-12 per 100,000 people in Europe and 3.84-5.56 per 100,000 people in the United States. About 90-95% of cases are sporadic (sALS), and the remaining 5-10% are hereditary or familial (fALS). The estimated risk of developing ALS is 1:350 in men and 1:400 in women [Berry et al., 2023]. Diagnosis primarily relies on phenotypic assessment as it is based on signs of upper motor neurons (UMN) or lower (LMN) motor neuron dysfunction in patients

Azucena Muñoz: am2288@hw.ac.uk, 0009-0003-9337-6153
Vasco Oliveira: vo2003@hw.ac.uk, 0009-0007-1866-0022
Marta Vallejo: m.vallejo@hw.ac.uk, 0000-0001-9957-954X

with progressive muscle weakness with no alternative explanation. There is no standard diagnostic protocol approved by doctors or specific diagnostic tools, and revised criteria of El Escorial [Costa et al., 2012], Awaji's algorithm [Costa et al., 2012] and Gold Coast criteria [Shen et al., 2021] are mainly used for patients' inclusion in clinical trials or scientific studies. The lack of consensus and tools causes a delay of up to one year in the diagnosis, which worsens the prognosis of the disease.

Only one drug has been approved by the European Medicines Agency (EMA) against ALS, Riluzole, a glutamate antagonist that offers a marginal yet statistically significant extension of survival [Bensimon et al., 1994]. Despite scientific advances, clinical trials for new drugs continue to face challenges because ALS is treated as a single disease without considering the cause or the mechanisms involved, which vary greatly depending on the case [Masrori and Van Damme, 2020].

As with most neurodegenerative pathologies, ALS is believed to occur as a combination of ageing-associated dysfunctions, genetic predispositions and environmental factors. At the genetic level, there is great pathological diversity across more than 20 associated genes. About 15% of cases stem from autosomal dominant factors due to mutations that affect protein degradation pathways and may favour the accumulation of TDP-43 (TBK1, OPTN, SQSTM1, UBQLN2, C9orf72 and VCP), metabolism pathways of RNA (TARDBP, FUS, MATR3, TIA1, hnRNPA1 and ATXN2), and dynamics of the cytoskeleton and axonal transport (TUBA4A, PFN1, KIF5A and DCTN1), among other genes involved (SOD1). Besides genetic factors, ageing and male sex increase the risk of ALS [Masrori and Van Damme, 2020]. Environmental risk factors include smoking, body mass index, physical exercise, occupational and environmental exposure to metals, pesticides, β-methylamino-l-alanine, head injuries, and viral infections. Despite these associations, the definitive causal relationship of these factors with ALS has not been established [Masrori and Van Damme, 2020].

Clinically, the onset of ALS starts with weakness and focal muscle atrophy that subsequently spreads as the disease progresses. There is great variability regarding the initial site of manifestation (usually limbs), the age of onset (58-63 years for sALS and 40-60 years for fALS) and the rate of progression. The average survival period after the onset of symptoms is three years, where death is caused by respiratory failure associated with disease progression. Beyond motor difficulties, about 50% of patients suffer from extra motor manifestations, which are limited to mild behavioural and/or cognitive changes in 35-40% of cases. However, 10-15% experienced an additional diagnosis of frontotemporal dementia (FTD). This complex condition involves degeneration of frontal and anterior temporal lobes that causes behaviour changes, executive functioning impairment, and/or language impairment by molecular mechanisms underlying both FTD and ALS [Masrori and Van Damme, 2020].

The pathogenic profile of ALS includes loss of neuromuscular connection, axonal retraction, and subsequent cell death of UMN and LMN, surrounded by astrogliosis and microgliosis, with ubiquitinated inclusions in surviving neurons with TDP-43 as the major component in more than 95% of ALS cases [Neumann et al., 2006]. Multiple molecular pathways are implicated in the pathogenesis, such as failures in proteostasis, neuroinflammation, excitotoxicity, mitochondrial dysfunction and oxidative stress, oligodendrocyte dysfunction, cytoskeletal abnormalities and axonal transport defects, RNA metabolism abnormalities, nucleocytoplasmic transport deficits, and impaired DNA repair [Masrori and Van Damme, 2020, Brown and Al-Chalabi, 2017]. In brief, ALS mainly affects the processes of protein control and degradation, RNA metabolism, and cytoskeletal and axonal transport, making the cytoplasmic aggregation of TDP-43 the prevailing neuropathological hallmark as it is present in more than 95% of patients [Masrori and Van Damme, 2020, Neumann et al., 2006].

1.2 TDP-43

TDP-43 (TAR DNA-binding protein 43) is a 414 amino acid RNA-binding protein. Therefore, it is involved in multiple RNA biogenesis and processing steps, such as transcription, splicing, microRNA maturation, and RNA transport [Bhardwaj et al., 2013].

On the one hand, TDP-43's two RNA recognition motifs (RRM1 and RRM2) allow the binding to GU repeats predominantly located in long introns and at the 3'UTR end of the mRNA [Bhardwaj et al., 2013, Lukavsky et al., 2013, Kuo et al., 2014]. Furthermore, TDP-43 regulates the splicing of many RNAs, both coding and non-coding, including mRNAs encoding proteins involved in

neuronal survival and various proteins relevant to neurodegenerative diseases [Tollervey et al., 2011, Wang et al., 2018]. It also stabilises the mRNA by recruiting CNOT7/ CAF1 deadenylase to the 3'UTR end, leading to deadenylation of the poly(A) tail and thus shortening [Fukushima et al., 2019]. It also regulates the processing of mitochondrial transcripts, so it is involved in the maintenance of mitochondrial homeostasis [Izumikawa et al., 2017].

In response to oxidative stress, TDP-43 associates with ribosomes by forming stress granules that promote cell survival [Higashi et al., 2013, Colombrita et al., 2009]. Likewise, TDP-43 participates in the formation and regeneration of skeletal muscle through the formation of cytoplasmic myo-granules and the binding to mRNA encoding sarcomeric proteins [Vogler et al., 2018]. On the other hand, TDP-43 regulates the expression of HDAC6, ATG7 and VCP in a PPIA/ CYPA-dependent manner [Vogler et al., 2018] and downregulates CDK6 expression [Kirby et al., 2010]. In addition, TDP-43 contributes to the maintenance of circadian clock periodicity by stabilising CRY1 and CRY2 proteins in an FBXL3-dependent manner [Hirano et al., 2016].

Considering its nuclear and cytoplasmic functions, TDP-43 can move between the nucleus and the cytoplasm, the cell nucleus its main location. In addition, there is a small concentration of TDP-43 located in the mitochondria, and in cases of oxidative stress, it is located in the stress granules. TDP-43 is encoded by the TARDBP gene, which has 18,185 base pairs and is located at 1p36.22 in the positive or sense strand of chromosome 1 between base pairs 11,012,344 and 11,030,528. It has 38 transcripts, and 5,371 single nucleotide polymorphisms (SNPs) that affect one or more of the protein's functions have been identified [Ensembl].

1.3 The TDP-43 proteinopathy

The TDP-43 proteinopathy is characterised by cytoplasmic aggregation and nuclear depletion. Its prevalence is observed in 95% of ALS cases, in 40-50% of FTD cases[Younes and Miller, 2020], in 20-50% of patients with Alzheimer's disease [Jo et al., 2020] and in the limbic structures of most patients with predominantly limbic age-related TDP-43 encephalopathy (LATE) [Nelson et al., 2019, Prasad et al., 2019]. These findings substantiate the association of TDP-43 pathology with neurodegeneration, particularly when explored within the framework of ALS.

The reasons for this suboptimal localisation can be multifaceted. Primarily, at least 48 pathogenic dominant variants of the TARDBP gene have been associated with ALS, most of them being nonsense mutations in the glycine-rich region of the carboxy-terminal region [Lattante et al., 2013]. In these cases, the causal participation of this protein can be established in the pathogenesis of the disease. However, there are TARDBP mutations in only 3% of fALS cases and 1.5% of sALS cases [Prasad et al., 2019, Grad et al., 2016], meaning a significant proportion of patients exhibit TDP-43 proteinopathy in the absence of TARDBP gene mutations, necessitating the involvement of other genetic factors. TDP-43 proteinopathy has already been associated with C9ORF72 expansions, VCP (Valosin-containing protein) and TBK1 mutations, and hnRNPA1 and hnRNPA2B mutations in multisystem proteinopathies (MSPs) [Jo et al., 2020, Prasad et al., 2019, Gitcho et al., 2009, Kim et al., 2013, Taylor, 2015].

However, although associated genes have been identified, the pathogenic mechanism leading from genetic mutation to the formation of cellular inclusion and the subsequent onset of ALS remains not understood yet. Three possible hypotheses that could give rise to this type of proteinopathy are discussed: the acquisition of a gain-of-function, the emergence of a novel deleterious function and the manifestation of a loss-of-function scenario [Broeck et al., 2014].

Ensuring the optimal functioning of TDP-43 mandates a meticulous balance in its expression levels. This homeostasis depends on a process of autoregulation, nucleocytoplasmic transport, correct liquid-liquid phase separation (LLPS), correct mitochondrial function, autophagy and RNA binding, so any of these processes' alteration can lead to TDP-43 proteinopathy [Versluys et al., 2022]. In addition to mislocalisation and aggregation, the positive and negative involvement of certain post-translational modifications (PTMs), such as ubiquitination, proteolytic cleavage, phosphorylation, and others less identified, such as acetylation, sumoylation, and disulfide bond formation should be considered [Versluys et al., 2022]. The pathological role of these MPTs has not been fully studied, with phosphorylation and ubiquitination being the most understood.

Clinically, TDP-43 proteinopathy impairs the motor system and exacerbates cognitive decline, which explains its association with the FTD that sometimes accompanies ALS. At the microscopic

level, gliosis and microvacuoles accompany synaptic and neuronal loss [Younes and Miller, 2020]. Four morphological patterns of TDP-43 inclusions have been established to classify four types of FTD: type A is characterised by having small intracytoplasmic inclusions in cortical layers and intranuclear inclusions in superficial cortical layers II and III; type B presents round neuronal inclusions in the cortex; type C has long neurites immunoreactive to TDP-43; and in type D, intranuclear inclusions and cytoplasmic inclusions are distinguished in neurons [Younes and Miller, 2020, Versluys et al., 2022]. However, these morphological patterns have not been specifically associated with ALS. Instead, cytoplasmic aggregates are discussed in a general and less detailed manner.

The involvement of TDP-43 proteinopathies beyond neurodegeneration in myopathies has also been studied. TDP-43 aggregates have been found in inclusion body myopathies such as sporadic inclusion body myositis and inclusion body myopathy associated with PCV multisystem proteinopathy [Küsters et al., 2009, Olivé et al., 2009]. The role of TDP-43 in muscle physiology and related diseases continues to be investigated [Vogler et al., 2018, Militello et al., 2018].

Currently, there are monoclonal antibodies designed to detect phosphoserines 409 and 410 of the C-terminal region of TDP-43 [Inukai et al., 2008] as phosphorylation of these residues is only found in patients with FTD or ALS and therefore can be considered a pathological marker of TDP-43 proteinopathy. However, this marker is only one of the many traits that can occur and cannot be used as a diagnostic tool for proteinopathy. For this reason, it is necessary to develop tests or diagnostic tools for TDP-43 proteinopathy that can be used for population screening [Turner, 2022, ALS Association, 2023].

1.4 Deep Learning

Deep Learning (DL) is a branch of Machine Learning (ML) at the same time that ML is a subset of Artificial Intelligence (AI). DL is considered the core technology of the Fourth Industrial Revolution, and in its supervised version, it seeks to overcome challenges related to classification and regression in multiple fields thanks to its ability to learn through training [Sarker, 2021]. These models can be applied to very diverse fields, including but not limited to healthcare [Yang et al., 2021], linguistics [Lauriola et al., 2022], visual recognition and cybersecurity [Xin et al., 2018].

The foundational architecture of DL is based on artificial neural networks, a type of model comprising an assembly of interconnected elementary processing units known as neurons that generate a series of real-value outputs processed through non-linear activation functions to obtain an objective final result (outcome). While the training phase of a DL model demands substantial time investment due to its considerations of myriad parameters, it requires little time to run in testing time [Xin et al., 2018]. Its main advantage over other classical ML algorithms is its ability to handle large amounts of data.

One crucial factor in the final performance reached by DL models is the quality and quantity of the data it learns from. This data can be sequential (with a relevant order), 2D (digital images that are numerical matrices, symbols, expressions or pixels arranged in rows and columns) or tabular (headed rows and columns organised as a dataset, following a logical and systematic arrangement of the data in the form of rows and columns that are based on the properties or characteristics of the data) [Sarker, 2021]. Furthermore, the models can be combined, forming complex architectures able to deal with more intricate data types, such as videos.

DL can be classified into three main categories according to the existence of labelled information (ground truth) to guide the training. These types are (1) supervised learning, where label data exists; (2) unsupervised learning, in which the absence of labelling data only allows the characterisation of high-order correlation properties and the analysis of patterns; and (3) hybrid learning, where label data is partial and requires a combination of both methods. The choice of a model will depend on the objective or problem to be solved and the availability of ground truth information [Sarker, 2021].

Among the most widely used supervised learning models to handle 2D data are Convolutional Neural Networks (CNNs). CNNs are designed to handle 2D data, so different variants are used for visual recognition, medical image analysis, or image segmentation [Sarker, 2021, Lecun et al., 1998]. These models learn directly from the input without the need for human intervention to extract features, elaborating multiple convolutions and pooling layers so that each layer considers

the optimal parameters to obtain a significant output and reduce the complexity of the model. In addition, overfitting, the problem of excessive adjustment to training data under limited sample sizes, can be ameliorated with tools like dropout, regularisation, early stopping and data augmentation [Gavrilov et al., 2018].

1.5 Goals

We aimed to develop CNN models capable of detecting the presence of ALS or TDP-43 protein pathology. We used a dataset of high-resolution microscopy images These images were obtained from post-mortem brain samples using immunofluorescence techniques to detect nuclei (DAPI) and super-resolution microscopy to highlight TDP-43 protein (pTDP-43$^{409/410}$ and Apt-1) as described in Twiddy [2021]'s work.

We subsequently processed these images to develop the CNN models based on the classification of samples by the presence or absence of ALS. However, initially, these models did not obtain substantial discriminative results. Our main objective of this paper is to understand and optimise the image data to support the model through a series of statistical analyses. By assessing the samples to be used for classification, we expect to identify any variable that may be useful for guiding the model. As a secondary objective, we seek to discern whether there are significant differences between the samples of the different degrees of protein pathology established by clinicians from the University of Edinburgh and assess the patient inter-variability.

Therefore, samples were analysed for five variables: cytoplasmic TDP-43 (red pixels), nuclear TDP-43 (pink pixels), the total amount of TDP-43 (red plus pink pixels), and cytoplasmic and nuclear ratios (red between red and pink sum or pink between red and pink sum respectively). These features were used to relate our samples with the existing bibliography, determine the most balanced classification, and understand variables with significant differences between groups. Some problematic samples were detected and deleted to repeat the statistical analysis and generate more significant conclusions. Two new datasets were created, compared and tested. Finally, the reduced dataset was used as input to a hierarchical clustering algorithm using Euclidean distance as the distance metric. The resulting dendrogram can provide insights into the structure and relationships within the data and help tailor classification models using the generated clusters of the data. Ultimately, the aim is to improve the performance and interpretability of binary classification tasks.

2 Methodology

2.1 Samples

Statistical analyses were conducted on the image dataset provided by the University of Edinburgh for the development of our CNN models. These samples were sourced from the Edinburgh Brain and Tissue Bank (EBTB), established by the UK Medical Research Council, as detailed in Twiddy [2021]. Originally, these samples were collected to facilitate the creation of a super-resolution microscopy-compatible probe aiming to characterise individual TDP-43 aggregates. This endeavour resulted in the creation of Apt-1, an RNA aptamer capable of binding to the pathologic TDP-43 protein in vitro. Apt-1 demonstrated superior discriminatory capability between ALS cases and healthy controls in post-mortem samples compared to pTDP-43$^{409/410}$. Furthermore, the extensively characterised primary antibody also exhibited binding affinity to pathologic TDP-43 [Twiddy, 2021].

Our study utilised post-mortem tissue samples from Brodmann area 4 (motor cortex) obtained from 48 individuals. Among these, 31 were identified as having ALS based on genetic and psychological assessments conducted through the ECAS (Edinburgh Cognitive and Behavioural ALS Screen) test [Twiddy, 2021].

Each ALS sample was dewaxed and rehydrated to undergo immunohistochemistry (IHC) with the primary rabbit polyclonal antibody pTDP-43$^{409/410}$ and staining with haematoxylin. These samples were recorded as Zeis.czi files using the fluorescence imaging mode of the ZEISS Axio Scan.Z1 scanner coupled to the brightfield microscope. Ten areas of interest were sectioned from each sample using Napari to evaluate the number of neurons and chromogen-positive glial cells (See

Table 1). Subsequently, the dataset passed a thresholding process with Python and Fiji, resulting in images with a black background, green signal for the antibody, red for the aptamer, and blue for the cell nuclei [Twiddy, 2021].

Based on this data, four grades of TDP-43 pathology were clinically assigned as follows: 0 (None or control), 1 (Mild), 2 (Moderate) and 3 (Severe), using the total number of chromogen-positive cells in the sample. See Table 2 and the corresponding assigned grades [Twiddy, 2021].

ALS Case ID	Crop 1 N, G	Crop 2 N, G	Crop 3 N, G	Crop 4 N, G	Mean N, G
Case 1	1, 1	1, 1	1, 1	1, 1	1, 1
Case 2	1, 2	1, 3	1, 2	0, 3	1, 2
Case 4	0, 1	1, 2	1, 2	1, 1	1, 2
Case 7	0, 1	1, 0	0, 0	0, 0	1, 1
Case 10	0, 1	0, 1	0, 0	0, 1	0, 1
Case 11	0, 0	0, 0	0, 0	0, 0	0, 0
Case 14	2, 2	3, 3	2, 2	3, 3	3, 3
Case 15	0, 1	0, 1	1, 0	0, 0	1, 0
Case 16	0, 1	0, 1	1, 0	0, 1	0, 1
Case 17	0, 0	0, 0	0, 1	0, 1	0, 1
Case 18	3, 1	3, 2	3, 2	3, 2	3, 2
Case 20	0, 1	1, 1	0, 1	1, 0	1, 1
Case 21	1, 0	1, 1	1, 1	1, 0	1, 1
Case 22	1, 0	1, 0	1, 0	1, 0	1, 0
Case 25	1, 1	1, 1	1, 2	1, 2	1, 2
Case 26	0, 1	0, 1	0, 0	0, 1	0, 1
Case 27	0, 0	1, 0	1, 1	1, 1	1, 1
Case 28	1, 3	1, 2	1, 2	1, 3	1, 3
Case 29	0, 0	0, 1	0, 1	0, 0	0, 1
Case 30	0, 0	0, 0	0, 1	0, 1	0, 1
Case 31	0, 0	0, 0	0, 0	0, 0	0, 0
Case 32	1, 1	1, 2	1, 2	1, 1	1, 2
Case 33	1, 1	1, 1	1, 2	1, 1	1, 2
Case 34	0, 2	2, 3	2, 3	1, 3	2, 3
Case 35	0, 0	0, 0	1, 0	1, 1	1, 0
Case 36	0, 3	0, 3	1, 2	1, 1	1, 2
Case 37	0, 3	1, 3	0, 3	1, 1	2, 3
Case 38	1, 1	1, 2	1, 1	1, 2	1, 2
Case 39	1, 1	1, 0	1, 0	1, 1	1, 1
Case 40	1, 3	1, 3	1, 2	1, 2	1, 3
Case 41	1, 0	0, 0	0, 0	1, 0	1, 0

Table 1: Grade of TDP-43 Pathology assigned to each area of interest for neurons (N), glial cells (G) and mean for each ALS sample [Twiddy, 2021].

Grade of TDP-43 Pathology	Number of TDP-43 positive cells in an area of interest
0 (Control or None)	0
1 (Mild)	1-5
2 (Moderate)	6-14
3 (Severe)	+15

Table 2: Number of TDP-43 positive cells in an area of interest related to each grade of TDP-43 pathology [Twiddy, 2021].

Images belonging to sections 1 and 2 of each sample, ten images per section, were used in this study. These were processed to eliminate the signal corresponding to pTDP-43$^{409/410}$ and keep only the DAPI nuclear signal and Apt-1 signal based on the results from the University of

Edinburgh indicating the superior ability of Apt-1 versus pTDP-43$^{409/410}$ to distinguish ALS cases and controls.

For the initial development of the Deep Learning models, these samples were used after a data augmentation process to balance the number of healthy samples with those with ALS, using non-destructive geometrical transformations. However, the set of images prior to the amplification process will be used for the statistical analysis in this work since we seek to understand the original samples and identify possible differences between classes.

2.2 Dataset and data collection

A dataset was generated in Microsoft Excel, including information related to each image (See Figure 1). First, data related to the patient (ID, sex and age) extracted from a table provided by the University of Edinburgh were included. Then, the sections (1-2) and the images (1-10) were specified in each case.

Figure 1: Dataset structure: (A) ID. (B) Sex. (C) Age. (D) Section (1-2). (E) Image number (1-10). Classifications of (F) TDP-43 Pathology (Yes/No), (G) Grades of TDP-43 pathology (0-Control, 0-No, 1-Mild, 2-Moderate, 3-Severe) and (H) ALS. Number of (I) red and (J) pink pixels and (K, L) ratios. (M) Total amount of TDP-43. Proportion of (N) cytoplasmic and (O) nuclear TDP-43. (P) Ratio of proportions.

Subsequently, three classification problems were considered. The first one referred to the presence or absence of TDP-43 pathology according to Table 1, where each section with a grade of 0 and every control was assigned with the absence of pathology. The second classification was related to the grade of TDP-43 pathology based on the provided clinical grading. The grade establishes five groups (0-Control, 0-No, 1-Mild, 2-Moderate, 3-Severe). The last classification was based on the presence or absence of ALS, so the samples in Table 1 are considered as pathological cases, and the rest are controls.

Next, the number of red and pink pixels in each sample was calculated and included. Red pixels indicate the signal emitted by the Apt-1, meaning the presence of TDP-43. Pink pixels are due to the combination of the blue nuclear signal (DAPI) and the red signal (Apt-1). Therefore, red pixels represent cytoplasmic TDP-43, while pink pixels indicate nuclear TDP-43. A Python self-made code generated these numerical data from the images to understand the samples' relation to the existing literature on TDP-43 proteinopathy. Ratios between these two variables and their sum, which corresponds to the total amount of TDP-43, were then calculated. In addition, the proportions of cytoplasmic (number of red pixels between the sum of red and pink pixels) and nuclear (number of pink pixels between the sum of red and pink pixels) TDP-43 were calculated, as well as the ratio between these variables.

2.3 Initial analysis

Statistical analysis was carried out using XLSTAT Software and Microsoft Excel. First, descriptive analysis was carried out based on the three classifications of interest (TDP-43 Pathology, TDP-43 Pathology Grades and ALS) in order to analyse pixels' frequencies and assess the most balanced and viable classification sets for training the deep learning models. The number of cases and controls is a key concern aspect for deep learning classification problems. The number of samples per class should be balanced to avoid possible generalisation problems that would cause the model to classify all the samples in the majoritarian group [Buda et al., 2018].

Subsequently, the samples for the three classification problems were analysed based on three variables: (1) the total amount of TDP-43, (2) the cytoplasmic proportion, and (3) the nuclear proportion. Sampling distribution was visually compared between groups of each classification scenario and for each variable using scattergrams.

The existence of statistically significant differences between groups for the three variables was evaluated with t-student tests for the classifications of the presence or absence of TDP-43 pathology

and ALS and with ANOVA for the classification by grades of TDP-43 pathology. Additionally, all grades were compared with each other using t-student tests to understand the established differences.

2.4 Dataset curation

Given the inconclusive results obtained in the initial statistical analysis (see Section 3.1), it was decided to clean the dataset and repeat the analysis in order to clarify and/or confirm the different conclusions. As a criterion to reduce the dataset, the sampling distribution based on the proportion of cytoplasmic TDP-43 was taken into account, in which a greater grouping of data was observed around the values 0 and 1 due to different possible reasons (See Figures 3, 5 and 12).

Considering the methodology employed for image acquisition, instances of a value of 1 for a particular variable may potentially be attributed to temporal or material oversights that could impede the concurrent binding of DAPI and Apt-1 to identical regions, resulting in reduced colocalisation. Conversely, instances of a value of 0 might be ascribed to challenges in binding Apt-1, stemming from lapses in aptamer preservation or during the execution of the experimental protocol.

Therefore, the 297 samples with these characteristics (values around 0 and 1) were deleted and annotated (See Table 3). We hypothesised that most of these samples were problematic due to accidental isolated errors, although some systematic errors could also happened. In total, 70% of the deleted samples had a value of 1 for the cytoplasmic ratio, suggesting a potential problem in the Apt-1 and DAPI colocalisation protocol.

2.5 Final analysis

The analysis performed for the initial dataset was repeated for the simplified dataset. Therefore, frequencies were analysed to assess the most balanced classification sample sets. In addition, scattergrams were used to compare sampling distribution between groups of each classification for three variables: total amount of TDP-43, cytoplasmic proportion, and nuclear proportion.

The existence of statistically significant differences between groups with TDP-43 pathology and ALS diagnosis for the three variables were studied using t-student tests. Later, amounts of red and pink pixels (cytoplasmic and nuclear TDP-43, respectively) were evaluated. Since the groups do not follow a normal distribution, verified by Saphiro-Wilk tests for all groups, differences for the five variables were evaluated with Mann-Whitney tests. Finally, statistically significant differences between grades of TDP-43 pathology were also evaluated with the ANOVA test and independently with the t-student test.

2.6 Datasets' testing

Both datasets were compared in order to understand the sampling distribution between patients. For this purpose, box plots were created for each dataset for five variables: the number of red and pink pixels (cytoplasmic and nuclear TDP-43, respectively), the total amount of TDP-43 (red and pink pixels sum) and cytoplasmic and nuclear proportions of TDP-43.

Two datasets distinguishing controls and TDP-43 pathological samples were created according to both the initial and the simplified datasets. Samples were grouped for training (80%) and testing (20%). Both datasets were tested for training the deep learning model to determine if the model could better differentiate the classes and if the guidance given was useful. Finally, the mean and standard deviation for each of the five variables were calculated for each patient in the reduced dataset.

We were interested in analysing subject inter-variability in the dataset. For such a task, we implement a hierarchical clustering approach, grouping all samples per individual from the reduced dataset. The resulting dendrogram would improve the understanding of the dataset and highlight issues in the distribution of each class.

Deleted Samples				
Patient	Deleted Samples	ALS	TDP-43	Grade
Case 1	1.1; 1.2; 1.3; 1.4; 1.5; 1.6; 1.7; 1.8; 1.9; 1.10; 2.1: 2.2; 2.4; 2.5	Yes	Yes	1-Mild
Case 2	1.1; 1.2; 1.3; 1.4; 1.5; 1.6; 1.7; 1.8; 1.9; 1.10; 2.1; 2.2; 2.3; 2.4; 2.5; 2.6; 2.7; 2.8; 2.9; 2.10	Yes	Yes	2-Moderate
Case 3	1.2; 1.5; 2.5	No	No	0-Control
Case 4	1.8	Yes	Yes	2-Moderate
Case 5	1.1; 1.2; 1.3; 1.10; 2.4; 2.5	No	No	0-Control
Case 6	1.6; 1.8; 2.2; 2.4; 2.6; 2.7; 2.8; 2.10	No	No	0-Control
Case 7	1.1; 1.2; 1.3; 1.4; 1.5; 1.6; 1.7; 1.8; 1.9; 1.10; 2.1; 2.2; 2.3; 2.4; 2.5; 2.6; 2.7; 2.8; 2.9; 2.10	Yes	Yes	1-Mild
Case 8	1.1; 1.7; 2.8	No	No	0-Control
Case 9	2.1	No	No	0-Control
Case 10	1.1; 1.2; 1.3; 1.4; 1.5; 1.6; 1.7; 1.8; 1.9; 1.10; 2.1; 2.2; 2.3; 2.4; 2.5; 2.6; 2.7; 2.8; 2.9; 2.10	Yes	Yes	1-Mild
Case 11	1.1; 1.4; 1.5; 1.6; 1.7; 1.8; 1.9; 1.10; 1.11; 2.1; 2.2; 2.3; 2.4; 2.5; 2.6; 2.7; 2.8; 2.9; 2.10	Yes	No	0-None
Case 12	1.1; 1.2; 1.3; 1.4; 1.5; 1.6; 2.3; 2.5; 2.8; 2.10	No	No	0-Control
Case 13	1.7; 2.2; 2.5	No	No	0-Control
Case 14	1.8	Yes	Yes	3-Severe
Case 15	1.5	Yes	Yes	1-Mild
Case 16	1.1; 1.2; 1.3; 1.4; 1.5; 1.6; 1.7; 1.8; 1.9; 1.10; 2.1; 2.2; 2.3; 2.4; 2.5; 2.6; 2.7; 2.8; 2.9; 2.10	Yes	Yes	1-Mild
Case 17	1.3; 1.5; 1.6; 1.7; 1.8; 1.9; 1.10; 2.1; 2.3; 2.4; 2.5; 2.7; 2.8; 2.10	Yes	No	1-Mild
Case 18	1.2; 1.4; 1.5; 1.6; 1.10; 2.6; 2.7; 2.8; 2.9; 2.10	Yes	Yes	3-Severe
Case 19	2.2; 2.10	No	No	0-Control
Case 20	1.1; 1.3; 1.4; 1.6; 1.9	Yes	Yes	1-Mild
Case 21	1.8; 2.6; 2.8; 2.9; 2.10	Yes	Yes	1-Mild
Case 22	1.1; 1.3; 1.4; 1.7; 2.3	Yes	Yes	1-Mild
Case 23	2.1; 2.2; 2.4; 2.5; 2.7; 2.8; 2.10	No	No	0-Control
Case 24	2.1; 2.4; 2.8; 2.10	No	No	0-Control
Case 25	1.1; 1.2; 1.3; 1.4; 1.5; 1.6; 1.7; 1.8; 1.9; 1.10; 2.1; 2.2; 2.3; 2.4; 2.5; 2.6; 2.7; 2.8; 2.9	Yes	Yes	2-Moderate
Case 26	2.1; 2.2; 2.4; 2.5; 2.6; 2.7, 2.8	Yes	Yes	1-Mild
Case 27	1.1; 1.3; 1.4; 1.5; 1.6; 1.7; 1.8; 1.9; 2.1; 2.2; 2.3; 2.5; 2.7; 2.8	Yes	No	1-Mild
Case 28	1.1; 1.2; 1.3; 1.4; 1.5; 1.6; 1.7; 1.8; 1.9; 1.10; 2.1; 2.2; 2.5; 2.6; 2.8; 2.9	Yes	Yes	3-Severe
Case 29	1.2; 1.4; 1.5; 1.6; 1.7; 1.9; 2.8	Yes	No	1-Mild
Case 30	1.1; 1.2; 1.3; 1.4; 1.7; 1.10; 2.1; 2.2; 2.3; 2.4; 2.6; 2.7; 2.9; 2.10	Yes	No	1-Mild
Case 31	1.2; 1.3; 1.4; 1.5; 1.9; 1.10; 2.1; 2.2; 2.3; 2.4; 2.5; 2.6; 2.7; 2.8; 2.9; 2.10	Yes	No	0-None

Table 3: Deleted samples in the dataset curation process per subject and their characteristics (ALS, TDP-43 and grade).

3 Results

3.1 Initial analysis

First, a descriptive analysis of the initial dataset containing all the available samples was performed. Groups' frequencies were extracted from this analysis (See Figure 2). According to the classification

by TDP-43 proteinopathy (See Figure 2A), 479 samples presented pathology (53%) compared to 420 healthy samples (47%). In the ALS classification (See Figure 2B), 609 samples (68%) showed pathology compared to 290 healthy samples (32%). For the classification by grades of TDP-43 (See Figure 2C) pathology, 290 controls (32%), 41 cases without protein pathology (5%), 309 mild cases (34%), 139 moderate cases (15%) and 120 severe cases (13%) were registered.

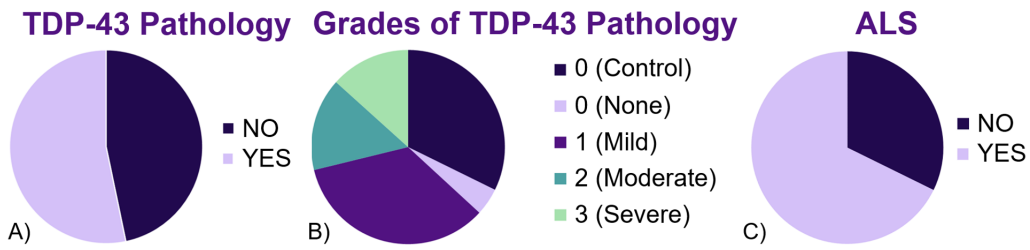

Figure 2: Frequency graphs of the complete dataset for the classifications of (A) TDP-43 Pathology, (B) TDP-43 Pathology Grades, and (C) ALS.

In the same way, it was possible to compare the sampling distribution for the cytoplasmic and nuclear proportion variables with the use of scattergrams for the classifications by TDP-43 pathology, see Figure 3, depicting grades of TDP-43 pathology in Figure 4, and distinguishing ALS in Figure 5.

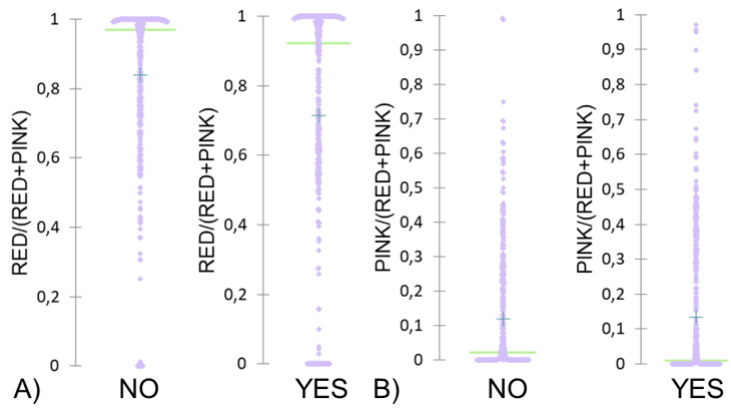

Figure 3: Scattergrams for the classification by TDP-43 pathology from the complete dataset for the variables of (A) cytoplasmic and (B) nuclear proportion.

T-student tests were performed with a confidence interval of 95%, so that α=0.05 and any p-value less than α is considered statistically significant, this significance being greater as the p-value decreases in comparison to α. For the classification by TDP-43 pathology (See Figure 6), significant differences were detected for the total amount of TDP-43 (p=0.003), with pathological samples having more protein. Significant differences were also detected for the cytoplasmic proportion (p<0.0001), which turned out to be lower for the pathological samples but not for the nuclear proportion (p=0.266).

On the other hand, for the classification by ALS (See Figure 7), no significant differences were detected for the total amount of TDP-43 (p=0.216). However, significant differences were found for the cytoplasmic (p<0.0001) and nuclear (p= 0.046) proportions, as pathological samples have smaller values for both variables.

Afterwards, ANOVA was used with a confidence interval of 95% for comparing the different protein pathology grades. It was determined that there were significant differences between groups for the three variables (total amount of TDP-43 and the cytoplasmic and nuclear ratios) as p<0.0001. In addition, it was possible to graphically compare the differences of the means for the three variables (See Figure 8).

Additionally, grades of TDP-43 pathology were compared independently with t-student tests. Results were grouped in tables, assessing the level of significance of each comparison (See Figure 9).

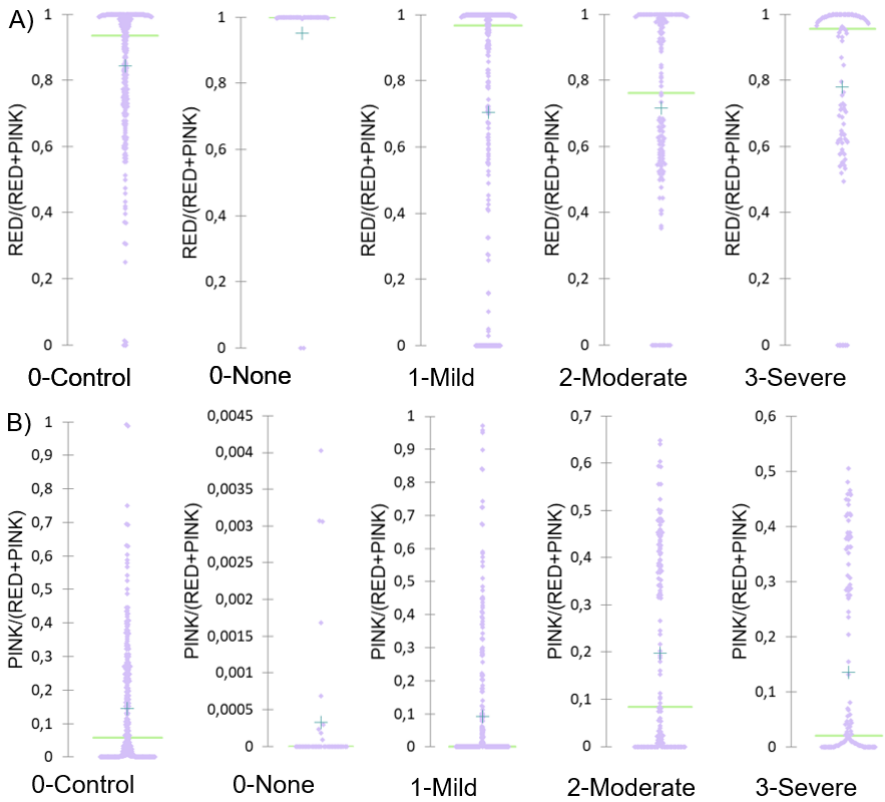

Figure 4: Scattergrams for the classification by grades of TDP-43 pathology from the complete dataset for the variables of (A) cytoplasmic and (B) nuclear proportion.

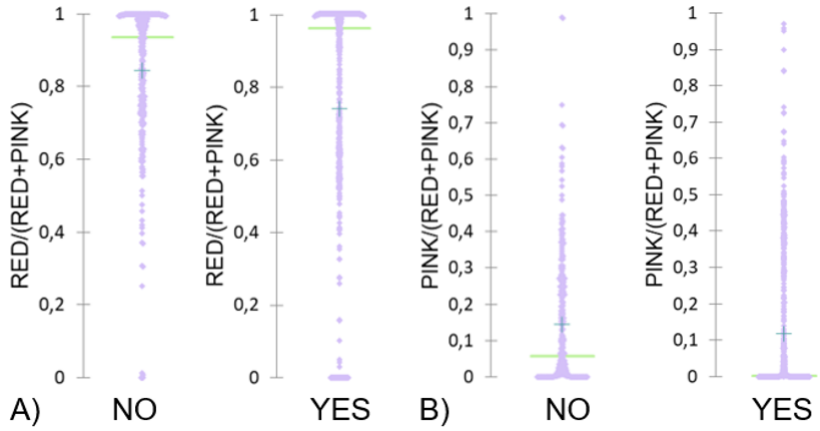

Figure 5: Scattergrams for the classification by ALS from the complete dataset for the variables of (A) cytoplasmic and (B) nuclear proportion.

3.2 Final analysis

The dataset was cleaned up based on the previous results, and the same procedures were repeated on this new reduced dataset. Initially, multiple descriptive analyses were performed, from which groups' frequencies were extracted (See Figure 10). According to the classification by TDP-43 pathology, 303 samples presented pathology (50.33%) compared to the 299 healthy cases (49.67%). In the ALS classification, 359 samples (60%) showed pathology compared to 243 healthy ones (40%), for the classification by grades of TDP-43 pathology, 243 controls (40%), 9 cases without protein pathology (1%), 158 mild cases (26%), 99 moderate cases (16%) and 93 severe cases (15%)

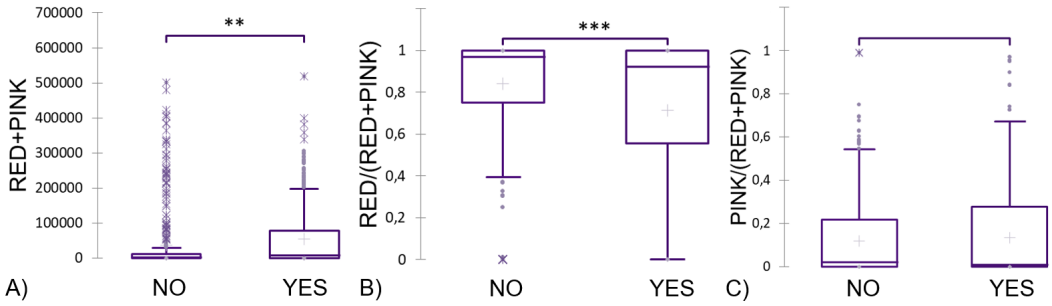

Figure 6: Comparative plots of means for the (A) total amount of TDP-43 and the (B) cytoplasmic and (C) nuclear proportion in the classification of TDP-43 pathology from the complete dataset. (*) Significant (p<0.05), (**) quite significant (p<0.01) or (***) very significant (p<0.001) difference at a statistical level verified with t-student (α= 0.05).

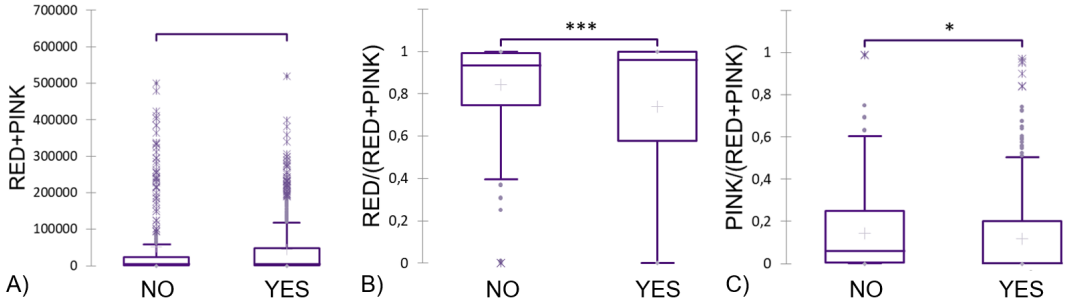

Figure 7: Comparative plots of means for the (A) total amount of TDP-43 and the (B) cytoplasmic and (C) nuclear proportions in the ALS classification of the complete dataset. (*) Significant (p<0.05), (**) quite significant (p<0.01) or (***) very significant (p<0.001) difference at a statistical level verified with t-student (α= 0.05).

were registered. In the same way, the sampling distribution for the cytoplasmic and nuclear proportion variables was compared for the classifications by TDP-43 pathology (See Figure 11), by grades of TDP-43 pathology (See Figure 12) and by ALS (See Figure 13).

Afterwards, differences between groups were verified using t-student tests with α=0.05 so that any p-value lower than α was considered statistically significant. The significance level is greater as the p-value decreases with respect to α. For the classification by ALS (See Figure 14), no significant differences were detected neither for the total amount of TDP-43 (p=0.137) nor for the cytoplasmic (p=0.107) or nuclear (p=0.107) proportions. On the other hand, for the classification by TDP-43 pathology (See Figure 15), significant differences were detected for the total amount of TDP-43 (p<0.0001), the cytoplasmic proportion (p<0.009), and the nuclear proportion (p=0.009). In this case, pathological samples showed a lower cytoplasmic TDP-43 proportion and higher values for the nuclear proportion and total TDP-43.

In order to better understand these results for the classification by TDP-43 pathology, significant differences for both cytoplasmic (p=0,001) and nuclear (p<0,0001) concentrations of TDP-43 (red and pink pixels, respectively) were found according to t-student tests (See Figure 16), having pathological samples higher values for both variables. For this dataset, the outcomes from ANOVA determined that there were significant differences between groups for the three variables since p<0.0001 for all three variables. In addition, it was possible to graphically compare the differences of the means for the three variables (See Figure 17).

The different grades of TDP-43 pathology were compared independently using t-student tests. Results were grouped in tables and analysed by assessing the levels of significance of each classification problem (See Figure 18).

Finally, as initial analysis showed that samples did not follow a normal distribution, verified by Saphiro-Wilk tests for all groups, differences for the classification by TDP-43 pathology were verified with Mann-Whitney tests (See Figure 19). Pathological samples revealed higher values for the

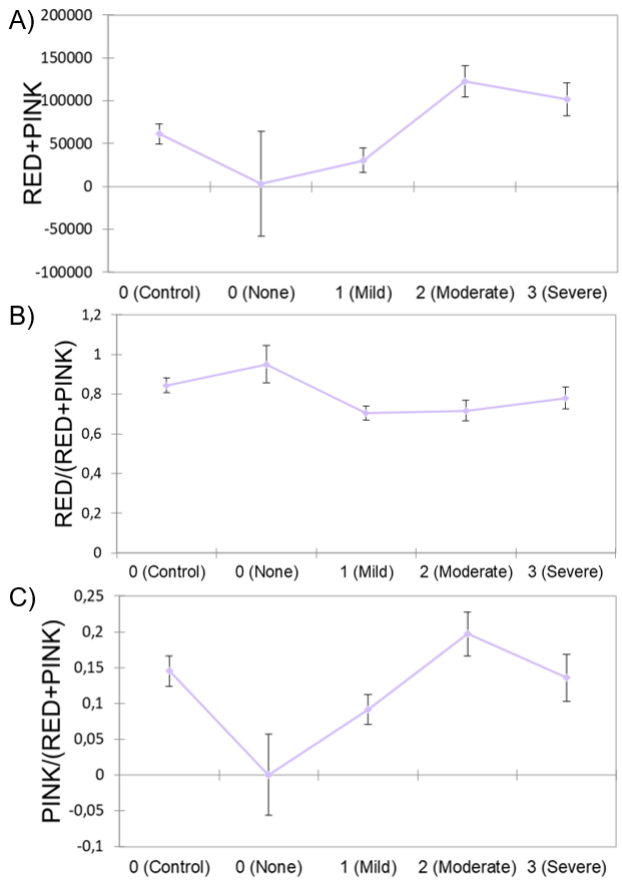

Figure 8: Comparative graphs of the means of the different grades of TDP-43 pathology from the complete dataset for the (A) total amount of TDP-43 and the (B) cytoplasmic and (C) nuclear ratio.

A)

p-value	0-Control	0-None	1-Mild	2-Moderate	3-Severe
0-Control	1	0,003**	<0,0001***	0,000***	0,014*
0-None	0,003**	1	0,087	<0,0001***	<0,0001***
1-Mild	<0,0001***	0,087	1	<0,0001***	<0,0001***
2-Moderate	0,000***	<0,0001***	<0,0001***	1	0,433
3-Severe	0,014*	<0,0001***	<0,0001***	0,433	1

B)

p-value	0-Control	0-None	1-Mild	2-Moderate	3-Severe
0-Control	1	0,002**	<0,0001***	<0,0001***	0,011*
0-None	0,002**	1	0,000***	<0,0001***	0,001**
1-Mild	<0,0001***	0,000***	1	0,753	0,060
2-Moderate	<0,0001***	<0,0001***	0,753	1	0,081
3-Severe	0,011*	0,001**	0,060	0,081	1

C)

p-value*	0-Control	0-None	1-Mild	2-Moderate	3-Severe
0-Control	1	<0,0001***	0,000***	0,008**	0,630
0-None	<0,0001***	1	0,003**	<0,0001***	<0,0001***
1-Mild	0,000***	0,003**	1	<0,0001***	0,030*
2-Moderate	0,008**	<0,0001***	<0,0001**	1	0,011*
3-Severe	0,630	<0,0001***	0,030*	0,011*	1

Figure 9: P-values comparing the means of the different grades of TDP-43 pathology from the complete dataset for the (A) total amount of TDP-43 and the (B) cytoplasmic and (C) nuclear ratio. (*) Significant ($p<0.05$), (**) quite significant ($p<0.01$) or (***) highly significant ($p<0.001$) difference at a statistical level verified with t-student ($\alpha = 0.05$).

total amount of protein (p<0,0001), the nuclear (p=0,038) proportion and nuclear (p<0,0001) and cytoplasmic (p<0,0001) amounts of TDP-43, despite a lower cytoplasmic proportion (p=0,038).

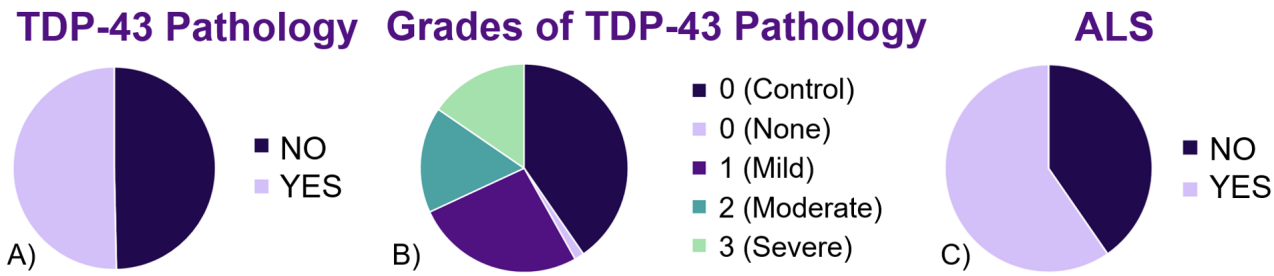

Figure 10: Frequency graphs of the simplified dataset for the classifications of (A) TDP-43 Pathology, (B) TDP-43 Pathology Grades, and (C) ALS.

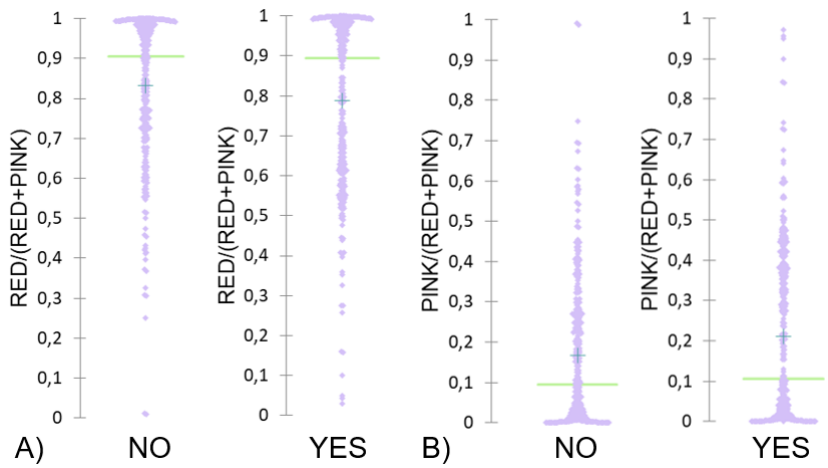

Figure 11: Scattergrams for the classification by TDP-43 pathology from the simplified dataset for the variables of (A) cytoplasmic and (B) nuclear proportion.

3.3 Datasets' testing

The distribution of each patient's values for each variable was compared between the initial and the reduced datasets (See Figure 20). Afterwards, a ResNet-50 deep learning model [He et al., 2016] was trained and tested 15 times with the two datasets. Training information is detailed in Table 4. Violin plots were created for accuracy, specificity, sensitivity, Matthew's correlation coefficient [Matthews, 1975] and F1-score for both training and testing with both datasets (See Figure 21) by training and testing the model 15 times to assess robustness and variability of the results.

Table 4: Training Details for ResNet-50

Training Strategy	Early Stopping
Pretraining	None
Optimizer	Adam
Learning Rate	0.001
Pooling	Average Pooling
Image Size	224×224 pixels
Scale	Automatic

The dendrogram resulting from the hierarchical clustering of the reduced dataset was linked to the subject information to improve the understanding (See Figure 22). Even if the dendrogram did not show the full distinction between groups, there are some trends that can be mentioned. The dendrogram showed two main clusters, separating 30 (on the left) from the other 11 subjects (on the right). The first group formed a three-subcluster division that was able to discriminate a

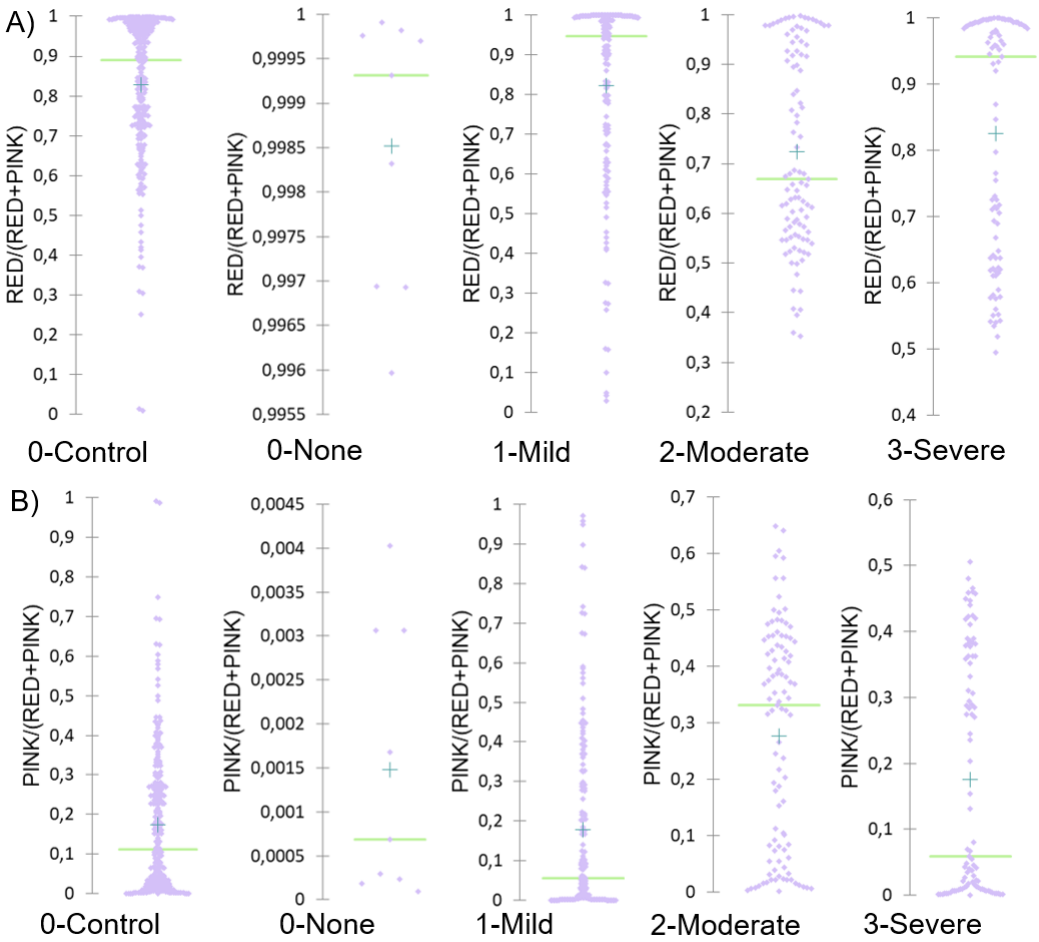

Figure 12: Scattergrams for the classification by grades of TDP-43 pathology from the simplified dataset for the variables of (A) cytoplasmic and (B) nuclear proportion.

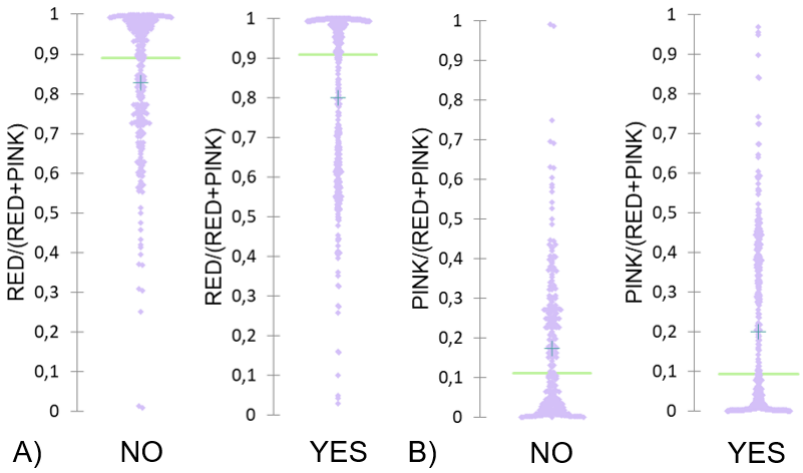

Figure 13: Scattergrams for the classification by ALS from the simplified dataset for the variables of (A) cytoplasmic and (B) nuclear proportion.

first group with ALS and no TDP-43 proteinopathy, a second group with ALS and TDP-43 and a third group with no ALS and no proteinopathy. The second group on the right showed a much clearer structure, with a subgroup with no ALS/TDP-43 and a group with severe/moderate levels

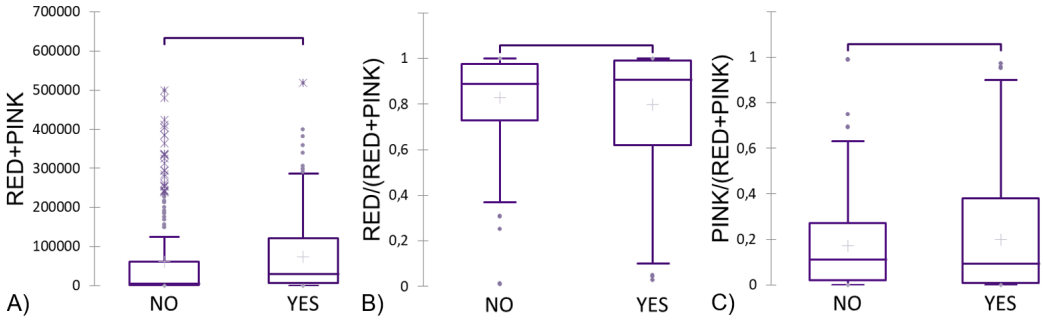

Figure 14: Comparative plots of means for the (A) total amount of TDP-43 and the (B) cytoplasmic and (C) nuclear proportions in the ALS classification of the simplified dataset. (*) Significant (p<0.05), (**) quite significant (p<0.01) or (***) very significant (p<0.001) difference at a statistical level verified with t-student (α= 0.05).

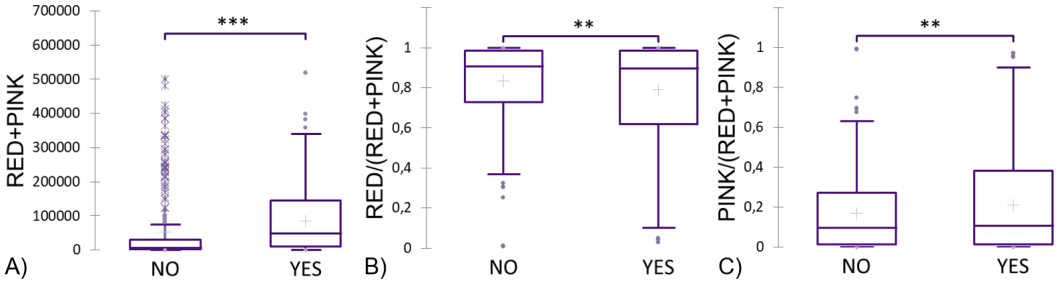

Figure 15: Comparative plots of means for the (A) total amount of TDP-43 and the (B) cytoplasmic and (C) nuclear proportions in the classification of TDP-43 pathology from the simplified dataset. (*) Significant (p<0.05), (**) quite significant (p<0.01) or (***) very significant (p<0.001) difference at a statistical level verified with t-student (α= 0.05).

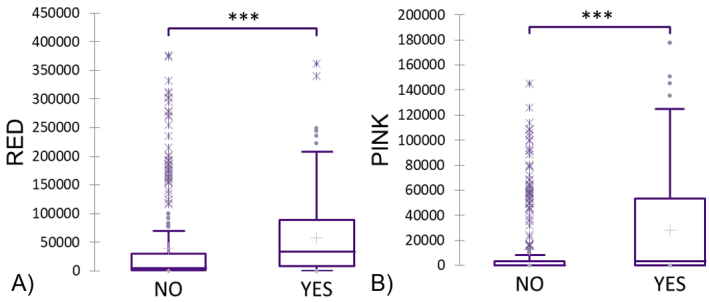

Figure 16: Comparative plots of means for the amount of (A) cytoplasmic and (B) nuclear amount of TDP-43 in the classification of TDP-43 pathology from the simplified dataset. (*) Significant (p<0.05), (**) quite significant (p<0.01) or (***) very significant (p<0.001) difference at a statistical level verified with t-student (α= 0.05).

of the disease.

4 Discussion

The initial frequency analysis of sample types demonstrated a strong lack of balance between cases and controls for both ALS classification and pathology classification of TDP-43 (See Figures 2). This fact justifies the need for a process of oversampling the underrepresented group prior to training the models to balance the sample sets, followed by a general process of data augmentation to increase the total number of images and reach a higher generality. In addition, a clear correlation is observed between the samples with grade 0 of TDP-43 pathology and the control subjects of

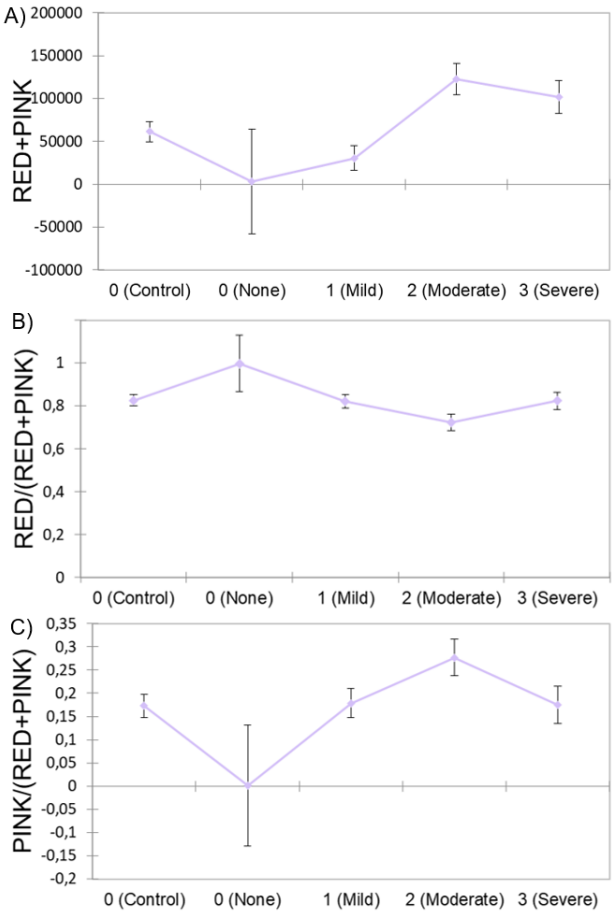

Figure 17: Comparative graphs of the means of the different grades of TDP-43 pathology from the simplified dataset for the (A) total amount of TDP-43 and the (B) cytoplasmic and (C) nuclear proportions.

A)

p-value	0-Control	0-None	1-Mild	2-Moderate	3-Severe
0-Control	1	0,118	0,001**	<0,0001***	0,002**
0-None	0,118	1	0,195	<0,0001***	0,004**
1-Mild	0,001**	0,195	1	<0,0001***	<0,0001***
2-Moderate	<0,0001***	<0,0001***	<0,0001***	1	0,118
3-Severe	0,002**	<0,004**	<0,0001***	0,118	1

B)

p-value	0-Control	0-None	1-Mild	2-Moderate	3-Severe
0-Control	1	0,005**	0,784	<0,0001***	0,925
0-None	0,005**	1	0,029*	<0,0001***	0,005**
1-Mild	0,784	0,029*	1	0,001**	0,897
2-Moderate	<0,0001***	<0,0001***	0,001**	1	0,000***
3-Severe	0,925	0,005**	0,897	0,000***	1

C)

p-value	0-Control	0-None	1-Mild	2-Moderate	3-Severe
0-Control	1	0,005**	0,784	<0,0001***	0,925
0-None	0,005**	1	0,029*	<0,0001***	<0,005**
1-Mild	0,784	0,029*	1	<0,001**	0,897
2-Moderate	<0,0001***	<0,0001***	<0,001**	1	0,000***
3-Severe	0,925	<0,005**	0,897	0,000***	1

Figure 18: P-values comparing the means of the different grades of TDP-43 pathology from the simplified dataset for the (A) total amount of TDP-43 and the (B) cytoplasmic and (C) nuclear proportions. (*) Significant ($p<0.05$), (**) quite significant ($p<0.01$) or (***) very significant ($p<0.001$) difference at a statistical level verified with t-student($\alpha = 0.05$).

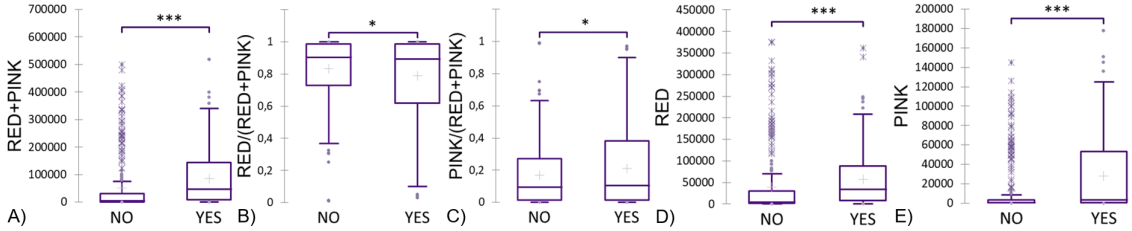

Figure 19: Comparative plots of means for the (A) total amount of TDP-43, (B) cytoplasmic and (C) nuclear proportions, and (D) cytoplasmic and (E) nuclear amount of protein in the classification of TDP-43 pathology from the simplified dataset. (*) Significant ($p<0.05$), (**) quite significant ($p<0.01$) or (***) very significant ($p<0.001$) difference at a statistical level verified with Mann-Whitney test ($\alpha= 0.05$).

the ALS, noting that the final clinical diagnosis assigned is independent of the individual grades of each image sample. However, these controls do not correlate with controls for the classification by pathology of TDP-43 as this classification has taken into account each section's grade according to Table 1. It is more objective to consider each section's grade separately and, therefore, the existent intra-patient variability. When assigning a unique classification label to all samples from a patient, a higher risk of generalisation can occur.

At this point, certain limitations related to the sample set must be considered. First, the initial sample size is limited, as often happens when dealing with low-incidence diseases. Secondly, samples of ALS controls were not clinically assessed for aggregated TDP-43, which can lead to certain biases by assuming that the controls do not have protein pathology. It is known that this proteinopathy may exist in the absence of ALS as it is compatible with other pathologies. On the other hand, the assignment of discrete grades (non, mild, moderate, and severe) is problematic. The initial annotation process associated the number of chromogen-positive cells to the assigned grade (See Table 2). However, the grades are not equal (the number of cells does not increase proportionally), and it is not explained why those grades are established with those assignments [Twiddy, 2021]. Furthermore, the number of chromogen-positive cells was not provided or published, and hence, we could not use it in this study to establish a more equitable association with the grades.

Samples with TDP-43 pathology were shown to have a superior total TDP-43 than controls but a lower proportion of protein in the cytoplasm (See Figure 6), suggesting that there is indeed an increased proportion of protein located in the nucleus although it is not significant, so results are not entirely conclusive. Considering these relevant results would be contradictory since TDP-43 proteinopathy has been characterised by a pattern of cytoplasmic aggregation and nuclear depletion, and the results obtained describe an opposite pattern.

For ALS classification, lower proportions of cytoplasmic and nuclear TDP-43 are observed in pathological samples without significantly altering the total amount of protein (See Figure 14). It could be said that these results are not conclusive as they are significant when considering only part of the information, but a minimum level of significance is not reached when considering all the information.

On the other hand, for the classification by grades of TDP-43 pathology, significant differences were determined for the three variables (total amount of TDP-43, cytoplasmic proportion, and nuclear proportion) according to the ANOVA results. If we compare the means of the two groups with grade 0, it is observed that controls have more total protein and higher nuclear proportion, while ALS samples have less total protein but a higher cytoplasmic proportion and a lower nuclear ratio (See Figure 17). When comparing the grades of pathology, an increase in the three variables is observed in grade 2 compared to grade 1, while these decrease in grade 3 with the exception of the cytoplasmic proportion, which keeps rising (See Figure 17). Therefore, differences will depend on which variable we measure and which grades we compare, varying the significance of these differences in each case (See Figure 18).

Due to the lack of conclusive results to guide the deep learning model, it was proposed to reduce the dataset by removing problematic samples, being also aware that this action would also harden the generalisation capability of the deep learning models. Sampling distribution of the complete dataset was analysed for the nuclear and cytoplasmic TDP-43 proportions variables for the three

Figure 20: Comparison of (A) the initial dataset and (B) the reduced dataset for five variables: the amount of cytoplasmic TDP-43 (red pixels), amount of nuclear TDP-43 (pink pixels), the total amount of TDP-43 (red+pink) and cytoplasmic and nuclear ratios (red between red and pink sum or pink between red and pink sum respectively).

classifications (See Figures 3, 4 and 5), and a greater accumulation of samples was observed at values 0 and 1. Since at both physiological and pathological levels, there must be protein in both the nucleus and cytoplasm. It can be interpreted that samples that apparently do not show protein or only have it in one location are due to errors in materials or obtaining protocols and, therefore, do not represent reality. Thanks to this idea, the dataset was reduced by deleting these samples.

On the reduced dataset, an almost perfect balance was observed between samples with TDP-43 pathology and their controls, but not for ALS classification (See Figure 10). The balanced dataset aimed at the classification of TDP-43 pathology is more optimal for training the deep learning models. As for the initial analysis, controls for the classifications by grades and by

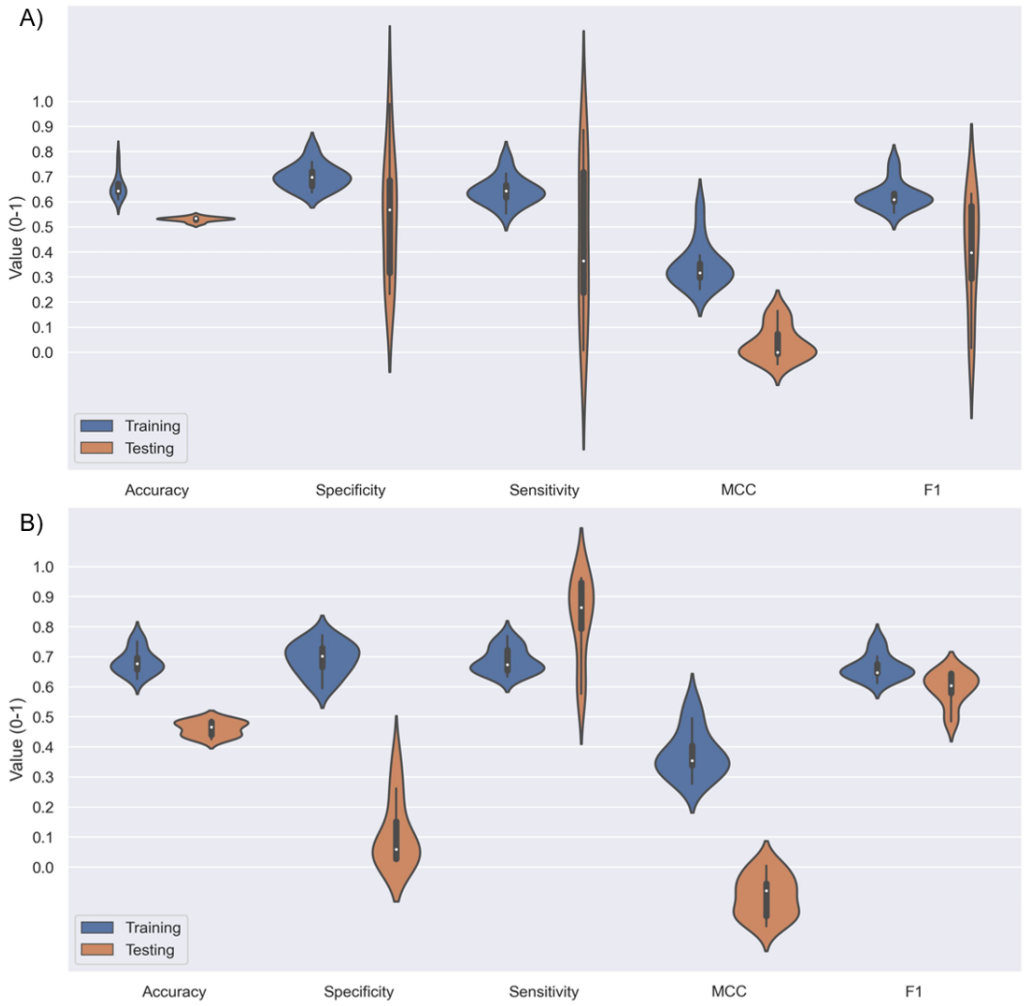

Figure 21: Violin plots for accuracy, specificity, sensitivity, Matthew's correlation coefficient and F1 score resulting from training and testing the CNN model 15 times with (A) the complete and (B) the reduced datasets.

ALS are correlated because they share criteria, which is not the case for the classification by protein pathology, so the limitations of the samples and the differences in criteria must always be considered.

When visualising the sampling distribution of the images from this reduced dataset for the cytoplasmic and nuclear proportions (See Figures 11, 12 and 13), a greater TDP-43 cytoplasmic proportion is still observed in all groups. Samples tend to group around values very close to 0 and 1, but since they are real values calculated based on the pixels, further reducing the dataset would be incorrect since some of the valid information would be lost.

For the classification by TDP-43 pathology of this reduced dataset, significant differences were established for the three variables (See Figure 15) as controls had a higher proportion of cytoplasmic TDP-43 and, therefore, a lower proportion of nuclear protein, as well as less protein. This suggests that pathological samples generally have an excess of protein, so even with a lower cytoplasmic proportion, there would still generally be a greater amount of protein in the cytoplasm. Differences for the amounts of cytoplasmic and nuclear TDP-43 (red and pink pixels respectively) were analysed again with t-student tests (See Figure 16) in order to verify that both amounts of TDP-43 were higher for pathological samples.

Samples with TDP-43 pathology had a general excess of protein with a significant increase in both cytoplasmic and nuclear protein concentration. This characteristic corroborates the literature in terms of the cytoplasmic TDP-43 aggregation pattern and the pathological profile of TDP-

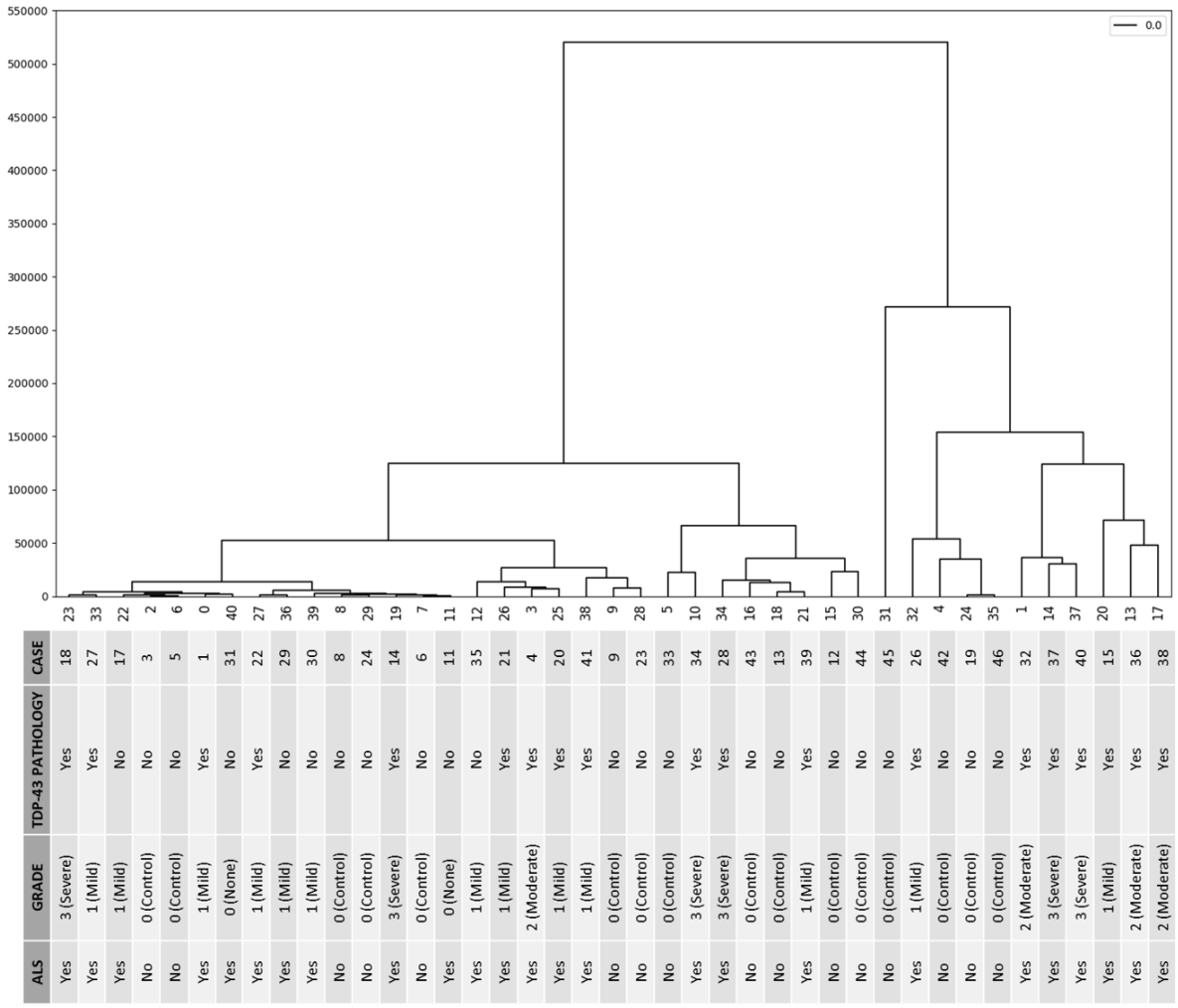

Figure 22: Hierarchical clustering graph between patients in the reduced dataset.

43 proteinopathy as the cytoplasmic levels of protein increase. However, cytoplasmic TDP-43 proportion decreases in pathological samples [Masrori and Van Damme, 2020, Neumann et al., 2006]. Nevertheless, it is impossible to report nuclear depletion as both the amount and proportion of nuclear TDP-43 increase in pathological samples.

As samples do not follow a normal distribution, which can be seen in the scattergrams and was verified by Saphiro-Wilk tests for all groups, these differences were verified with Mann-Whitney tests (See Figure 19). TDP-43 pathological samples showed a generalised protein excess with a significant increase of both cytoplasmic and nuclear amounts of protein, verifying the pattern of cytoplasmic aggregation [Masrori and Van Damme, 2020, Neumann et al., 2006] with no nuclear depletion although there is no increase of cytoplasmic proportion. Results for the classification by ALS (See Figure 14) determined that there are no differences between controls and pathological samples for any of the variables analysed, which would partially explain the results of the initial results of the deep learning model based on this dataset.

Considering these previous points, it could be said that training the model based on the classification by pathology of TDP-43 could be more promising. In addition, the possibility of including some of the analysed variables in the current or new model could be studied in order to guide the model to the protein locations in the images.

Finally, for the classification by grades of TDP-43, significant differences were detected for the three variables. As for the complete dataset, the means of the groups vary in one way or another depending on the variable and the groups that are compared (See Figure 17), and in the same way, the significance of these differences varies according to the grades that are compared and the

variable that is analysed in each case (See Figure 18).

The dataset clean-up process deleted most of the outliers of every patient for the five analysed variables, as can be seen in Figure 20. It also allowed us to see a more homogeneous pattern for the variables of cytoplasmic and nuclear proportions. Training with the reduced dataset generated slightly higher metric values for accuracy, specificity, sensitivity, Matthew's correlation coefficient and F1 score. Testing the model with this dataset got more consistent results, although there is still a margin for improvement.

Regarding the dendrogram, it can be observed that there is no clear distinction between groups of any classification according to the clusters (Figure 22). This fact verifies that it is hard to distinguish samples according to analysed features and that the number of samples for some categories is limited.

5 Conclusion

In conclusion, the analysis of the information from all the samples was inconclusive in guiding the deep learning model, but it facilitated the dataset-cleaning process. Based on the reduced dataset, it was determined that the sample size for each class in the classification of TDP-43 pathology is almost completely balanced, being an ideal scenario for training the deep learning models. In addition, there were significant differences between cases and controls for the total amount of TDP-43, its cytoplasmic and nuclear amounts and its cytoplasmic and nuclear proportions.

In this way, it was possible to guide the deep learning model based on the classification by TDP-43 pathology by ruling out the classification by ALS as a possibility. A new model possibility is proposed considering variables for which cases and controls have significant differences. On the other hand, significant differences between the varied grades of protein pathology for the three variables have been clarified for both datasets. However, it must be understood that there is some variability in terms of these differences when individualising them so that for each comparison between two groups, more or less significant differences will be found depending on the variable that is used. The reduced dataset has fewer outliers and resulted in better and more consistent results when training the model than the complete dataset, so the cleanup process was useful and efficient. However, the results were still insufficient as the clustering process did not group samples consistently. High-quality and more abundant samples are needed in order to produce better results. Thus, the goal of developing an effective TDP-43 pathology detection tool remains to be met.

Acknowledgments

We would like to thank Dr Mathew Horrocks from the University of Edinburgh for providing the data used in this study. Their contributions were invaluable to the completion of this research.

References

ALS Association. ALS symptoms and diagnosis. Available at: https://www.als.org/understanding-als/symptoms-diagnosis (Accessed on 2024-02-06), 2023. 4

G Bensimon, L Lacomblez, VALS Meininger, and ALS/Riluzole Study Group. A controlled trial of riluzole in amyotrophic lateral sclerosis. *New England Journal of Medicine*, 330(9):585–591, 1994. 2

James D Berry, Marie Blanchard, Kerina Bonar, Emma Drane, Molly Murton, Uffe Ploug, Kristen Ricchetti-Masterson, Natasa Savic, Emma Worthington, and Terry Heiman-Patterson. Epidemiology and economic burden of amyotrophic lateral sclerosis in the united states: A literature review. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, pages 1–13, 2023. 1

Amit Bhardwaj, Michael P Myers, Emanuele Buratti, and Francisco E Baralle. Characterizing TDP-43 interaction with its RNA targets. *Nucleic acids research*, 41(9):5062–5074, 2013. 2

Lies Vanden Broeck, Patrick Callaerts, and Bart Dermaut. Tdp-43-mediated neurodegeneration: towards a loss-of-function hypothesis? *Trends in molecular medicine*, 20(2):66–71, 2014. 3

Robert H Brown and Ammar Al-Chalabi. Amyotrophic lateral sclerosis. *New England Journal of Medicine*, 377(2):162–172, 2017. 2

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018. 7

Claudia Colombrita, Eleonora Zennaro, Claudia Fallini, Markus Weber, Andreas Sommacal, Emanuele Buratti, Vincenzo Silani, and Antonia Ratti. Tdp-43 is recruited to stress granules in conditions of oxidative insult. *Journal of neurochemistry*, 111(4):1051–1061, 2009. 3

João Costa, Michael Swash, and Mamede de Carvalho. Awaji criteria for the diagnosis of amyotrophic lateral sclerosis: a systematic review. *Archives of neurology*, 69(11):1410–1416, 2012. 2

Ensembl. Gene: Tardbp ensg00000120948. Available at: https://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000120948;r=1:11012344-11030528 (Accessed on 2024-02-06). 3

Makoto Fukushima, Nao Hosoda, Kotaro Chifu, and Shin-ichi Hoshino. Tdp-43 accelerates deadenylation of target mRNA s by recruiting caf1 deadenylase. *Febs Letters*, 593(3):277–287, 2019. 3

Andrei Dmitri Gavrilov, Alex Jordache, Maya Vasdani, and Jack Deng. Preventing model overfitting and underfitting in convolutional neural networks. *International Journal of Software Science and Computational Intelligence (IJSSCI)*, 10(4):19–28, 2018. 5

Michael A Gitcho, Eileen H Bigio, Manjari Mishra, Nancy Johnson, Sandra Weintraub, Marsel Mesulam, Rosa Rademakers, Sumi Chakraverty, Carlos Cruchaga, John C Morris, et al. Tardbp 3-utr variant in autopsy-confirmed frontotemporal lobar degeneration with tdp-43 proteinopathy. *Acta neuropathologica*, 118:633–645, 2009. 3

Leslie I Grad, Guy A Rouleau, John Ravits, and Neil R Cashman. Clinical spectrum of amyotrophic lateral sclerosis (als). *Cold Spring Harbor perspectives in medicine*, page a024117, 2016. 3

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 14

Shinji Higashi, Tomohiro Kabuta, Yoshitaka Nagai, Yukihiro Tsuchiya, Haruhiko Akiyama, and Keiji Wada. Tdp-43 associates with stalled ribosomes and contributes to cell survival during cellular stress. *Journal of neurochemistry*, 126(2):288–300, 2013. 3

Arisa Hirano, Tomoki Nakagawa, Hikari Yoshitane, Masaaki Oyama, Hiroko Kozuka-Hata, Darin Lanjakornsiripan, and Yoshitaka Fukada. USP7 and TDP-43: pleiotropic regulation of cryptochrome protein stability paces the oscillation of the mammalian circadian clock. *PLoS One*, 11(4):e0154263, 2016. 3

Yuki Inukai, Takashi Nonaka, Tetsuaki Arai, Mari Yoshida, Yoshio Hashizume, Thomas G Beach, Emanuele Buratti, Francisco E Baralle, Haruhiko Akiyama, Shin-ichi Hisanaga, et al. Abnormal phosphorylation of ser409/410 of TDP-43 in FTLD-U and ALS. *FEBS letters*, 582(19):2899–2904, 2008. 4

Keiichi Izumikawa, Yuko Nobe, Harunori Yoshikawa, Hideaki Ishikawa, Yutaka Miura, Hiroshi Nakayama, Takashi Nonaka, Masato Hasegawa, Naohiro Egawa, Haruhisa Inoue, et al. Tdp-43 stabilises the processing intermediates of mitochondrial transcripts. *Scientific reports*, 7(1):7709, 2017. 3

Myungjin Jo, Shinrye Lee, Yu-Mi Jeon, Seyeon Kim, Younghwi Kwon, and Hyung-Jun Kim. The role of tdp-43 propagation in neurodegenerative diseases: integrating insights from clinical and experimental studies. *Experimental & molecular medicine*, 52(10):1652–1662, 2020. 3

Hong Joo Kim, Nam Chul Kim, Yong-Dong Wang, Emily A Scarborough, Jennifer Moore, Zamia Diaz, Kyle S MacLea, Brian Freibaum, Songqing Li, Amandine Molliex, et al. Mutations in prion-like domains in hnrnpa2b1 and hnrnpa1 cause multisystem proteinopathy and als. *Nature*, 495(7442):467–473, 2013. 3

Janine Kirby, Emily F Goodall, William Smith, J Robin Highley, Rudo Masanzu, Judith A Hartley, Rachel Hibberd, Hannah C Hollinger, Stephen B Wharton, Karen E Morrison, et al. Broad clinical phenotypes associated with tar-dna binding protein (tardbp) mutations in amyotrophic lateral sclerosis. *Neurogenetics*, 11:217–225, 2010. 3

Pan-Hsien Kuo, Chien-Hao Chiang, Yi-Ting Wang, Lyudmila G Doudeva, and Hanna S Yuan. The

crystal structure of tdp-43 rrm1-dna complex reveals the specific recognition for ug-and tg-rich nucleic acids. *Nucleic acids research*, 42(7):4712–4722, 2014. 2

Benno Küsters, Bas JA van Hoeve, Helenius Jurgen Schelhaas, Henk Ter Laak, Baziel GM van Engelen, and Martin Lammens. Tdp-43 accumulation is common in myopathies with rimmed vacuoles. *Acta neuropathologica*, 117:209–211, 2009. 4

Serena Lattante, Guy A Rouleau, and Edor Kabashi. Tardbp and fus mutations associated with amyotrophic lateral sclerosis: summary and update. *Human mutation*, 34(6):812–826, 2013. 3

Ivano Lauriola, Alberto Lavelli, and Fabio Aiolli. An introduction to deep learning in natural language processing: Models, techniques, and tools. *Neurocomputing*, 470:443–456, 2022. 4

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 4

Peter J Lukavsky, Dalia Daujotyte, James R Tollervey, Jernej Ule, Cristiana Stuani, Emanuele Buratti, Francisco E Baralle, Fred F Damberger, and Frédéric HT Allain. Molecular basis of ug-rich rna recognition by the human splicing factor tdp-43. *Nature structural & molecular biology*, 20(12):1443–1449, 2013. 2

Pegah Masrori and Philip Van Damme. Amyotrophic lateral sclerosis: a clinical review. *European Journal of Neurology*, 27(10):1918–1929, 2020. 1, 2, 21

Bruce C. Matthews. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 35(1):7–10, 1975. https://doi.org/10.1177/001316447503500101. 14

Giuseppe Militello, Mohammed Rabiul Hosen, Yuliya Ponomareva, Pascal Gellert, Tyler Weirick, David John, Sajedah Mahmoud Hindi, Kamel Mamchaoui, Vincent Mouly, Claudia Döring, et al. A novel long non-coding rna myolinc regulates myogenesis through tdp-43 and filip1. *Journal of molecular cell biology*, 10(2):102–117, 2018. 4

Peter T Nelson, Dennis W Dickson, John Q Trojanowski, Clifford R Jack, Patricia A Boyle, Konstantinos Arfanakis, Rosa Rademakers, Irina Alafuzoff, Johannes Attems, Carol Brayne, et al. Limbic-predominant age-related tdp-43 encephalopathy (late): consensus working group report. *Brain*, 142(6):1503–1527, 2019. 3

Manuela Neumann, Deepak M Sampathu, Linda K Kwong, Adam C Truax, Matthew C Micsenyi, Thomas T Chou, Jennifer Bruce, Theresa Schuck, Murray Grossman, Christopher M Clark, et al. Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Science*, 314(5796):130–133, 2006. 2, 21

Montse Olivé, Anna Janué, Dolores Moreno, Josep Gámez, Benjamín Torrejón-Escribano, and Isidre Ferrer. Tar dna-binding protein 43 accumulation in protein aggregate myopathies. *Journal of Neuropathology & Experimental Neurology*, 68(3):262–273, 2009. 4

Archana Prasad, Vidhya Bharathi, Vishwanath Sivalingam, Amandeep Girdhar, and Basant K Patel. Molecular mechanisms of tdp-43 misfolding and pathology in amyotrophic lateral sclerosis. *Frontiers in molecular neuroscience*, 12:25, 2019. 3

Iqbal H Sarker. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN Computer Science*, 2(6):420, 2021. 4

Dongchao Shen, Xunzhe Yang, Yanying Wang, Di He, Xiaohan Sun, Zhengyi Cai, Jinyue Li, Mingsheng Liu, and Liying Cui. The gold coast criteria increases the diagnostic sensitivity for amyotrophic lateral sclerosis in a chinese population. *Translational Neurodegeneration*, 10(1): 1–8, 2021. 2

J Paul Taylor. Multisystem proteinopathy: intersecting genetics in muscle, bone, and brain degeneration. *Neurology*, 85(8):658–660, 2015. 3

James R Tollervey, Tomaž Curk, Boris Rogelj, Michael Briese, Matteo Cereda, Melis Kayikci, Julian König, Tibor Hortobágyi, Agnes L Nishimura, Vera Župunski, et al. Characterizing the rna targets and position-dependent splicing regulation by tdp-43. *Nature neuroscience*, 14(4): 452–458, 2011. 3

Martin R Turner. Diagnosing als: the gold coast criteria and the role of EMG. *Practical Neurology*, 2022. 4

Victoria Twiddy. Development of a novel tdp-43 binding aptamer. Master's thesis, University of Edinburgh, 2021. 5, 6, 18

Lauren Versluys, Pedro Ervilha Pereira, Nika Schuermans, Boel De Paepe, Jan L De Bleecker, Elke Bogaert, and Bart Dermaut. Expanding the tdp-43 proteinopathy pathway from neurons

to muscle: physiological and pathophysiological functions. *Frontiers in Neuroscience*, 16:815765, 2022. 3, 4

Thomas O Vogler, Joshua R Wheeler, Eric D Nguyen, Michael P Hughes, Kyla A Britson, Evan Lester, Bhalchandra Rao, Nicole Dalla Betta, Oscar N Whitney, Theodore E Ewachiw, et al. Tdp-43 and rna form amyloid-like myo-granules in regenerating muscle. *Nature*, 563(7732): 508–513, 2018. 3, 4

Ailin Wang, Alexander E Conicella, Hermann Broder Schmidt, Erik W Martin, Shannon N Rhoads, Ashley N Reeb, Amanda Nourse, Daniel Ramirez Montero, Veronica H Ryan, Rajat Rohatgi, et al. A single n-terminal phosphomimic disrupts TDP-43 polymerization, phase separation, and RNA splicing. *The EMBO journal*, 37(5):e97452, 2018. 3

Yang Xin, Lingshuang Kong, Zhi Liu, Yuling Chen, Yanmiao Li, Hongliang Zhu, Mingcheng Gao, Haixia Hou, and Chunhua Wang. Machine learning and deep learning methods for cybersecurity. *Ieee access*, 6:35365–35381, 2018. 4

Sijie Yang, Fei Zhu, Xinghong Ling, Quan Liu, and Peiyao Zhao. Intelligent health care: Applications of deep learning in computational medicine. *Frontiers in Genetics*, 12:607471, 2021. 4

Kyan Younes and Bruce L Miller. Frontotemporal dementia: neuropathology, genetics, neuroimaging, and treatments. *Psychiatric Clinics*, 43(2):331–344, 2020. 3, 4