Medical Data Analytics and Beyond

Weichung Wang
Institute of Applied Mathematical Sciences and MeDA Lab
National Taiwan University

International Workshop on the Integration of (Simulation + Data + Learning): Toward Society 5.0 by h3-Open-BDEC, 2021/11/30

Prelude

0

People

Unmet Needs

People

Unmet Needs

Data Driven Discovery

Multimodality Data

Intelligent Analytics

Medical Workflow

People

Unmet Needs

Framework

Data Driven Discovery

Multimodality Data

Intelligent Analytics

Medical Workflow

Medical AI

High Value & High Impact

People

Unmet Needs

Framework

Ecosystem

Solutions

Data Driven Discovery

Multimodality Data

Intelligent Analytics

Medical Workflow

Medical AI

High-Value & High-Impact

Data

Model

Medicine

Data

Multimodality Data

Medical Records | Temporal Data | Medical Imaging | Whole Slide Imaging | Genomic Data

Model

Intelligent Analytics

Medical Knowledge | Deep/Machine Learning | Mathematical Insight | Statistical Analysis | Computational Method

Medicine

Medical Workflow

Segmentation | Detection | Classification | Prediction | Recommendation

Solution Landing

Hospital | Device | Insurance | Research | Education

Data

Multimodality Data

Medical Records | Temporal Data | Medical Imaging | Whole Slide Imaging | Genomic Data

Simulation

Intelligent Analytics

Medical Knowledge | Deep/Machine Learning | Mathematical Insight | Statistical Analysis | Computational Method

Learning

Medical Workflow

Segmentation | Detection | Classification | Prediction | Recommendation

Society 5.0

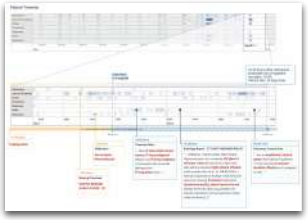
Solution Landing

Hospital | Device | Insurance | Research | Education

MeDA Framework

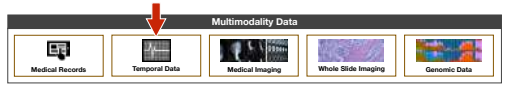
Macroscopic → Microscopic

- **Electronic healthcare records**: encounters, lab data, medications, vital signs, procedures, notes, diagnoses, outcomes, etc.
- **FHIR**: Fast Healthcare Interoperability Resources
- **Various** data, **missing** data
- Deep learning and natural language processing



http://hl7.org/fhir/us/mcode/2019Sep/

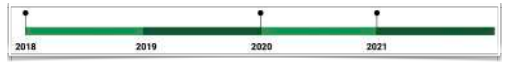
Rajkomar et al., NPJ Digital Medicine, 2018. https://www.nature.com/articles/s41746-018-0029-1



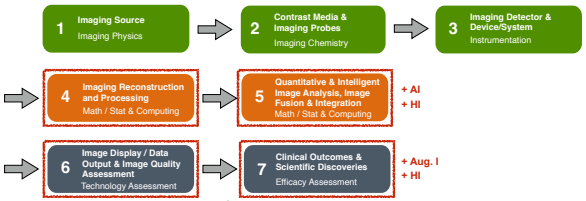
Time series and longitudinal study

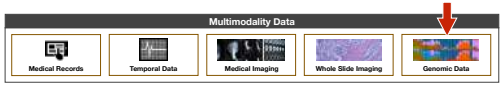
- Risk of atrial fibrillation
- Time for the next colonoscopy
- Morality in the next 24, 48, or 72 hours in ICU
- Brain metastasis velocity prediction for treatment decision

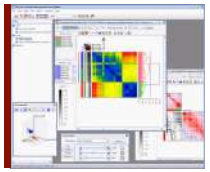
https://en.wikipedia.org/wiki/Electrocardiography



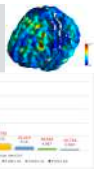
Slide credit: Chien-Min Kao, ATAT2018



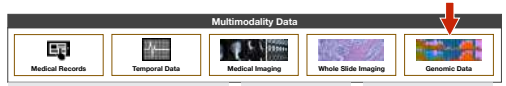


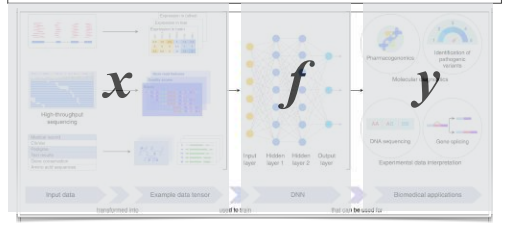


Genome-Wide Association Study
36 billion linear regressions
(653291SNPs x 55749 brain voxels)



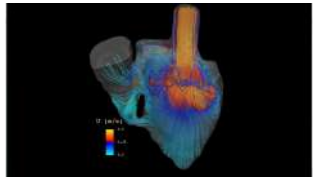
Wu and Chen, Generalized Association Plots, 2012. http://gap.stat.sinica.edu.tw/Software/GAP/

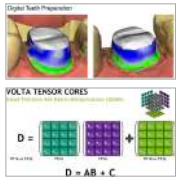




Esteva et al., Nature Medicine, 2019. https://www.nature.com/articles/s41591-018-0316-z

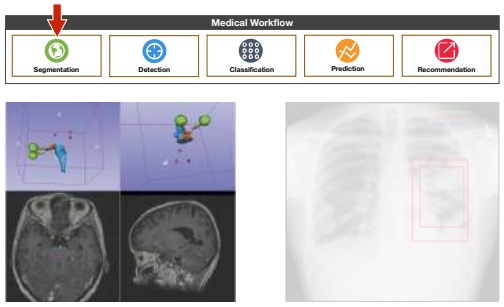
MeDA Framework





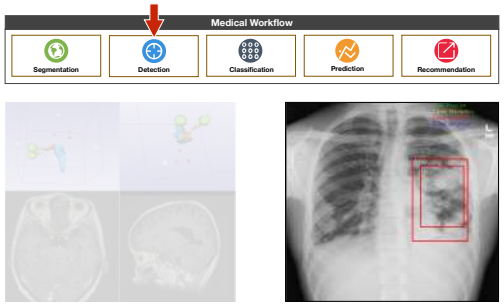
https://youtu.be/Y-gqQXeqJvY?t=138 (2:18)

MeDA Framework



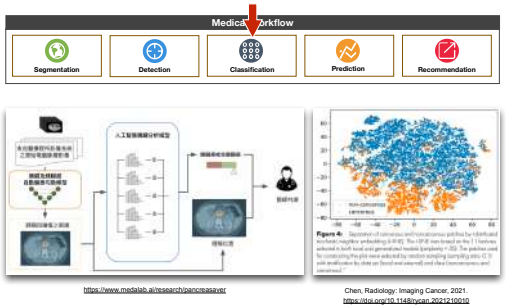
https://www.medalab.ai/research/pca

https://www.medalab.ai/research/smarted



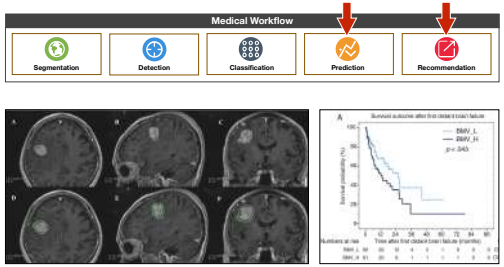
https://www.medalab.ai/research/pca

https://www.medalab.ai/research/smarted



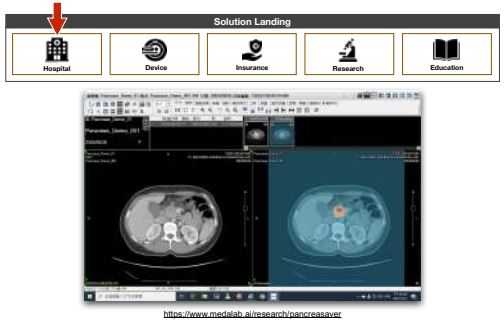
https://www.medalab.ai/research/pancreasaver

Chen, Radiology: Imaging Cancer, 2021.
https://doi.org/10.1148/rycan.2021210010

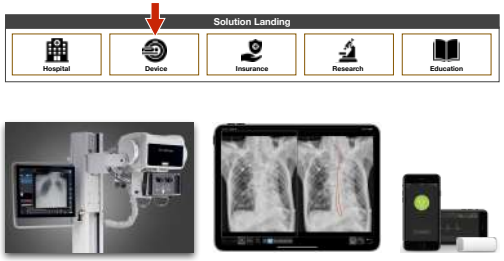


Hsu et al., Neuro-Oncology Advances, 2020. https://academic.oup.com/noa/article/2/1/vdaa100/5897083

MeDA Framework



https://www.medalab.ai/research/pancreasaver

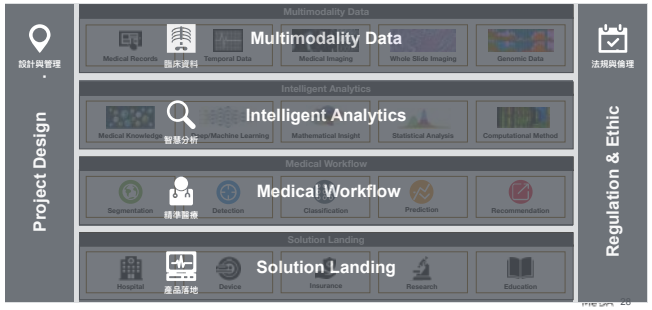


https://www.medalab.ai/research/smarted

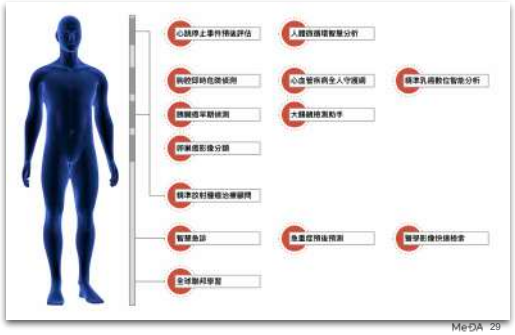
https://www.medalab.ai/research/bodypart

https://www.medaschool.ai/activity

Medical Data Analytics (MeDA) Framework



MeDA Lab
https://medalab.ai/



1. **Make Invisible Visible** — Pancreatic Cancer Early Detection
2. **Global Generalization** — Large-scale Federated Learning
3. **See the Unseen** — Prognosis for Brain Metastases
4. **Find the Tumors** — Segmentation for Brain Metastases
5. **MeDA School** — Medical AI Education

Make Invisible Visible

Pancreatic Cancer Early Detection

1

Needs / Pain Points | Solutions | Impacts

40%
PCs < 2cm missed on computed tomography (CT)

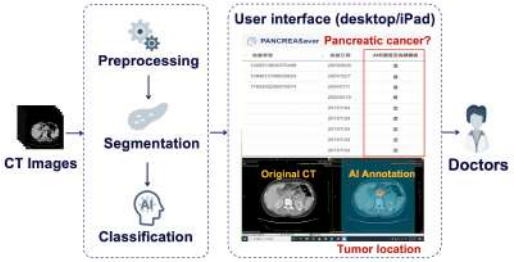
< 10%
5-year survival after surgery

#2
Second leading cause of cancer deaths in US by 2030

US$ 8K
Per month for treatment

Needs / Pain Points | Solutions | Impacts

PANCREASaver



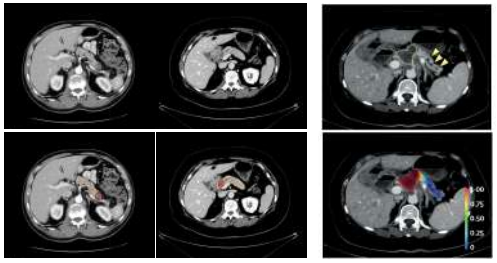
Needs / Pain Points | Solutions | Impacts

85%
Detect 85% PCs < 2 cm (~60% by radiologists) in a NTUH dataset

11/12
Detect 11/12 PCs missed by radiologists in a NTUH dataset

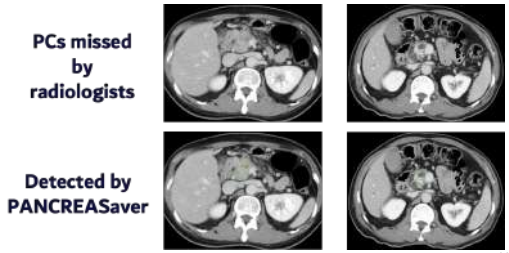
91%
91.4% accuracy in nationwide testing (National Health Insurance)

Needs / Pain Points | Solutions | Impacts



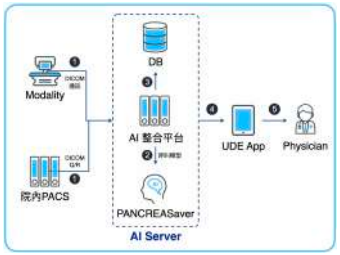
Needs / Pain Points | Solutions | Impacts

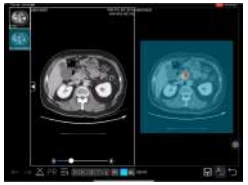
Invisible to human, visible with PANCREASaver



Needs / Pain Points | Solutions | Impacts

PANCREASaver





MeDA NTU Cancer Center 台大癌醫中心 FDA ebm

Global Generalization

Large-scale Federated Learning

2

>16,000 cases across 5 continents

Mass General Brigham 20 client-sites global

MOST

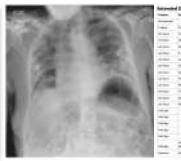
COVID-O2 Int'l Federated Learning

MAHC NVIDIA

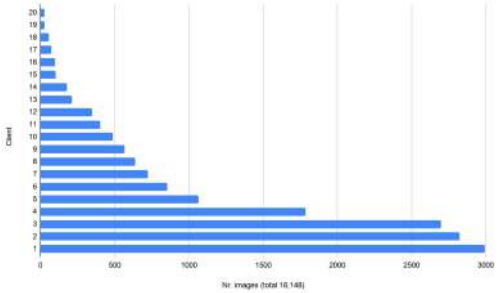
Needs / Pain Points | Solutions | Impacts

COVID-19 pandemic caused a shortage of medical resources

Needs international community to leverage AI to fight the pandemic

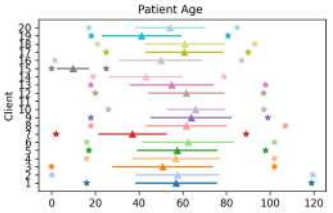


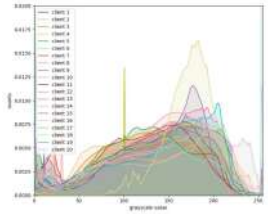
Needs / Pain Points | Solutions | Impacts



Federated Learning used for predicting outcomes in SARS-COV-2 patients, Mona Flores et al., Preprint, 2021

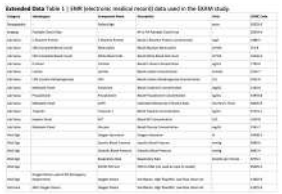
Needs / Pain Points | Solutions | Impacts

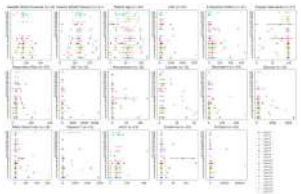




Federated Learning used for predicting outcomes in SARS-COV-2 patients, Mona Flores et al., Preprint, 2021

Needs / Pain Points | Solutions | Impacts



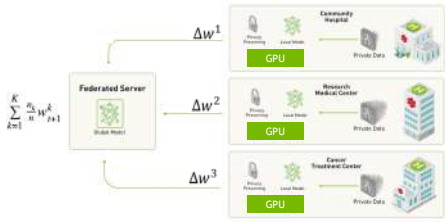


Min. and max. values (asterisks) and mean and standard deviation (length of bars) for each EMR feature used as an input to the model. n specifies the number of sites that had this particular feature available. Missing values were imputed using a MissedForest algorithm.

Federated Learning used for predicting outcomes in SARS-COV-2 patients, Mona Flores et al., Preprint, 2021

Needs / Pain Points | Solutions | Impacts

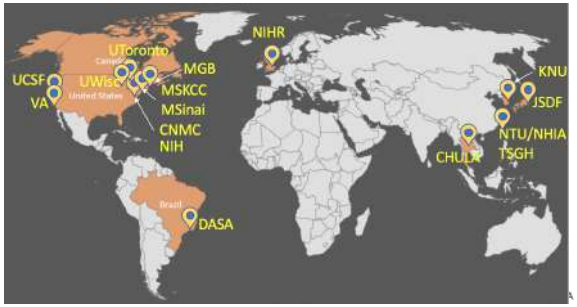
$$\min_{\phi} \mathcal{L}(X;\phi) \quad \text{with} \quad \mathcal{L}(X;\phi) = \sum_{k=1}^{K} w_k \mathcal{L}_k(X_k;\phi)$$



Li, W., et al (2019, October). Privacy-preserving Federated Brain Tumour Segmentation. In International Workshop on Machine Learning in Medical Imaging (pp. 133-141). Springer, Cham.
Rieke, N. et al (2020). The Future of Digital Health with Federated Learning. npj Digital Medicine DOI: 10.1038/s41746-020-00323-1

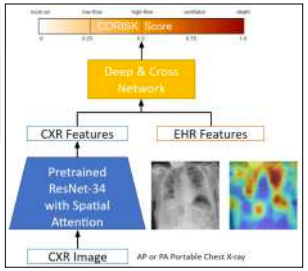
Federated Learning, Dr. Nicola Rieke I Sr. Solution Architect – Healthcare NVIDIA, EMEA, 10.09.2020

Needs / Pain Points | Solutions | Impacts

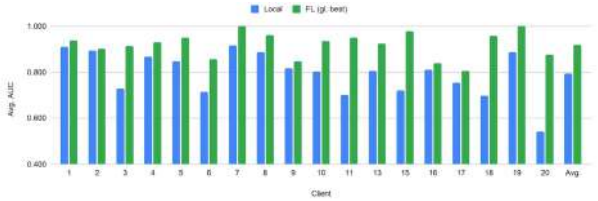


Needs / Pain Points | Solutions | Impacts

https://doi.org/10.1038/s41591-021-01506-3



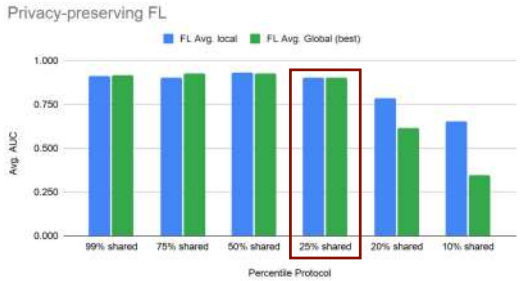
Needs / Pain Points | Solutions | Impacts



Test performance of models predicting 24h oxygen treatment trained on local data only (Local) versus the performance of the best global model available on the server (FL (gl. best)).

Federated Learning used for predicting outcomes in SARS-COV-2 patients, Mona Flores et al., Preprint, 2021

Needs / Pain Points | Solutions | Impacts



Federated Learning used for predicting outcomes in SARS-COV-2 patients, Mona Flores et al., Preprint, 2021

Needs / Pain Points | Solutions | Impacts

Outlook: Functionality

- Streamline data access
- Real-time model inference and processing
- Hyperparameter engineering

Outlook: Data Curation

- Data domain shifts
- Socio-economic status or ethnicity

Consortia Endeavors

- To predict each client-site's contribution for client-site selection and prioritizing data acquisition and annotation
- To capture diversity rather than sheer quantity of data samples

全球大規模聯邦學習

Federated Learning

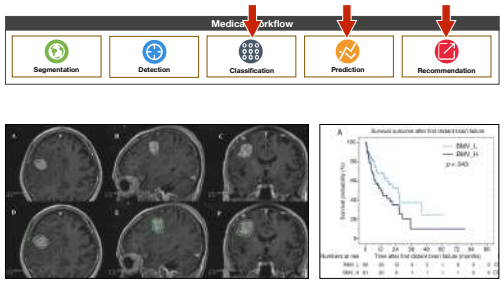
MeDA

https://medalab.ai/research/fl

See the Unseen

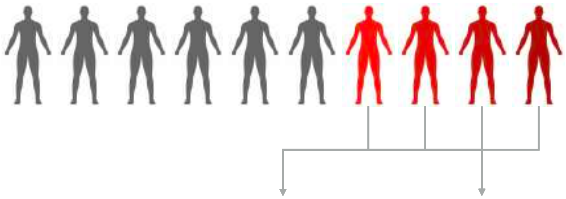
Prognosis for Brain Metastases

3



Hsu et al., Neuro-Oncology Advances, 2020. https://academic.oup.com/noa/article/2/1/vdaa100/5897083

Precision Medicine for Brain Metastases

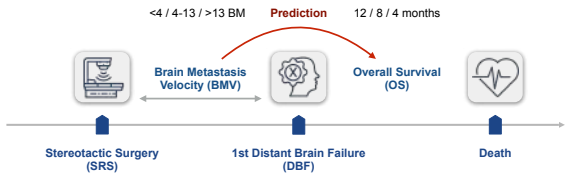


20-40%
Brain Metastases

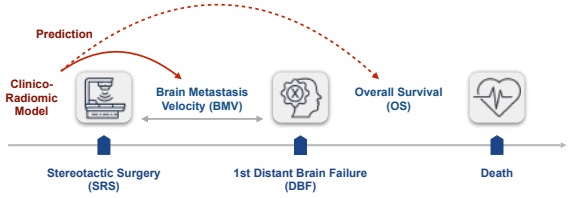
WBRT
Whole Brain Radio Therapy

SRS
Stereotactic Radiosurgery

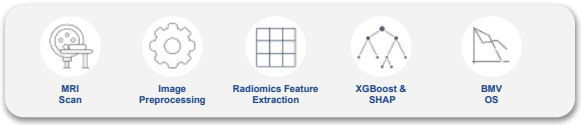
Lung, Breast, Gastrointestinal System Origin Brain Metastasis



Lung, Breast, Gastrointestinal System Origin Brain Metastasis



Clinico-Radiomic Model



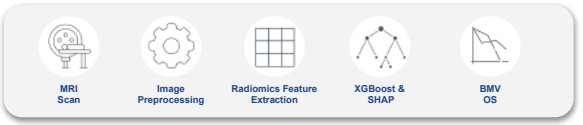
Neuro-Oncology ADVANCES
https://doi.org/10.1093/noajnl/vdaa100

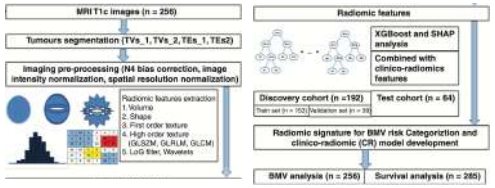
Radiomic analysis of magnetic resonance imaging predicts brain metastases velocity and clinical outcome after upfront radiosurgery

Che-Yu Hsu, Furen Xiao, Kao-Lang Liu, Ting-Li Chen, Yueh-Chou Lee, Weichung Wang

Neuro-Oncology Advances, Volume 2, Issue 1, January-December 2020, vdaa100, https://doi.org/10.1093/noajnl/vdaa100

Clinico-Radiomic Model





Tumor Segmentation

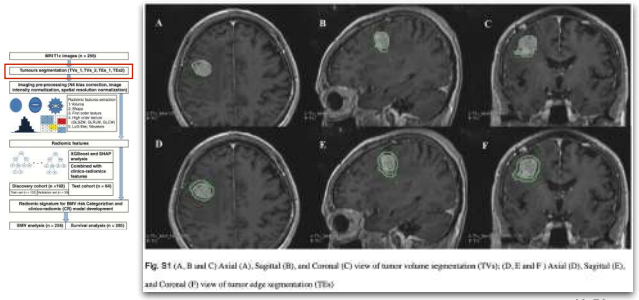
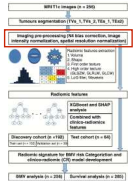
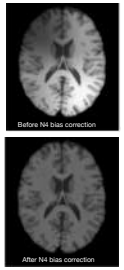


Fig. S1 (A, B and C) Axial (A), Sagittal (B), and Coronal (C) view of tumor volume segmentation (TVs); (D, E and F) Axial (D), Sagittal (E), and Coronal (F) view of tumor edge segmentation (TEs)

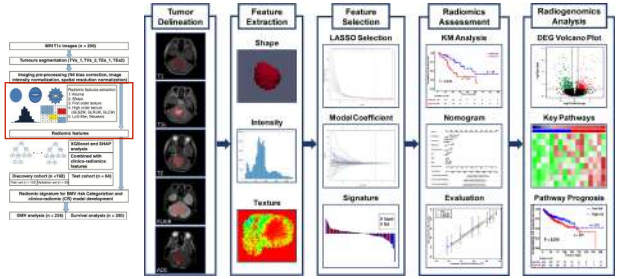
Image Preprocessing



- **N4 bias correction**
  to remove unwanted low-frequency intensity non-uniformity
- **Image intensity normalization**
  to transform arbitrary MR imaging signal intensity values within 5 mm from the TV surface into standardized intensity ranges from 0 to 1024 Gy.
- **Spatial resolution normalization**
  by linear interpolation to make the voxel size isotropic (1×1×1 mm³)

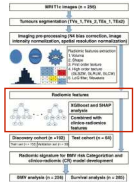


Radiomics: Quantitative Radiographic Phenotyping



Yan et al., EBioMedicine, 2020. https://www.thelancet.com/journals/ebiom/article/PIIS2352-3964(20)30469-2/fulltext

Radiomic Score via XGBoost



- **Intraclass correlation coefficient (ICC)**
  Only radiomic features from TVs_1 and TEs_1 and with ICC values higher than 0.9 were selected
- **Model**

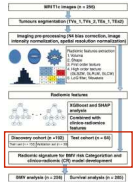
$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F$$

$$L(\phi) = \sum_i l_{\text{loss}}(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2$$

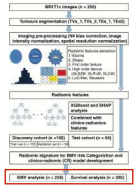
- **Hyper-parameter tuning**
  Grid search, cross-validation, and early stopping for fitness (e.g. learning_rate, scale_pos_weight, and base_score) and regularity (e.g. min_child_weight, max depth, gamma, subsample, reg_alpha, and colsample_bytree).
- **Radiomic score**
  The output of the XGBoost model was converted into a probability score, i.e. the radiomic score (R-score), indicating the probability for the patient to belong to the BMV-H group.

Clinico-Radiomic Model



- **Univariate analysis**
  To determine significant clinical predictors (risk coefficients) via odds ratio and p value
- **Logistic regression**
  To build the Clinical (C) Model and Clinico-Radiomic (CR) Model
- **Model discrimination performance**
  Harrell's C-index, corrected C-index
- **Decision curve analysis (DCA)**
  The net benefit of the Radiomic Score in the CR model was estimated by using the decision curve with the difference between the true-positive and false-positive rates, weighted by the odds of the selected threshold probability of risk.
- **Model Interpretation via SHapley Additive exPlanations (SHAP)**
  The mean absolute SHAP values of radiomic features can represent their impact on the R-score. For the radiomic features with the top 5 SHAP values, we further used **student t-test** and analysis of variance to evaluate their **correlation** with putative clinical factors, where a P value of less than .01 was considered statistically significant.

Statistical Analysis of Clinical Outcomes



- **Median Follow-up**
  From the beginning of SRS to the most recent follow-up or event of interest
- **Time-to-Event Outcomes**
  Via Kaplan-Meier analysis and log-rank test

Clinical and Radiomic Factors for the Multimodality Model

Clinical Factors
- Age
- Gender
- Primary malignancy (Lung, Breast, GI, others)
- Number of bran metastases
- Systemic disease burden
- Extracranial progression (EP)
- Diagnosis-specific graded prognostic assessment (DS-GPA)
- Central necrosis
- Prilesional edema (PE)
- EGFR mutant status (lung)
- Maximum tumor volume

Radiomic Factors
- Redi_Tumor(a radiomic score)
- Redi_Edge (a radiomic score)

Risk coefficients estimated by univariate analysis (Odds ratio and p value)

BMV_H or BMV_L

Clinical Predictors
- Number of Brain metastasis (BM)
- Presence of perilesional edema (PE)
- Status of extracranial progression (EP)
- Redi_Tumor
- Redi_Edge

Discrimination Performance

Clinic Model: Y = -2.50 + 0.53 BM_numbers + 0.79 PE + 1.08 EP

Clinico-Radiomic Model: Y = -4.8 + 0.02 BM_numbers + 0.31 PE + 1.21 EP + 8.78 Radi_Tumor + 4.00 Radi_Edge

影像體學機器學習模型預測顱內復發結果，較傳統臨床因子陣列圖準確提升 **25% (53.1% => 79.7%)**

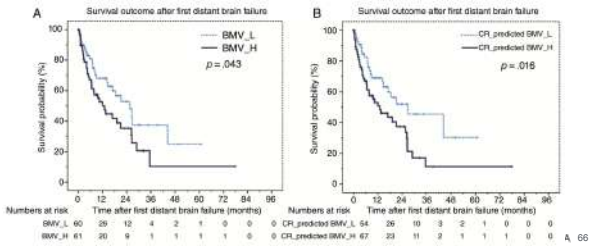
台大醫院 NTUH
2008 ~2018
285 cases

靈敏度 (Sensitivity) 92.3%
特異度 (Specificity) 76.5%
準確度 (Accuracy) 79.7%

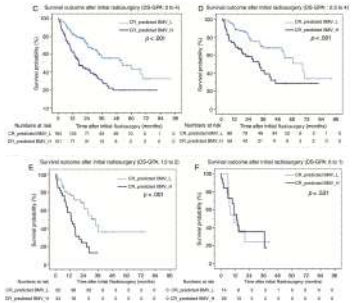
	Test Cohort						
	BMs	PE	EP	C Model	Radi_Tumor	Radi_Edge	CR Model
C-index	0.517	0.679	0.604	0.721	0.713	0.764	0.833
95% CI	0.335–0.698	0.515–0.844	0.421–0.787	0.581–0.861	0.571–0.855	0.598–0.930	0.698–0.967
Corrected C-index	0.453	0.676	0.581	0.7192	0.709	0.770	0.832
Cutoff value	1.5	0.5	0.5	0.275	0.275	0.272	0.215
Sensitivity	46.2%	69.2%	38.5%	69.2%	61.5%	61.5%	92.3%
Specificity	54.9%	66.7%	82.4%	70.6%	64.7%	82.4%	76.5%
PPV	20.7%	34.6%	35.7%	37.5%	30.8%	47.1%	50.0%
NPV	80.0%	89.5%	84.0%	90.0%	86.8%	89.4%	97.5%
Accuracy	0.531	0.672	0.734	0.703	0.641	0.781	0.797

Overall Survival Curves via Kaplan-Meier Analysis

- BMV_L: 25.6 (15.9-35.2)
- BMV_H: 13.0 (7.7-18.4) (month)
- CR_predicted BMV_L: 26.7 (8.1-45.5)
- CR_predicted BMV_H: 13.0 (4.8-21.3)



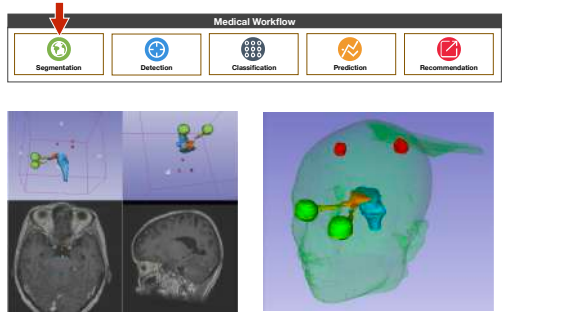
Overall Survival Curves via Kaplan-Meier Analysis



Find the Tumors

Segmentation for Brain Metastases

4



https://www.medalab.ai/research/pca

Tumor Segmentation

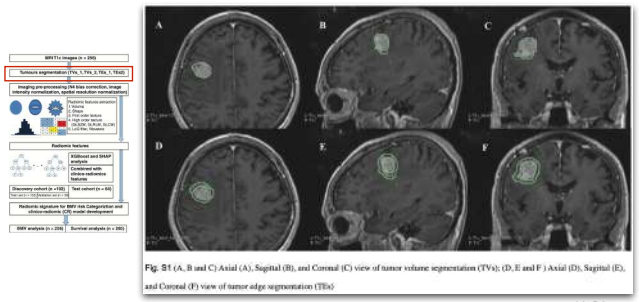
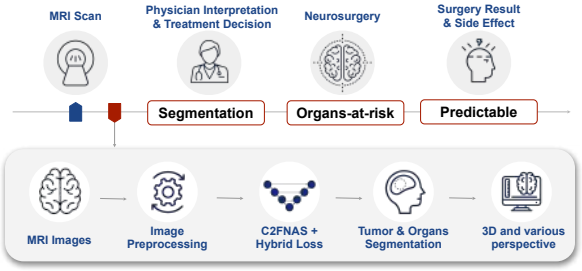
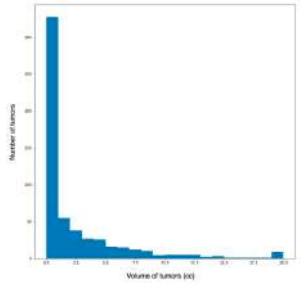


Fig. S1 (A, B and C) Axial (A), Sagittal (B), and Coronal (C) view of tumor volume segmentation (TVs); (D, E and F) Axial (D), Sagittal (E), and Coronal (F) view of tumor edge segmentation (TEs)

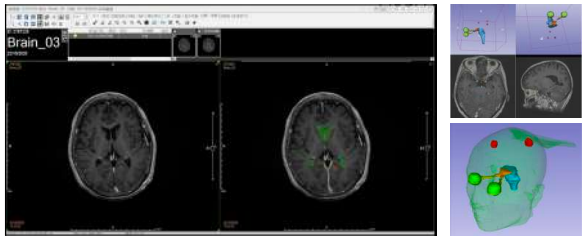
Automatic BM Segmentation AI



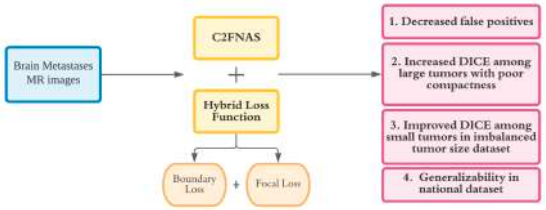
Histogram of the Tumor Volumes in NTUH Dataset



PCA-Segmentation



PCA-Segmentation



Needs / Pain Points | Solutions | Impacts

台大醫院 NTUH

2008 ~2018
67 cases

Large tumors (>1cc):
Precision/Recall 0.962/0.927
False Positive Average: 0.05
False Negative Average: 0.10

Small tumors (<1cc) :
Precision/Recall 0.656/0.701
False Positive Average: 0.54
False Negative Average: 0.44

衛生福利部 中央健康保險署

2018/2 ~ 2019/7
142 cases

DICE score (similarity): 81.8%

Large tumors (>1cc)
FPA 0.569, FNA 0.314

Small tumors (<1cc)
FPA 0.872, FNA 1.064

精準腫瘤治療顧問

Precision Cancer Advisor

MeDA

https://medalab.ai/research/pca

MeDA School

Medical AI Education

5

理工電資醫學院．大學生與研究生 MeDA School 國立臺灣大學

MeDA x 國立臺灣大學 · 人工智慧在醫學影像的分析與應用 · 學期課程

醫師、研發人員、大學生 MeDA School 中研院統計所

MeDA x 中研院統計所 · 2020/8/9-14 · 智慧醫療影像分類 · 短期課程

醫院、醫學院 MeDA School

MeDA School goes to Medical School

MeDA x 慈濟醫院與慈濟大學 · 2021/07/12 - 2021/09/12 · 遠距課程

高中生 MeDA School AI4kids

MeDA x AI4kids · 高中生 AI 醫療專題實作營 · 2021、2022 年 / 台北 / 新竹

線上課程 MeDA School YouTube teach:able

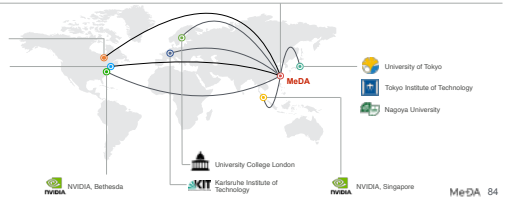
MeDA School

為台灣與全球的智慧及精準醫療人才庫，建立永續經營之厚實基礎

https://medaschool.ai

National Taiwan University
National Taiwan University Hospital
Ministry of Science and Technology
Academia Sinica
Taipei Medical University Hospital
Changhua Christian Medical Foundation
Taipei Veterans General Hospital
Chang-Geng Memorial Hospital
MacKay Memorial Hospital
NTU Cancer Center
Tzu Chi Hospital
Chi Mei Medical Center
Taiwan Food and Drug Administration
National Health Insurance Administration
Institute for Information Industry
Industrial Technology Research Institute
EBM Technologies
InnoCare
Asus
Advantech
Cohesion Data
Quanta Computer
NVIDIA,Taipei
Acer Healthcare

University of Chicago
Harvard Medical School
Massachusetts General Hospital
IBM T. J. Watson Research Center
Mount Sinai Health Center
New York University
Rhino Health



Help Doctors to Help People

MeDA 國立臺灣大學 http://meda.ai