Vincenzo Schiano Di Cola

Tutors: Nicola Mazzocca, Francesco Piccialli

XXXIV Cycle - III year presentation

Data Science for predictive analysis

FSE
POR CAMPANIA
2014 - 2020

Università degli Studi di Napoli
Federico II

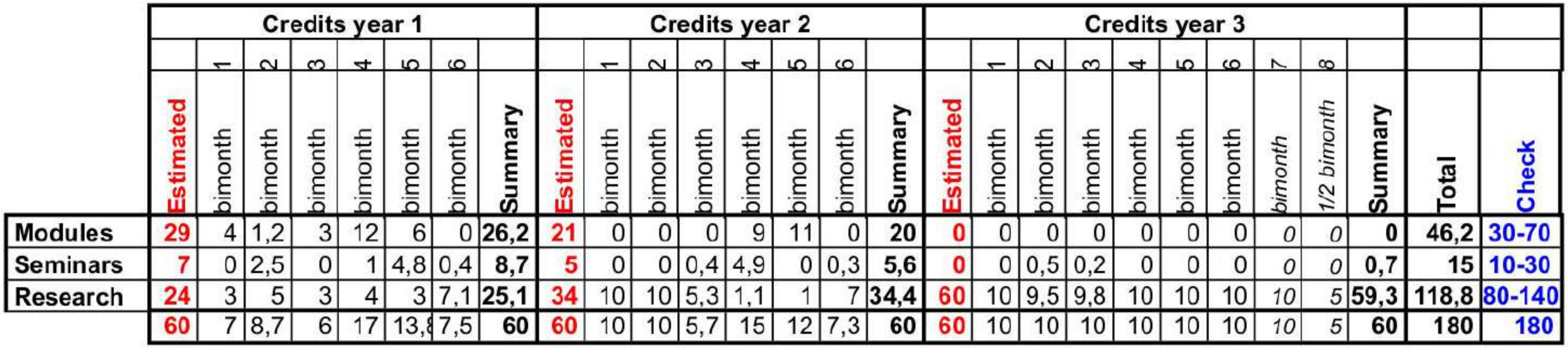
Background

- *Graduation*:
  MS in Mathematics with thesis in Numerical Analysis

- *Cooperations*:
  DATABOOZ ITALIA, Incheon National University (Korea)

- *Fellowship Type*:
  P.O.R. entitled: **Data Scientist for Predictive Analytics**; according to the "Dottorati di ricerca con Caratterizzazione Industriale" - DGR n. 156 del 21/03/2017 - DD n. 155 del 17/05/2018 - within the POR Campania FSE 2014/2020 - Obiettivo Specifico 14 - Azione 10.4.5.

Credits summary



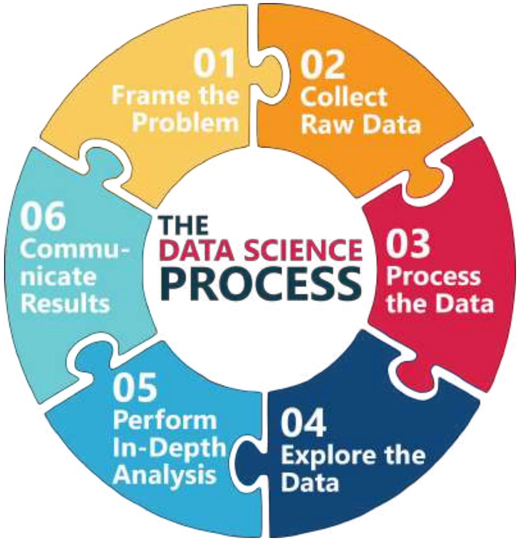
	Credits year 1								Credits year 2								Credits year 3											
		1	2	3	4	5	6			1	2	3	4	5	6			1	2	3	4	5	6	7	8			
	Estimated	bimonth	bimonth	bimonth	bimonth	bimonth	bimonth	Summary	Estimated	bimonth	bimonth	bimonth	bimonth	bimonth	bimonth	Summary	Estimated	bimonth	bimonth	bimonth	bimonth	bimonth	bimonth	*bimonth*	*1/2 bimonth*	Summary	Total	Check
Modules	**29**	4	1,2	3	12	6	0	**26,2**	**21**	0	0	0	9	11	0	**20**	**0**	0	0	0	0	0	0	*0*	*0*	**0**	**46,2**	**30-70**
Seminars	**7**	0	2,5	0	1	4,8	0,4	**8,7**	**5**	0	0	0,4	4,9	0	0,3	**5,6**	**0**	0	0,5	0,2	0	0	0	*0*	*0*	**0,7**	**15**	**10-30**
Research	**24**	3	5	3	4	3	7,1	**25,1**	**34**	10	10	5,3	1,1	1	7	**34,4**	**60**	10	9,5	9,8	10	10	10	*10*	*5*	**59,3**	**118,8**	**80-140**
	60	7	8,7	6	17	13,8	7,5	**60**	**60**	10	10	5,7	15	12	7,3	**60**	**60**	10	10	10	10	10	10	*10*	*5*	**60**	**180**	**180**

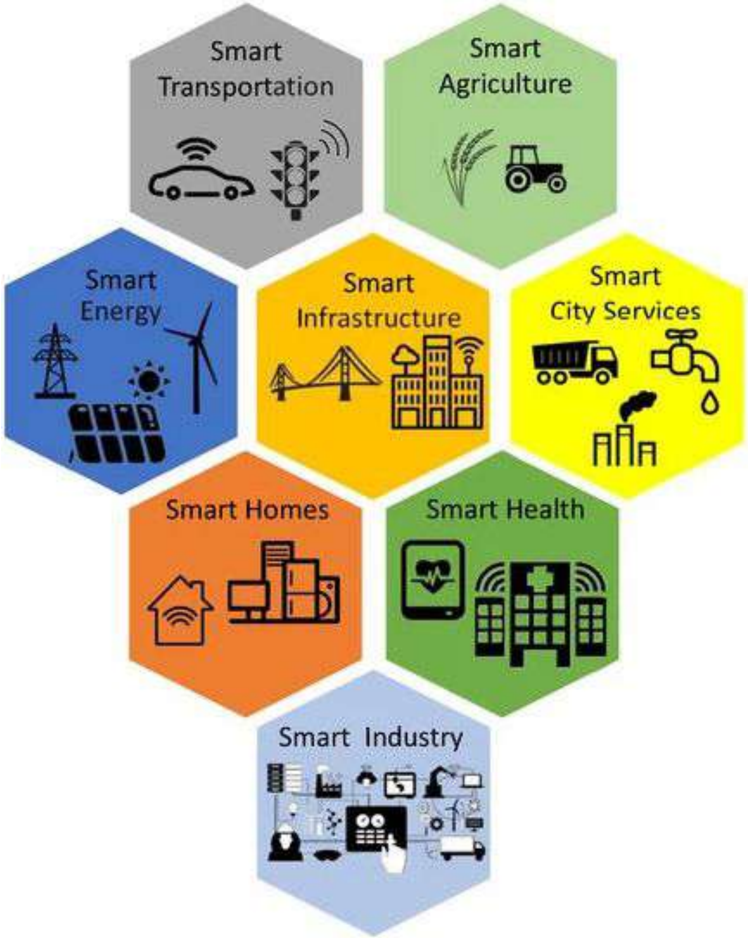
Table of training

- 7 Phd Schools, in particular:
  - eBISS 2019
  - Lipari School 2019: Data Science
  - 16th Reasoning Web Summer School – RW 2020
  - ACDL 2020
  - OxML 2020

- 7 coursers
- 16 Seminars

Research context -
Data Science for Smart Cities

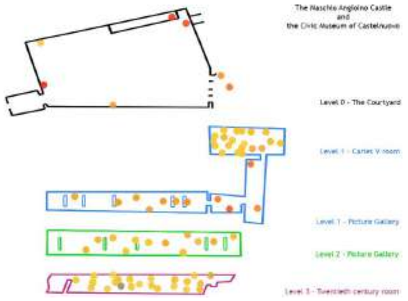




Research problems: Where?

Domains:

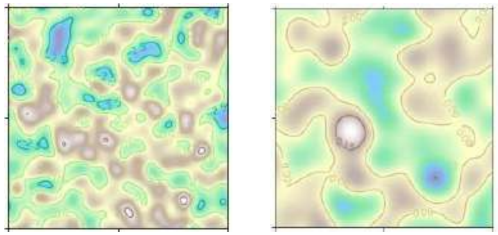
- Cultural Heritage:
  - Clustering of visitor’s paths behavior
  - Give insights of visitor’s behavior to museums’ decision-makers
  - Predict visitors movement


The Maschio Angioino Castle
and
the Civic Museum of Castelnuovo
Level 0 - The Courtyard
Level 1 - Carlet V room
Level 1 - Picture Gallery
Level 2 - Picture Gallery
Level 3 - Twentieth century room

- E-health:
  - link prediction through a “Knowledge Graph” data structure, of medical prescriptions and booking appointments
  - biosensors for Point-of-Care Tests (POCT)

ASL
REGIONE CAMPANIA
Azienda Sanitaria Locale Napoli 1

- Geoscience:
  - local uncertainty estimation for diffusive models



Vincenzo Schiano Di Cola

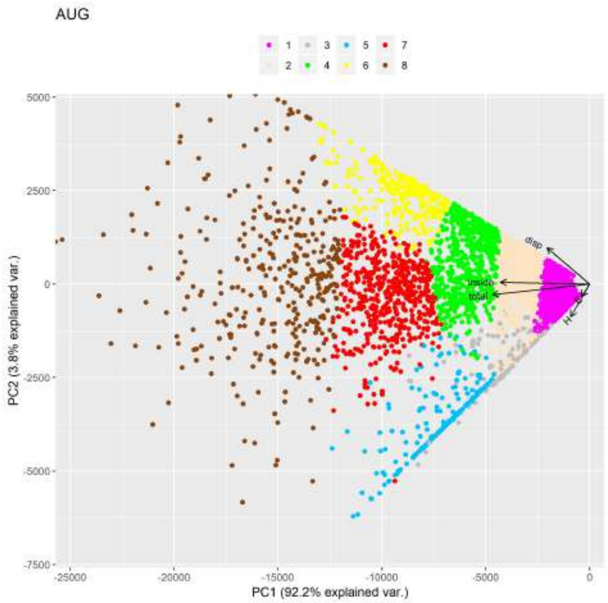
Research problem: How?

1. **Data ETL**
    - accessing the data source
    - basic investigation of the data source
    - changing the data, so it can be readily worked with, using some Data Cleansing
    - make the data available to downstream exploration and analytics processes
2. **Data Exploration**
    i. Load data
    ii. Explore data
    iii. In-depth statistical measures and visualizations

Data Exploration gives statistics and visualization on Data Set to identify appropriate columns for modelling, data quality issues (e.g. missing values, ...) and anticipate potential feature changes that may be required. Assess how important is a certain measurement (e.g. utilising correlation matrix).

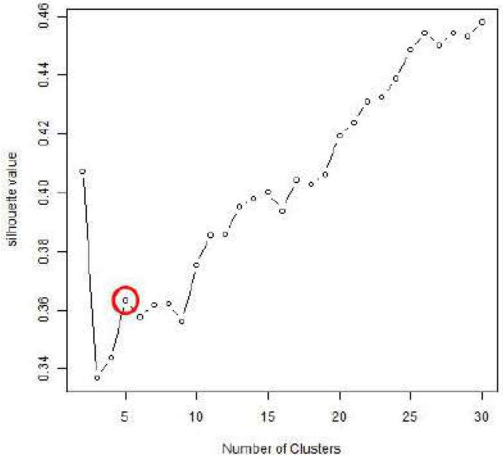
3. **Feature Engineering**
    Depends on data and model. E.g. time, strings, graph, images properties, etc.
3. **Modeling (model Definition, Training, Evaluation and Deployment)**
    Eg. ensemble model such as combining Logistic Regression model and NN model.
3. **Result export and data visualization**

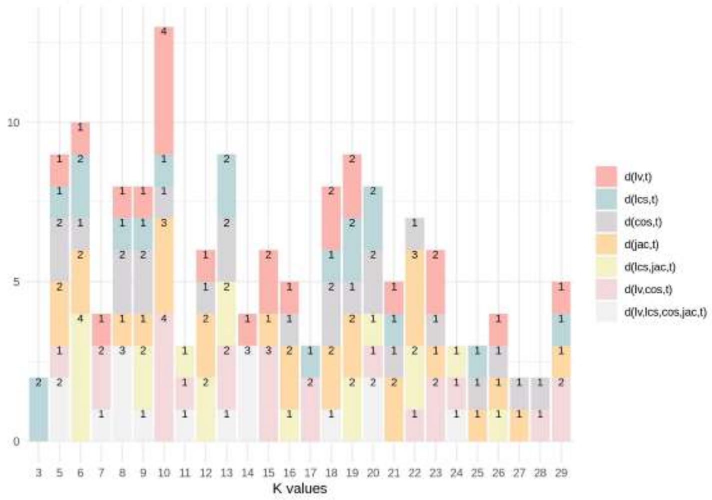
Research problem: How?



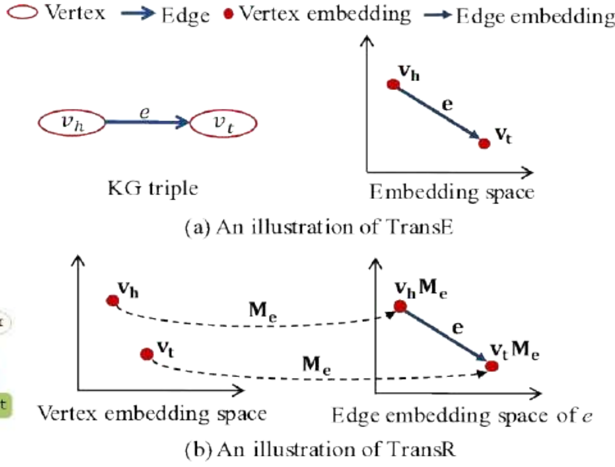
Cultural Heritage:

- ○ Clustering (K-means, D-SCAN, HC)
- ○ Elbow, Silhouette method, majority rule
- ○ PCA





Research problem: How?

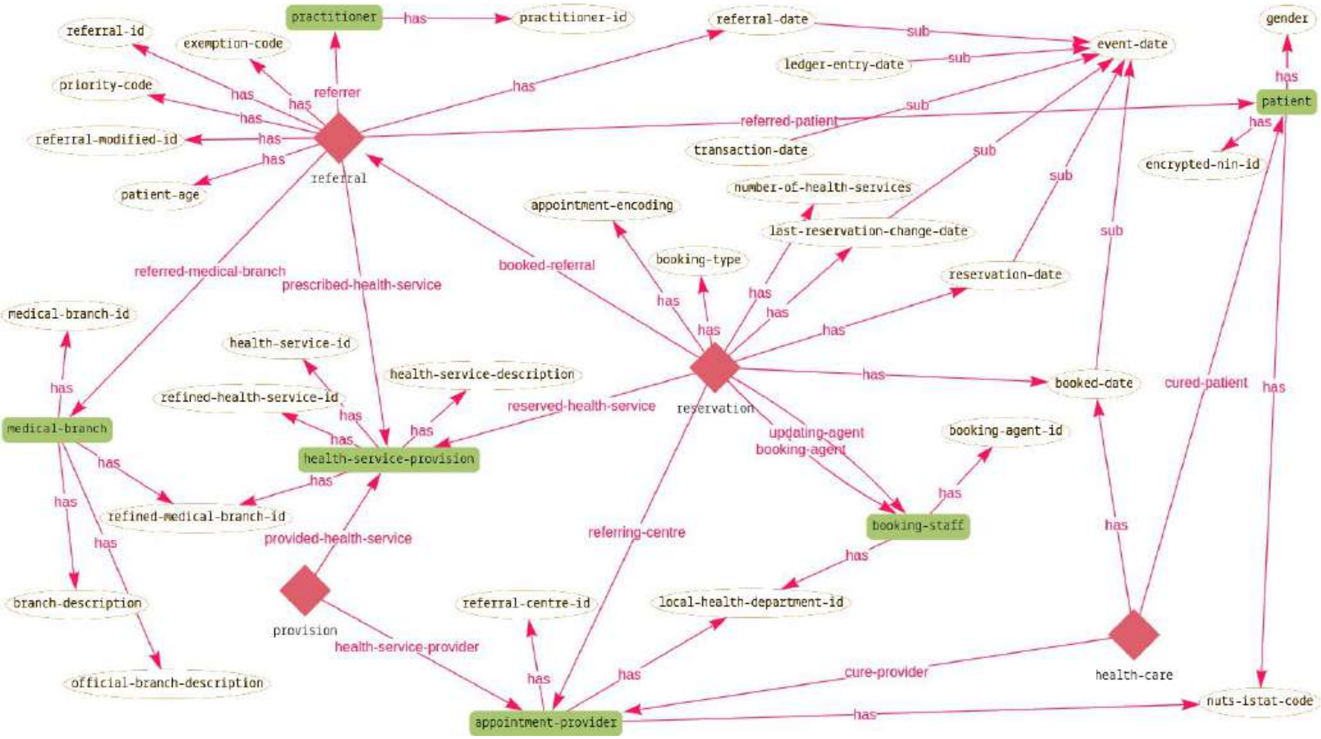


(a) An illustration of TransE

(b) An illustration of TransR

E-Health:

- ○ Knowledge Graphs and Graph Embedding
- ○ Classification
- ○ Object Identification (for images)



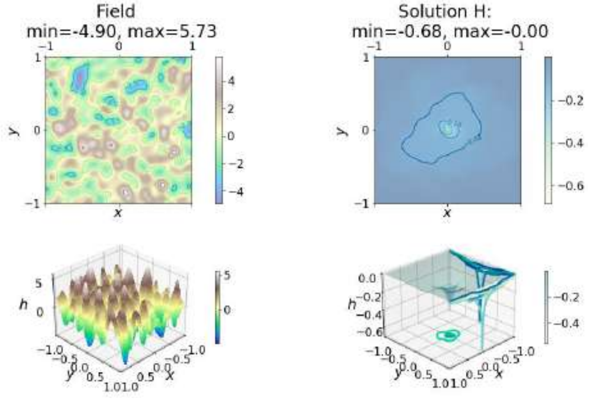
Research problem: How?

Geoscience:

- ○ Monte Carlo simulation
- ○ Diffusion simulations
- ○ Physics Informed Neural Networks

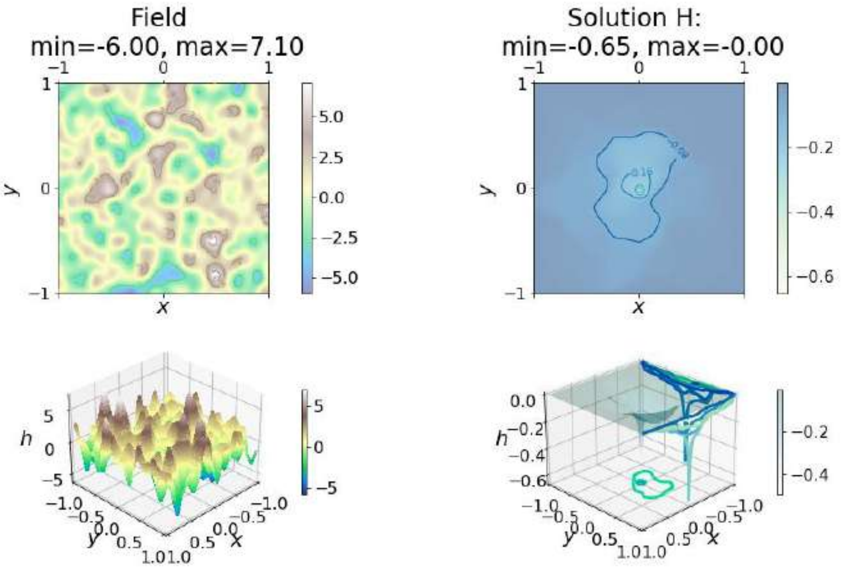
$\nabla \cdot (e^Y \nabla h) = delta^{2*-cos}(x-0), x \in [-1, 1]^2 \subset R^2$

Dirichlet=0; filed Y: IntegralScale:0.1, var:3.79; mesh:512



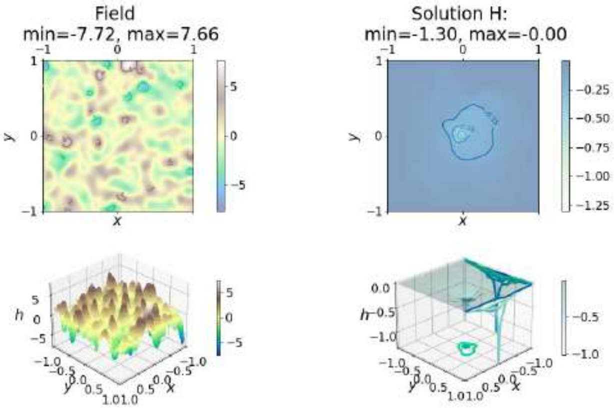
$\nabla \cdot (e^Y \nabla h) = delta^{2*-cos}(x-0), x \in [-1, 1]^2 \subset R^2$

Dirichlet=0; filed Y: IntegralScale:0.1, var:3.79; mesh:512

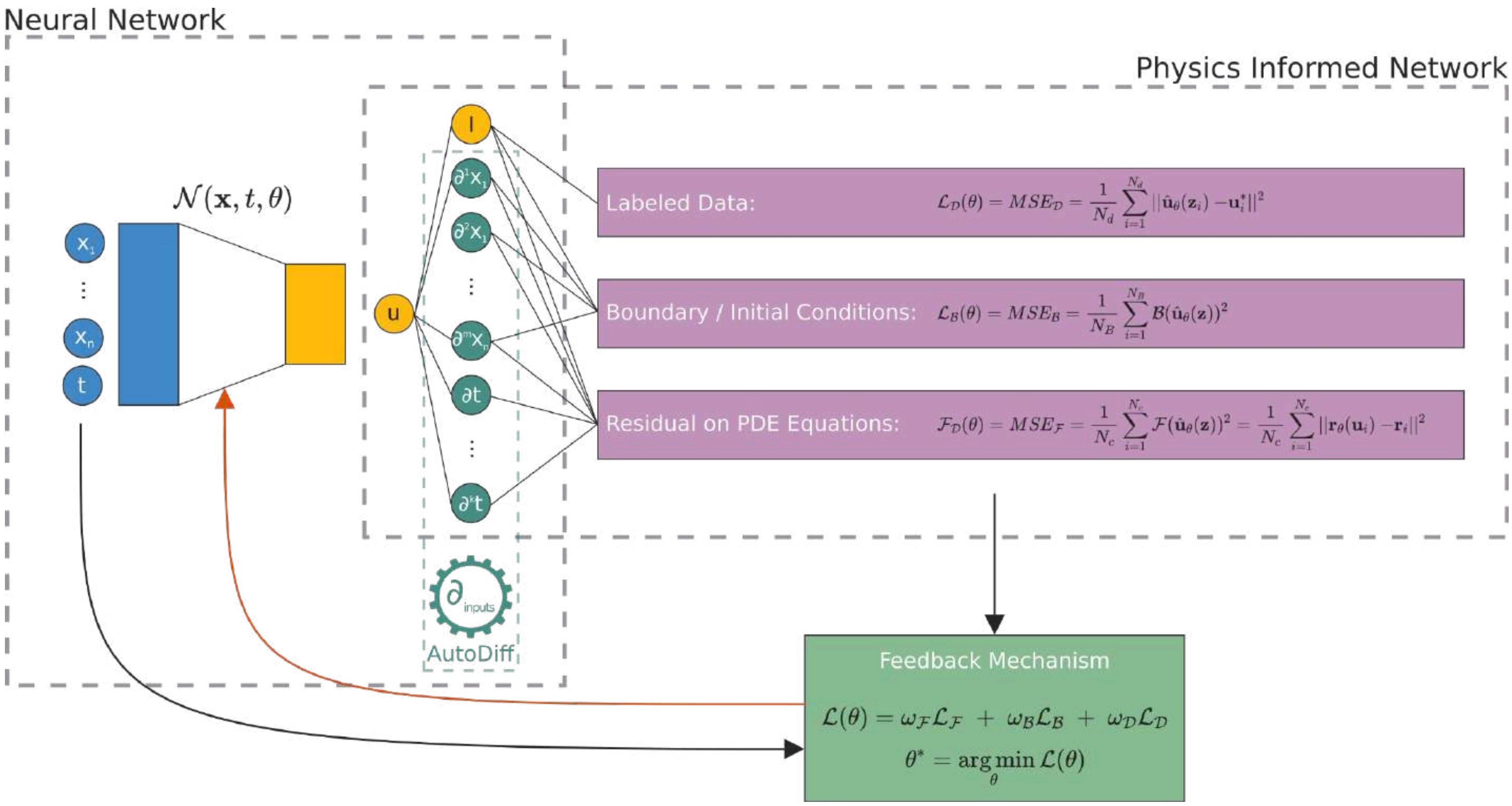


$\nabla \cdot (e^Y \nabla h) = delta^{2*-cos}(x-0), x \in [-1, 1]^2 \subset R^2$

Dirichlet=0; filed Y: IntegralScale:0.1, var:3.79; mesh:512



Physics Informed Neural Networks



Labeled Data: $\mathcal{L}_{\mathcal{D}}(\theta) = MSE_{\mathcal{D}} = \frac{1}{N_d}\sum_{i=1}^{N_d} ||\hat{\mathbf{u}}_\theta(\mathbf{z}_i) - \mathbf{u}_i^*||^2$

Boundary / Initial Conditions: $\mathcal{L}_{\mathcal{B}}(\theta) = MSE_{\mathcal{B}} = \frac{1}{N_B}\sum_{i=1}^{N_B} \mathcal{B}(\hat{\mathbf{u}}_\theta(\mathbf{z}))^2$

Residual on PDE Equations: $\mathcal{F}_{\mathcal{D}}(\theta) = MSE_{\mathcal{F}} = \frac{1}{N_c}\sum_{i=1}^{N_c} \mathcal{F}(\hat{\mathbf{u}}_\theta(\mathbf{z}))^2 = \frac{1}{N_c}\sum_{i=1}^{N_c} ||\mathbf{r}_\theta(\mathbf{u}_i) - \mathbf{r}_i||^2$

Feedback Mechanism

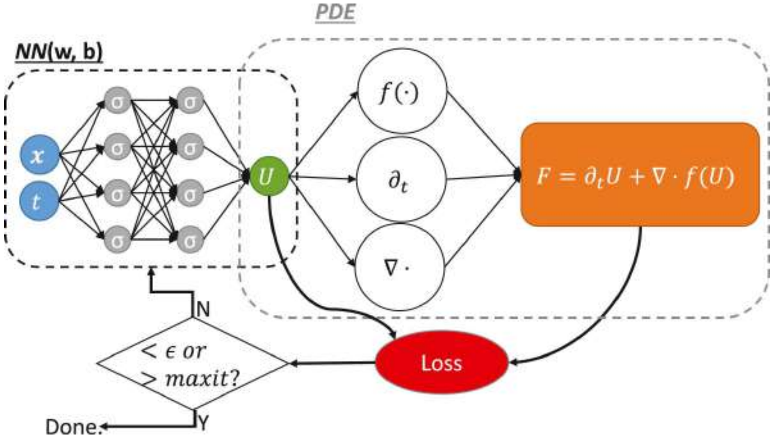
$$\mathcal{L}(\theta) = \omega_{\mathcal{F}}\mathcal{L}_{\mathcal{F}} + \omega_{\mathcal{B}}\mathcal{L}_{\mathcal{B}} + \omega_{\mathcal{D}}\mathcal{L}_{\mathcal{D}}$$

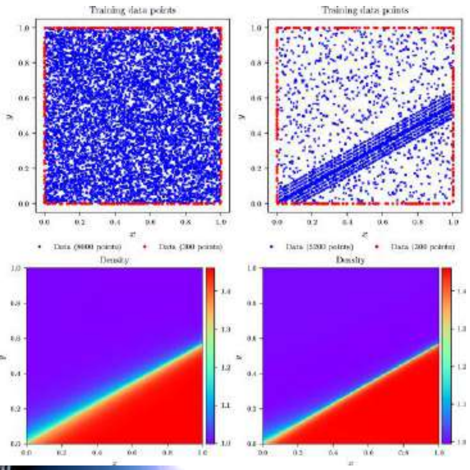
$$\theta^* = \arg\min_\theta \mathcal{L}(\theta)$$

PINNs in literature

high-speed flows

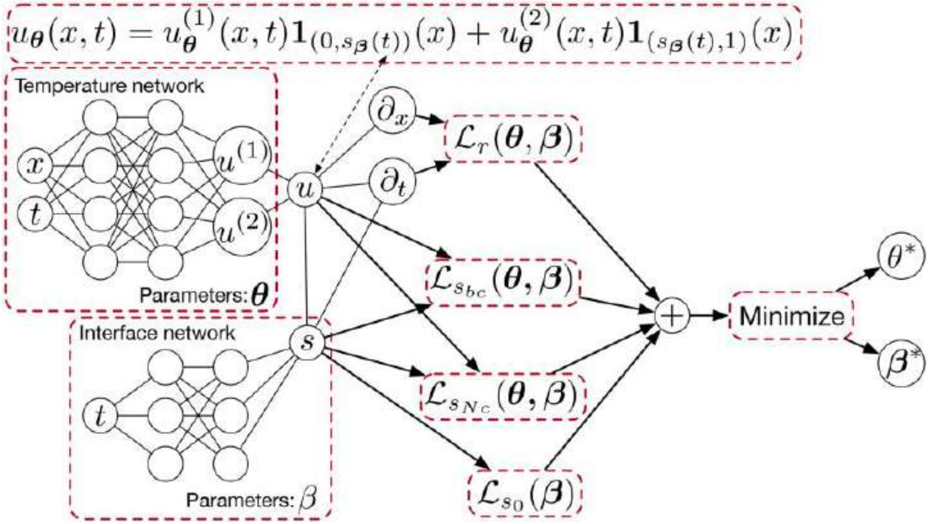
Physics-informed neural networks for high-speed flows Z Mao · 2020
https://doi.org/10.1016/j.cma.2019.112789





Two-phase Stefan problem

Physics-Informed Neural Networks for Heat Transfer Problems, S Cai · 2021
https://doi.org/10.1115/1.4050542

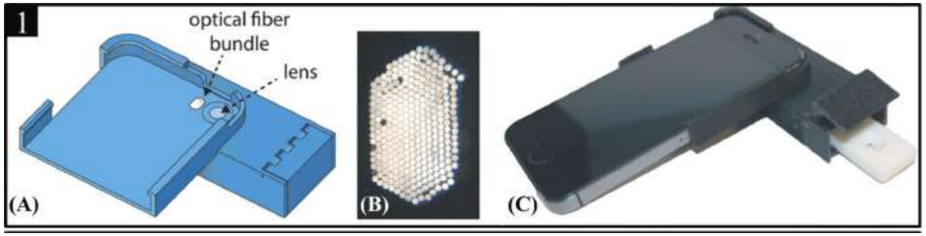

$u_{\boldsymbol{\theta}}(x,t) = u_{\boldsymbol{\theta}}^{(1)}(x,t)\mathbf{1}_{(0,s_{\boldsymbol{\beta}}(t))}(x) + u_{\boldsymbol{\theta}}^{(2)}(x,t)\mathbf{1}_{(s_{\boldsymbol{\beta}}(t),1)}(x)$
Temperature network
x
t
$u^{(1)}$
$u^{(2)}$
Parameters: $\boldsymbol{\theta}$
u
Interface network
t
s
Parameters: β
∂_x
∂_t
$\mathcal{L}_r(\boldsymbol{\theta}, \boldsymbol{\beta})$
$\mathcal{L}_{s_{bc}}(\boldsymbol{\theta}, \boldsymbol{\beta})$
$\mathcal{L}_{s_{Nc}}(\boldsymbol{\theta}, \boldsymbol{\beta})$
$\mathcal{L}_{s_0}(\boldsymbol{\beta})$
$+$
Minimize
θ^*
$\boldsymbol{\beta}^*$

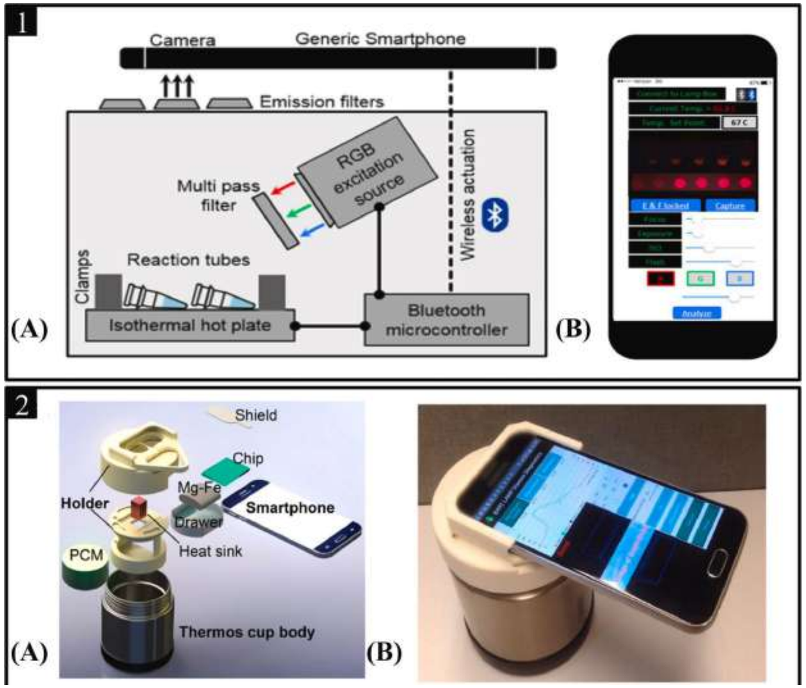
Correct PDE	$\frac{\partial u_1}{\partial t} - 2\frac{\partial^2 u_1}{\partial x^2} = 0$	$\frac{\partial u_1}{\partial t} - \frac{\partial^2 u_1}{\partial x^2} = 0$
Identified PDE (original)	$\frac{\partial u_1}{\partial t} - 1.712\frac{\partial^2 u_1}{\partial x^2} = 0$	$\frac{\partial u_1}{\partial t} - 1.137\frac{\partial^2 u_1}{\partial x^2} = 0$
Identified PDE (adaptive)	$\frac{\partial u_1}{\partial t} - 2.003\frac{\partial^2 u_1}{\partial x^2} = 0$	$\frac{\partial u_1}{\partial t} - 1.000\frac{\partial^2 u_1}{\partial x^2} = 0$

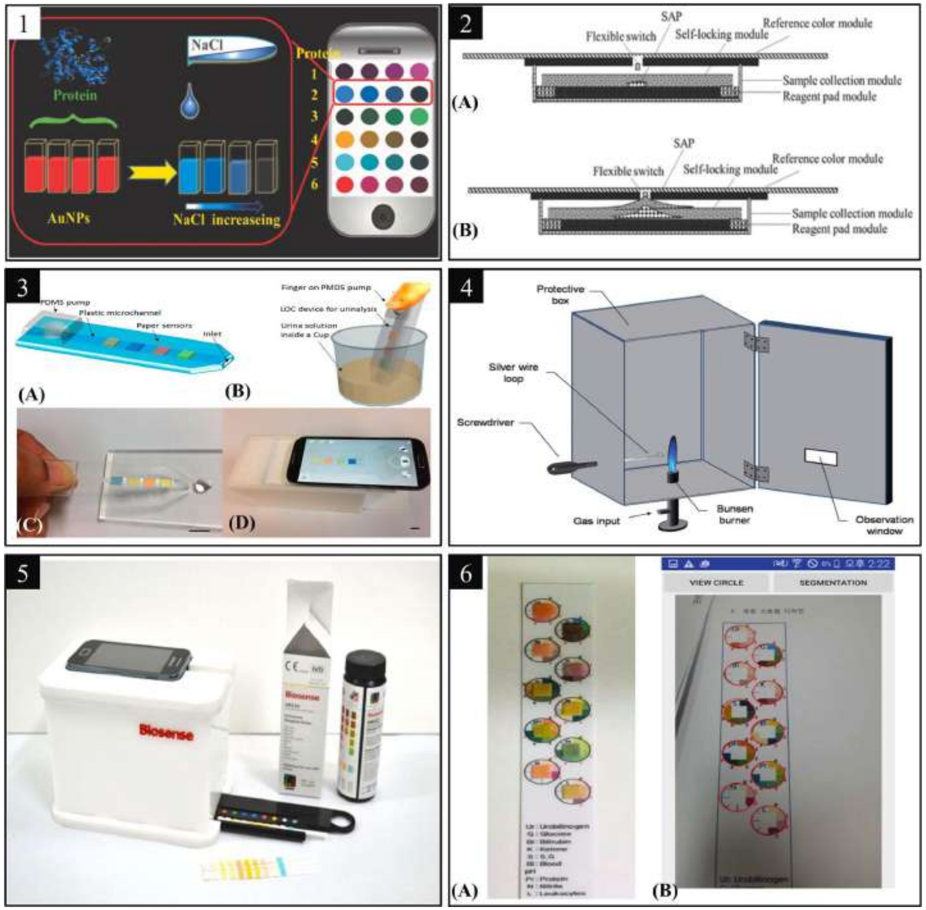
with or without adaptive learning rate annealing

Lab on Hand devices examples

In literature, most research in on extending hardware capabilities, and simple regression software.

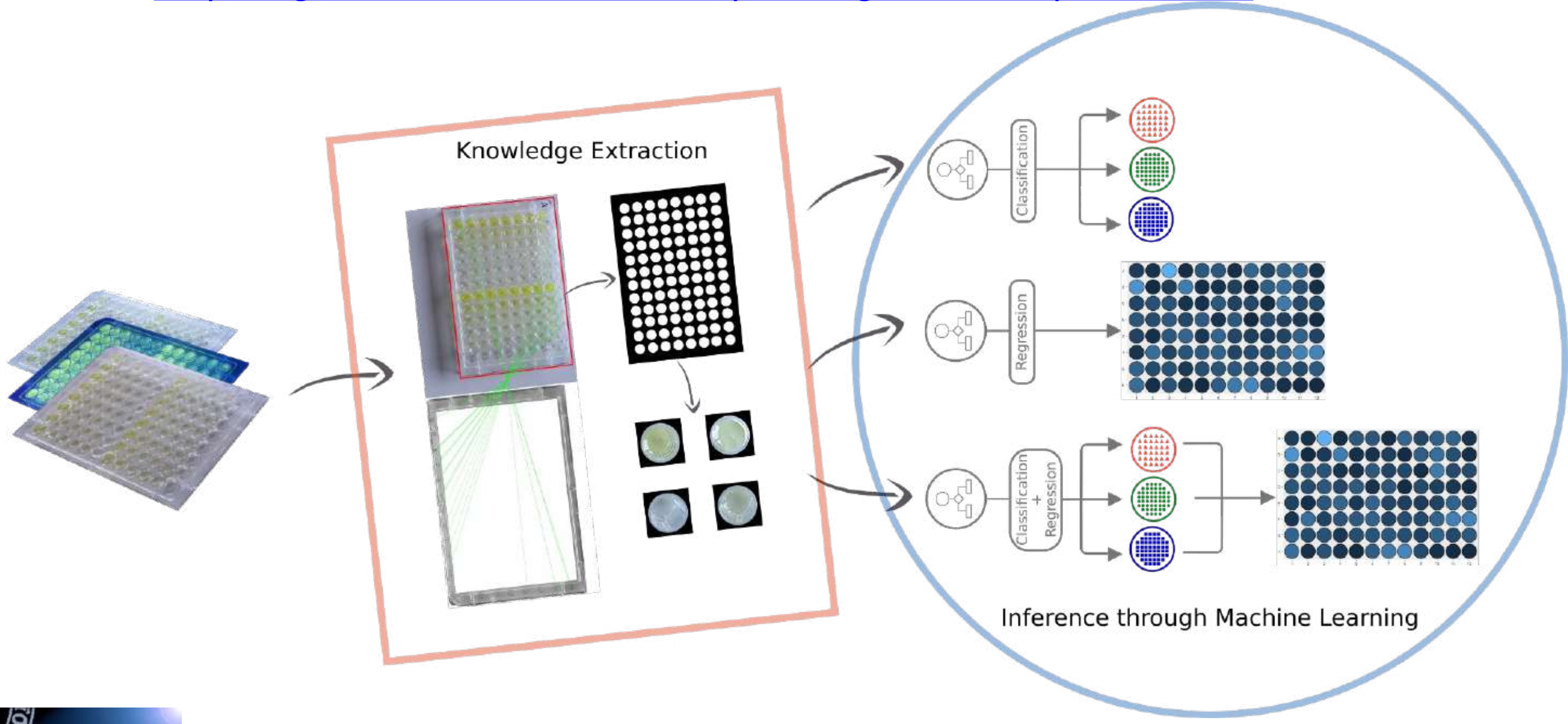






Proof of Concept for a Lab on Hand

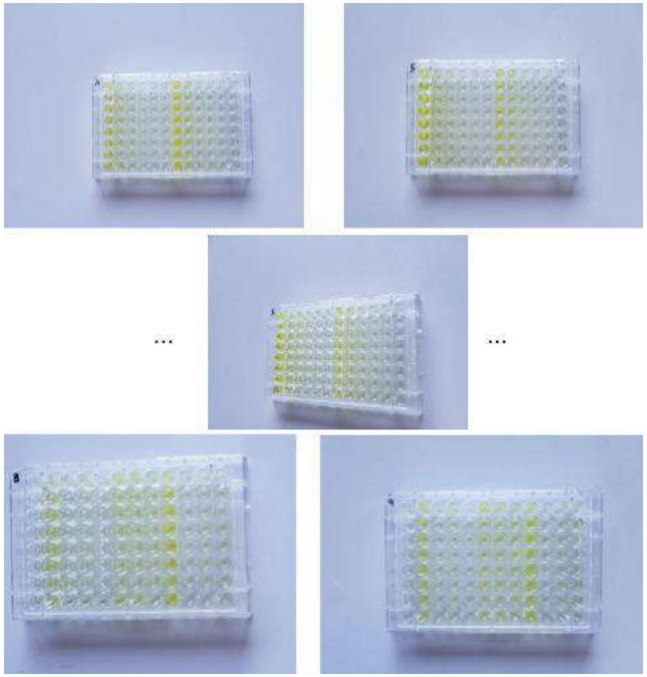
Create a software that automatically recognizes point of interest in an image, and use ML to predict the compound concentration based on color intensity level
Code: https://github.com/MthBr/well-plate-light-driven-predictions

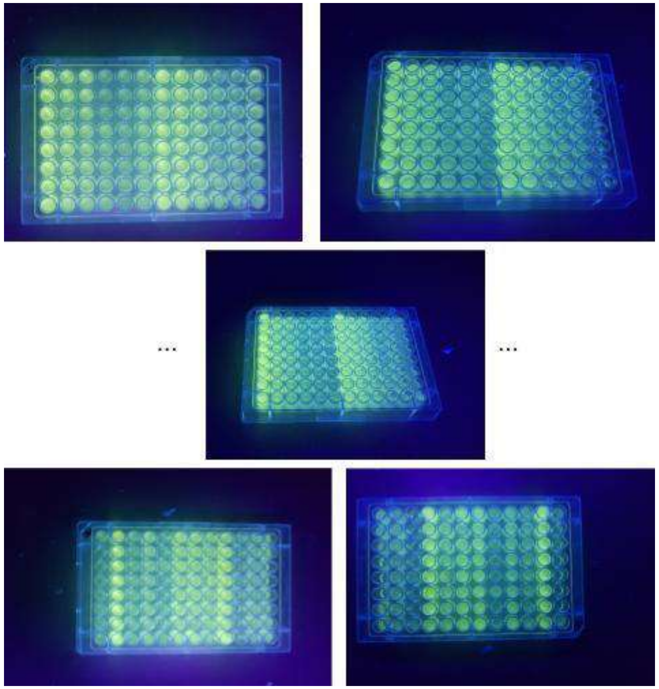

Knowledge Extraction
Classification
Regression
Classification + Regression
Inference through Machine Learning

Challenges

Extract features using a Smartphone:

- automatically identify the position of a well in well plate
- be able to take photos, while holding the smartphone NOT parallel to the surface
- be able to be camera independant
- adaptable to different light setting





Recognize an object in a photo

Tested multiple techniques to extract ROI from well plates, as segmentation techniques, among all:

- k-means clustering
- Region Adjacency Graph (RAG) segmentation
- felzenszwalb segmentation
- and edge detector for contouring (Canny),

applied on multiple combinations of color spaces channels extracted from the images.

Eventually relied on **SIFT** (scale-invariant feature transformation) algorithm that is invariant to translation, scale, and rotation, and is robust to affine distortion, change in lighting, and change in 3D point of view.

The SIFT algorithm identifies and describes keypoints, i.e. points in a picture that are interesting or stand out. Each detected keypoint has a descriptor that is associated with it, and these descriptors are invariant to affine transformation or distortion.

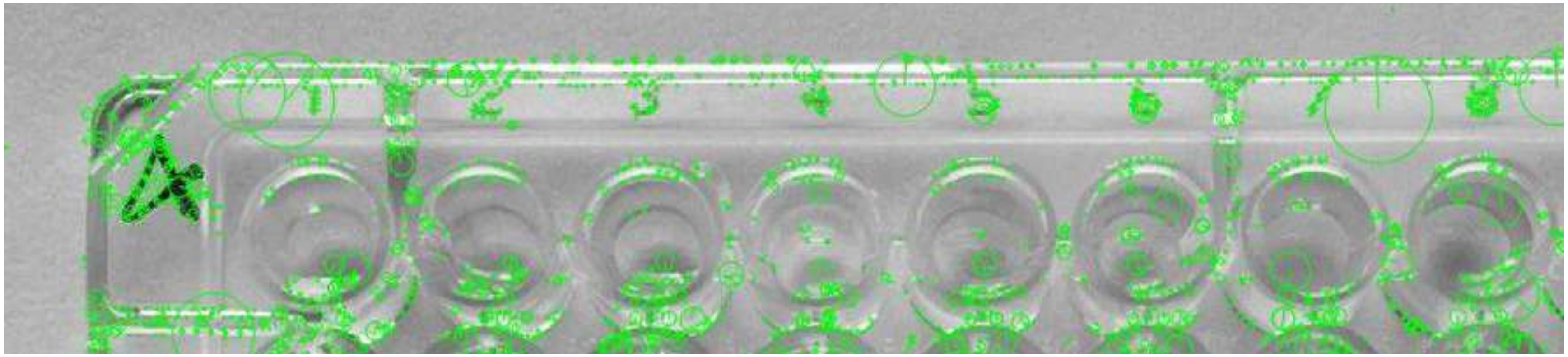
SWIFT algorithm

SIFT generate is a set of key point-descriptor, (p,s,r,f), where

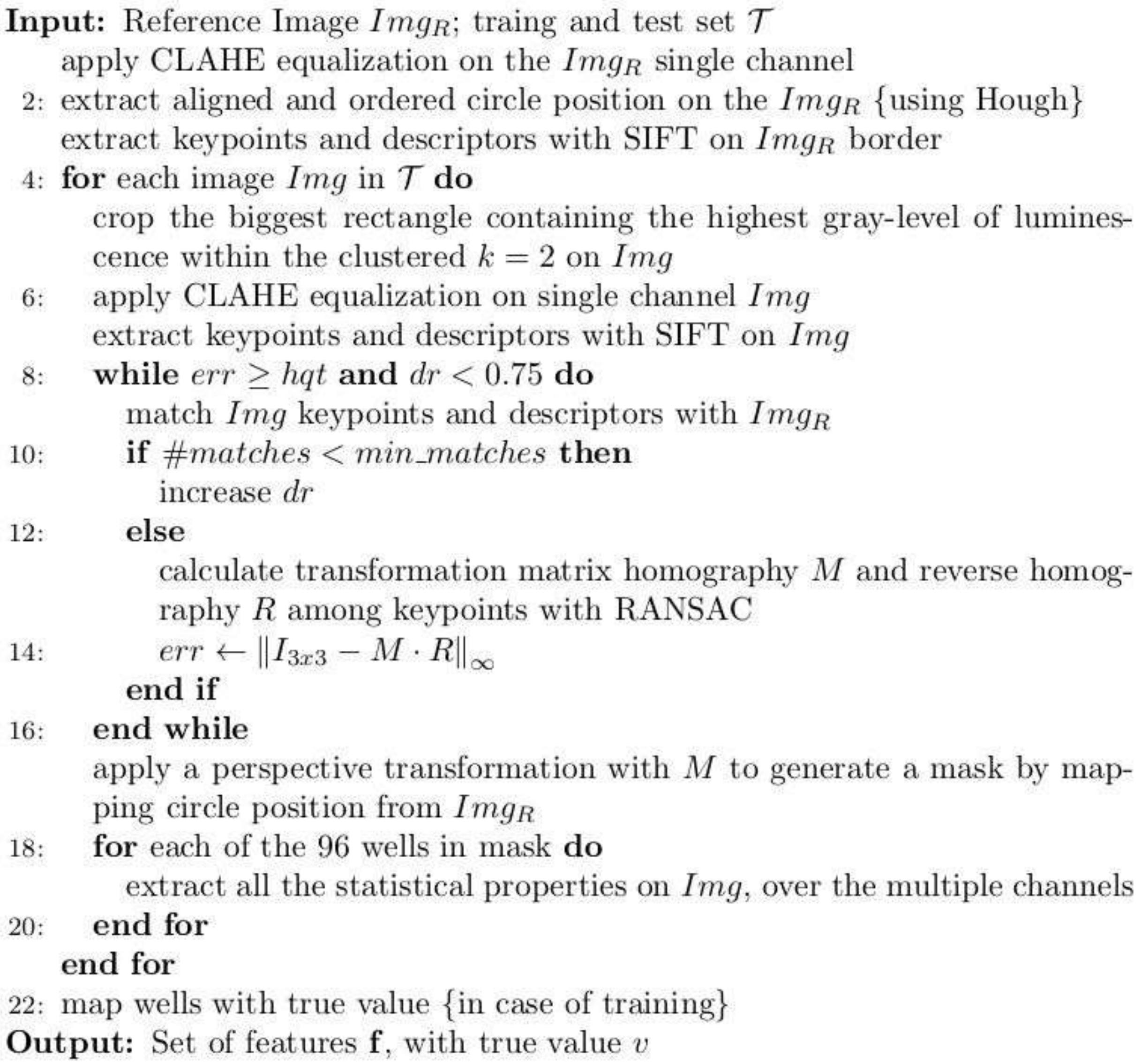
- p=(x,y) is the location of the key point pixel on the image,
- s the scale,
- r the orientation,
- f a 128-dim descriptor generated from local image gradients

Steps:

1. Identifying the candidate keypoint positions
2. Keypoint filtering
3. Keypoint Orientation Assignment
4. Keypoint Descriptor Generation



Extracting wells features



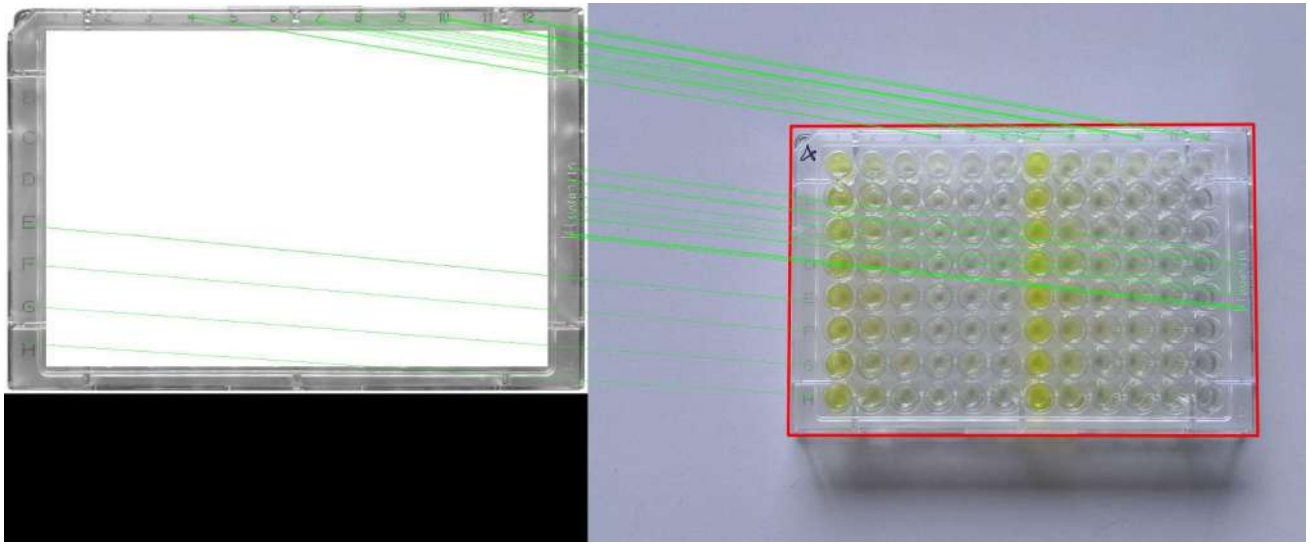
```
Input: Reference Image $Img_R$; traing and test set $\mathcal{T}$
  apply CLAHE equalization on the $Img_R$ single channel
  extract aligned and ordered circle position on the $Img_R$ {using Hough}
  extract keypoints and descriptors with SIFT on $Img_R$ border
  for each image $Img$ in $\mathcal{T}$ do
    crop the biggest rectangle containing the highest gray-level of luminescence within the clustered $k = 2$ on $Img$
    apply CLAHE equalization on single channel $Img$
    extract keypoints and descriptors with SIFT on $Img$
    while $err \geq hqt$ and $dr < 0.75$ do
      match $Img$ keypoints and descriptors with $Img_R$
      if $\#matches < min\_matches$ then
        increase $dr$
      else
        calculate transformation matrix homography $M$ and reverse homography $R$ among keypoints with RANSAC
        $err \leftarrow \|I_{3x3} - M \cdot R\|_\infty$
      end if
    end while
    apply a perspective transformation with $M$ to generate a mask by mapping circle position from $Img_R$
    for each of the 96 wells in mask do
      extract all the statistical properties on $Img$, over the multiple channels
    end for
  end for
  map wells with true value {in case of training}
Output: Set of features $\mathbf{f}$, with true value $v$
```

Match homography among keypoints

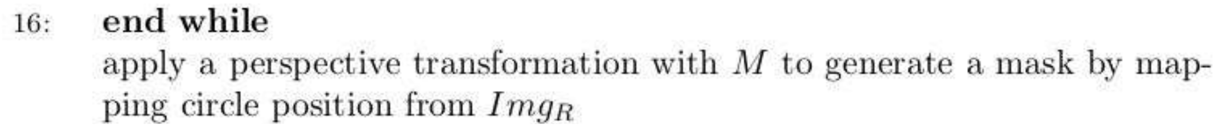
```
crop the biggest rectangle containing the highest gray-level of luminescence within the clustered $k = 2$ on $Img$
apply CLAHE equalization on single channel $Img$
extract keypoints and descriptors with SIFT on $Img$
while $err \geq hqt$ and $dr < 0.75$ do
  match $Img$ keypoints and descriptors with $Img_R$
  if $\#matches < min\_matches$ then
    increase $dr$
  else
    calculate transformation matrix homography $M$ and reverse homography $R$ among keypoints with RANSAC
    $err \leftarrow \|I_{3x3} - M \cdot R\|_\infty$
  end if
end while
```

Nearest Neighbor Distance Ratio **(NNDR)** check, i.e. dr is the ratio among the nearest and second nearest neighbor distances

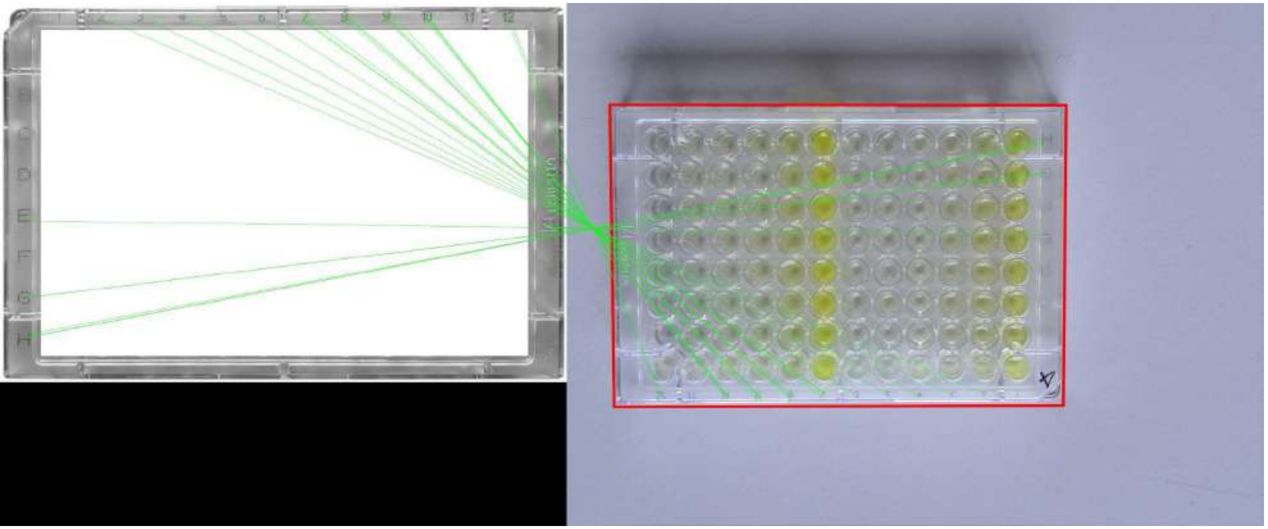
RANSAC (RANdom SAmple Consensus)

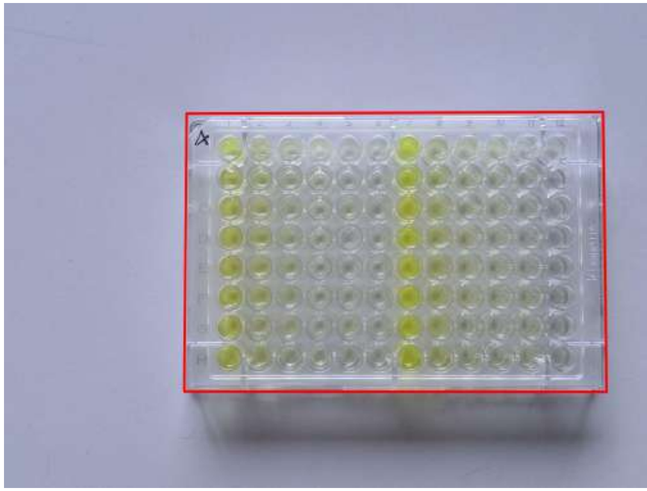


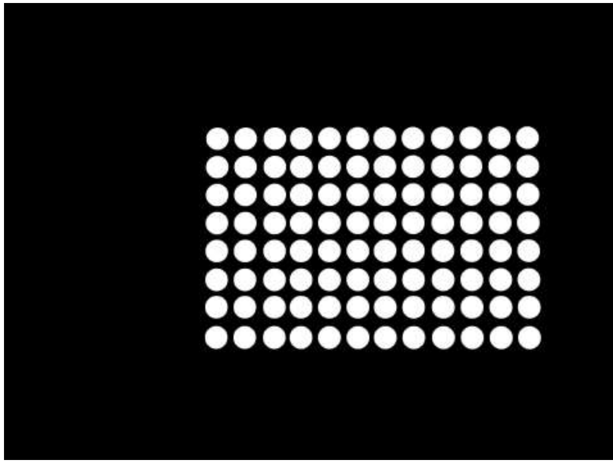
Perspective transformation



```
end while
    apply a perspective transformation with $M$ to generate a mask by mapping circle position from $Img_R$
```







The 96-well microplate has standard dimensions and has 12 columns (1-12) and 8 rows (A-H) for easy well recognition.

Data Generation

More than 700 mobile shots of a 96-well microplate (also known as a microtiter plate, MTP, or multiwell) were taken in a laboratory under three different lighting conditions.

Photos were taken using the smartphone's automatic exposure setting in three light configurations:

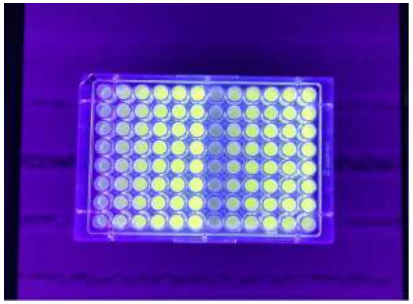
- ambient light (AB)
- portable UVc (PUVc)
- UVc

During the tests, two smartphones were used:

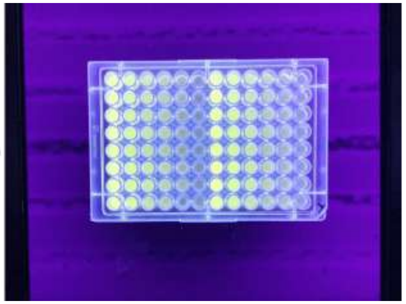
- Samsung Galaxy S7
- Huawei Mate 10 Pro

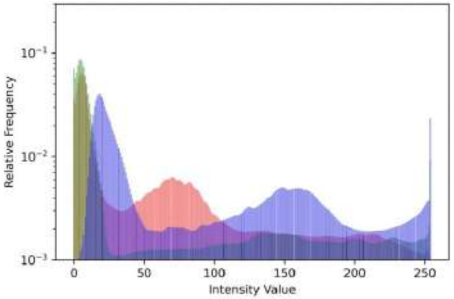
Different Light setting and device

Huawei
Uniform UV

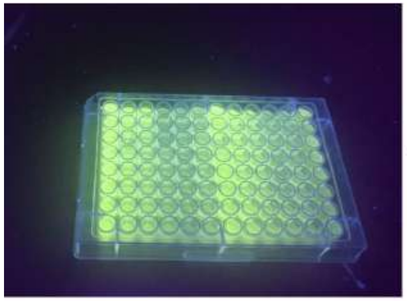


...

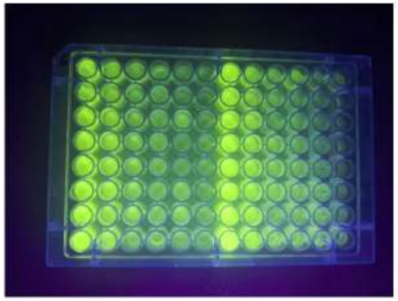


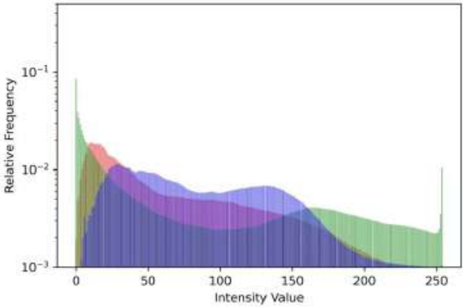


Huawei
Portable UV

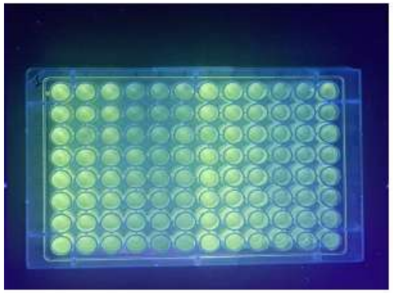


...

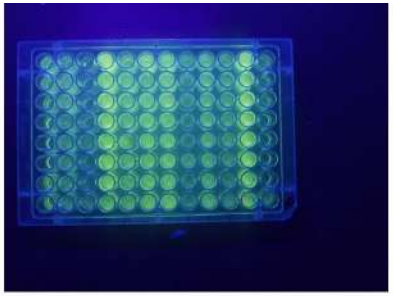


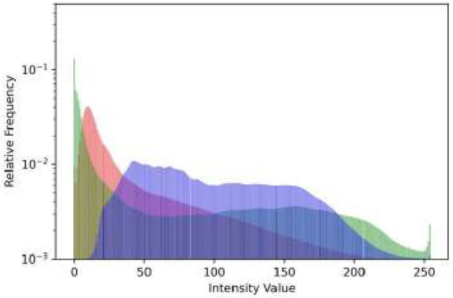


Samsung
Portable UV



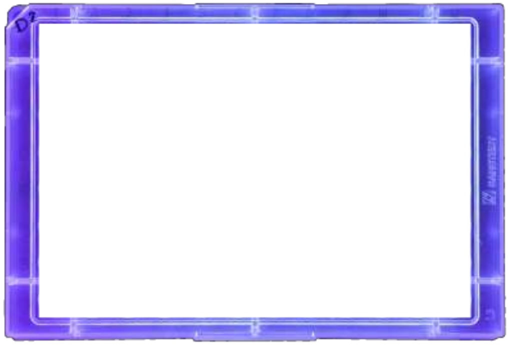
...



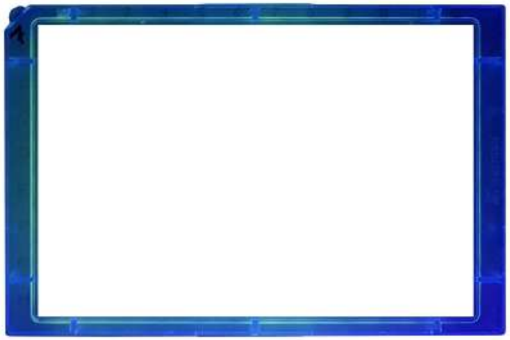


Matching dictionary

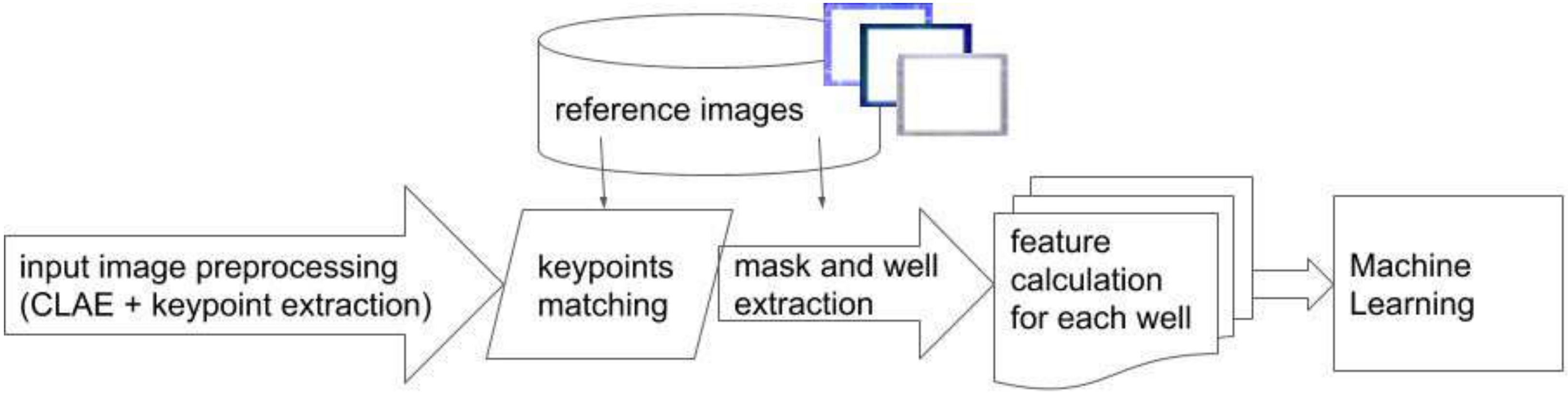
Uniform UV

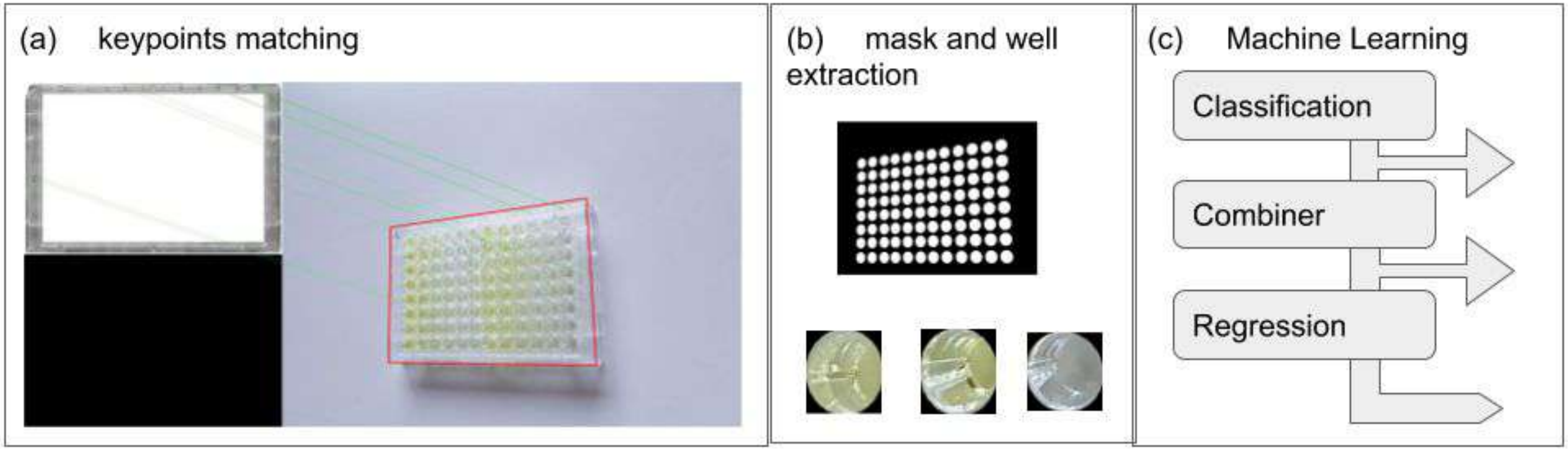


Portable UV

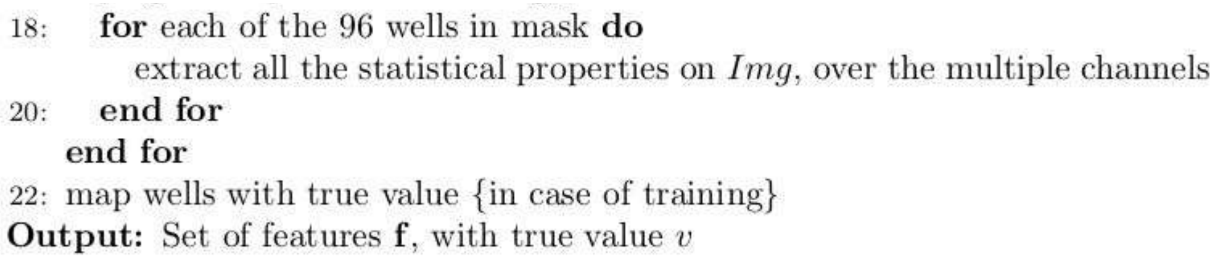


Development Summary


reference images
input image preprocessing (CLAE + keypoint extraction)
keypoints matching
mask and well extraction
feature calculation for each well
Machine Learning



Feature Extraction



```
    for each of the 96 wells in mask do
      extract all the statistical properties on $Img$, over the multiple channels
    end for
  end for
  map wells with true value {in case of training}
Output: Set of features $\mathbf{f}$, with true value $v$
```

For each well, extracted multiple features:
the mean and truncated mean at 30% for each channel color
standard deviation, skewness, and entropy for each channel color,
consisting in gray, RGB, LAB, and HSV.

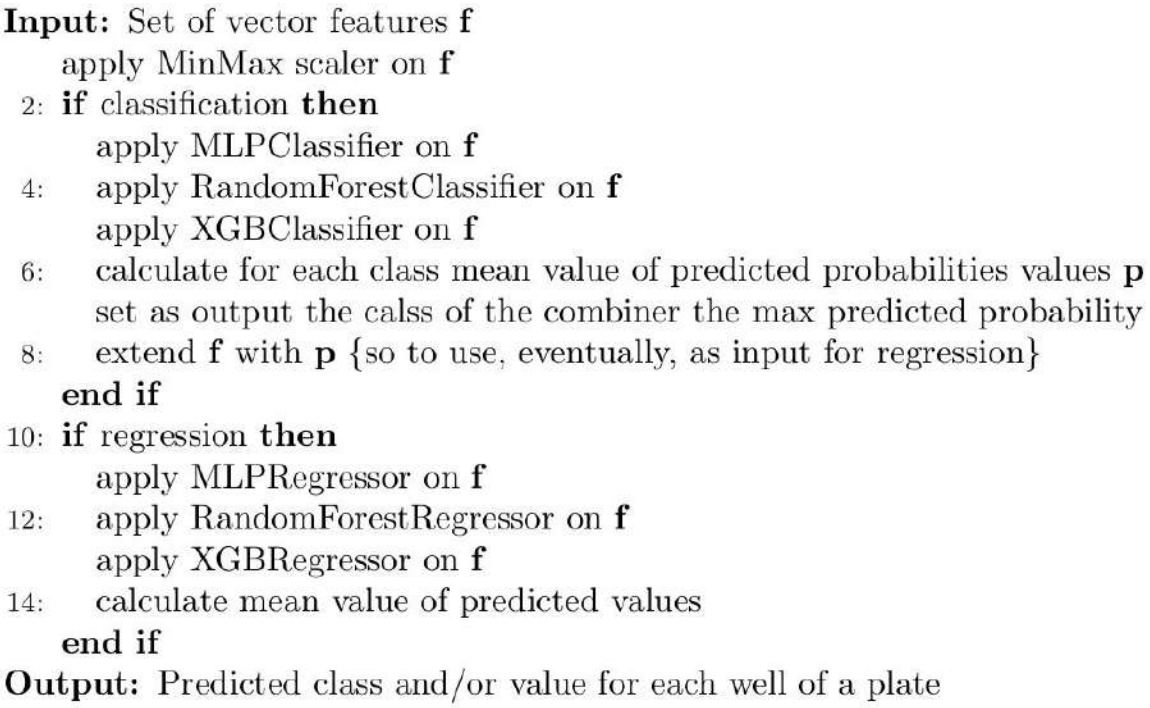
Each well could have more than 50 features, plus 7 feature of properties coming from the metadata of the phone camera, selecting the tags more commonly available, i.e. ShutterSpeedValue, ApertureValue, BrightnessValue, ExposureBiasValue, MaxApertureValue, FocalLength, ExposureTime.

Machine Learning

In the case of a training process, we know what class the well belongs to and what luminescence value it has.

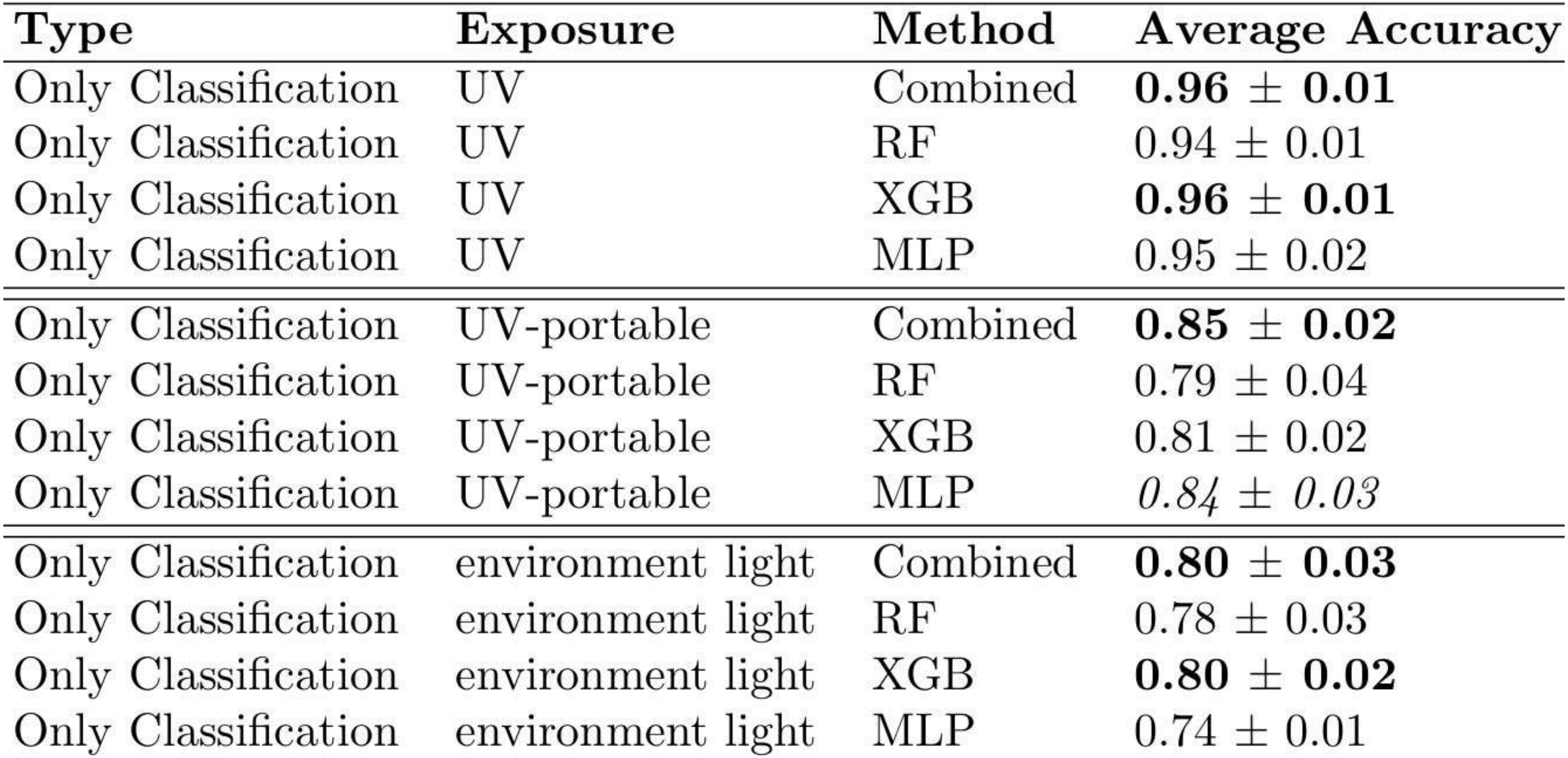
This information is then applied to the final data collection, ready to be transferred to the supervised learning algorithms.

Next, all the features are transferred to the combiner for classification and regression.



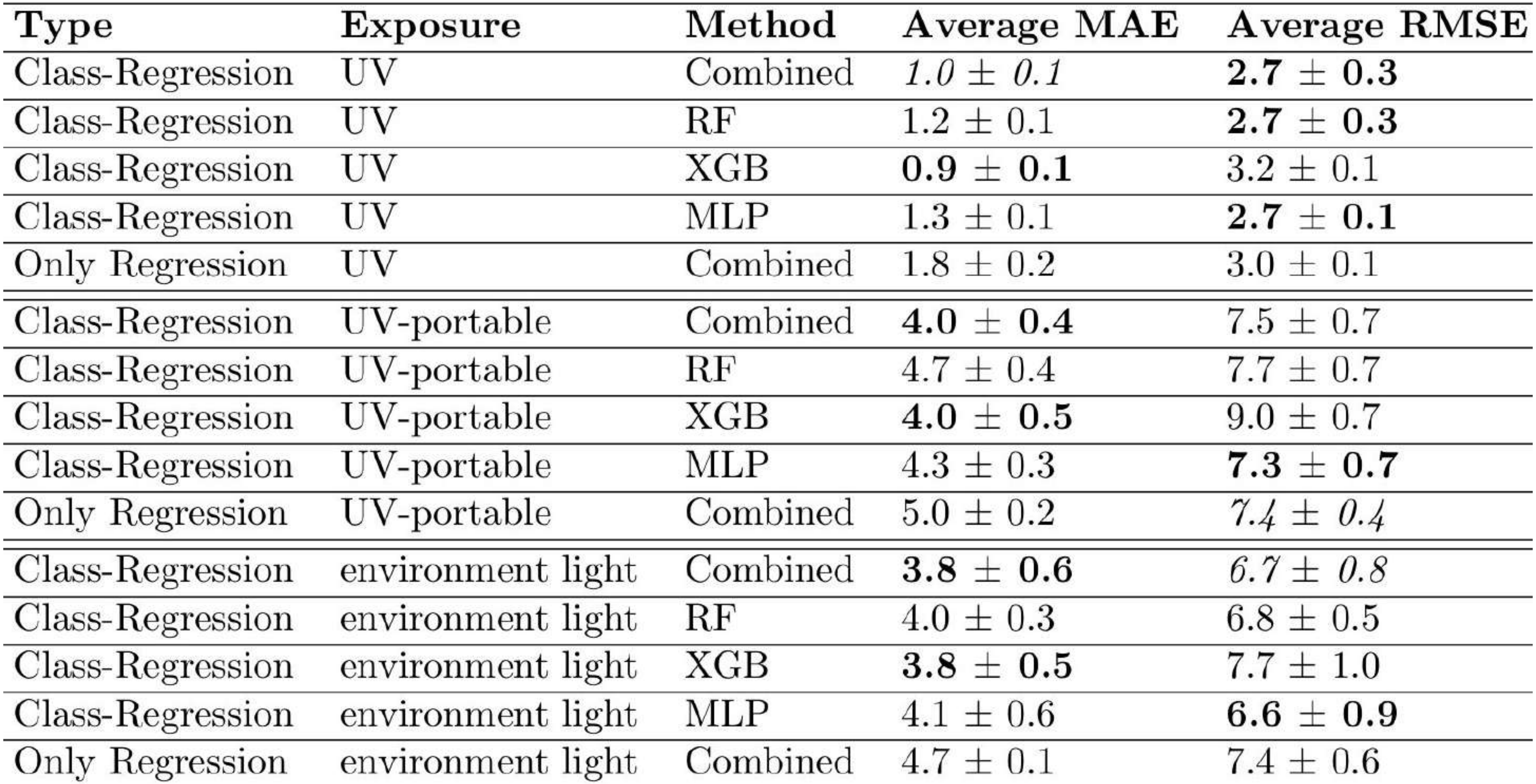
```
Input: Set of vector features f
    apply MinMax scaler on f
    if classification then
        apply MLPClassifier on f
        apply RandomForestClassifier on f
        apply XGBClassifier on f
        calculate for each class mean value of predicted probabilities values p
        set as output the calss of the combiner the max predicted probability
        extend f with p {so to use, eventually, as input for regression}
    end if
    if regression then
        apply MLPRegressor on f
        apply RandomForestRegressor on f
        apply XGBRegressor on f
        calculate mean value of predicted values
    end if
Output: Predicted class and/or value for each well of a plate
```

Classification Results



Type	Exposure	Method	Average Accuracy
Only Classification	UV	Combined	**0.96 ± 0.01**
Only Classification	UV	RF	0.94 ± 0.01
Only Classification	UV	XGB	**0.96 ± 0.01**
Only Classification	UV	MLP	0.95 ± 0.02
Only Classification	UV-portable	Combined	**0.85 ± 0.02**
Only Classification	UV-portable	RF	0.79 ± 0.04
Only Classification	UV-portable	XGB	0.81 ± 0.02
Only Classification	UV-portable	MLP	*0.84 ± 0.03*
Only Classification	environment light	Combined	**0.80 ± 0.03**
Only Classification	environment light	RF	0.78 ± 0.03
Only Classification	environment light	XGB	**0.80 ± 0.02**
Only Classification	environment light	MLP	0.74 ± 0.01

Results and Validation



Type	Exposure	Method	Average MAE	Average RMSE
Class-Regression	UV	Combined	*1.0 ± 0.1*	**2.7 ± 0.3**
Class-Regression	UV	RF	1.2 ± 0.1	**2.7 ± 0.3**
Class-Regression	UV	XGB	**0.9 ± 0.1**	3.2 ± 0.1
Class-Regression	UV	MLP	1.3 ± 0.1	**2.7 ± 0.1**
Only Regression	UV	Combined	1.8 ± 0.2	3.0 ± 0.1
Class-Regression	UV-portable	Combined	**4.0 ± 0.4**	7.5 ± 0.7
Class-Regression	UV-portable	RF	4.7 ± 0.4	7.7 ± 0.7
Class-Regression	UV-portable	XGB	**4.0 ± 0.5**	9.0 ± 0.7
Class-Regression	UV-portable	MLP	4.3 ± 0.3	**7.3 ± 0.7**
Only Regression	UV-portable	Combined	5.0 ± 0.2	*7.4 ± 0.4*
Class-Regression	environment light	Combined	**3.8 ± 0.6**	*6.7 ± 0.8*
Class-Regression	environment light	RF	4.0 ± 0.3	6.8 ± 0.5
Class-Regression	environment light	XGB	**3.8 ± 0.5**	7.7 ± 1.0
Class-Regression	environment light	MLP	4.1 ± 0.6	**6.6 ± 0.9**
Only Regression	environment light	Combined	4.7 ± 0.1	7.4 ± 0.6

Conclusion

- The work provides a Proof of Concept (PoC) for developing a Point of Diagnosis (PoD) by studying the luminescence of a compound interacting acquired with a smartphone.
- Downaladable at https://github.com/MthBr/well-plate-light-driven-predictions

- Application examples: measuring the concentration of mercury in water directly on a boat.

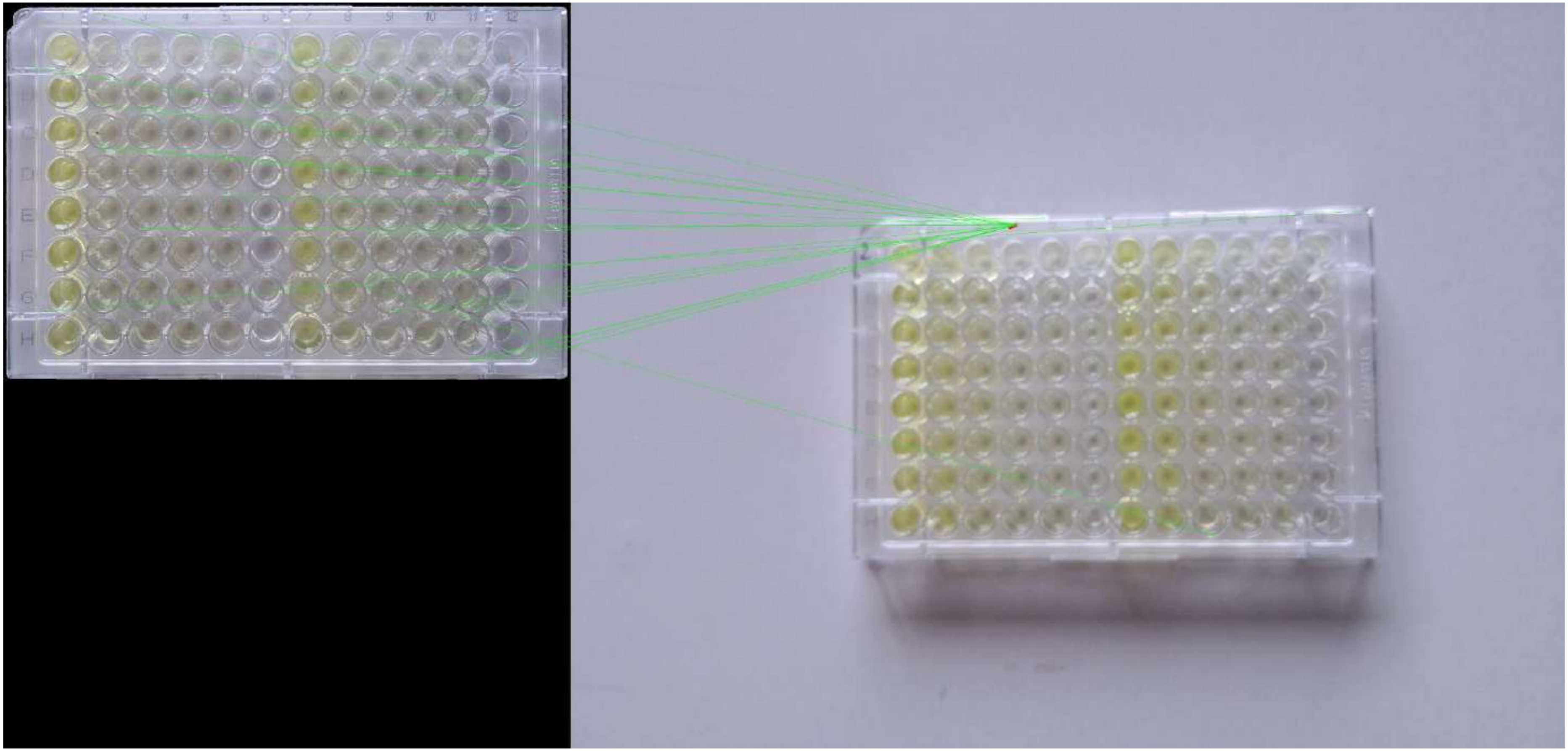
- Approach is feasible since the reaction of a compound with a reacting luminescent agent can be determined, in general, providing support for extensive training data.

- Extensions to Android application are possible, and training on several light natural light settings or possibly the use of a homogeneous dark camera to improve performance are options for improving the tool's accuracy.
- Future studies would include the possibility of improving overall classification accuracy and reducing regression errors using various light settings, camera phones, and other types of well plates.

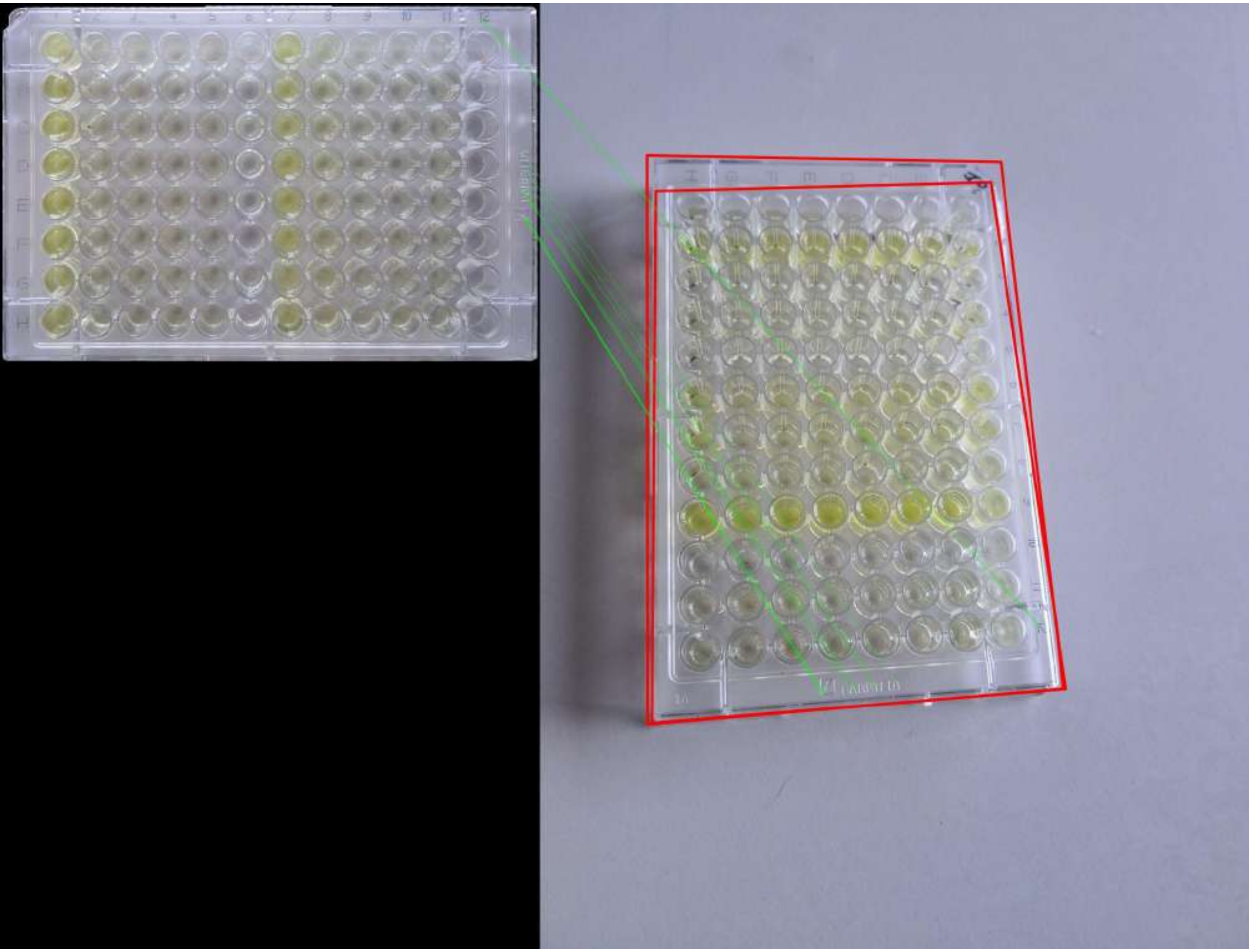
Publications

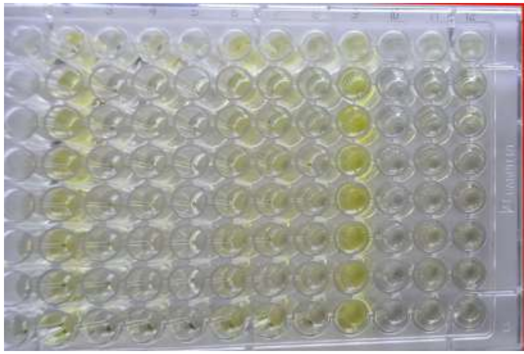
- F. Piccialli, S. Cuomo, V. S. Di Cola, G, Casolla, "A machine learning approach for IoT cultural data", *Journal of Ambient Intelligence and Humanized Computing*, pp. 1-12, Sep. 2019, DOI: 10.1007/s12652-019-01452-6.
- G. Casolla, S. Cuomo, V. S. Di Cola, F Piccialli, "Exploring unsupervised learning techniques for the Internet of Things", *IEEE Transactions on Industrial Informatics*, vol. 16 (4), pp. 2621-2628, Apr. 2020, DOI: 10.1109/TII.2019.2941142.
- F. Piccialli, G. Casolla, S. Cuomo, F. Giampaolo, V. S. Di Cola, "Decision making in IoT environment through unsupervised learning", *IEEE Intelligent Systems*, vol. 35 (1), pp. 27-35, Jan.-Feb. 2020, DOI: 10.1109/MIS.2019.2944783.
- F Piccialli, F Giampaolo, G Casolla, V. S. Di Cola, K Li, "A deep learning approach for path prediction in a location-based IoT system", *Pervasive and Mobile Computing*, vol. 66, 101210, July 2020, DOI: 10.1016/j.pmcj.2020.101210.
- F. Piccialli, S. Cuomo, F. Giampaolo, G. Casolla, V. S. Di Cola, "Path prediction in IoT systems through Markov Chain algorithm", *Future Generation Computer Systems*, vol. 109, pp. 210-217, Aug. 2020, DOI: 10.1016/j.future.2020.03.053.
- F. Piccialli, G. Casolla, S. Cuomo, F. Giampaolo, E. Prezioso, V. S. Di Cola, "Unsupervised learning on multimedia data: a Cultural Heritage case study", *Multimedia Tools and Applications*, vol. 79 (45), pp. 34429-34442, Dec. 2020, DOI: 10.1007/s11042-020-08781-1.
- S. Cuomo, G. Colecchia, V. S. Di Cola, U. Chirico, "A virtual assistant in cultural heritage scenarios", *Concurrency and Computation: Practice and Experience*, vol. 33 (3), pp. e5331, May 2019, DOI: 10.1002/cpe.5331.
- V. S. Di Cola, S Cuomo, G Severino, "Remarks on the numerical approximation of Dirac delta functions", *Results in Applied Mathematics*, vol. 12, pp. 100200, Nov. 2021, DOI: 10.1016/j.rinam.2021.100200.
- F. Piccialli, V. S. di Cola, F. Giampaolo, S. Cuomo, "The role of artificial intelligence in fighting the COVID-19 pandemic", *Information Systems Frontiers*, vol. 23 (6), pp. 1467-1497, Dec. 2021, DOI: 10.1007/s10796-021-10131-x.
- S. Cuomo, V. S. Di Cola, F. Giampaolo, G. Rozza, M. Raissi, F. Piccialli, "Scientific Machine Learning through Physics-Informed Neural Networks: Where we are and What's next", *arXiv preprint*, arXiv:2201.05624.

Bad match

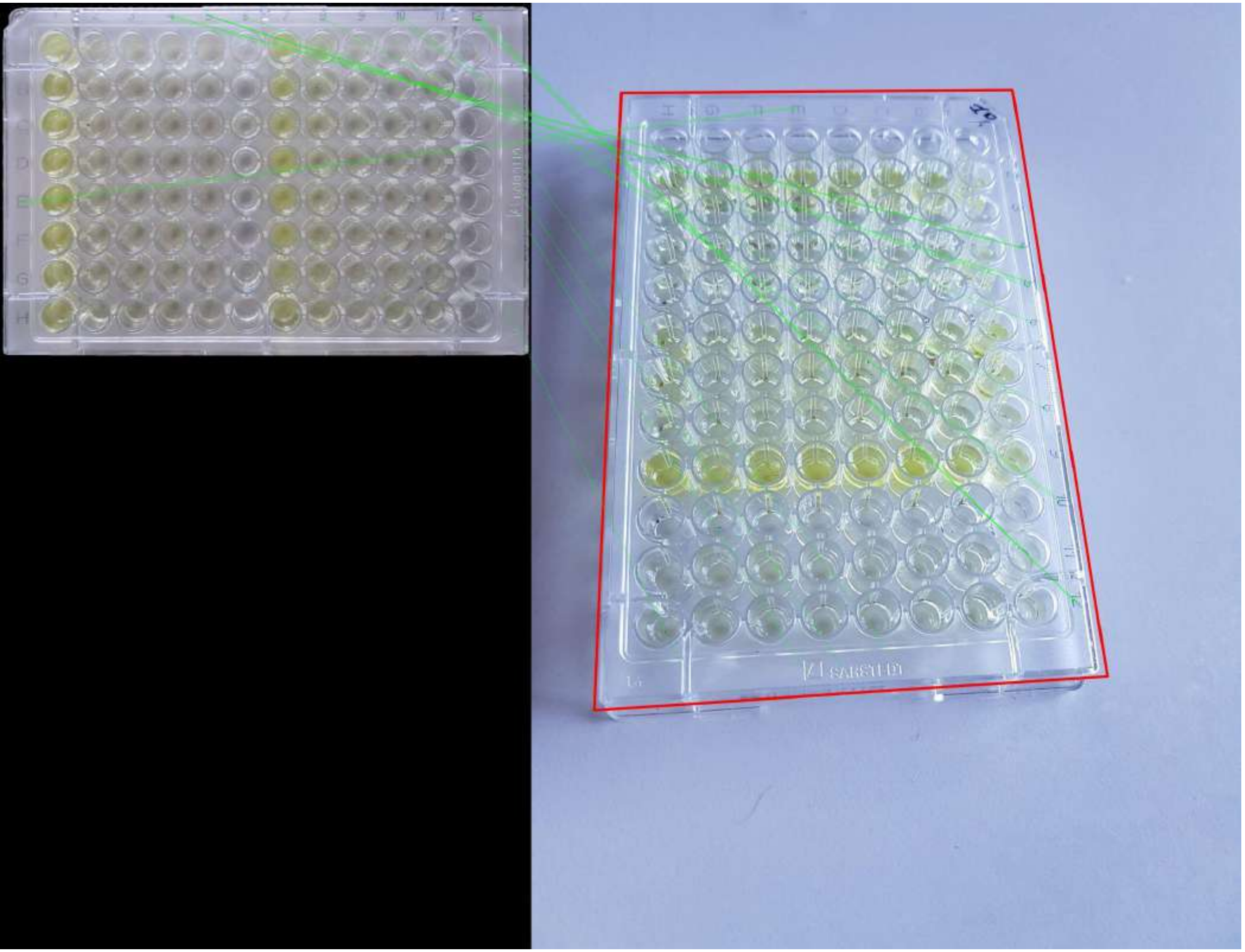


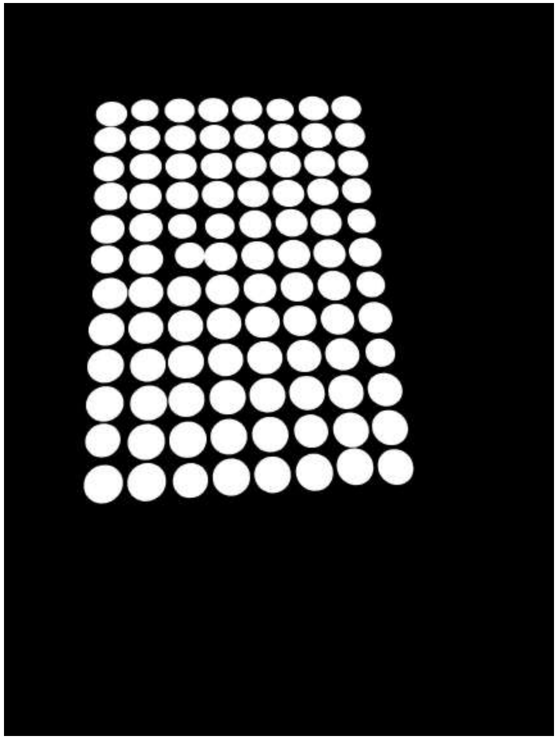
Bad match



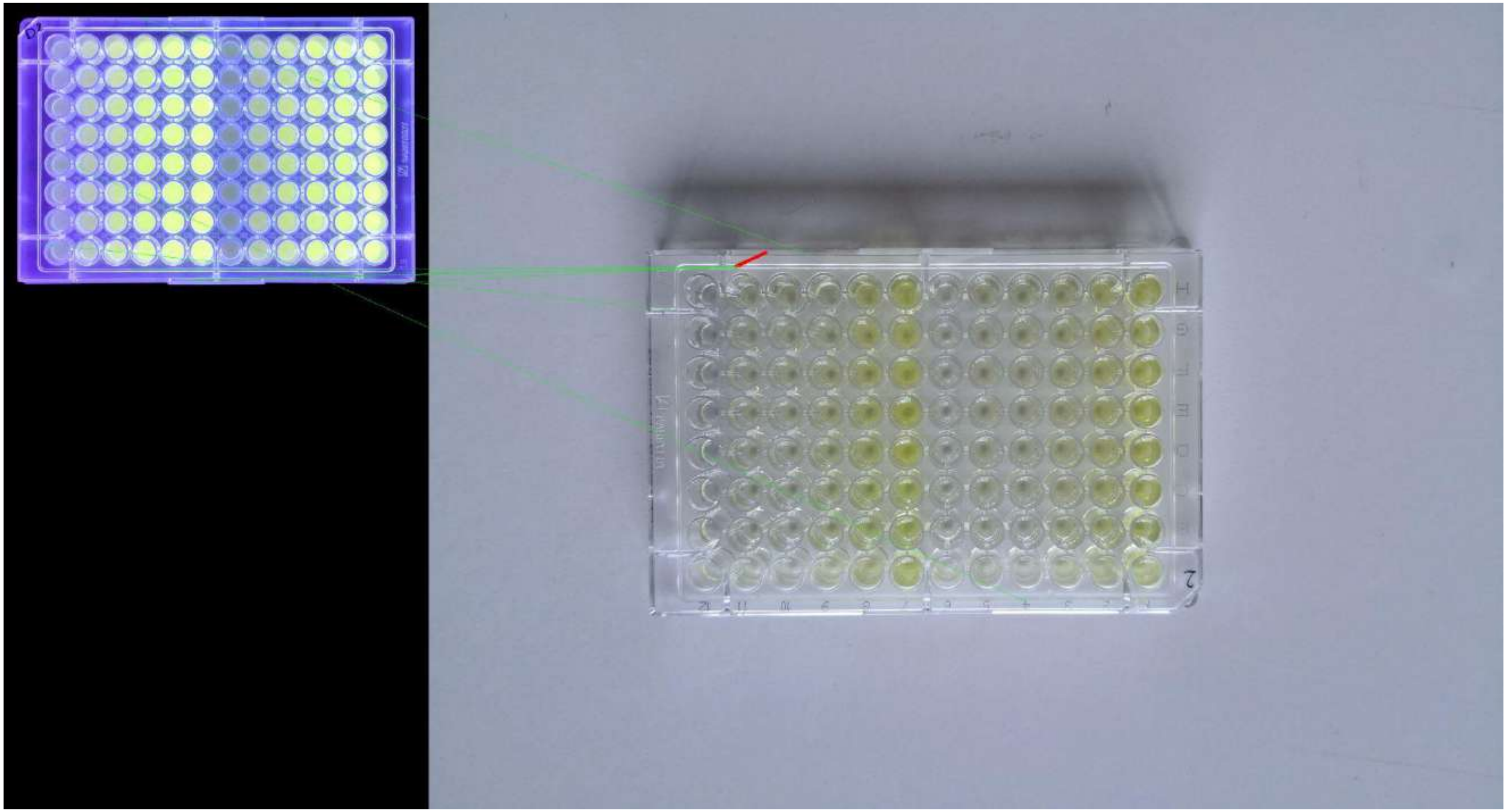


Good match

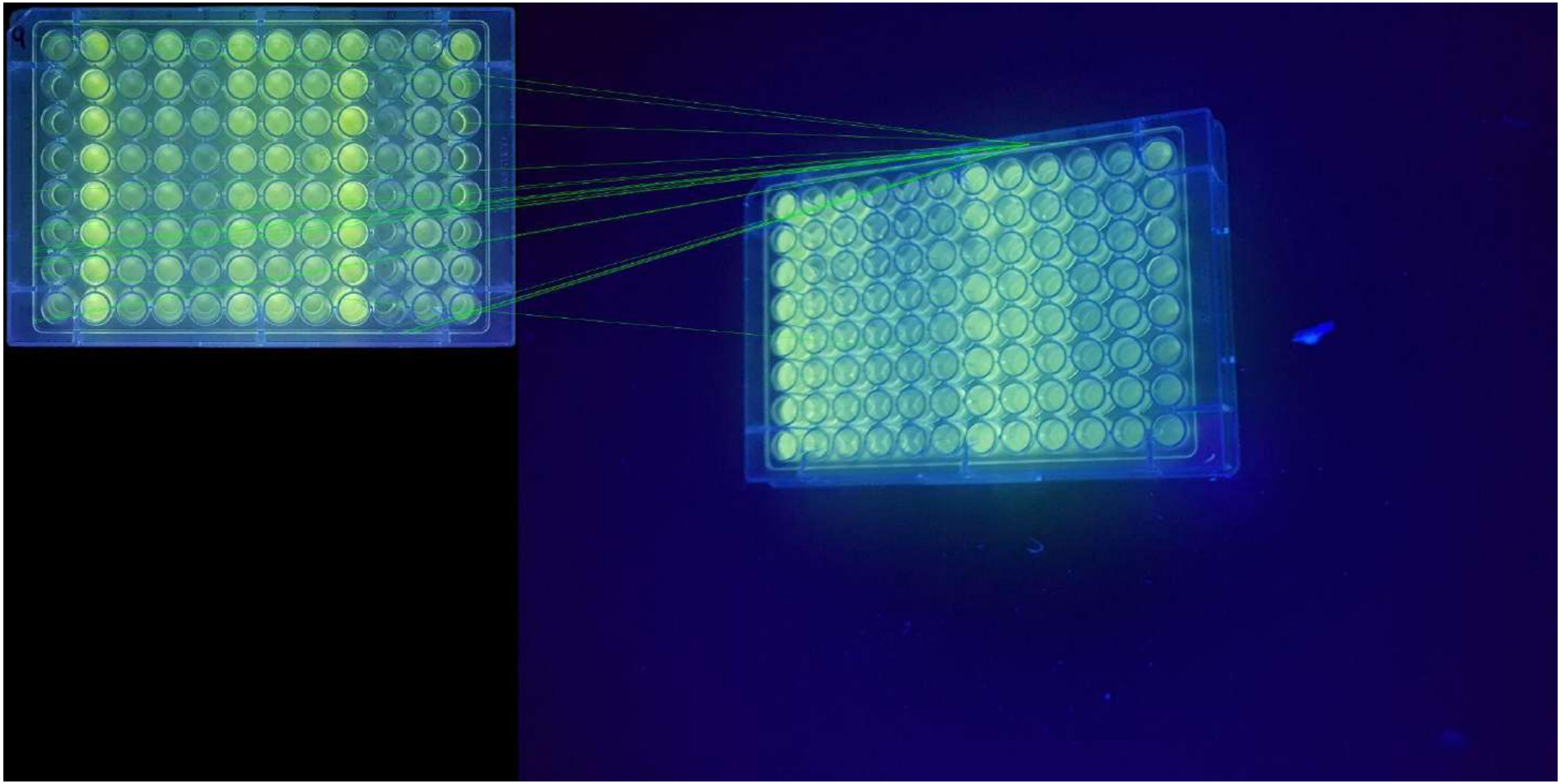




Bad match



Bad match



Main Features

ISO speed (sensitivity of the CMOS sensor toward light)

Mean values and truncated mean at 30% of…
Gray channel, B and R from RGB, and a from Lab