# Using Deep Learning to Automate Osteocyte Viability Quantification

Vanessa Ma

Mark Reynolds, Du Huynh, Jacob Kenny
Computer Science and Software Engineering
University of Western Australia

Sam Withers
CEED Client: Australian Institute of Robotic Orthopaedics

### Abstract

*In surgical orthopaedics, the oscillating saw is an antiquated tool that could be replaced by novel, more precise laser-cutting techniques to improve patient outcomes. Proving the laser's safety on human bone involves a standardised analysis of osteocyte viability before and after the laser has been applied, which can be compared with existing mechanical cutting tools. Osteocytes comprise between 90 – 95% of all bone cells and are considered indicative of bone health and the bone's ability to recover after orthopaedic surgery. They can be classified as viable, apoptotic or necrotic, which are then counted to quantify osteocyte viability. Currently, osteocyte viability quantification is a manual process that is laborious and costly but advances in deep learning suggest that automation is feasible. This paper explores whether two state-of-the-art object detection models, Faster R-CNN and RetinaNet can achieve human-level accuracy. Results show that both models have a high potential to automate osteocyte viability quantification, with comparable performance against each other. When tested on a full Whole Slide Image while using an Intersection over Union (IoU) threshold of 0.3 to account for inexact ground truth boxes, Faster R-CNN achieved a mAP*$_{30}$ *score of 79.24% while RetinaNet achieved a mAP*$_{30}$ *score of 80.06%.*

## 1. Introduction

In Australia, there is a growing need for orthopaedic surgeries due to an ageing and increasingly sedentary population, with 1 in 2 adults not meeting the physical activity requirements in 2017– 2018, a figure which is projected to rise (Australian Institute of Health and Welfare, 2020; Australian Bureau of Statistics, 2019). There has also been a rise in chronic diseases such as obesity and younger, more active patients who require joint replacement surgeries. Consequently, there is a greater prevalence of end-stage osteoarthritis (OA), which is the progressive degradation of a synovial joint. By end-stage arthritis, a joint is no longer functional. To restore joint function, arthroplasty is needed. Two common types are total hip and total knee replacements.

The current gold standard tool in osteotomy (bone cutting) is the mechanical oscillating saw which suffers known limitations in precision and exposes the cut site to high thermal, vibration and mechanical stress which could lead to osteonecrosis (when bone tissue dies causing the bone to collapse) (Pandey and Panda, 2013). Less accurate cuts can also increase the likelihood of post-operative complications for the patient, such as aseptic loosening, with 53% of revision surgeries arising from intra-operative issues (Australian Orthopaedic Association and National Joint Replacement Registry, 2019). One promising alternative is the laser, which could lead to improvements including (Kreisler et al., 2002):

1. Reducing operating time by minimising duration of the cutting process
2. Offers high precision and usability to avoid damage to adjacent tissue, and minimizes the removal and dispersion of excess tissue
3. Minimise thermal damage to the site area

While it has been established that lasers can achieve more precise cuts to a micrometre range, it has yet to be confirmed that thermal necrosis (death) of bone tissue caused by the laser is at worst, equal to the damage caused by the mechanical oscillating saw. This can be achieved through a standardised analysis of osteocyte viability before and after the tool has been applied. Osteocytes are cells found in the bone matrix that are considered key in bone remodelling due to their ability to recruit osteoclasts and osteoblasts to replace and repair damaged bone (Bonewald, 2011). They comprise over 90% to 95% of all bone cells and can be classified as either necrotic (dead), viable (alive) or apoptotic (currently alive, but guaranteed to die in the future). They are thus considered representative of general bone health.

Research investigating thermal effects of osteotomy technology has stagnated due to the manual, laborious process of quantifying osteocyte viability of bone. This process includes preparation of the sample bone, preserving its state, and undergoing a process to reveal osteocyte viability. One method involves staining a bone slide with Cleaved Caspase 3 before counterstaining with Haematoxylin. A histologist, the trained expert, must then identify and count each osteocyte. For the Australian Institute of Robotic Orthopaedics, 216 slides lead to a cost of ~$8500 for around 54 hours. Considering that many more slides must be analysed to conclusively examine the damage caused by a tool, this cost and time become unsustainable and sluggish. However, advances in convolutional neural networks (CNN) and digital pathology suggest that an object detector capable of detecting and classifying osteocytes is feasible with proven cases of successful small object detection and nuclei/cell detection (Van der Laak, Litjens and Ciompi, 2021). Two such detectors that have been used successfully before are Faster R-CNN and RetinaNet. Faster R-CNN is a two-stage object detection model comprising a regional proposal network sharing convolutional layers with the classification CNN layer (Ren et al., 2015). RetinaNet is a one-stage object detector, that used a novel loss function, focal loss, to reduce the effect of class imbalance against the background (Lin, 2020).

Having a tool to automate this process could be a valuable resource for any surgical tool novel development. Although human intervention is likely to be required due to limitations in ground truth and reason justification being desired, automating osteocyte viability quantification will increase the development speed of novel orthopaedic surgical tools to enhance surgical outcomes (Campanella et al., 2019). It will also allow a histologist to spend their time on other research and development areas, while reducing costs associated with osteocyte viability quantification.

## 1.2 Objectives

The main objective of this research is to determine whether osteocyte viability quantification can be automated, and how well could it operate without human oversight. This will be assessed through training an existing deep learning model to classify and locate osteocytes as necrotic, apoptotic or viable to a human-level accuracy. We aim to achieve human-level accuracy, with an emphasis on classifying accurately rather than predicting the precise location, given that annotations are not always guaranteed to be perfect. We also aim to examine factors in the model and data preparation that may affect its effectiveness. Hence out specific objectives can be outlined as:

1. Train two models to detect osteocyte classes and compare their performance
2. Explore reasons for why the model may confuse or miss different osteocyte classes
3. Explore how using individual thresholds can lead to an overall more accurate result
4. Explore the effect of noise in training an effective model

# 2. Process

The main programming language used was Python with Jupyter Notebook and Visual Studio Code used as the Integrated Development Environment (IDE).

## 2.1 Data Preparation and Processing

Ground truth was obtained by AIRO consisting of offcuts from joint replacement surgery obtained from the theatre under ethics approval HPH435. These offcuts have then been preserved and stained with Cleaved Caspase 3 to reveal apoptosis, and Haematoxylin to reveal necrotic and viable cells. They were labelled by a histologist with available data including: the original whole slide image (WSI), the annotated WSI and coordinates of centre-points. The labels were live for viable cells, empty for necrotic cells, and caspase for apoptotic cells.

To process the data, it was first necessary to generate bounding boxes that would closely encapsulate the small, oblong osteocytes. The best size found was 20px x 20px, as this captured longer osteocytes, and those that did not have an accurate centre-point label. However, these bounding boxes comprised under 0.01% of the total WSI (resolution of 2448 x 1920). It was thus necessary to reduce the size of the images that would be used to train and test the models. Image tiling was ideal. Tile size was chosen to be 300 x 300 pixels, with an overlap of 50% to prevent cells on the boundary being missed. Tiles at the boundary of the image were padded by a black border.

All osteocytes, which are the objects of interest, are found in the purple-stained regions. However, these stained regions can be sparse, particularly in WSIs of off-cuts that were preserved at a later time point. This led to many tiles that were considered background, i.e. not containing any objects of interest. Non-background tiles are tiles that contain at least one object of interest. To avoid training the model with background noise, two subsets of training data were created. One training subset contained only non-background tiles, while the other contained background tiles that comprised 2761 out of 13183 total tiles. The background tiles included in this set was selected randomly. This would allow the effect of background tile noise to be examined. Two evaluation holdout sets were also created, one with non-background tile only and the other with all tiles.

## 2.2 Model Training

RetinaNet and Faster R-CNN were chosen as the most appropriate of all the CNN-based object detectors. RetinaNet was selected as its loss function, focal loss could have additional benefit in identifying small osteocytes in comparison to the overwhelming size of the background. Faster R-CNN, on the other hand, is the latest iteration of the R-CNN family of detectors. RetinaNet and Faster R-CNN were initialised with a ResNet-50 backbone that was pre-trained on ImageNet.

Data was then split into training, testing and validation sets. As it was desirable to test models on whole slides rather than tiles, holdout sets were generated by randomly selecting 41 WSIs, which was ~20% of the data. The remaining 173 WSIs were used for training and validation. These WSIs were tiled before the tiles were split into a 90:10 training and validation set. The 90:10 split was chosen as training the model on a greater variety of data was prioritised. This was due to there being only six bio-replicates in the dataset.

Evaluation was made on whole slides rather than tiles. Mean Average Precision (mAP) was used as a general measure of object detection accuracy. mAP is calculated as the area under the Precision-Recall curve, which was generated using the Pascal VOC method of 11-point interpolation. $F_1$-score was also evaluated to examine the ideal confidence threshold for accurately predicting osteocytes, where precision and recall have equal importance.

# 3. Results and Discussion

| Model | FRCNN-NB | RetinaNet-NB | FRCNN-AT | RetinaNet-AT |
|---|---|---|---|---|
| IoU | **mAP scores** (%) | | | |
| 0.01 | 80.28 | 81.89 | 77.28 | 79.29 |
| 0.1 | 79.95 | 81.52 | 76.92 | 78.72 |
| 0.3 | 79.24 | 80.06 | 76.34 | 77.68 |
| 0.5 | 72.76 | 72.59 | 70.13 | 70.52 |
| 0.7 | 40.10 | 33.74 | 38.14 | 32.61 |
| 0.9 | 3.78 | 2.06 | 3.49 | 1.98 |
| | **$F_1$ score** *(threshold)* | | | |
| 0.3 | 0.7613 (0.78) | 0.7615 (0.4) | 0.7407 (0.79) | 0.7435 (0.4) |
| 0.5 | 0.7166 (0.79) | 0.7177 (0.4) | 0.6971 (0.79) | 0.7014 (0.41) |

**Table 1** Faster R-CNN and RetinaNet overall performance on osteocyte detection and classification. NB stands for Non-Background and AT stands for All Tiles. FRCNN is Faster R-CNN e.g. FRCNN-NB is the Faster R-CNN model evaluated. on only non-background tiles.

Table 1 shows the performance of Faster R-CNN and RetinaNet over all osteocyte classes. Intersection over Union (IoU) is generally measured from .5:0.5:0.95. However, due to the imprecise location of some annotations in WSIs, IoU was measured from 0.1 to 0.95 with 0.5 increments, in addition to at 0.01. An IoU of 0.01 implies predicted boxes only needed to touch the ground truth boxes. mean Average Precision (mAP) across the classes was highest at the 0.01 IoU threshold which FRCNN-NB scoring 80.28% and RetinaNet-NB scoring 81.89%. mAP scores were also similar, albeit marginally lower, at IoU thresholds 0.1 and 0.3. While a higher mAP score at 0.01 is likely to be influenced by accepting more bounding boxes found in the background to be ground truth, the similar score at 0.3 implies most accepted bounding boxes remain similar at these lower thresholds. As precision would have been reduced if most of these predictions had been ground truths from other nearby osteocytes, there is a good chance that many of these predictions were correct. Nevertheless, it is likelier that a false positive at higher thresholds that was accepted as ground truth for a 0.3 threshold is from human error as opposed to a 0.01 threshold.

However, it is impossible to ascertain that a 0.3 IoU would be an acceptable threshold without verification from a histologist. $mAP_{50}$ still gives an acceptable score with overall scores for FRCNN-NB and RetinaNet-NB at 72.76% and 72.59% respectively. Faster R-CNN can be observed to overtake RetinaNet's performance at higher thresholds as seen in the $mAP_{70}$ scores. This suggests that Faster R-CNN may have been better than RetinaNet at locating osteocytes where the histologist has placed ground truths, but not necessarily the better location predictor.

F-measure scores were taken using IoU thresholds of 0.5. They further confirm the lack of performance difference between Faster R-CNN and RetinaNet, regardless of the holdout set they were evaluated on, with minimal differences seen between scores in Table 1.

| **Model** | **FRCNN-NB** | **RetinaNet-NB** | **FRCNN-AT** | **RetinaNet-AT** |
|---|---|---|---|---|
| **Class** | **$AP_{50}$ scores (%)** | | | |
| Empty | 74.64 | 79.90 | 71.91 | 77.85 |
| Live | 74.37 | 75.30 | 71.28 | 70.78 |
| Caspase | 69.27 | 65.14 | 67.19 | 62.92 |

**Table 2** Faster R-CNN and RetinaNet performance on osteocyte detection and classification for individual classes

In Table 2, RetinaNet-NB has the best performance in identifying Empty osteocytes with RetinaNet-NB having a $mAP_{50}$ of 79.90%. In contrast, FRCNN-NB obtained a $mAP_{50}$ of 74.64%, which is around 5 points lower. Conversely, RetinaNet is also worse at identifying Caspase osteocytes with RetinaNet-NB having a $mAP_{50}$ of 65.14% compared to FRCNN-NB's of 69.27%, which is around 4 points lower. RetinaNet-NB and FRCNN-NB have comparable scores when predicting the Live class. This makes sense as this balances their overall performance as seen in Table 1. Both models struggle comparatively more when predicting Caspase cells, with mAP scores that are consistently lower than their Live and Empty predictions. This could be because caspase cells are harder to identify. There are multiple instances where apoptotic cells, or caspase as they are labelled, are difficult to identify due to debris, cell fragmentation and weak DAB staining.

In general, models that were evaluated on the non-background data subset had higher scores than their all-tiles counterpart. This difference is seen primarily in a 2-3 percent performance reduction for both their mAP and F-measure scores. This is likely due to predictions formed in the background leading to more background false positives. It is interesting to note the performance difference is smaller at $mAP_{90}$ which makes sense given the large number of bounding boxes that are now considered false positives.

# 4. Conclusions and Future Work

The main goal of this project was to determine the suitability of state-of-the-art object detectors for automating osteocyte viability detection. It also sought to explore the effect of imperfect data and how they could affect results. Overall, both RetinaNet and Faster R-CNN have the potential to locate and classify osteocytes, with comparable results. Further evaluation is required to confirm their true accuracy. While asserting that either model has achieved human-level accuracy is unlikely due to their lower $\text{mAP}_{50}$ scores, there are potential cases where the model may have outperformed the expert by identifying osteocytes the expert missed and predicting bounding boxes that are more on the centre than the original annotations. Therefore, it is impossible to measure how much worse the models perform in comparison to the human assessor. The models may still be able to accelerate the process of osteocyte viability quantification, but they require further training to increase their accuracy. Continual verification with a human evaluator would be necessary. Future work could examine whether using an imperfect model can still boost efficiency of a human expert.

# 5. Acknowledgements

I would like to thank Brett Robertson, Richard Chipper and AIRO for sponsoring this research. I would also like to acknowledge Zeyi Wen, Tim French, Jeremy Leggoe, Amanda Bolt and Kimberlie Hancock for their ongoing support in coordinating my honours year. Finally, I would like to thank Chris Bording for teaching me how to set up and use Kaya.

# 6. References

Australian Bureau of Statistics. (2019). *Twenty years of population change.* https://www.abs.gov.au/ausstats/abs@.nsf/0/1cd2b1952afc5e7aca257298000f2e76.

Australian Institute of Health and Welfare. (2020). *Insufficient physical activity.* https://www.aihw.gov.au/reports/risk-factors/insufficient-physical-activity/contents/insufficient-physical-activity.

Australian Orthopaedic Association and National Joint Replacement Registry. (2019). *Hip, Knee & Shoulder Arthroplasty.* https://aoanjrr.sahmri.com/annual-reports-2019

Bonewald, L. F. (2011). The amazing osteocyte. *Journal of Bone and Mineral Research: The Official Journal of the American Society for Bone and Mineral Research, 26*(2), 229–238, https://doi.org/10.1002/jbmr.320

Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S. and Fuchs, T. J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine, 25*(8), https://doi.org/10.1038/s41591-019-0508-1.

Kreisler, M., Kohnen, W., Marinello, C. , G¨otz, H., Duschner, H., Jansen, B., and d'Hoedt, B. (2002). Bactericidal effect of the er:yag laser on dental implant surfaces: An in vitro study. *Journal of periodontology, 73*(11), 1292–1298. https://doi.org/10.1902/jop.2002.73.11.1292.

Lin, T.-Y. , Goyal, P., Girshick, R., He, K. and DollÅLar, P. (2020). Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(2), https://doi.org/10.1109/TPAMI.2018.2858826

Pandey, R. K. and Panda, S. S. (2013). Drilling of bone: A comprehensive review. *Journal of Clinical Orthopaedics and Trauma, 4*(1), 15–30, https://doi.org/10.1016/j.jcot.2013.01.002.

Ren, S., He, K., Girshick, R. and Sun, J. (2015) Faster R-CNN: Towards real-time object detection with region proposal networks, https://arxiv.org/abs/1506.01497

Van der Laak, J., Litjens, G. and Ciompi, F. (2021) Deep learning in histopathology: the path to the clinic. *Nature Medicine*, *27*(5), https://doi.org/10.1038/s41591-021-01343-4.