USING MACHINE LEARNING TO AUTOMATE CERVICAL PRE-CANCER SCREENING

DEMELZA ROBINSON,[1] KEVIN HOONG,[1] W. BASTIAAN KLEIJN,[1] ALEXANDER DORONIN,[1] JEAN REHBINDER,[2] JEREMY VIZET,[2] ANGELO PIERANGELO,[2] AND TATIANA NOVIKOVA[2]

PROBLEM

DETECTING CERVICAL CANCER IN EARLY PRE-MALIGANT PHASE (CIN3) CAN IMPROVE SURVIVAL.

CURRENT SCREENING MEASURES LACK SENSITIVITY AND RELY ON HUMAN EYE.

SOLUTION

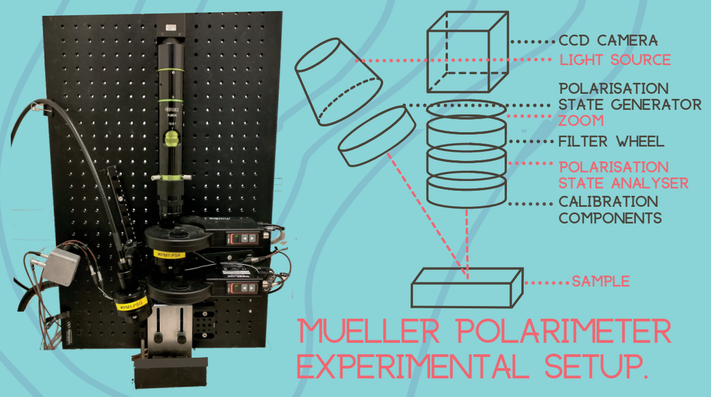
NOVEL OPTICAL TECHNIQUE, MUELLER POLARIMETRY, CAN DETECT MICRO-STRUCTURAL TISSUE CHANGES ASSOCIATED WITH PRE-CANCER.

USE MACHINE LEARNING TO RELATE THESE CHANGES TO GOLD STANDARD HISTOPATHOLOGY LABELS.

DATASET

2 DATA TYPES TO RELATE:

- HISTOPATHOLOGICAL LABELS HEALTHY VS. PRECANCEROUS (CIN3).
- MUELLER MATRICES CONTAINING TISSUE-LIGHT INTERACTION INFO.



MUELLER POLARIMETER EXPERIMENTAL SETUP.

DATA SPLITTING

2 APPROACHES COMPARED:

- **90:10 TRAIN:TEST SPLIT** RANDOMLY SHUFFLE ALL PIXELS AND SPLIT INTO TRAIN (90%) + TEST (10%) SETS. REPEATED 30 TIMES THEN AVERAGED.
- **LEAVE-ONE-OUT CROSS-VALIDATION** EACH TISSUE SAMPLE WITHELD FROM TRAINING AS TEST SET, RESULTS AVERAGED.

ML MODELS

3 METHODS COMPARED:

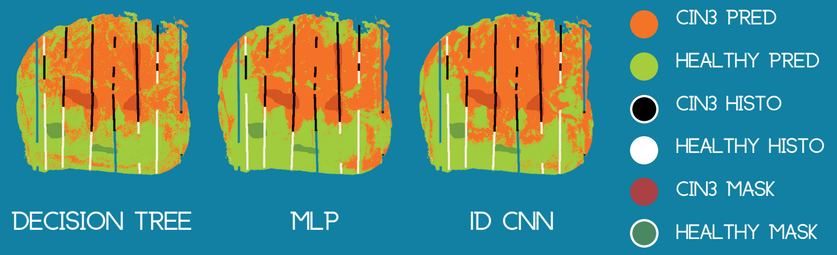
- DECISION TREE **DT**
- MULTI-LAYER PERCEPTRON **MLP**
- 1D CONVOLUTIONAL NEURAL NETWORK **1D CNN**

QUANTITATIVE RESULTS

	AUC OR ACCURACY	SPECIFICITY (TNR)	SENSITIVITY (TPR)
	90:10 TRAIN-TEST SPLIT		
DT	0.962±0.00	0.977±0.00	0.946±0.00
MLP	0.986±0.00	0.991±0.00	0.981±0.01
ID CNN	0.964±0.00	0.978±0.00	0.952±0.01
	LEAVE-ONE-OUT CROSS-VALIDATION (LIOCV)		
DT	0.762±0.20	0.822±0.16	0.697±0.23
MLP	0.803±0.23	0.846±0.22	0.756±0.25
ID CNN	0.812±0.21	0.851±0.20	0.763±0.25

MEAN + STANDARD DEVIATIONS ACROSS THE THREE ML TECHNIQUES AND TWO SPLITTING APPROACHES.

QUALITATIVE RESULTS



WHOLE SAMPLE EVALUATION ON UNSEEN DATA. HISTOPATHOLOGY LINES INDICATE CONFIRMED UNDERLYING DISEASED OR HEALTHY TISSUE TO COMPARE MODEL PREDICTIONS WITH.

[1] VICTORIA UNIVERSITY OF WELLINGTON TE HERENGA WAKA
CATALYST FUND Advancing global science partnerships for New Zealand
CAMPUS FRANCE campusfrance.org
[2] ÉCOLE POLYTECHNIQUE UNIVERSITÉ PARIS-SACLAY

CONCLUSION

RESULTS ILLUSTRATE HOW THE CONVENTIONAL 90:10 SPLITTING APPROACH CAN YIELD DECEPTIVELY HIGH PERFORMANCE.

THE MORE REALISTIC LIOCV STILL OBTAINS PROMISING PERFORMANCE BUT SUGGESTS MORE WORK NEEDED BEFORE CLINICAL DEPLOYMENT.

NEXT STEPS

A SHORTCOMING OF THE CURRENT ML APPROACH IS THAT IT CAN'T INDICATE WHEN NEW SAMPLES ARE DISSIMILAR TO THE DATA IT TRAINED ON. WE CAN ADDRESS THIS PROBLEM USING UNCERTAINTY ESTIMATION.

BAYESIAN STATISTICS

WE WILL USE A BAYESIAN APPROACH TO OUTPUT PREDICTIONS AS PROBABILITY DISTRIBUTIONS RATHER THAN POINT ESTIMATES LIKE IN THIS WORK.

- **MODE** = PREDICTED OUTPUT.
- **SPREAD** = MODEL UNCERTAINTY.

THIS WILL REFLECT HOW CONFIDENT OR FAMILIAR THE MODEL IS WITH THE DATA IT IS BEING APPLIED TO WHICH WILL MAKE CLINICAL DEPLOYMENT MORE FEASIBLE.