Actief lerend model voor de classificatie van celkernen

Sam Vanmassenhove
Universiteit Gent, Faculteit Ingenieurswetenschappen en Architectuur
Mei 2018

Promotoren: Dr. Saskia Lippens, Prof. dr. ir. Wilfried Philips
Begeleiders: Dr. ir. Jan Aelterman, Dr. ir. Evelien Van Hamme, Joris Roels

***Abstract* – In de biologie gebeurt het nog vaak dat detectie- en classificatieproblemen handmatig opgelost worden. Zulke problemen zouden met machine learning kunnen worden opgelost, maar deze methoden vereisen vaak honderden of duizenden geannoteerde samples. Onderzoekn ondernemen niet snel zo'n grote manuele taak. In deze thesis stellen we een aanpak voor met *active learning* om het vereiste aantal gelabelde samples drastisch te reduceren.**

We tonen aan dat het in bepaalde gevallen mogelijk is om het aantal samples te verminderen met een factor zes.

***Kernwoorden* – actief leren, machinaal leren, classificatie, celkernen**

I. Introductie

A. *Probleem*

Onderzoekers aan het Vlaams Instituut voor Biotechnologie (VIB) werken aan verschillende classificatieproblemen van cellen en celkernen in biologische miscroscoopbeelden. Deze problemen kunnen opgelost worden door de *machine learning* technieken die in de meeste commerciële softwarepakketten aanwezig zijn.

Deze technieken zijn vaak niet toepasbaar in de praktijk omdat ze een groot aantal manueel geannoteerde *samples* vereisen. Zo'n grote inspanning door een onderzoeker is vaak de moeite niet waard voor problemen die niet vaak voorkomen. In deze gevallen wordt machine learning in de praktijk niet toegepast en wordt de classificatie volledig manueel gedaan. Het is duidelijk dat er veel ruimte voor verbetering is in dit domein.

In deze thesis stellen we een actief lerend model voor als oplossing voor dit probleem. Hiermee wordt het aantal training samples dat nodig is drastisch verminderd.

De structuur van deze paper is als volgt. Eerst wordt een algemeent overzicht van het voorgestelde framework gegeven, waarna de verschillende vormen van *active learning* worden uitgelegd. Ten slotte worden twee experimenten op reële data besproken, waarbij zowel de biologische context wordt uitgelegd als de resultaten geanalyseerd.

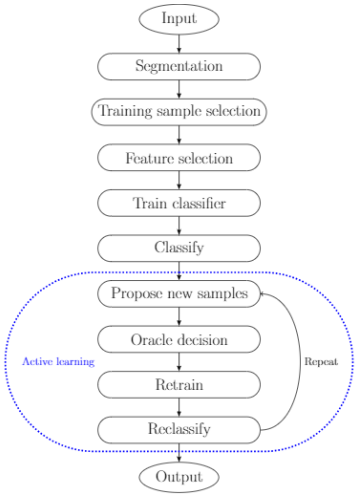


Figure 1 - Active learning workflow. De feedback loop wordt in het blauw aangeduid.

B. *Software*

Er bestaan verschillende softwarepakketten om biologische microscoopbeelden te analyseren. De meest populaire pakketten zijn ImageJ en haar uitgebreide versie Fiji, maar ook CellProfiler and QuPath. De methode die in deze thesis wordt beschreven werd als QuPath extensie geïmplementeerd. De source code voor deze extensie is te vinden op https://github.com/savmasse/qupath-extension-tunel.

Het grote voordeel van QuPath over de andere pakketten is dat QuPath ontworpen werd om grote datasets te verwerken en om de gebruiker volledige slides in één afbeelding te laten bekijken en analyseren zonder deze in meerdere stukken te moeten knippen.

QuPath bevat functionaliteit voor zowel licht- als fluorescentiemicroscopie, maar geen *tools* voor de analyse van 3D- en 4D-beelden (3D + tijd).

II. Voorgestelde framework

A. *Overzicht framework*

De classificatie van cellen en celkernen in een microscoopbeeld zijn afhankelijk van veel meer dan de *classifier* zelf. De volledige workflow die in deze thesis wordt voorgesteld wordt in **Figuur 1** geïllustreerd. De eerste stap in

het diagram is de segmentatie van celkernen uit het beeld. Dit wordt gevolgd door de selectie van de initiële trainingsamples en de relevante features door de gebruiker. Deze stappen worden hieronder verder uitgelegd.

B. Segmentatie

Vooraleer de classificatie kan worden uitgevoerd moeten de celkernen eerst gesegmenteerd en van elkaar gescheiden worden. Er zijn verscheidene bestaande technieken hiervoor, waarvan de *watershed transform* de meest gebruikte is voor biologische beelden.

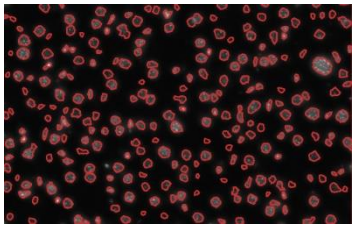


Figure 2 - Watershed segmentation in QuPath on a fluorescence image.

De watershedtechniek is gebaseerd op het natuurlijk fenomeen van verscheidene aanéén grenzende waterbassins.

De randen die de watermassa's scheiden kunnen gevonden worden door de bassins te vullen van onder uit. De plaats waar het rijzende water van twee verschillende bassins elkaar ontmoeten zal de rand zijn.

Watershedtechnieken zullen meestal een beeld oversegmenteren, dit wil zeggen dat er meer segmenten aangemaakt worden dan er werkelijkheid aanwezig zijn. Dit probleem kan worden aangepakt door zeer voorzichtig de minima te kiezen waaruit men begint te vullen. Er kan ook achteraf gefiltert worden op grootte om bepaalde kleine ruiselementen te verwerpen. Zie **Figure 2** als voorbeeld van een segmentatie in QuPath.

C. Selectie van training samples

Op dit punt in de workflow heeft de gebruiker nog steeds de volledige controle over de selectie van training samples. Dit zal niet het geval zijn eens de *feedback loop* bereikt wordt. Het is de verantwoordelijkheid van de gebruiker om minstens één correct voorbeeld van elke klasse te geven.

D. Feature selection

Een volgende stap voor de classificatie is de selectie van de correcte features voor de taak. In deze implementatie is de eindgebruiker verantwoordelijk voor het kiezen van de goede features. Populaire features in software packages zijn shape factors zoals oppervlakte, perimeter, compactheid en eccentriciteit van een segment, maar ook intensiteitsstatistieken zoals de mean en extrema van de pixelwaarden. Complexere features zoals randen kunnen berekend worden maar zullen waarschijnlijk niet gebruikt worden door eindgebruikers met weinig kennis van beeldverwerking.

Vooruitgang in deep learning heeft aangetoond dat autoencoders (AE) en convolutional neural networks (CNN) gebruikt kunnen worden om features uit beelden te extraheren [2]. Opmerkelijk is dat deze extractie volledig *unsupervised* uitgevoerd kan worden, zodat geen voorafgaandelijke manuele labelling vereist is. Hoewel feature extractie met het gebruik van autoencoders even wordt aangekaart in de volledige thesis,werd er geen conclusie getrokken over het nut ervan in deze context.

Toekomstig werk kan deze mogelijkheden verder onderzoeken, daar de automatische extractie van features de classifier minder afhankelijk kan maken van het goede oordeel van de eindgebruiker.

E. Classifier

Deep learning (DL) methodes steeds populairder voor classificatie taken. Dit is vooral het geval voor beeld classificatie problemen, waar CNN's bewezen hebben beter te presteren dan alle klassieke machine learning methodes [4]. Neurale netwerken hebben evenwel veel grotere annotated datasets nodig voor het trainen dan het geval is voor andere technieken, en zijn derhalve niet van toepassing in de context van deze thesis.

Random forests (RF) is een ensemble machine learning methode die uit een groot aantal decision trees bestaat. Deze methode wordt het meest toegepast in biologie. Recent werd deze methode toegepast op X-raybeelden. In vergelijking met DL heeft RF veel minder training data nodig en traint dus ook sneller.

III. Active learning

A. Introductie

Active learning (AL) is een tak van machine learning dat een feedback loop introduceert in het proces. Na de initiële classificatiestap worden nieuwe training samples toegevoegd om mogelijk ontbrekende informatie aan te vullen. Dit proces van het toevoegen van nieuwe samples ,retrainen en reclassificieren kan herhaald worden tot de performance convergeert of de gebruiker (hier het 'orakel' genoemd) tevreden is met het resultaat.

Om active learning te kunnen toepassen moet aan de volgende voorwaarden voldaan worden:

1. Het orakel moet altijd correct zijn.
2. Het orakel moet altijd beschikbaar zijn.
3. Er is maar één orakel.
4. De kost om het orakel te consulteren is steeds constant. Deze kost is normaal gezien een tijds- of computationele kost in academische papers maar in praktische applicaties kan dit ook een financiële kost zijn.

B. Sampling

Actief leren vereist dat het volgende sample steeds nieuwe informatie levert die de huidige classifier niet bevat. Het voorstellen van samples moet op een intelligente manier gebeuren zodat overbodige informatie vermeden wordt maar toch alle *outliers* vertegenwoordigd worden. Enkele populaire manieren van samplen worden hieronder besproken.

1. Random sampling

De meest eenvoudige manier van samplen is om willekeurig samples voor te stellen uit de training set. De enige voorwaarde die hier wordt opgelegd is dat elke sample uniek moet zijn en dus nog niet voorgesteld mag geweest zijn.

2. Least confidence sampling (LC)

Veel classifiers zullen een waarde teruggeven voor de zekerheid, ook de *label confidence* genoemd, van de voorspelling. Voor *random forests* is dit een waarschijnlijkheid gebaseerd op de hoeveelheid stemmen die iedere klasse krijgt; voor *support vector machines* (SVM) is de maat voor zekerheid de afstand van het datapunt tot de hyper-plane.

In *least confidence* (LC) sampling zal de volgende voorgestelde sample altijd die sample zijn waarvan de classifier het minst zeker is.

Voor classificatieproblemen waar de hoeveelheid classes M groter is dan twee zal de *cross-entropy* (Vegelijking(1) voornamelijk gebruikt worden om de label confidence te meten.

Hier is p_c de voorspelde waarschijnlijkheid dat het sample een class label *c*, heeft, en y_c is de binaire indicator die gelijk is aan één als c het ware label is.

$$-\sum_{c=1}^{M} y_c \log p_c \qquad (1)$$

3. Clustered sampling

Een andere methode van sampling is om de feature space in clusters onder te verdelen. Dit wordt gedaan om de te forceren dat de volledige feature space wordt beschouwd. Clustering zorgt er voor dat verschillende types van samples voorgesteld worden die anders niet zouden gebruikt worden.

Wanneer willekeurige sampling wordt toegepast worden enkel de dichtste gebieden goed vertegenwoordigd en worden outliers vooral genegeerd. Daarentegen worden bij het gebruik van LC enkel deze samples aan de grens van de klassen gesampeld.

Het algoritme dat hier gebruikt wordt voor de clustering is k-means clustering omdat dit een courant gebruikt algoritme is voor het subsamplen van een dataset.

4. Least confidence clustered sampling (LCC)

De methode die in de implementatie wordt gebruikt is een combinatie van de twee vorige methoden. De feature space wordt in clusters verdeeld waarna de samples binnen elke cluster volgens stijgende *confidence* worden gesorteerd.

Vanuit elke cluster wordt een sample voorgeschoteld tot de laatste cluster bereikt wordt. Dan worden opnieuw geclassificeerd en gesorteerd volgens de nieuwe zekerheden.

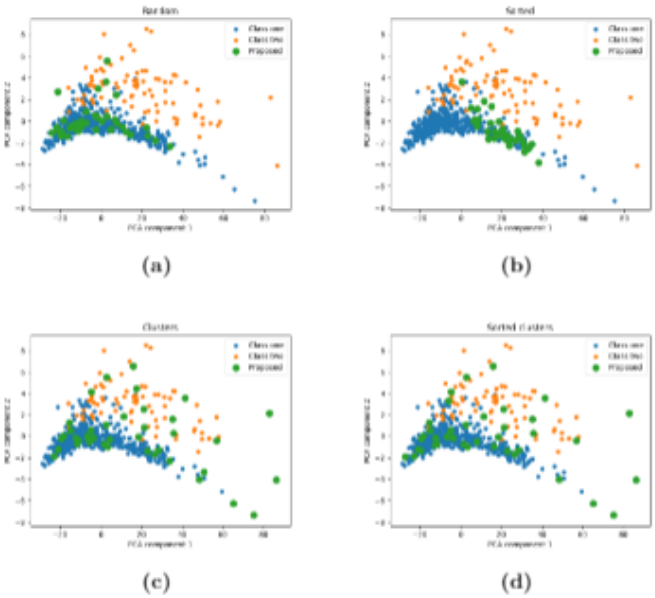


Figure 3 – Verschillende sampling methoden vergeleken voor een binair classificatieprobleem, waarbij de ware labels in blauw en oranje aangeduid staan, terwijl de 40 eerste voorgestelde samples in groen staan getekend.

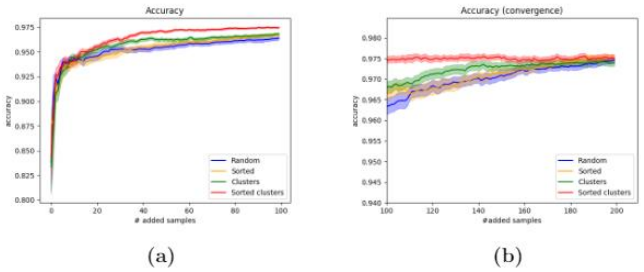


Figure 4 – Vergelijking van de accuracy van de vier sampling methoden op een synthetische dataset. Hier toont (a) de performance voor de 100 eerste samples en (b) de convergentie aan het einde van de training set. De error op deze grafieken is steeds het 95% betrouwbaarheidsinterval op 100 simulaties.

C. Performantie

Figure 3 toont de eerste 40 samples voor elk van de vier verschillende methode die hierboven warden beschreven. Random en LC sampling zijn zeer gefocust op dichte gebieden. De clusteringmethoden hebben wel samples die over de ganse feature space verspreid liggen.

Het is niet genoeg om enkel mooi verspreide samples te hebben als dit geen extra performance opbrengt. De resultaten in **Figure 4a** tonen aan dat de clusteringmethoden duidelijk sneller een betere accuracy leveren, maar het is duidelijk dat LCC de beste performantie geeft van alle methoden. In **Figure 4b** is te zien hoe de andere drie methoden pas op het einde van de training set convergeren naar het niveau van de LCC.

IV. Evaluatie en experimenten

Deze sectie zal uitleggen hoe de experimenten worden uitgevoerd and beschrijft de methode van evaluatie. Ook de biologische context van de experimenten wordt hier kort uitgedrukt. De werkelijke resultaten van de experimenten komen pas in de volgende sectie aan bod.

A. Evaluatie van de classifier

Classificatieproblemen in de biologische en medische wetenschappen zijn vaak zeer ongebalanceerd (*imbalanced*). Dit wil zeggen dat de klassen niet gelijk zijn verdeeld. Het is bekend dat classifiers slecht werken op imbalanced datasets en dat de accuracy daar geen goede metriek is voor de kwaliteit van de classificatie. Betere metrieken worden voorgesteld in vergelijkingen (2-(5).

De volgende termen moeten worden gedefinieerd: *True positives* (TP) zijn samples die positief geannoteerd en voorspeld werden, terwijl *false positives* (FP) negatief geannoteerd werden en positief voorspeld. Analoog werden *True negatives (TN)* negatief geannoteerd en voorspeld, terwijl *false negatives* (FN) positief werden geannoteerd maar negatief voorspeld.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad (2)$$

$$Precision = \frac{TP}{TP + FP} \qquad (3)$$

$$Recall = \frac{TP}{TP + FN} \qquad (4)$$

$$F1 = \frac{2 * Recall * Precision}{Recall + Precision} \qquad (5)$$

Beide experimenten die in deze sectie worden aangehaald worden op dezelfde manier uitgevoerd. De volledige active learning simulatie wordt 100 maal uitgevoerd waarna gemiddelden en betrouwbaarheidintervallen kunnen berekend worden. Elke simulatie wordt als volgt uitgevoerd.

De dataset wordt in training- en testset gesplitst. Vervolgens wordt een initïele trainingset willekeurig aangemaakt door een sample van elke klasse te kiezen.

Nadien worden samples een voor een toegevoegd terwijl voor elke stap de performantie wordt berekend. Merk op dat de x-as op de grafieken steeds op nul begint; in deze nul zitten reeds de initïtele samples.

B. *TUNEL experiment*

Terminal deoxynucleotidyl transferase dUTP nick end labeling (TUNEL) is een techniek voor de detective van fragmentatie van DNA in cellen [3]. Deze fragmentatie komt voor bij een bepaalde vorm van celdood (*apoptosis*), and kan gevisualiseerd worden door een chemische *marker* toe te voegen. Zo kunnen de dode celkernen in een populatie gedetecteerd worden. Een extra marker (DAPI) is toegevoegd waarvoor alle celkernen, dood of levend, oplichten. Elk van deze chemische markers vormen een kanaal van een fluorescentiebeeld (zie **Figuur 5**).

Om te beslissen of een cel dood of levend is wordt vaak enkel naar het rode kanaal gekeken, maar is een goede reden om naar beide kanalen te kijken. Het is namelijk zo dat een dode cellen niet langer oplicht in het DAPI-kanaal. Ook is het zo dat de vorm van een celkern veranderd bij de celdood, dus ook *shape factors* kunnen hier als features worden gebruikt.

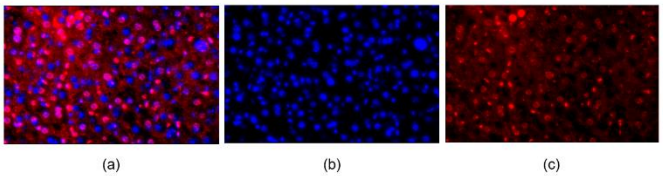


Figure 5 -Voorbeeld van een TUNEL beeld: (a) beide kanalen; (b) DAPI kanaal; (c) kanaal met TUNEL marker.

De segmentatie voor deze *use-case* is meestal niet ideaal, zodat meerdere celkernen in eenzelfde segment worden gegroepeerd. Dit probleem wordt hier aangepakt door het segmentatieprobleem in de classifier op te nemen: we maken vier mogelijke klassen aan voor een segment:

- Multiple-Negative: meerdere celkernen die allen levend zijn.
- Multiple-Positive: meerdere celkernen waarvan minstens één dood is.
- Single-Negative: enkele levende celkern.
- Single-Positive: enkele dode celkern.

C. *Microglia classification*

De tweede use-case die in deze thesis besproken werd is gebaseerd op een recent door VIB onderzoekers gepubliceerde paper [1]. Tijdens dit onderzoek moesten specifieke hersencellen, microglia genaamd, worden gedetecteerd en geteld. Deze cellen zijn te vinden in de hersenen en de ruggengraat waar ze een immuunfunctie vervullen.

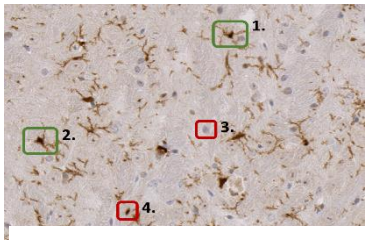


Figure 6 – RGB-beeld met de microglia in groen en de andere cellen in rood.

Onderzoekers moesten deze cellen manueel tellen omdat huidige technieken hier niet toepasbaar waren.

De moeilijkheid in dit geval is dat er bepaalde donkere cellen zijn die exact dezelfde kleur hebben als de microglia. Deze andere cellen hebben echter geen stervorm zoals de echte microglia hebben.

In **Figuur 6** zijn enkele voorbeelden aangeduid. De microglia (positief) zijn groen omcirkelt. Cel (3) is een goed voorbeeld van een blauwe/paarse hersencel die duidelijk negatief moet worden geclassificeerd, terwijl (4) de goede kleur heeft maar de kenmerkende stervorm ontbreekt hier.

V. Resultaten

A. *Hypothese*

Na de indentificatie van LCC als de beste methode voor het samplen tijdens de active learning in **Figure 4** wordt deze methode nu vergeleken met random sampling op reële data. We verwachten dat het resultaat analoog zal zijn als dat op de synthetische data maar dat het effect wellicht niet zo sterk zal zijn.

De LCC methode zal sneller convergeren naar het optimum van elke metriek.

B. *TUNEL*

De resultaten voor de accuracy en precision (andere metrieken werden in de thesis ook berekend) worden gegeven in **Figuur 7**. LCC convergeert duidelijk sneller dan random sampling: er zijn slechts 82 samples nodig om de maximale waarde te bereiken voor LCC terwijl de volledige training set (500) moet doorlopen worden bij random sampling.

Merk op dat de metrieken die hier geïllustreerd worden gewogen gemiddelden zijn van de metrieken voor elke van de vier klassen.

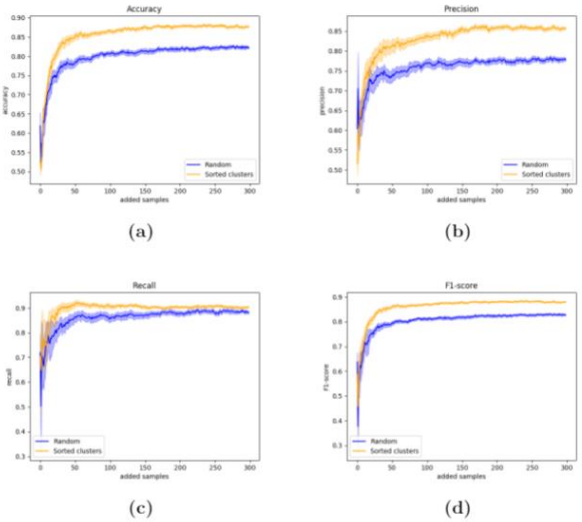
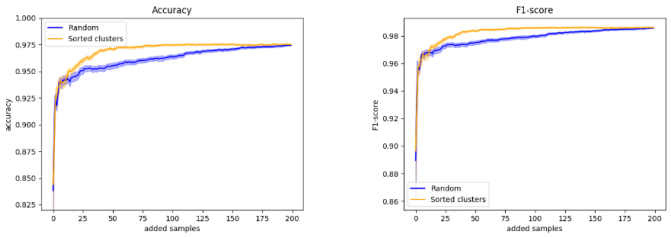


Figure 7 -TUNEL classificatie. Vergelijking van performantie van LCC (geel) en random (blauw) sampling. (a) Accuracy, (b) precision, (c) recall, en (d) f1-score.

C. *Classificatie van microglia*

Figure 8 illustreerd de resultaten van de accuracy en f1-score voor de microglia use-case. Hier is het nog duidelijker dat LCC

sneller convergeert dan random sampling.



(a) (b)

Figure 8 –Classificatie van microglia. Vergelijking van performantie van LCC (geel) en random (blauw) sampling. (a) Accuracy, (b) F1-score.

D. Conclusie

Het active learning sampling (LCC) heeft een betere performance dan random sampling voor beide experimenten. De maximale performantie wordt bij het *microglia* experiment vier maal sneller en bij TUNEL zes maal sneller dan random bereikt. De reductie in het aantal vereiste samples gaat gepaard met een evenredige reductie in de tijd die een onderzoeker aan het labellen moet spenderen.

VI. Toekomstig werk

Het blijkt dat moeilijke classificatieproblemen vaak voorafgegaan worden door een moeilijk segmentatieprobleem. Toekomstig werk moet dus niet focussen op een van de twee, zoals deze thesis heeft gedaan, maar moet de twee problemen samen bekijken.

Een extra vereiste voor een generisch framework is het wegnemen van verantwoordelijk van de eindgebruiker. Autoencoders kunnen toegepast worden om automatisch de beste features te extraheren. Selectietechnieken kunnen dan gebruikt wordenom enkel de meest revelante features voor de classificatie te behouden.

VII. Conclusie

In deze thesis hebben we actief leren voorgesteld als oplossing voor de grote manuele werklast aan annoteren die bij traditionele machine learning-technieken nodig is.

Verschillende sampling-methoden werden beschouwd, waarbij LCC duidelijk als de meest performante naar boven kwam, zowel bij testen op synthetische en reële data.

Twee verschillende experimenten werden uitgevoerd op reële data. Deze toonden eveneens aan dat active learning sneller convergeert dan random sampling, in één geval wel zes keer sneller.

Dit is een significant verschil dat waardevol is om machine learning praktisch bruikbaar te maken in deze context.

Referenties

[1] Voet, S., Guire, C. M., Hagemeyer, N., Martens, A., Schroeder, A., Wieghofer, P., . . . Loo, G. V. (2018). A20 critically controls microglia activation and inhibits inflammasome-dependent neuroinflammation. Nature Communications, 9(1). doi:10.1038/s41467-018-04376-5

[2] Rawat, W., & Wang, Z. (2017). Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. Neural Computation, 29(9), 2352-2449. doi:10.1162/neco_a_00990

[3] Labat-Moleur, F., Guillermet, C., Lorimier, P., Robert, C., Lantuejoul, S., Brambilla, E., & Negoescu, A. (1998). TUNEL Apoptotic Cell Detection in Tissue Sections: Critical Evaluation and Improvement. Journal of Histochemistry & Cytochemistry, 46(3), 327-334. doi:10.1177/002215549804600306

[4] Chandra, B., & Sharma, R. K. (2015). Exploring autoencoders for unsupervised feature selection. 2015 International Joint Conference on Neural Networks (IJCNN). doi:10.1109/ijcnn.2015.7280391