CCMCT

(Canine Cutaneous Mast Cell Tumor – 개 피부 비만 세포 종양)

분석, 학습 보고서

영우 글로벌 러닝
데이터기반 인공지능 시스템 엔지니어 양성 과정
2기 교육생
이현수

https://github.com/hslee4716/CCMCT.git

데이터셋 배경

Abstract

We introduce a novel, large-scale dataset for microscopy cell annotations. The dataset includes 32 whole slide images (WSI) of canine cutaneous mast cell tumors, selected to include both low grade cases as well as high grade cases. The slides have been completely annotated for mitotic figures and we provide secondary annotations for neoplastic mast cells, inflammatory granulocytes, and mitotic figure look-alikes. Additionally to a blinded two-expert manual annotation with consensus, we provide an algorithm-aided dataset, where potentially missed

Background & Summary

Microscopy image recognition has seen vast advances in recent years, fostered by the availability of high quality datasets as well as by the application of sophisticated deep learning pipelines. One of the most important topics in the field of microscopy imaging is the classification of cells, typically stained with hematoxylin and eosin (H&E) dye. In this area, one particularly challenging task is the detection of mitotic figures, i.e. cells undergoing division, in tumor tissue. It is commonly accepted that the quantity of mitotic figures is one of the most powerful prognosticators of biological behavior for many tumor types, both in humans[1,2] and animals[3,4,5]. In the field of automatic detection of those mitotic figures, there have been a number of competitions in recent years, e.g. the TUPAC16 challenge[6], the ICPR MITOS-2012[7] and ICPR MITOS-ATYPIA-2014 challenge[8].

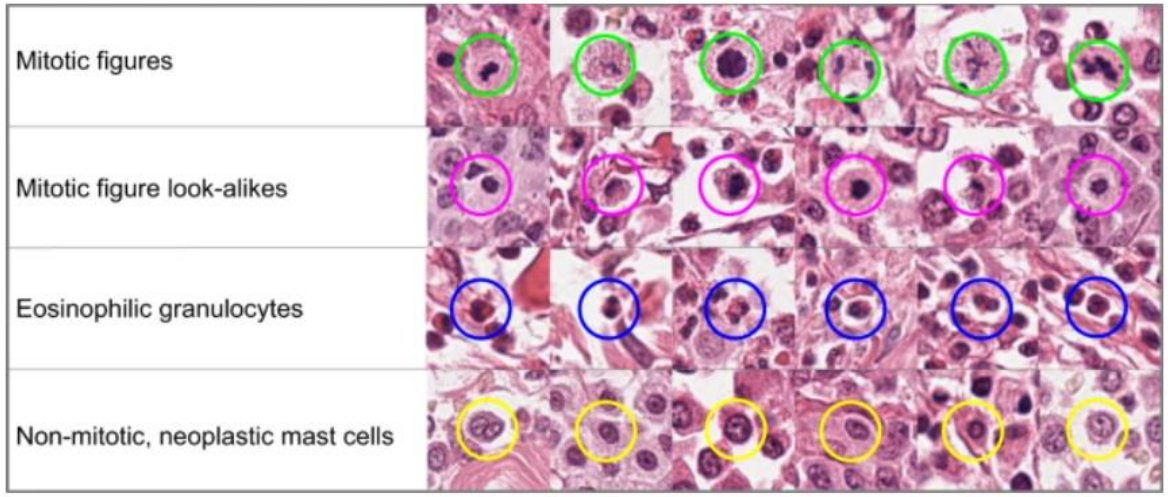
Abstract

Kaggle 데이터셋에는 훈련 세트로 21개의 slide image만 포함되어 있다.

논문에서, 모든 slides들은 모든 유사분열체 (뿐만 아니라 과립구, 암세포, 유사 분열체 의심 세포 등...)에 대해 annotated 되어 있다고 설명한다.

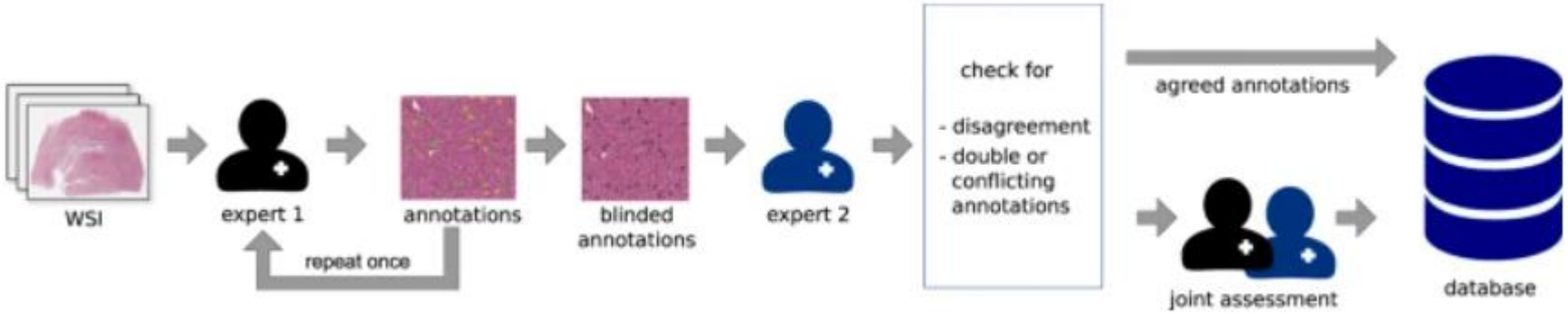
Background

유사 분열체의 개수는 사람, 동물 상관 없이 많은 종양 유형과 밀접한 관계가 있다고 설명하고 있다.



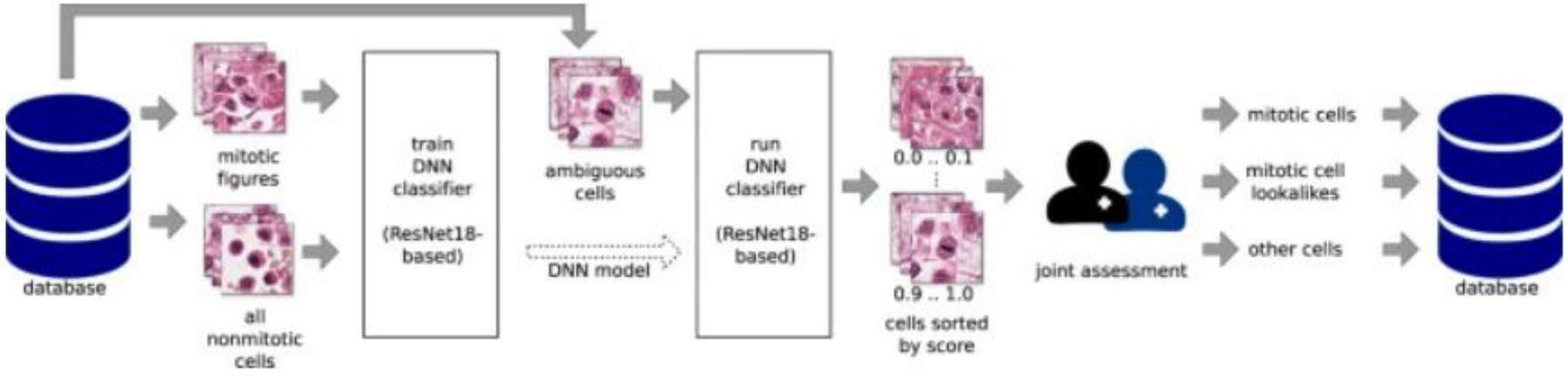
데이터셋 생성 과정

베이스 데이터셋 (MEL- manually, expert-labelled dataset – 전문가 수동 라벨 데이터셋)



모호 클래스 데이터셋 결정 과정(HEAL - Hard-example augmented expert labelled dataset variant)

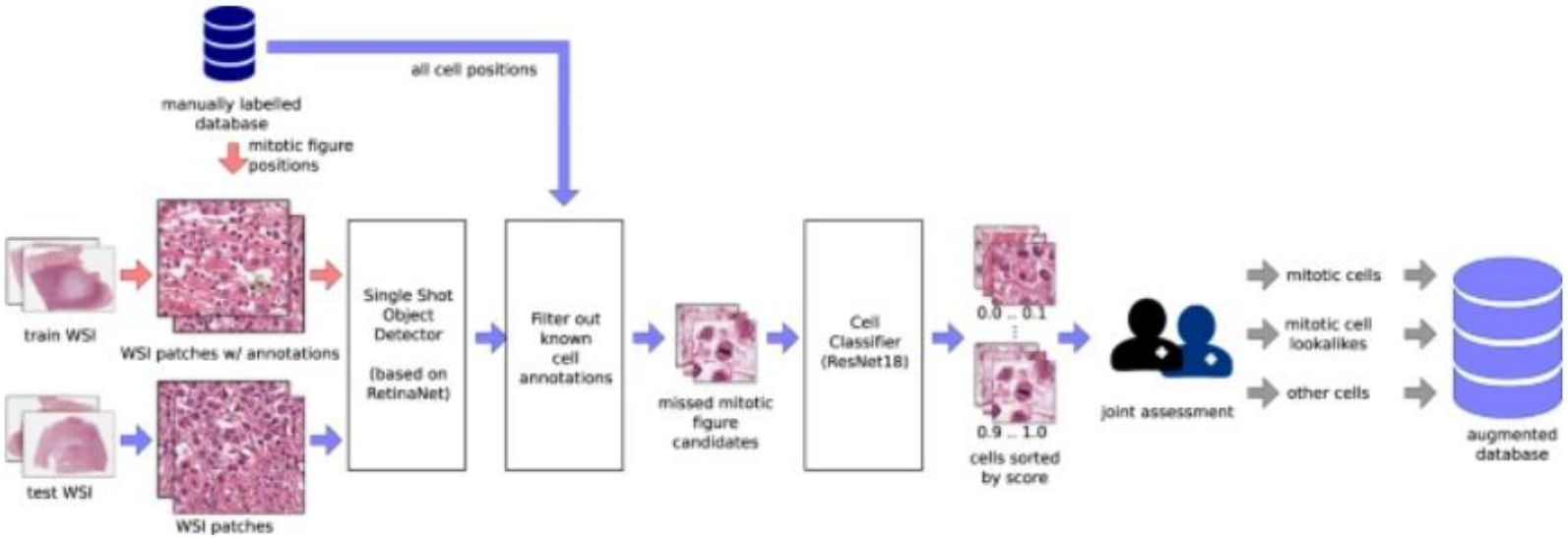
- AI활용 증강 데이터셋

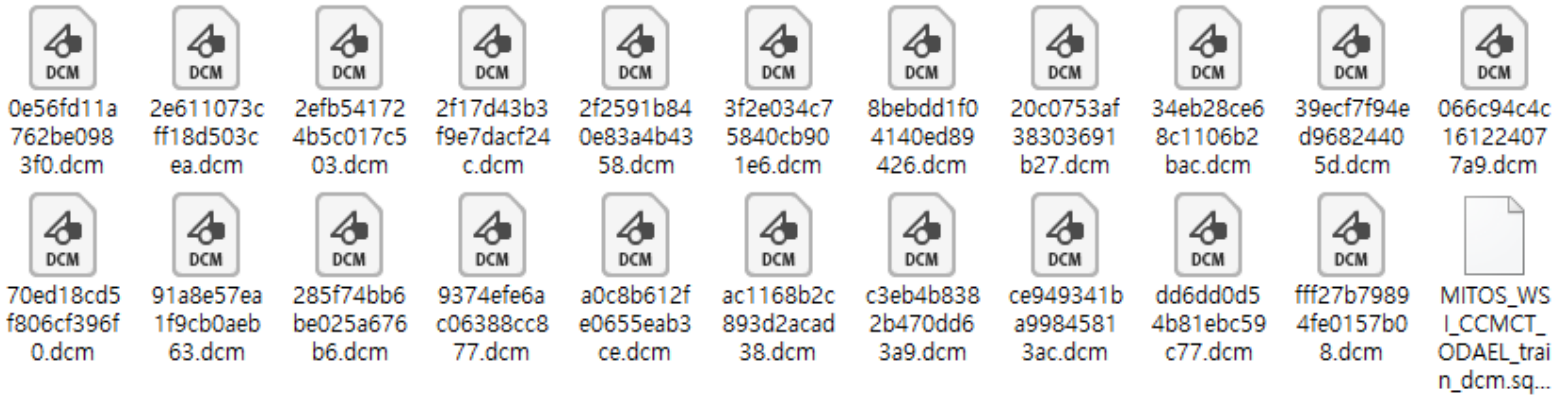


잠재적 유사분열 클래스 데이터셋 결정 과정

-(ODAEL - Object-detection augmented expert labelled dataset variant)

- 잠재적 유사 분열 figure을 DeepLearning을 이용해 추가로 제안, 전문가가 데이터 세트의 다른 그룹으로 등급을 지정하고 할당.
- Mitotic cell lookalikes, other cells 라벨 추가

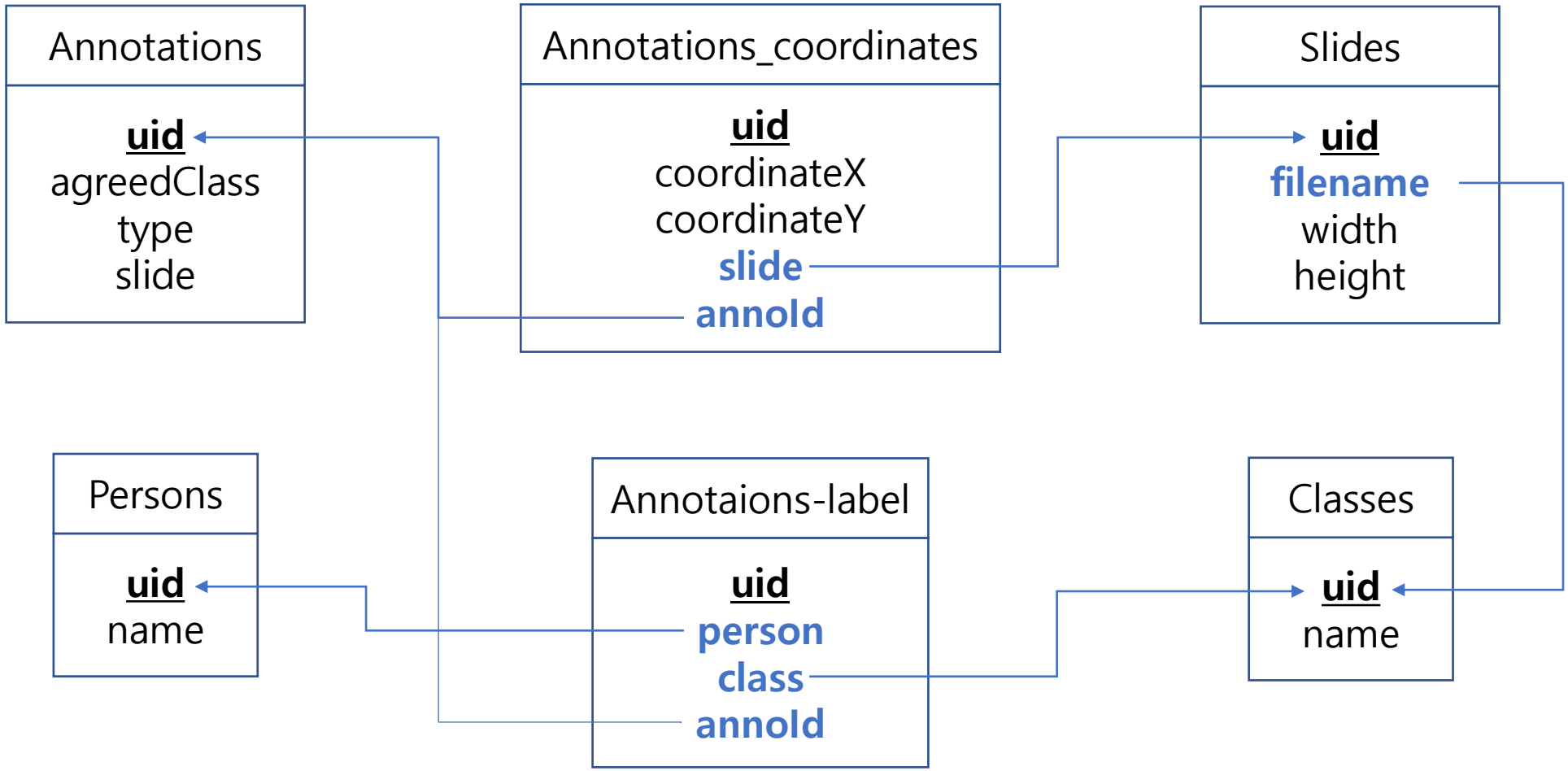




총 21개의 대용량 slide 이미지(.dcm) 와 annotation이 포함된 데이터 베이스 파일로 나뉘어진다.

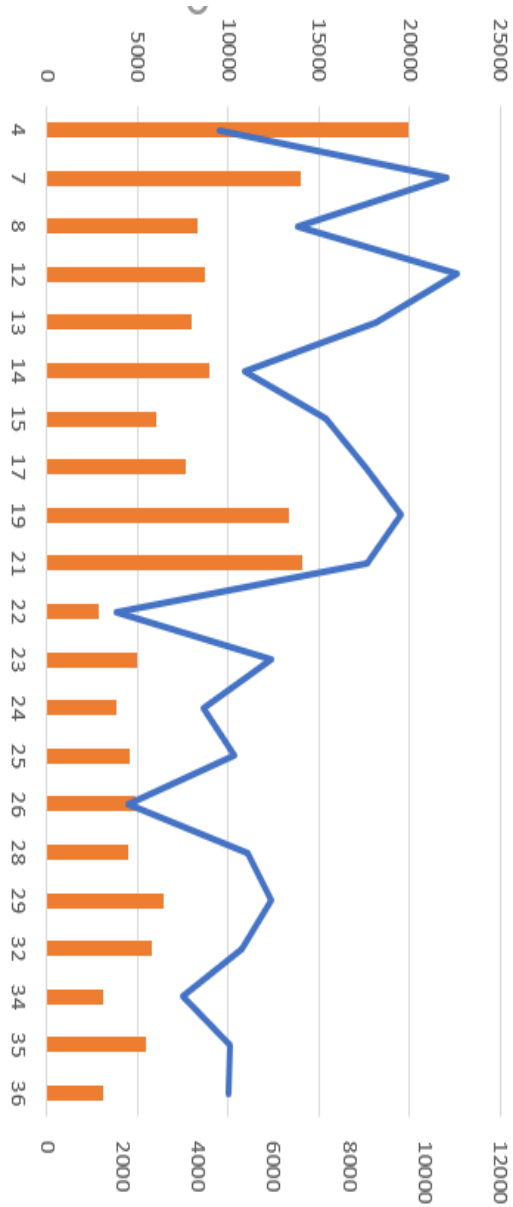
Database Schema

Unique key
Foreign Key



EDA

Slide 번호	width	height	Tiles(1000 x 1000)	Abnormal cells
4	82799	54996	4565	19936
7	113399	92597	10602	14040
8	80999	81620	6642	8314
12	127799	84057	10880	8717
13	104399	82678	8715	7991
14	77399	66558	5226	8986
15	88199	82404	7387	6017
17	113399	73675	8436	7656
19	100799	92632	9393	13390
21	109799	76571	8470	14057
22	23399	76964	1848	2901
23	89999	65400	5940	4997
24	82799	49066	4150	3864
25	55799	88262	4984	4622
26	52199	40646	2173	4889
28	88199	59153	5340	4541
29	77399	75215	5928	6457
32	68400	74760	5175	5776
34	70199	50863	3621	3153
35	62999	76054	4851	5454
36	71999	66251	4824	3154
			총	158912

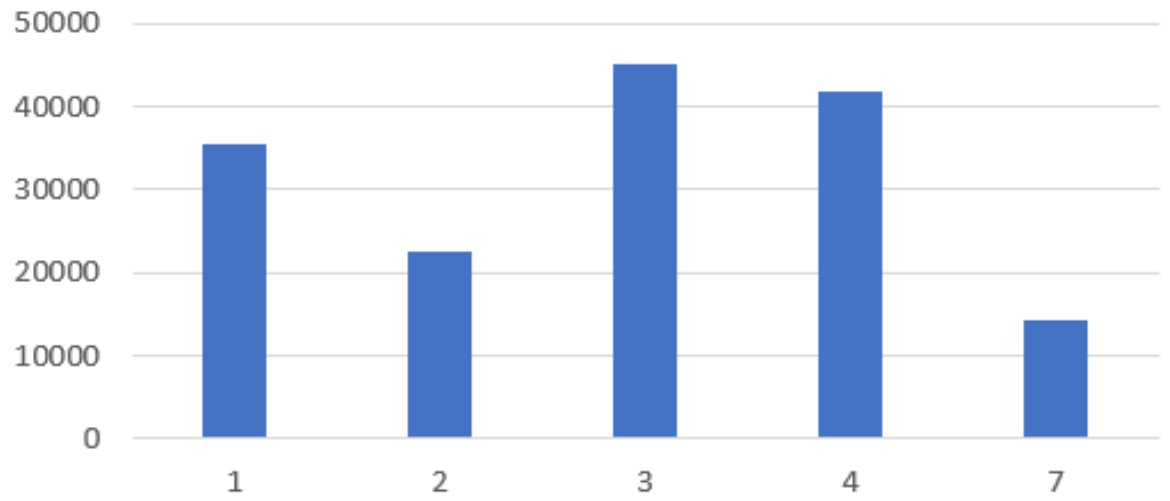


EDA

Classes	Cells
1	35331
2	22404
3	45179
4	41658
5	0
6	0
7	14340
합계	158912

1	granulocyte
2	mitotic figure
3	tumor cell
4	other/ambigous cells
5	binucleated cell
6	multinukleated cell
7	Mitotic figure lookalike

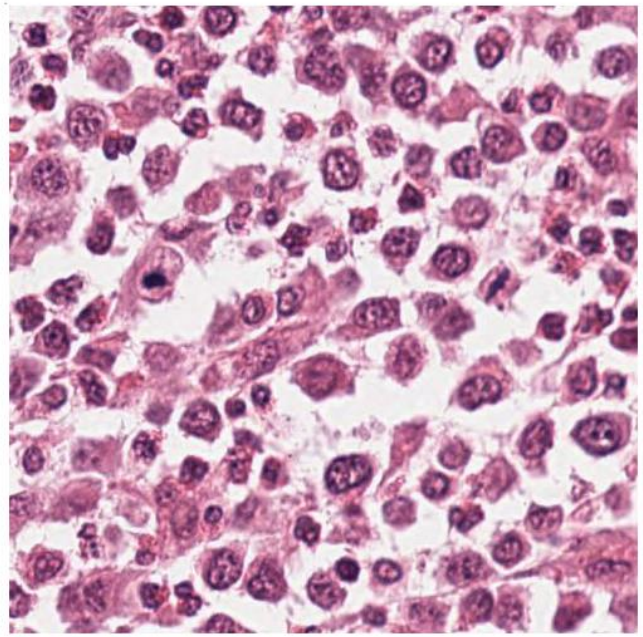
Binucleated cell, multinucleated cell 은 train set(제공받은 Kaggle data)에는 포함되어 있지 않다.



Tiling, Labeling

In [4]:
```
ds = ReadableDicomDataset('/kaggle/input/mitosis-wsi-ccmct-training-set/fff27b79894fe0157b08.dcm')
location=(69700,17100)
size=(500,500)
img = Image.fromarray(ds.read_region(location=location,size=size))
img
```

Out[4]:



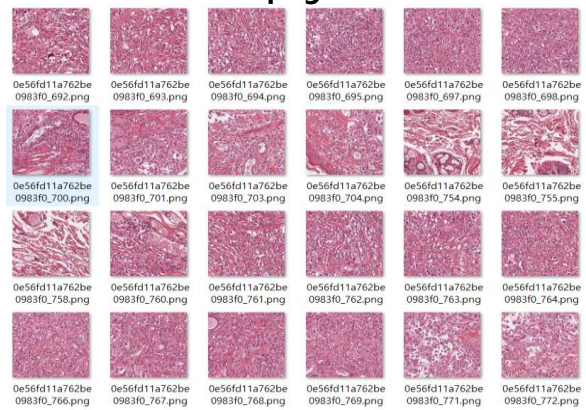
왼쪽 코드를 이용해 엄청나게 큰 dcm 포멧 파일에서 특정 타일을 조회할 수 있고, 아래 코드로 sqlite파일의 데이터셋에 접근해 정보를 얻을 수 있다.

In [2]:
```
import sqlite3
DB = sqlite3.connect('/kaggle/input/mitosis-wsi-ccmct-training-set/MITOS_WSI_CCMCT_ODAEL_train_dcm.sqlite')
cur = DB.cursor()
```

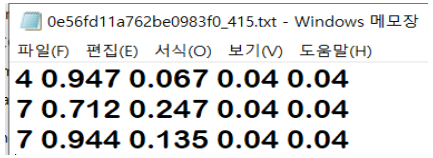
```
cells = cur.execute(f"""SELECT coordinateX-{location[0]}, coordinateY-{location[1]}, annoId
                        from Annotations_coordinates where slide=={slide[0]} and
                        coordinateX>{location[0]} and coordinateX<{location[0]+size[0]} and
                        coordinateY>{location[1]} and coordinateY<{location[1]+size[1]}""").fetchall()
```

Large Scale Images
.DCM

Tiled Images
.png



labels
.txt



전 슬라이드의 코드와 dcm 이미지, sqlite파일을 활용해 원하는 크기의 Yolo_v5 학습 데이터를 만드는 모듈을 제작했으며 여러 resolution(320 * 320, 640 * 640, 256 * 256)의 trainset을 만들어 model을 학습시켰다.

이 때, 이미지의 모든 tile을 저장한다면 320*320사이즈의 png파일이 약 124만장 가까이 나오며 annotation이 없는 이미지만 100만장이 넘게 나온다. 때문에 효율성을 위해 sql을 먼저 조회해 이상 세포가 존재하는 tile만 가져오도록 하였으며 12시간 가까이 하던 이미지 분할 시간을 2 ~ 3시간으로 단축시킬 수 있었다.

Train 결과

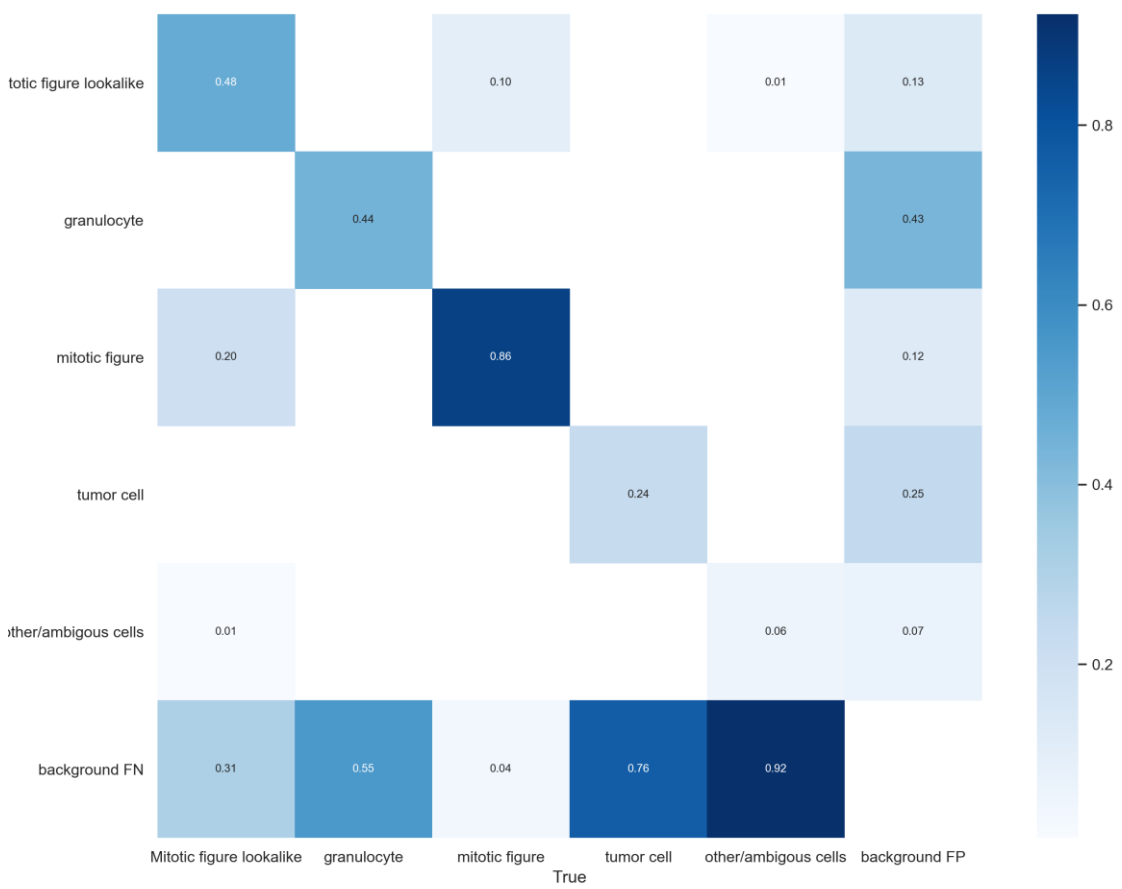
Actual	pred. mitotic fig.	pred. mitotic fig. look-alike	pred. granulocyte	pred. tumor cell
Mitotic figure	19478	2985	10	3
Mitotic figure look-alike	2942	10582	57	44
Granulocyte	1	66	16011	30
Tumor cell	3	92	53	20651

위 표는 데이터에 대한 설명이 포함된 논문에서 ImageNet으로 학습된 weights를 포함한 ResNet-18 베이스의 CNN 모델을 훈련시켜 나온 결과이며 91.390%의 Accuracy를 달성했다고 설명한다.

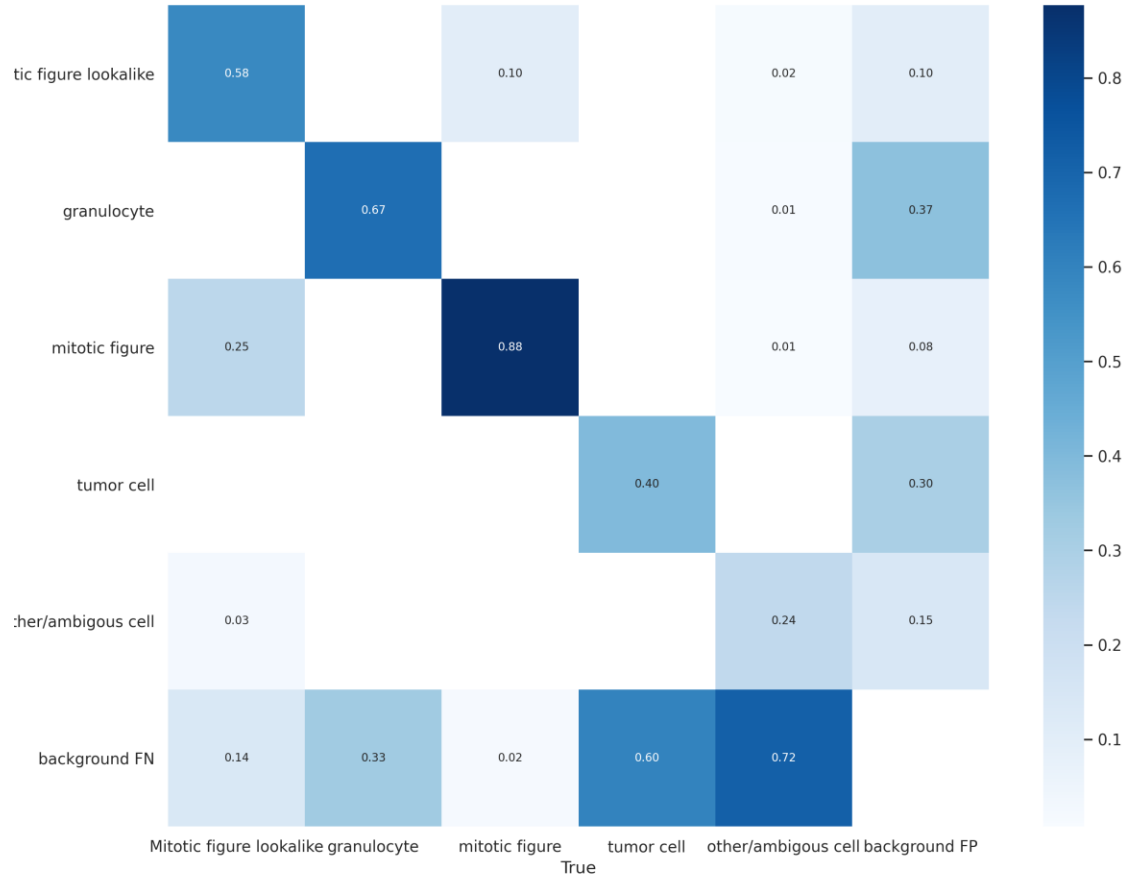
이번 프로젝트에서는 해당 수치를 목표로 두고 학습을 진행했다.

Train Result

- imgsz: 640 / train_imgsz: 640
- epochs: 100 / weight: yolov5m.pt
- 5Labels

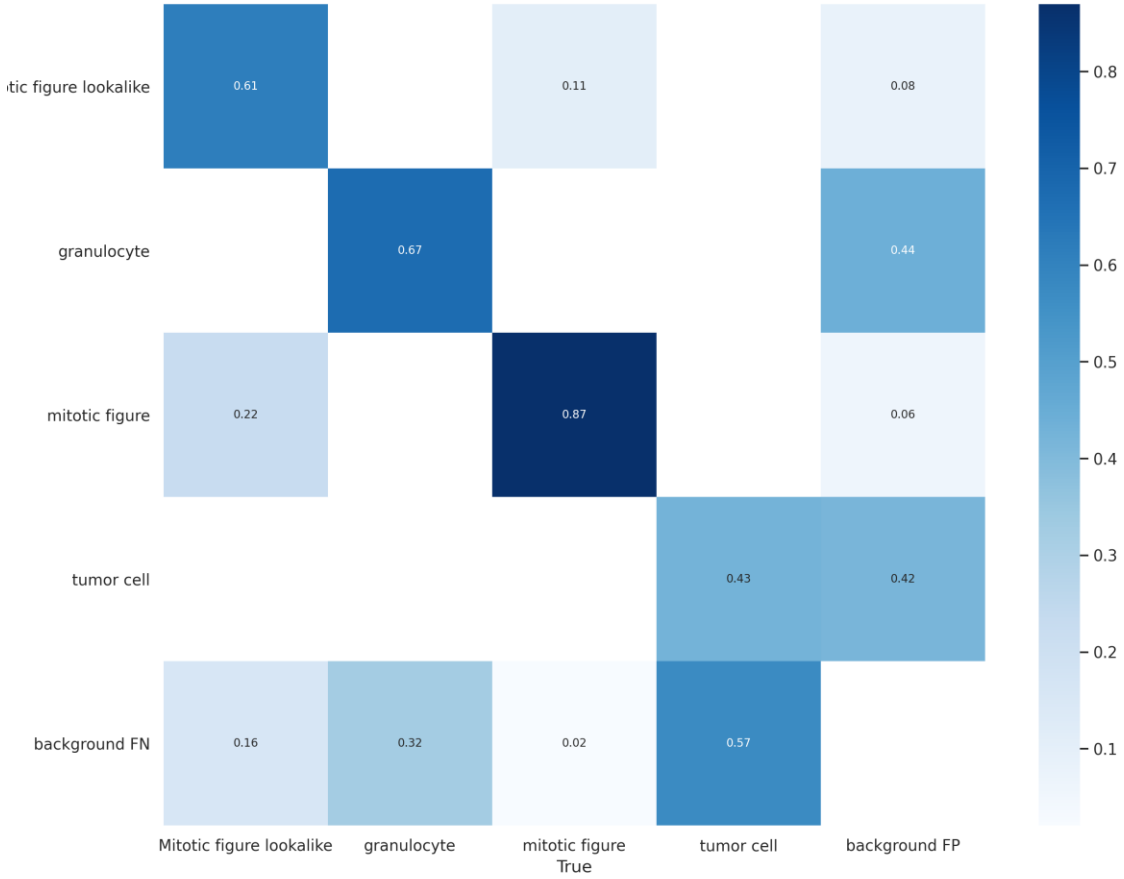


- imgsz: 320 / train_imgsz: 320
- epochs: 250 / weight: yolov5s.pt
- 5Labels

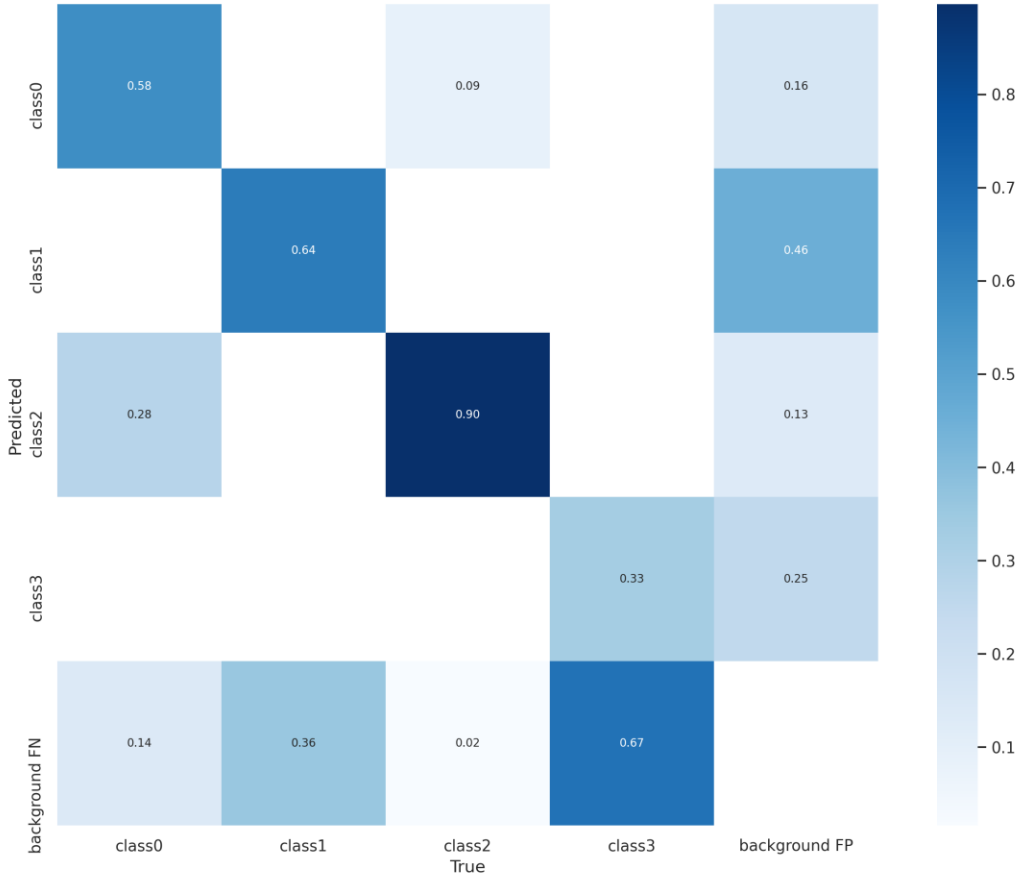


Train Result

- imgsz: 320 / train_imgsz: 640
- epochs: 200 / weight: yolov5s.pt
- 4Labels



- imgsz: 320 / train_imgsz: 320
- epochs: 120 / weight: yolov5l.pt
- 4Labels



Summary

- 총 10번 이상의 학습 중 라벨 수나, 원본 데이터셋에 변화를 주거나 했던, 큰 변화를 주었을 때 4가지 훈련의 confusion matrix이다.

- 이미지가 클수록 이미지 크기 대비 작은 객체 탐지에서 불리하다. 따라서 원본 이미지 사이즈를 640->320으로 줄였을 때 BackgroundFN(잘못된 바운딩 박스)비율이 확연히 줄었다.

- 320*320 사이즈의 이미지를 640*640이미지로 resizing해서 학습시킨 결과 그냥 학습 시킨 것과 큰 차이가 없었다.

- Annotation을 가지고 있지 않은 tile의 이미지를 포함해 학습시켰을 때, tumor cell의 FP rate는 확연히 줄었지만, 나머지에선 큰 변화가 없었고, 오히려 성능이 떨어졌다. -> 이때 epochs가 120이었는데, 더 늘려서 학습 시켜볼 필요성이 있다고 판단된다.

- 원래 target인 Mitotic cell의 경우 논문의 결과와 유사한 accuracy를 달성했지만 나머지의 경우는 크게 떨어진다. -> Background FN, FP를 줄이는 방향으로 간다면 도달할 수 있을 것이라 생각된다

회고

- 원본이 .dcm파일이었는데 해당 파일을 png파일로 tiling할때 짧으면 2시간 ~ 길면 12시간까지 소요됐다. 코드를 최적화 시키며 많이 나아 졌지만 초반에 이때문에 시간이 많이 소요됐고, 미리 train set 생성 모듈을 완벽하게 만들어 놓는 것에 대한 필요성을 많이 느낄 수 있었다.

- 이번 프로젝트에서는 한 모델만 가지고 하이퍼파라메터와 trainset만 바꿔가며 학습시켰는데 만족스런 결과는 나오지 않았다. 더 많은 모델을 가용해서 학습시켜볼 필요성이 있다.

- 익숙하지 않은 domain이어서 그런지 초반 갈피를 잡는데 시간이 많이 걸렸다.

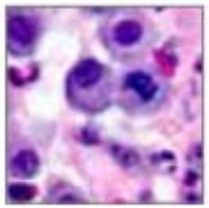
서비스화

개 비만 세포 종양 (CMCT)

예후, 기수 판별에 MC(Mitotic count)가 주요 요소로 작용

MC - 현미경 10배율 당 유사분열체 수
AgNOR - 세포주기 진행속도
Ni67 - 성장을 식별하는 면역조직화학(IHC)방법 분수
AG67 - AgNor, Ni67 측정에서 얻을수 있는 인자
…

Mitotic count
(previously referred to as "mitotic index")



Actively dividing neoplastic mast cell

- MC < 7: indicates a low-grade CMCT (Kiupel system):
  - MST: > 2 years
  - 5% of dogs died due to MCT-associated disease
  - 20% developed additional MCTs
- MC ≥ 7: indicates a high-grade CMCT (Kiupel system):
  - MST: < 4 months
  - 90% of dogs died due to MCT-associated disease
  - 70% developed metastasis

- MC > 5:
  - MST of approximately 2 months (compared to 70 months in case of CMCTs with MC lower than 5)
- 91% specificity of identifying aggressive CMCTs (with 79% diagnostic accuracy)

서비스화

현재 데이터에서 MC(Mitotic count) 측정

해당 데이터셋의 경우 640 * 640사이즈로 잘랐을 때 평균 몇 개의 Mitotic cell이 관측되는지 계산하면 MC를 얻을 수 있다.

was performed by a linear scanner (ScanScope CS2, Leica, Germany) in one focal plane by default settings at a magnification of 400x (image resolution: 0.25 *μm*/pixel), using an Olympus UPlanSAPO 20x lens (field number = 26.5, numerical aperture = 0.75).

Manually expert labelled (MEL) dataset

Area per high-power field for some microscope types:

- **Olympus** BX50, BX40 or BH2 or AO: 0.096 mm^2[1]
- **AO** with 10x eyepiece: 0.12 mm^2[1]
- **Olympus** with 10x eyepiece: 0.16 mm^2[1]
- **Nikon Eclipse E400** with 10x eyepiece and 40x objective: 0.25mm^2[2]
- **Leitz Ortholux**: 0.27 mm^2[1]
- **Leitz Diaplan**: 0.31 mm^2[1]

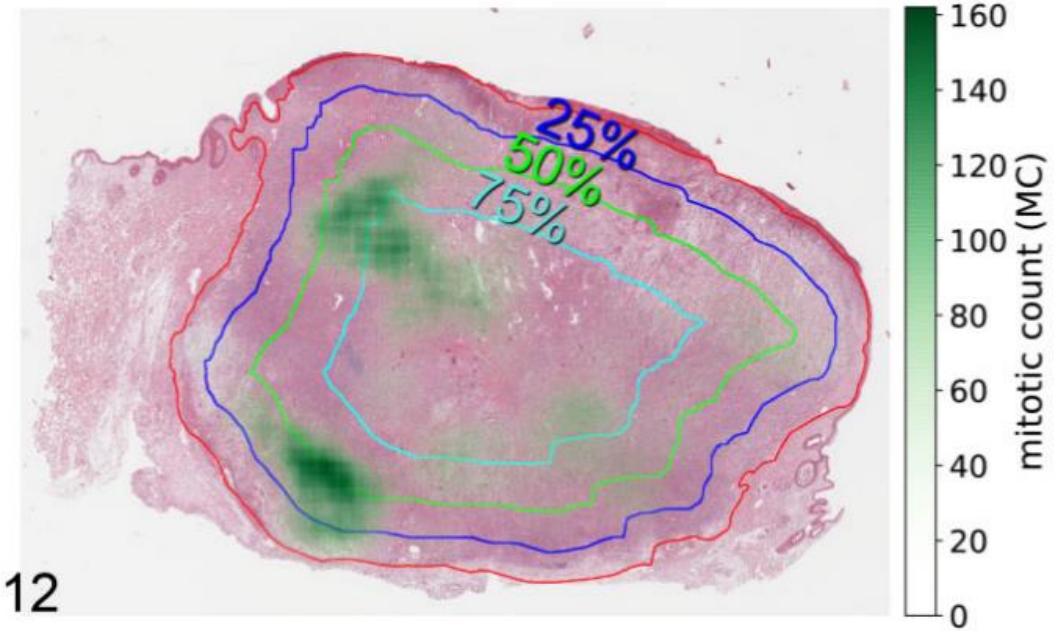
640 x 640당 평균 몇 개의 Mitotic cell 이 있는가?

서비스화

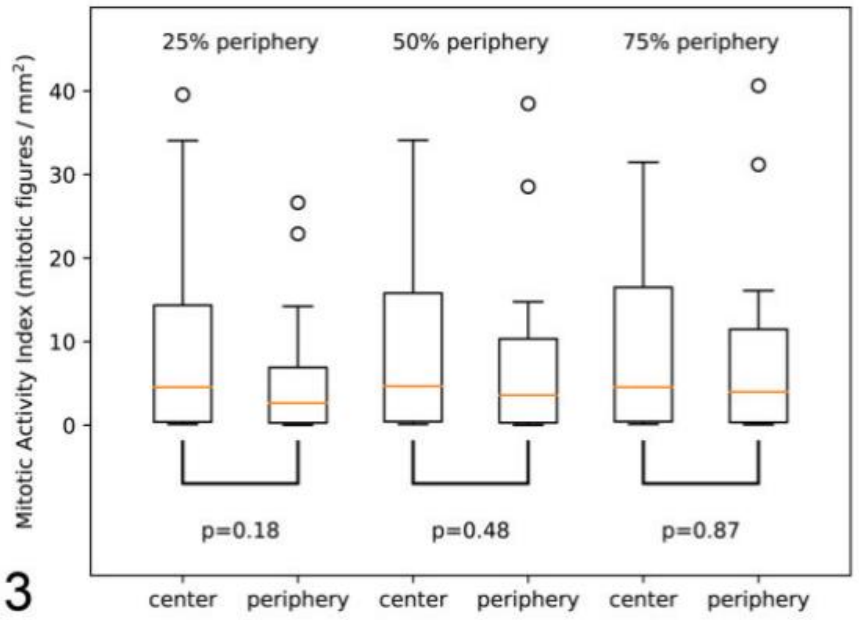
특정 범위에 Mitotic Cell이 얼마나 분포되어 있는지도 기수 판별의 주 요소

-> 모든 이미지를 tiling하여 detection시키고, Mitotic cell이 검출된 tile을 표시하여 아래 이미지와 같이 표시할 수 있다.

-> 의사가 각 tile을 직접 보고 하나하나 표시하지 않아도 MC와 분포를 알 수 있으며 종양의 상태를 확인하는데 중요한 정보를 제공할 수 있다.



12



13

출처

https://www.nature.com/articles/s41597-019-0290-4

https://www.kaggle.com/marcaubreville/first-steps-with-the-mitos-wsi-ccmct-data-set

논문

Histopathology and prognostic panels to aid in the diagnosis and management

canine-mct.pdf

Computerized Calculation of Mitotic Count Distribution in Canine Cutaneous Mast Cell Tumor Sections: Mitotic Count Is Area Dependent