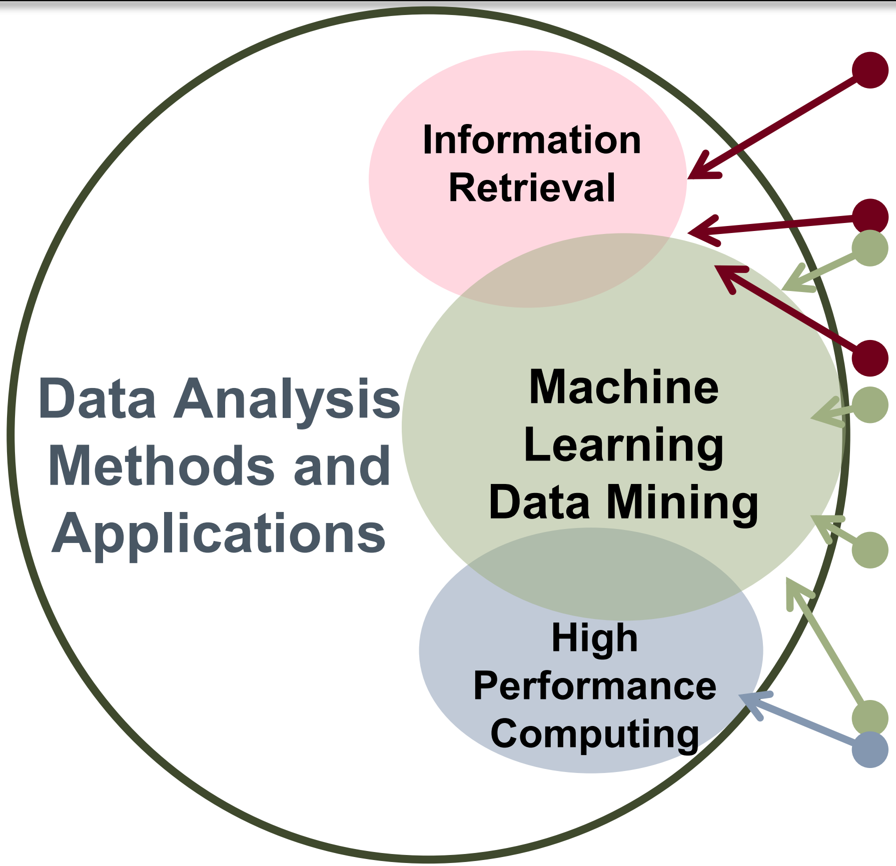
AI for the Greater Good

David C. Anastasiu

Research Summary



Improving Web search utility
[WWWJ'13, IEEE IC'2013, ACM CIKM'09, DMIN'09]

Clustering and pattern mining
[IEEE BDS'19, IEEE MC'19, StatsRef'17, SIP'14, CRC'13, IEEE CIKM'11, IEEE SIGIR'11, COLING'10]

Traffic analytics from video
[IEEE CVPR'23, IEEE CVPR'22, IEEE CVPR'21, BDAT'20, IEEE CVPR'20, IEEE CVPR'19, IEEE SOSE'19, IEEE MC'19, IEEE CVPR'18, IEEE SmartWorld'17]

Applications of machine learning and data mining
[IEEE BigData'23, AAAI'23, Bioinformatics'23, DataBrief'23, GHTC'22, ECTEL'22, Microbio'22, KDD-UC'22, iDSC'19, IEEE CIKM'18, GHC'18, IEEE SCI'17, IEEE ICDE'15]

Fast nearest neighbors computation
[JPDC'17, JDSA'17, iDSC'17, IEEE DSAA'16, IA3'16, ACM CIKM'15, IA3'15, IEEE ICDE'14]

Outline

- Hydrologic Flow Prediction
  - An Extreme-Adaptive Time Series Prediction Model Based on Probability-Enhanced LSTM Neural Networks
  - SEED: An Effective Model for Highly-Skewed Streamflow Time Series Data Forecasting
- Chronic Kidney Disease
  - On-Device Prediction for Chronic Kidney Disease
  - Color Constancy for Accurate CKD Prediction
- Other Current Projects
- References

An Extreme-Adaptive Time Series Prediction Model Based on Probability-Enhanced LSTM Neural Networks

Yanhong Li, Jack Xu, **David C. Anastasiu**

AAAI 2023

Problem: Univariate Time Series with Extreme Events Forecasting

$$[X_1, X_2, \ldots, X_T] \in R^T \rightarrow [X_{T+1}, \ldots, X_{T+H}], \in R^H$$

Challenges:

- A majority of normal values that significantly contribute to the overall prediction performance;
- A minority of extreme values that must be precisely forecasted to avoid disastrous events.

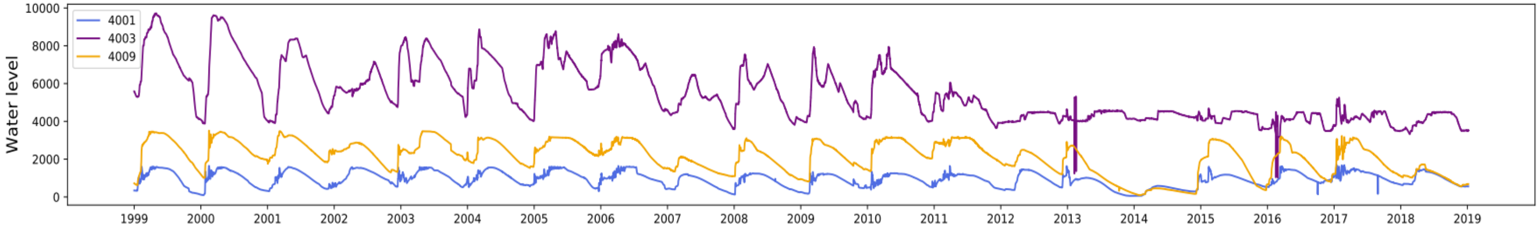
Goal:

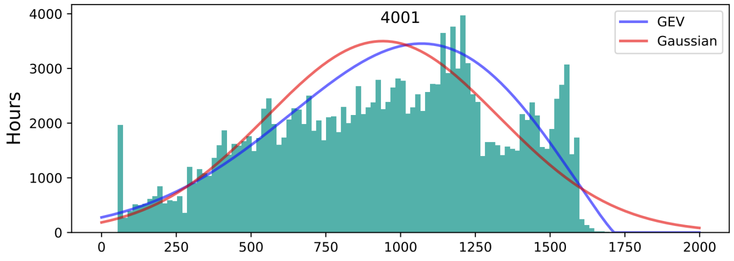
- A model concurrently learns extreme and normal prediction functions;
- Long sequence forecasting;
- Good generalization.

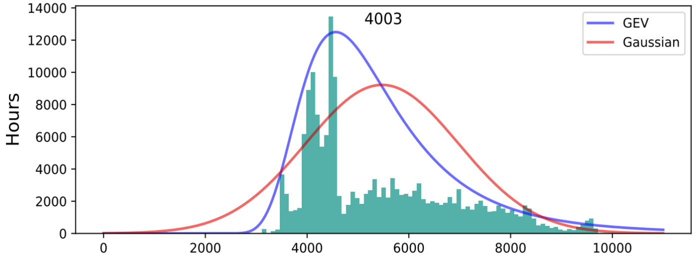
Dataset:

- We evaluate the proposed model on the difficult 3-day ahead hourly water level prediction task applied to 9 reservoirs in California.

Extreme Events:







GEV (Generalized Extreme Value) distribution provides a better fit, showing the presence of extreme values in our data,

$$F(x;\mu,\sigma,\xi) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, \quad (1)$$

where $\mu \in \mathbb{R}$, $\sigma > 0$, and ξ are the location, scale, and shape parameters, respectively, conditioned on $1+\xi(x-\mu)/\sigma > 0$.

Motivation: achieving the best overall prediction performance, without sacrificing either the quality of normal or of extreme predictions.

Root Mean Square Error ***(RMSE)***

Mean Absolute Percentage Error ***(MAPE)***

Proposed Methods:

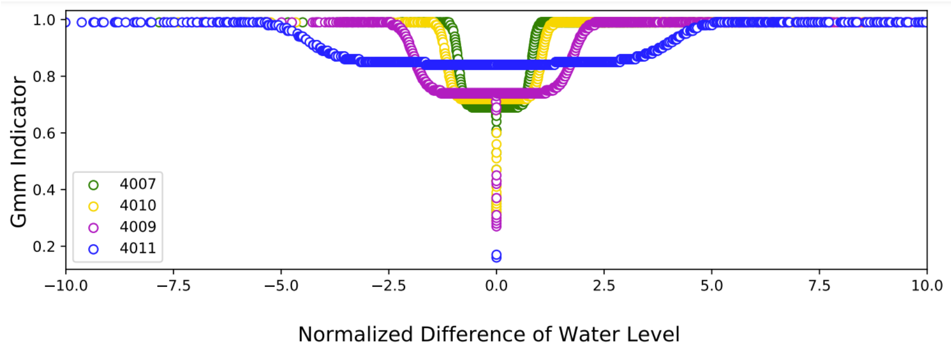
GMM Indicator : an unsupervised clustering approach to dynamically produce distribution indicators, which improves the model's robustness to the occurrence of severe events.

NEC framework : a framework to account for the distribution shift between normal and extreme values in the time series.

Selected Backpropagation : to help the models learn the positions and values of appropriate normal or extreme data better.

Parameterized Loss Function : a model concurrently learns extreme and normal prediction functions.

GMM Indicator :



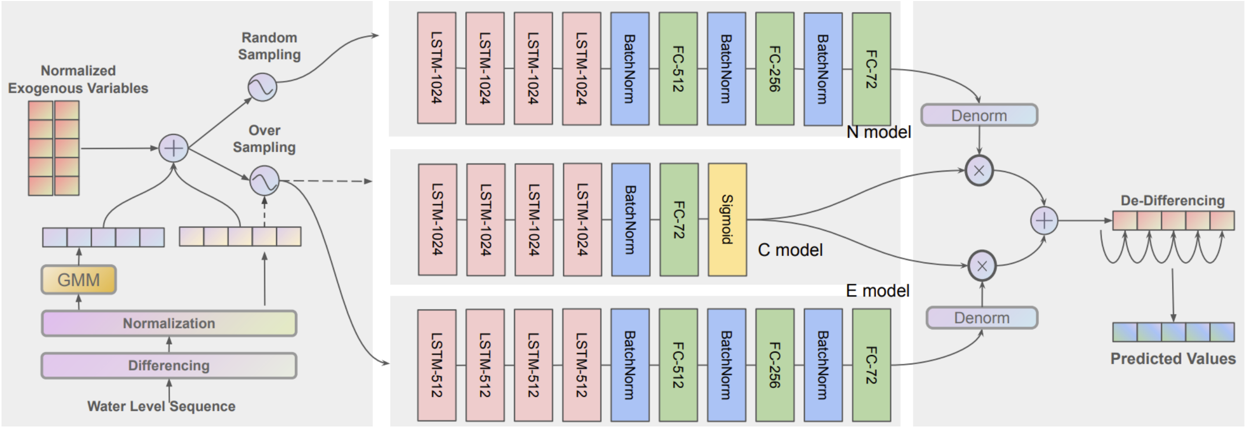
A Gaussian mixture model (GMM) is a weighted sum of M component Gaussian densities,

$$p(x|\lambda) = \sum_{i=1}^{M} w_i g(x|\mu_i, \sum i) \ ,$$

where x is a D-dimensional continuous-valued vector, w_i ∀i = 1, . . . , M are the mixture weights, and $g(x|\mu_i, \sum i)$ are the component Gaussian densities.

We compute an indicator feature as the weighted sum of all component probabilities, given the weights learned when fitting the GMM model.

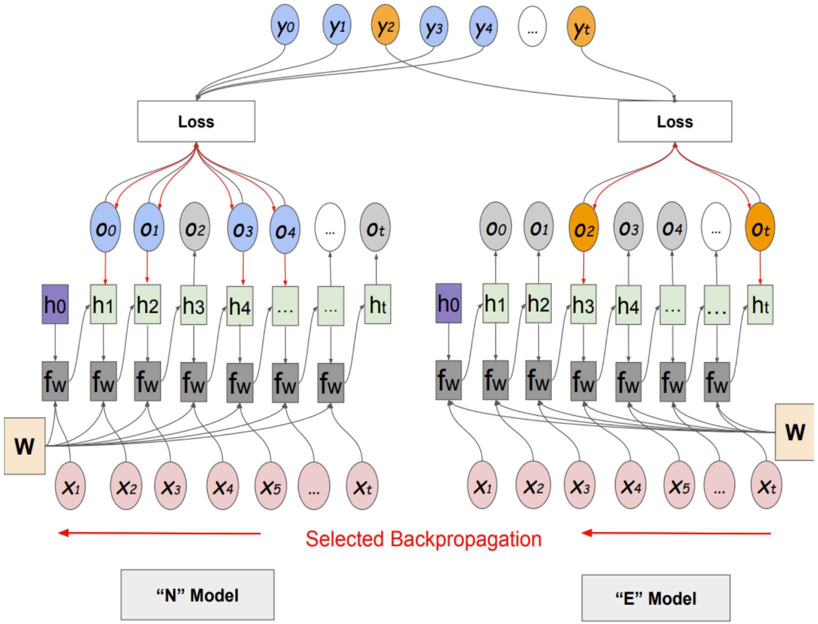
NEC framework :



NEC is composed of three separate models, which can be trained in parallel:

- The Normal (N) model is trained to best fit normal values in the time series;
- The Extreme (E) model is trained to best fit extreme time series values;
- The Classifier (C) model is trained to detect when a certain value may be categorized as normal or extreme.

Selected Backpropagation :



two-stage sampling policy:

1, randomly sample subsections of length h + f

from the series as samples to use in training our models,

2, perform stratified sampling of regions with and without extreme values, allowing the E and C models to oversample up to OS% samples with at least 1 extreme value in the prediction zone.

Selected Backpropagation:

N model: only normal values add to the loss;

E model: only extreme values add to the loss.

Parameterized Loss Function:

$$BCE(t, p) = -(t \times \log(p) + (1 - t) \times \log(1 - p))$$

$$L = \beta \times BCE(t, p^{\alpha}) + (1 - \beta) \times RMSE(t, p)$$

where α and β are parameters that can be tuned. Values $\alpha > 1$ cause the model to predict p values that are higher in general in order to minimize the distance between t and p^{α}.

Problem: for datasets with a high imbalance between the two classes, BCE will favor the prominent class.

- The BCE part: can be thought of as a blunt instrument that grossly exaggerates all miss-classifications in order to more accurately predict the obscure class;
- The RMSE part: allows for a more gentle penalty based on the distance between t and p.

Research Questions:

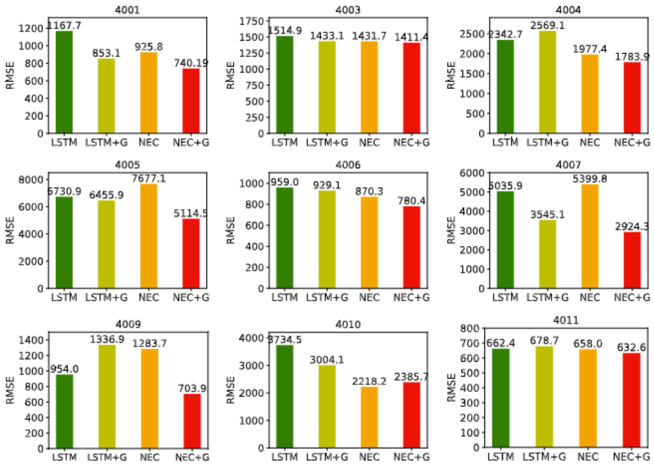
1. What is the effect of adding the GMM indicator to a model?
2. What is the effect of introducing exogenous features?
3. How do the loss function parameters affect performance?
4. How does NEC+ compare against state-of-the-art baselines?

Baselines:

- ARIMA
- Prophet
- LSTM
- DNN-U: univariate LSTM-based encoder-decoder hydrologic model.
- Attention-LSTM: a state-of-the-art hydrologic model used to predict stream-flow
- N-BEATS: a state-of-the-art time series prediction method that outperformed all competitors on the standard M3, M4 and TOURISM datasets.

Research Questions:

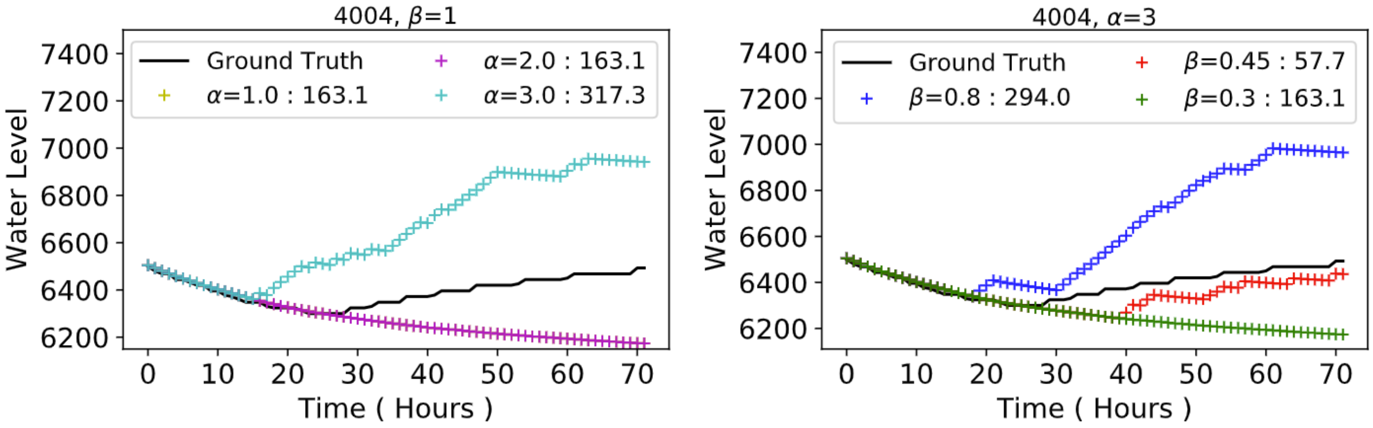
1. What is the effect of adding the GMM indicator to a model?
2. What is the effect of introducing exogenous features?



Model/Reservoir	4005	4007	4010
LSTM	6730.93	5035.91	3734.53
LSTM+W	7568.68	5728.30	4145.16
LSTM+G	6455.90	3545.19	3004.14
LSTM+G+W	9760.62	4128.37	2602.58
NEC+G	5114.49	2924.30	2385.77
NEC+G+W (NEC+)	**4352.74**	**2092.73**	**2275.48**

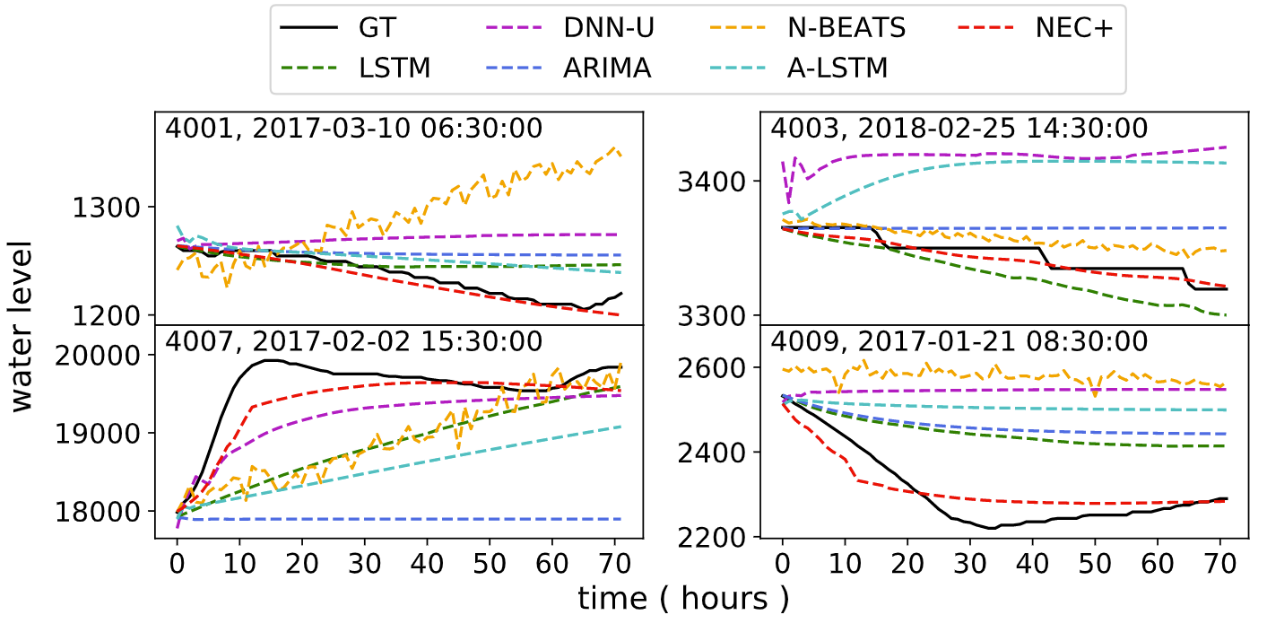
Research Questions:

3. How do the loss function parameters affect performance?



Research Questions:

4. How does NEC+ compare against state-of-the-art baselines?



Overall Evaluation:

Effectiveness Comparison (RMSE) of NEC+ Against Baselines for 9 Reservoirs

Model/Reservoir	4001	4003	4004	4005	4006	4007	4009	4010	4011
ARIMA	1016.32	1859.70	2501.97	9692.87	1039.38	5854.48	1060.05	3465.20	690.23
Prophet	8469.74	38827.22	95279.31	181607.50	20904.57	187603.80	28629.44	114115.4	2829.26
LSTM	1167.73	1514.90	2342.71	6730.93	959.05	5035.91	954.04	3734.53	662.48
DNN-U	1162.01	1597.72	3989.20	9878.41	983.27	4320.40	1411.63	4257.58	763.73
A-LSTM	878.71	1536.04	2548.56	8919.33	1638.65	13529.86	1064.15	2914.75	700.50
N-BEATS	937.24	1926.74	2280.83	7153.82	960.42	3153.76	1295.90	3162.17	**514.30**
NEC+	**740.19**	**1411.44**	**1783.92**	**4352.74**	**780.46**	**2092.73**	**703.93**	**2275.48**	632.61

MAPE of NEC+ vs. Baselines for 9 Reservoirs

Model/Reservoir	4001	4003	4004	4005	4006	4007	4009	4010	4011
ARIMA	1.3573	0.7626	0.8694	1.2560	1.5401	0.8517	0.9504	1.7871	3.2914
Prophet	16.7877	19.8559	38.9642	35.6662	56.0537	32.9152	31.8069	45.2579	15.3312
LSTM	1.6697	0.6153	0.7450	1.0092	1.3264	0.9253	0.9298	2.5520	3.1282
DNN-U	1.6509	0.6812	1.8738	1.9394	1.4551	0.6509	1.5604	2.1582	3.7131
A-LSTM	1.3533	0.6506	0.8424	1.2060	2.8017	2.1738	0.9705	1.3986	3.4137
N-BEATS	1.3346	0.7972	0.7882	1.1405	2.0061	0.4709	1.4580	1.7146	**2.3108**
NEC+	**1.0319**	**0.5687**	**0.6030**	**0.6350**	**1.0662**	**0.3316**	**0.5992**	**1.2894**	2.9237

SEED: An Effective Model for Highly-Skewed Streamflow Time Series Data Forecasting

Yanhong Li, Jack Xu, **David C. Anastasiu**

IEEE BigData 2023

Problem: predicting long-term streamflow values with rain off data

$$[x_1, x_2, \ldots, x_T] \in \mathbb{R}^T \rightarrow [x_{T+1}, \ldots, x_{T+H}] \in \mathbb{R}^H,$$

x_1 to x_T : the input sequence

x_{T+1} to x_{T+H} : the output sequence

*In our research: H = 3 * 24 * 4 = 288, with majority of normal values and much fewer extreme values which cause the data skewness to one side.*

Challenges:

- Long-range dependencies.
- Rare but important extreme values; very imbalanced data.

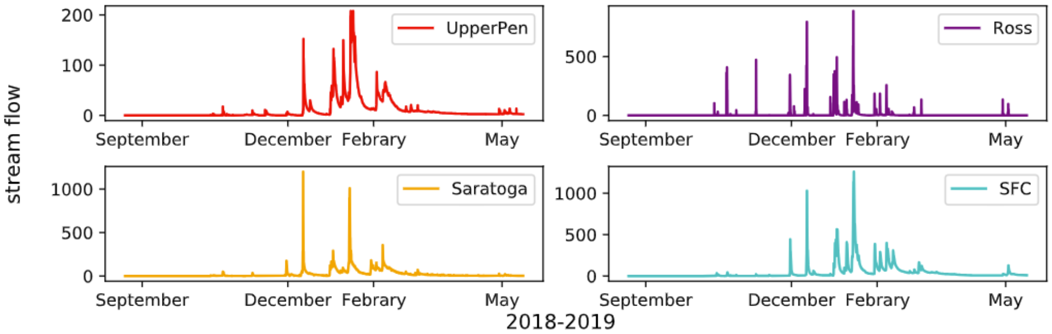
Goal:

- An end-to-end extreme-adaptive model;
- Long sequence forecasting *(predicted length = 288)*;

Dataset:

- Four groups of hydrologic datasets from Santa Clara County, CA.
- Namely Ross, Saratoga, UpperPen, and SFC, named after their respective locations.

Dataset with high skewness and kurtosis scores:



Four streams: Ross, Saratoga, UpperPen, and SFC.

Hydro year: from September to May.

High skewness and kurtosis scores indicate that there is significant deviation from a normal distribution in our data!

Statistic / Stream	Ross	Saratoga	UpperPen	SFC
mean	2.91	5.77	6.66	20.25
max	1440.00	2210.00	830.00	7200.00
min	0.00	0.00	0.00	0.00
median	0.17	1.00	3.20	1.20
variance	597.22	711.09	452.90	12108.14
skewness	19.84	19.50	13.42	18.05
kurtosis	523.16	697.78	262.18	555.18

Motivation: achieving the best overall prediction performance, without sacrificing either the quality of normal or of extreme predictions.

Root Mean Square Error ***(RMSE)***

Mean Absolute Percentage Error ***(MAPE)***

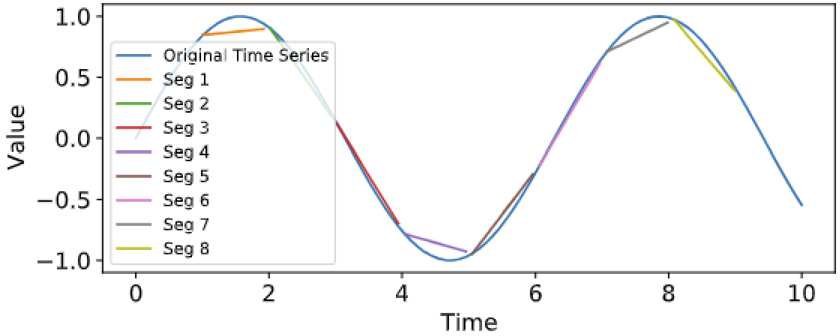
Proposed Methods:

Framework: Segment-Expandable Encoder-Decoder (SEED) model, which is the first to integrate segment representation learning with a multi-tiered encoder-decoder framework.

Importance-enhanced sampling strategy: embedded within the SEED model, allowing it to skillfully identify key features and trends in datasets.

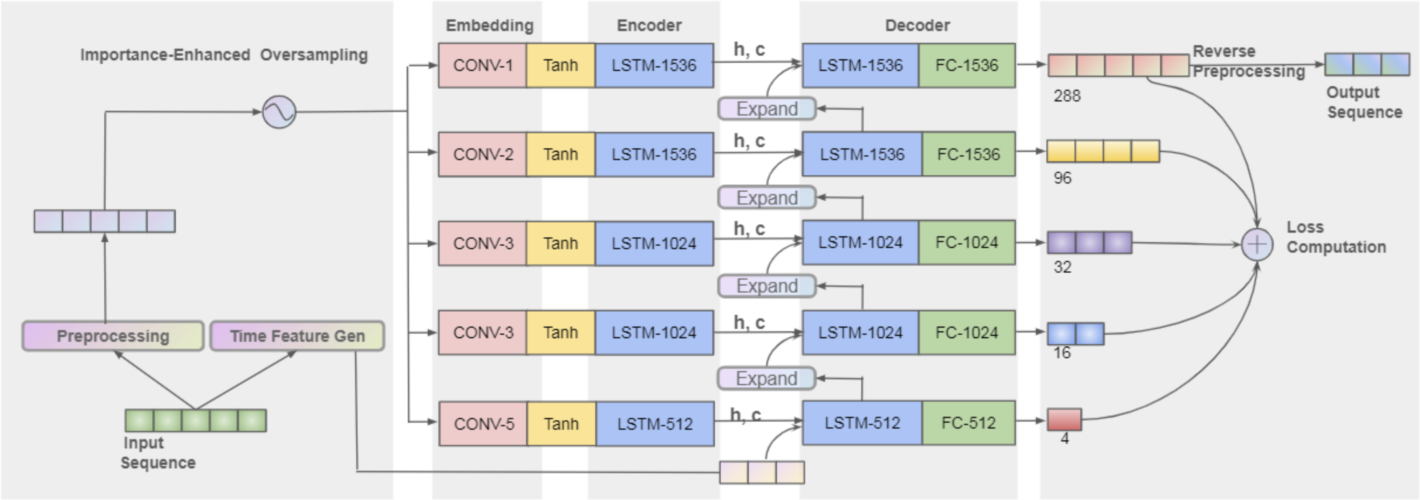
Representation Learning: A unique regularization strategy that incorporates a Kullback-Leibler divergence regularization loss term across multiple stacked layers, thereby increasing the model's robustness against anomalous events with divergent distributions.

Background: Piecewise Linear Representation (PLR)



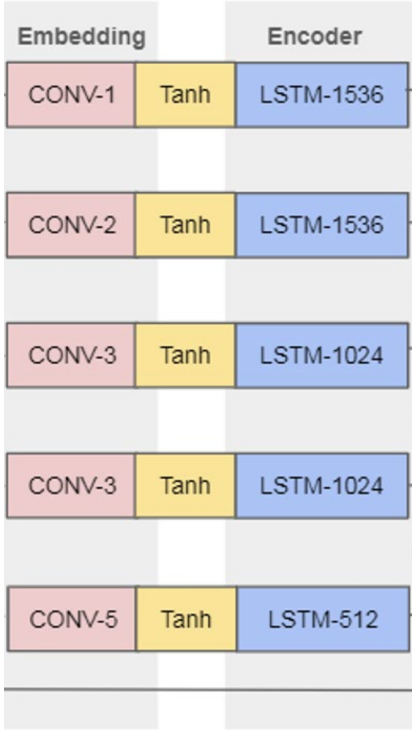
- PLR splits a series into several segments such that the maximum error of each segment does not exceed a threshold;
- **Prior work**: PLR describes the **linear** relationship of the multi-segment representation, mainly works as a preprocessing step to reduce both the space and computational cost of storing and transmitting time series.
- **Our work**: inspired by PLR, SEED learns **nonlinear segment representations** for heavily skewed long term time series.

SEED framework :



- Comprises three core components: embedding, encoder, and decoder.
- The encoder generates a unique hidden state and a cell state which serve as the initial values for the corresponding layers in the decoder.
- Each decoder layer is assigned a distinct task, as they represent the mean value distribution of different lengths of subsegments in the predicted sequence.

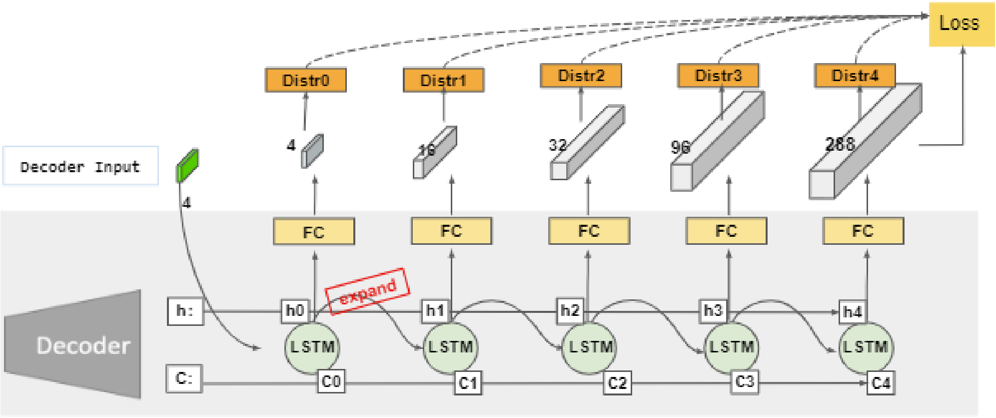
Convolutional Embedding Layers:



CNN Layers :

- different kernel sizes to extract features at different spatial scales.
- subsequent tanh activation function;
- lower level: larger kernel sizes, capturing broader patterns and global context.
- higher level: smaller kernel sizes, capturing local patterns and fine grained details.

Decoder Architecture:



- First level: the output length is 4, which is meant to predict the mean values of 4 segments, each of which contains 288/4 = 72 points in the forecasted series.
- Second level: the 4 outputs are expanded to16, each represents the mean value of 288/16 = 18 points.
- Expansion: ⟨a, b, c, d⟩ becomes ⟨a, a, a, a, b, b, b, b, , c, c, c, c, d, d, d, d⟩, in the **high-dimensional hidden space**.
- By predicting the mean value of different length sub-segments, **extreme values** are represented and spread across multiple levels in the hierarchy, leading to higher mean values in the segments containing them.

Multiple-Objective Loss Functions:

$$\mathrm{KL}(p \parallel q) = \sum p(x) \log \left(\frac{p(x)}{q(x)} \right),$$

$$\mathcal{L}_i = \mathrm{KL}\left(\mathrm{softmax}(p\_m_i), \mathrm{softmax}(g\_m_i)\right),$$

$$\mathcal{L} = RMSE(\hat{y}, y) + \lambda \times \left(\sum_{i=1}^{k} \mathcal{L}_i \right),$$

Motivations:

- Kullback-Leibler divergence loss acts as a regularization term that encourages the model to match the target distribution while balancing the sequence generation loss.
- p_m_i represents the predicted segment mean values in the *ith* layer, while g_m_i is the vector of computed ground truth mean values for the segments in the *ith* layer. .

Importance-Enhanced Oversampling Policy:

Input : Dataset with training and inference sequences
Output: Oversampled training set

Procedure Oversampling;

```
while training set size is not satisfied do
    Randomly sample a sequence including training and
      inference sections;
    if maximum value in inference section > T then
        Mark sequence as important;
        Move maximum value to the middle of the inference
          section of the sequence;
        foreach index I in the sequence with step size S do
            Sample starting at I;
            Add sampled sequence to oversampled training
              set;
        end
    end
    else
        Add sequence to oversampled training set;
    end
end
```

Steps:

- important sequences: maximum values in the inference section of the series exceed a threshold T;
- moving the maximum value to the middle of the inference section;
- multiple iterations of sampling from the beginning with a specified step size S.

Baselines:

- FEDFormer
- InFormer
- NLinear
- Dlinear
- NEC+
- EnDecoder, the common encoder-decoder model built with LSTM layers.

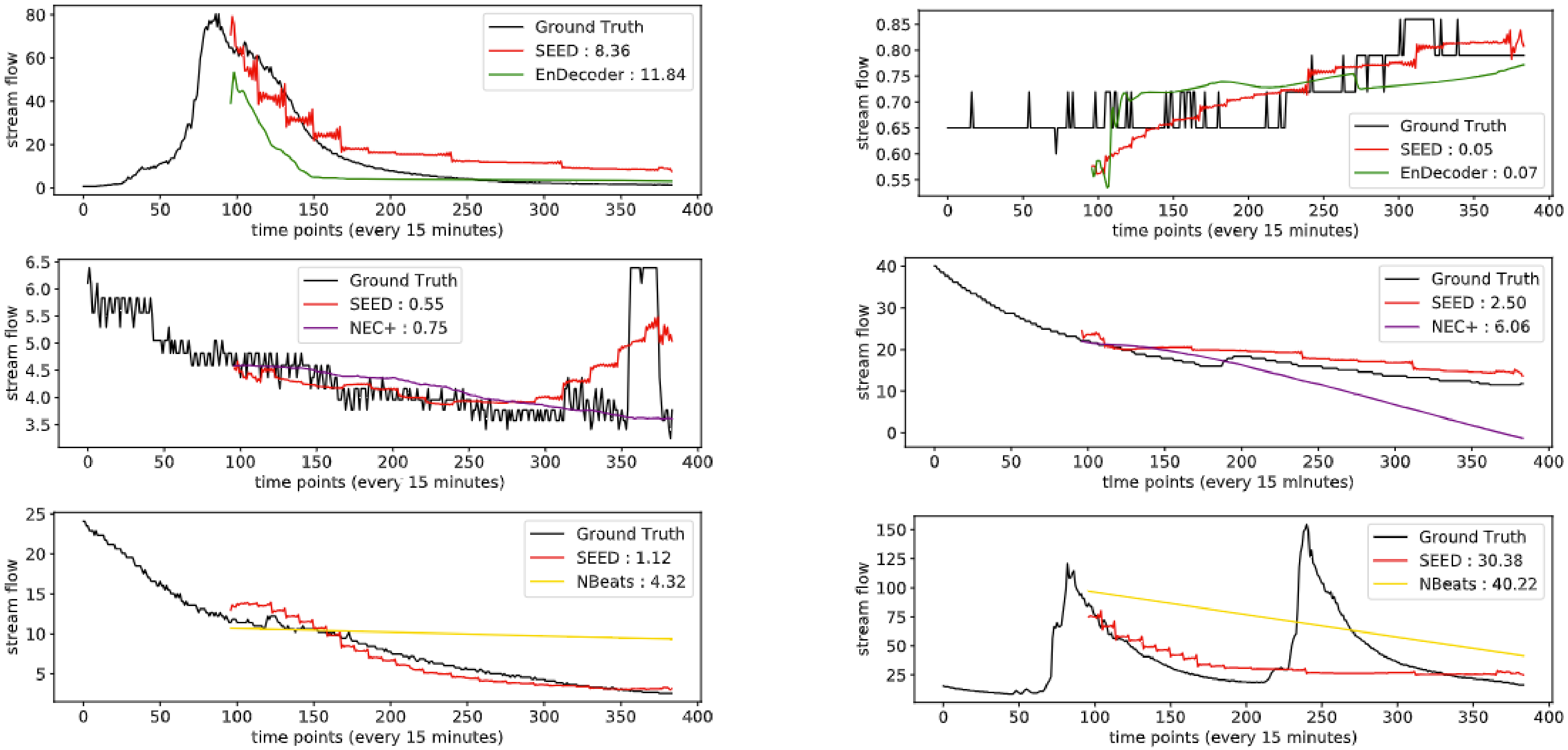
Main results:

Methods	Metric	Ross			Saratoga			Upperpen			SFC		
		All	High	Low	All	High	Low	All	High	Low	All	High	Low
FEDformer	RMSE	6.49	30.82	2.51	6.85	11.95	4.59	2.38	17.68	1.07	24.15	94.28	6.68
	MAPE	2.49	5.27	2.04	2.26	1.50	2.60	1.02	2.37	0.90	2.81	1.70	3.09
Informer	RMSE	9.14	31.00	5.56	4.89	13.64	1.01	5.33	16.26	4.40	19.00	85.40	2.46
	MAPE	5.45	5.80	5.39	0.73	1.40	0.43	4.21	2.70	4.34	0.54	0.71	0.49
Nlinear	RMSE	5.84	32.12	1.54	4.98	14.61	0.70	1.74	15.07	0.61	18.43	83.31	2.26
	MAPE	1.62	4.89	1.09	0.75	1.74	0.31	0.57	1.69	0.47	0.87	1.03	0.83
Dlinear	RMSE	6.90	30.96	2.97	4.06	7.63	2.48	3.25	14.01	2.33	23.64	79.76	9.65
	MAPE	2.79	4.03	2.58	1.31	0.85	1.51	2.04	1.68	2.07	4.02	1.04	4.76
NEC+	RMSE	9.33	38.34	4.58	1.95	5.55	0.35	1.94	13.92	0.92	16.39	76.63	1.38
	MAPE	4.53	8.33	3.91	0.21	0.30	0.17	0.80	0.84	0.80	0.55	0.61	0.54
NBeats	RMSE	5.16	30.09	1.08	3.60	9.44	1.01	1.23	13.20	0.21	31.47	95.33	15.55
	MAPE	1.25	3.17	0.94	0.70	1.21	0.47	0.25	0.78	0.20	3.24	0.88	3.83
EnDecoder	RMSE	5.58	30.81	1.45	1.93	5.69	0.26	2.95	16.33	1.80	17.46	79.04	2.11
	MAPE	1.62	3.72	1.28	0.16	0.29	0.11	1.81	2.34	1.76	0.84	0.96	0.80
SEED	RMSE	**4.23**	**29.74**	**0.05**	**1.67**	**5.14**	**0.12**	**1.07**	**12.83**	**0.07**	**14.44**	**70.04**	**0.59**
	MAPE	**0.11**	**0.53**	**0.04**	**0.09**	**0.19**	**0.05**	**0.10**	**0.57**	**0.06**	**0.20**	**0.42**	**0.14**

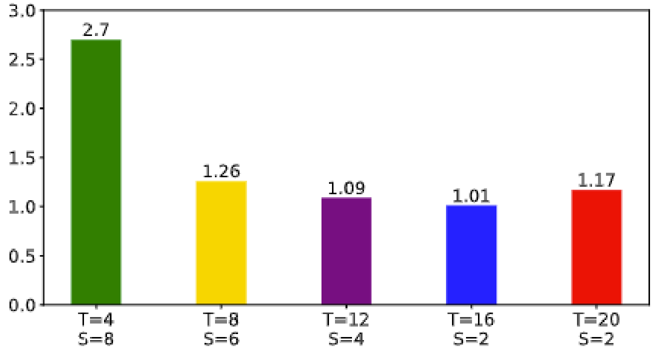
- Univariate Long-Term (h = 288) Series Forecasting Results.
- Over 1600 test points in the test set were inferenced on all datasets.
- The **best results** are in bold and the second best results are underlined.
- ``All‘’ represents the average RMSE of all test samples compared with the ground truth. `’High‘’ means larger than the mean value; ``Low‘’ includes test samples lower than the mean value.

In comparison to the three second-best models (NEC+, Nbeats and EnDecoder), SEED achieved, on average, relative RMSE reduction of 31.44%, 34.68%, and 29.67% across the datasets.

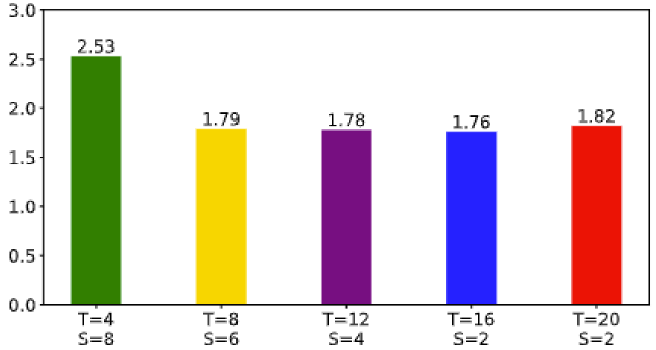
Example comparisons with the second best baselines:



Effect of the Importance-Enhanced Oversampling Policy:



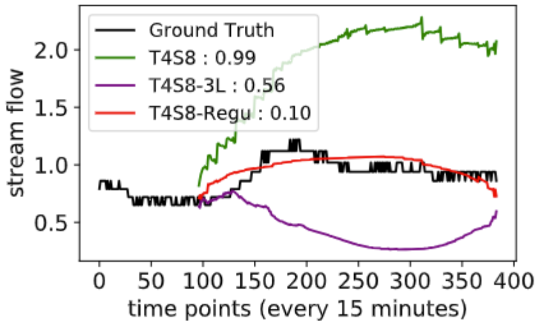
(a) RMSE of UpperPen

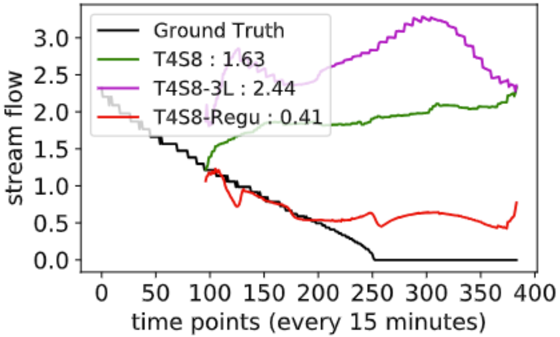


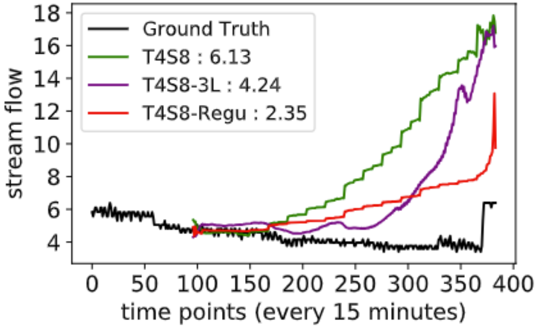
(b) RMSE of Saratoga

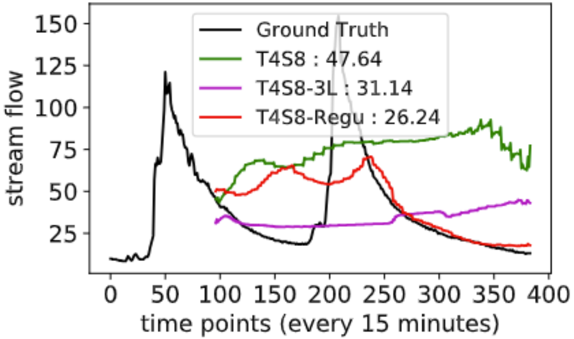
- To evaluate the impact of this policy, we increased the threshold T while simultaneously decreasing the step size S.
- Increasing the threshold T and decreasing the step size S had a positive impact on the results.
- There is an optimal threshold T value beyond which the policy's effectiveness plateaus.

Effect of the KL Regularization Terms & Segment Expanding:









We just use T=4, S=8 as an example:

- T4S8: 5-layer without regularization loss terms.
- T4S8-3L: 3-layer SEED with regularization loss terms.
- T4S8-Regu: 5-layer SEED with regularization loss terms, which gives the best result.

On-Device Prediction for Chronic Kidney Disease

Alex Whelan, Soham Phadke, **David C. Anastasiu**

IEEE GHTC 2022

Chronic Kidney Disease (CKD)

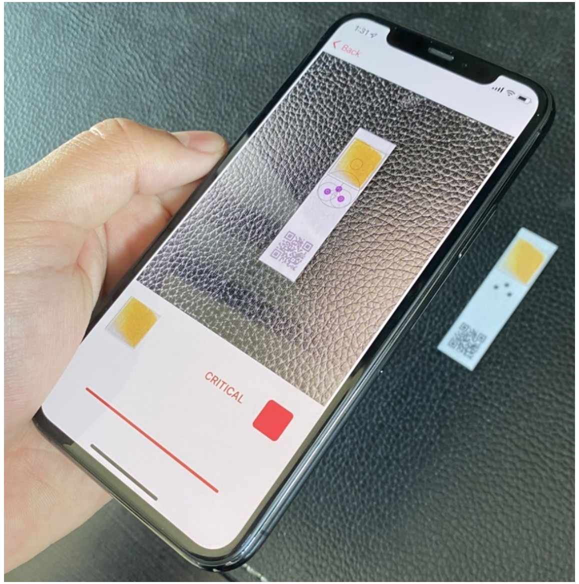
- Progressive decline of kidney function
- Approximately 15% of US population affected by CKD

Stage of CKD	STAGE 1	STAGE 2	STAGE 3A	STAGE 3B	STAGE 4	STAGE 5
eGFR	90 or greater	Between 60 and 89	Between 45 and 59	Between 30 and 44	Between 15 and 29	Less than 15
Level of kidney damage	Mild kidney damage	Mild kidney damage	Mild to moderate kidney damage	Mild to moderate kidney damage	Moderate to severe kidney damage	End-stage kidney disease. Kidneys are close to failure or have completely failed. You will need to start dialysis or have a kidney transplant.

https://www.ckdandt2d.com/managing-chronic-kidney-disease

Point of Care Testing (PoCT)

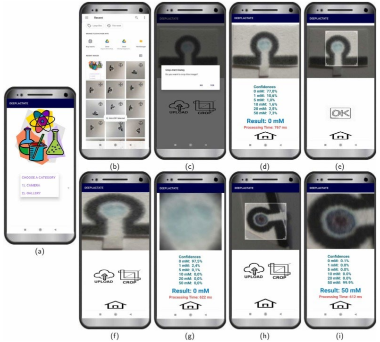
- Kidney Health Monitoring (KHM) System
  - Accessible
  - Fast & Reliable
  - Affordable
- Humanitarian assistance
- Alternative to LAB testing



Related Works

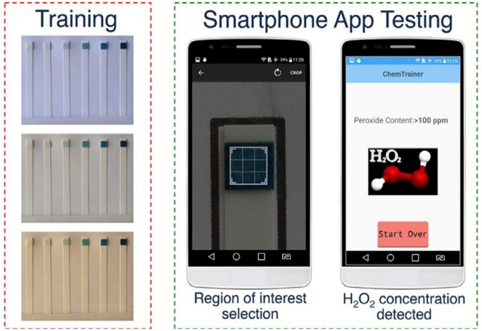
- SmartBioPhone [1]
- ChemTrainer [2]
- DeepLactate [3]

DeepLactate : https://ars.els-cdn.com/content/image/1-s2.0-S0925400522011315-gr3.jpg



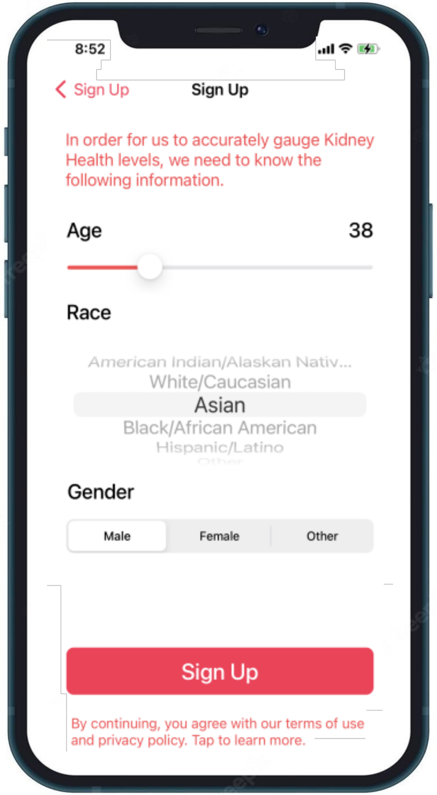
SmartBioPhone : https://pubs.rsc.org/en/Image/Get?imageInfo.ImageType=GA&imageInfo.ImageIdentifier.ManuscriptID=B902354M&imageInfo.ImageIdentifier.Year=2009

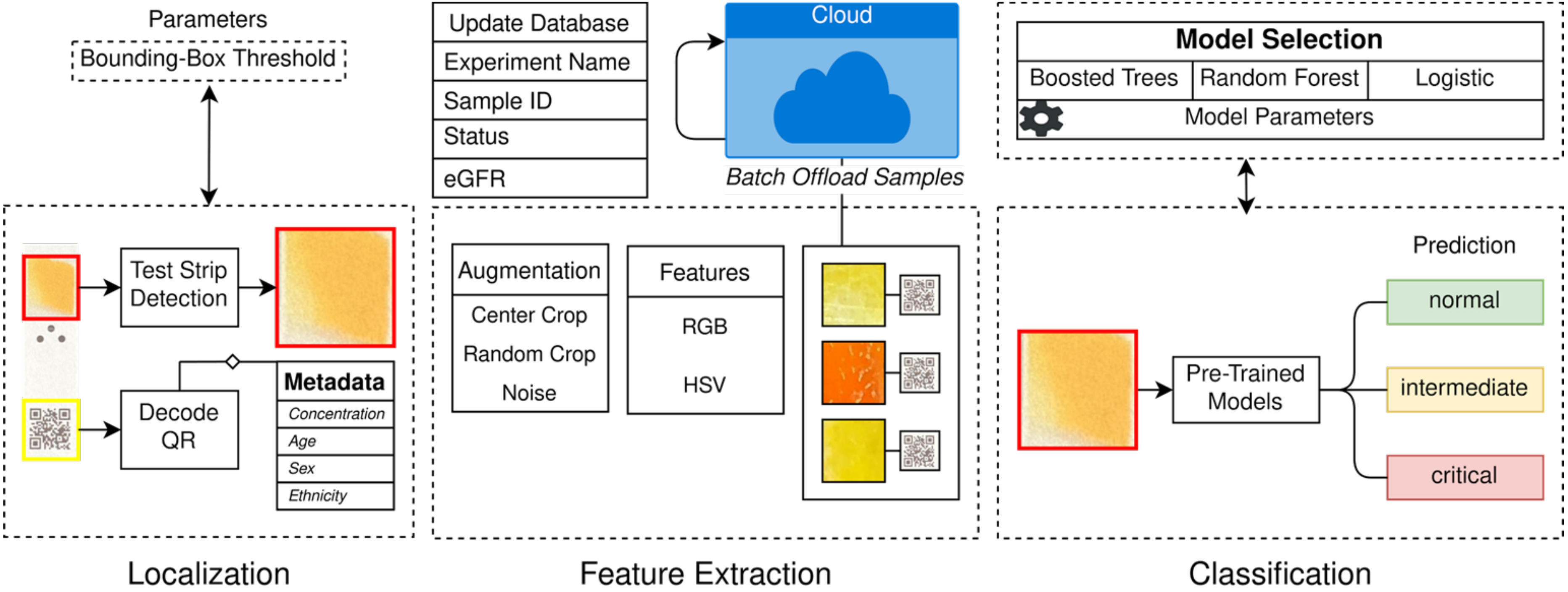
ChemTrainer : https://ars.els-cdn.com/content/image/1-s2.0-S0925400517316519-fx1.jpg



Application Setup

- **[Sign up]**
- Modes of Operation
  - User
  - Researcher

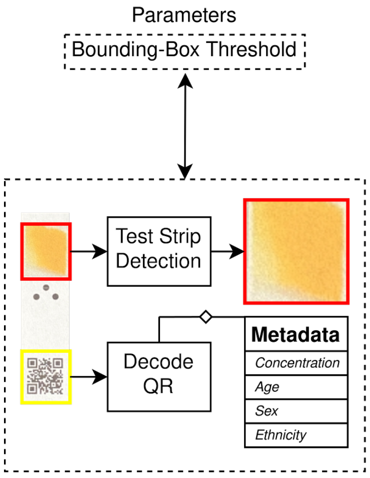




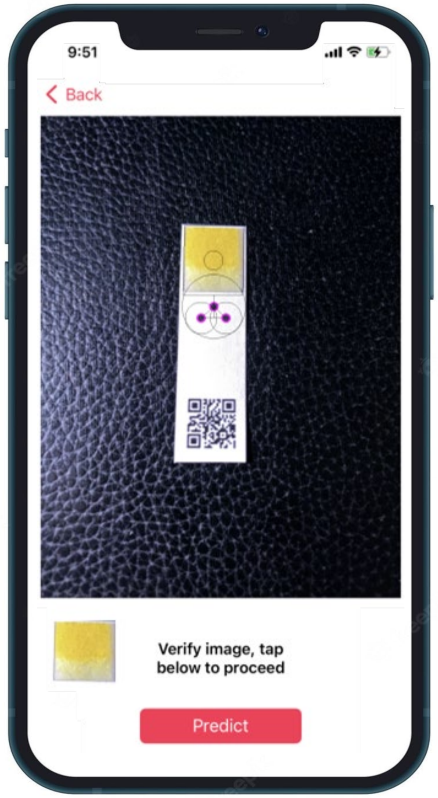
Application Workflow

Localization

- Hough Circle Transform Method
- Decode Metadata

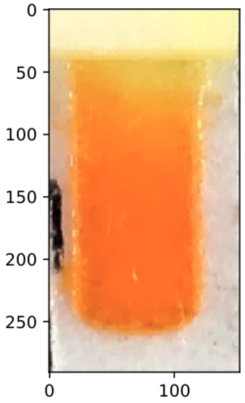


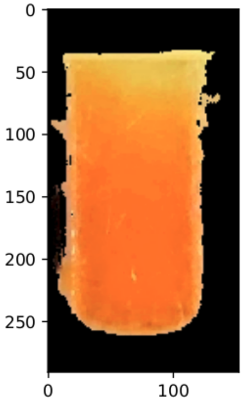
Localization

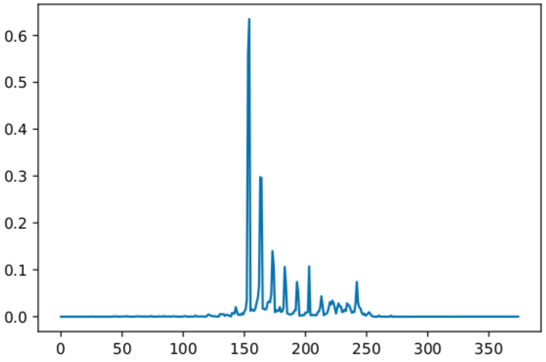


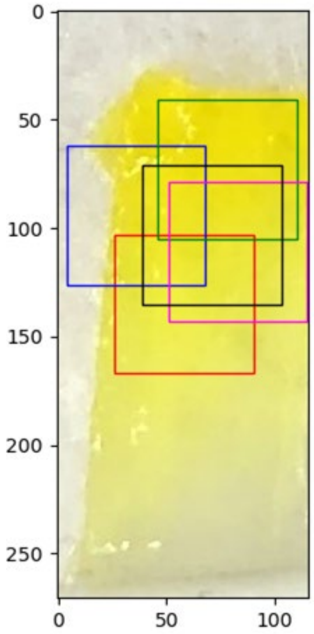
Feature Extraction

- HSV Feature Vector Construction (bottom)
- Randomized Crop (right)
- Dimensionality Reduction



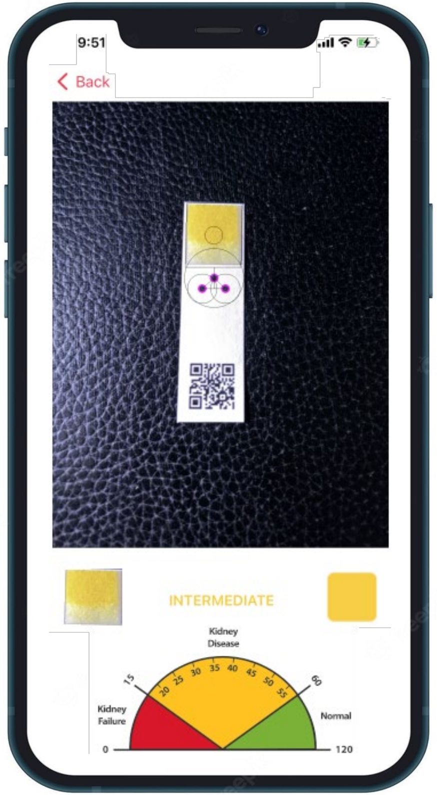






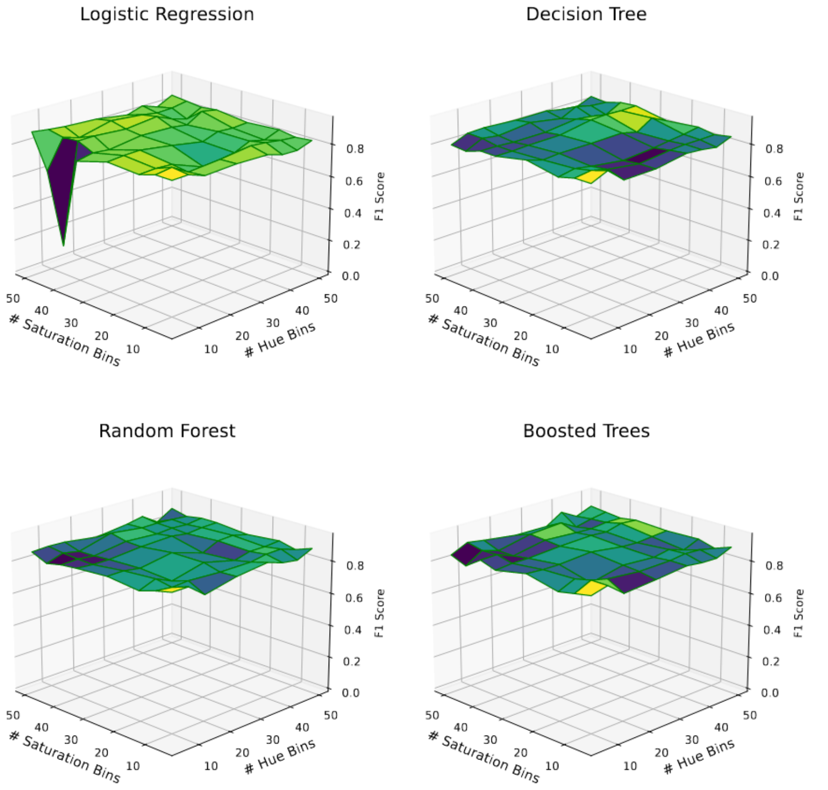
Classification

- Predictions using **eGFR** and **metadata**
- Update Models in Cloud Database



Evaluation

- 10-Fold Cross-Validation
- F1 evaluation metric
- Gridsearch (right)



Model Effectiveness

RGB Features

Augmentation/Dataset	**No Crop**			**Center Crop**			**Random Crop**		
Model	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Logistic Regression	83.32	83.08	83.17	75.52	75.77	75.31	80.99	81.08	81.00
Decision Tree	81.00	80.77	80.87	76.19	76.54	76.27	79.03	78.85	78.92
Random Forest	80.34	80.38	80.36	79.30	79.62	79.35	82.59	82.46	82.51
Boosted Trees	84.48	84.23	84.32	81.41	81.54	81.18	**85.11**	**85.00**	**85.04**

HSV Features

Augmentation/Dataset	**No Crop**			**Center Crop**			**Random Crop**		
Model	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Logistic Regression	83.74	83.85	83.85	78.50	78.55	78.46	82.23	82.15	82.18
Decision Tree	81.40	81.34	81.54	83.10	83.16	83.08	81.47	81.46	81.46
Random Forest	87.61	87.70	87.69	86.16	86.28	86.15	83.84	83.77	83.79
Boosted Trees	**90.34**	**90.37**	**90.38**	88.10	88.13	88.08	85.99	85.85	85.87

Citations

[1] J. M. Ruano-Lopez, M. Agirregabiria, G. Olabarria, D. Verdoy, D. D. Bang, M. Bu, A. Wolff, A. Voigt, J. A. Dziuban, R. Walczak, and J. Berganzo, “The smartbiophone™ , a point of care vision under development through two european projects: Optolabcard and labonfoil,” Lab Chip, vol. 9, pp. 1495–1499, 2009.

[2] M. E. Solmaz, A. Y. Mutlu, G. Alankus, V. Kılıc¸, A. Bayram, and N. Horzum, “Quantifying colorimetric tests using a smartphone app based on machine learning classifiers,” *Sensors and Actuators B: Chemical*, vol. 255, pp. 1967–1973, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925400517316519

[3] E. Y¨uzer, V. Do˘gan, V. Kılıc¸, and M. S¸en, “Smartphone embedded deep learning approach for highly accurate and automated colorimetric lactate analysis in sweat,” *Sensors and Actuators B: Chemical*, vol. 371, p. 132489, 2022. [Online]. Available: www.sciencedirect.com/science/article/pii/S0925400522011315

[4] O. Sidorov, “Conditional gans for multi-illuminant color constancy: Revolution or yet another approach?” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

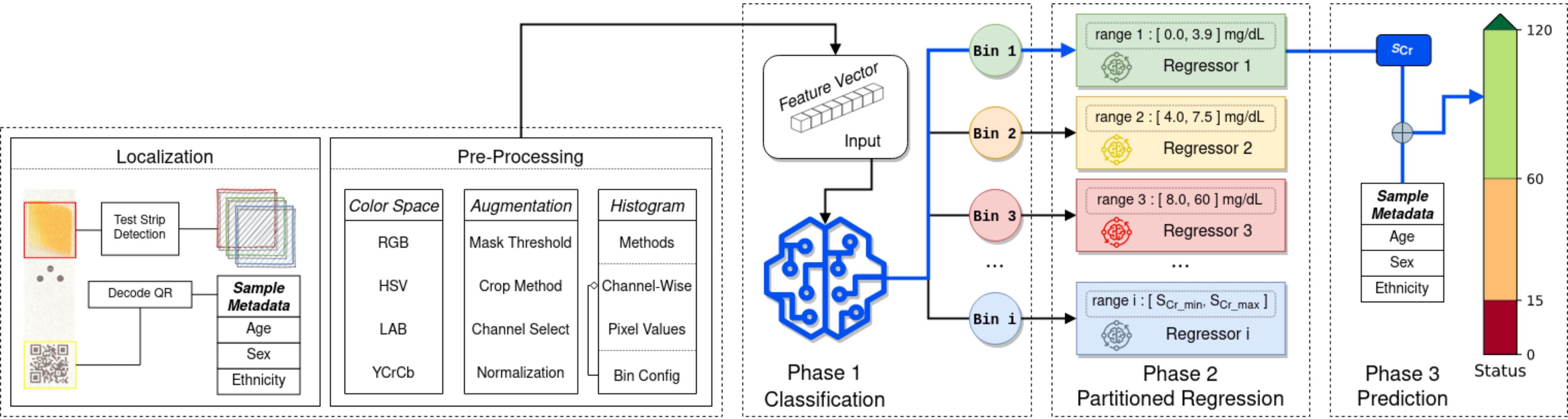
Selective Partitioned Regression for Accurate Kidney Health Monitoring

Alex Whelan, Ragwa Elsayed, Alessandro Bellofiore,

David C. Anastasiu

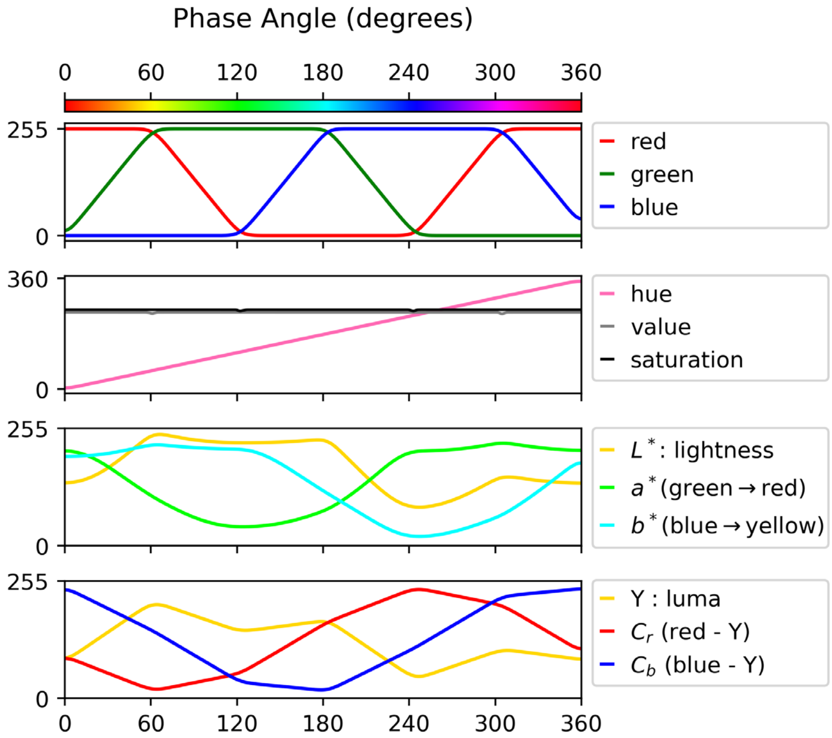
Under Submission

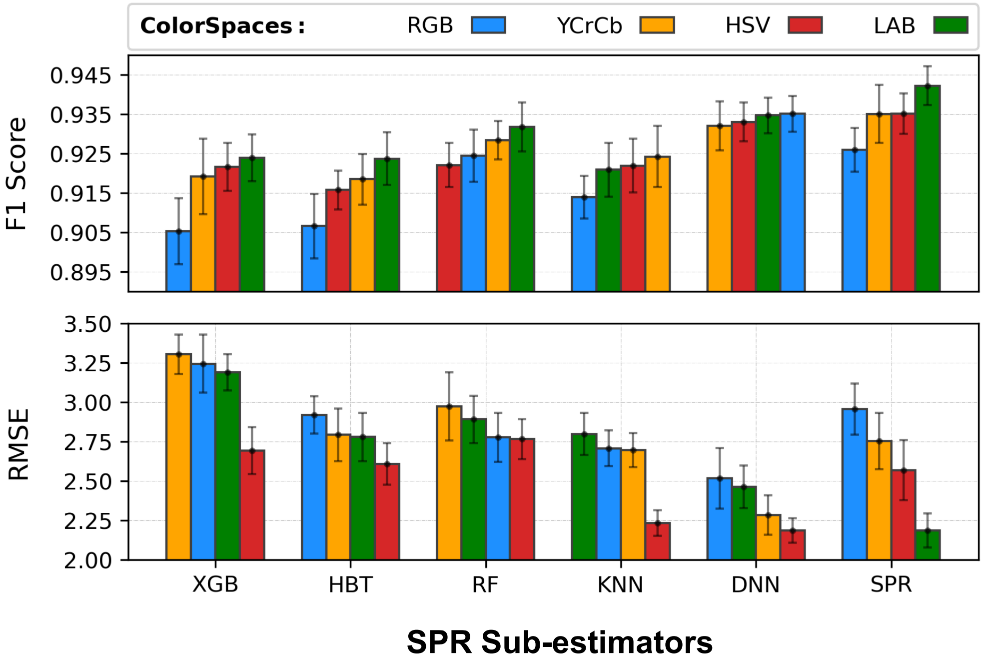
SPR Model



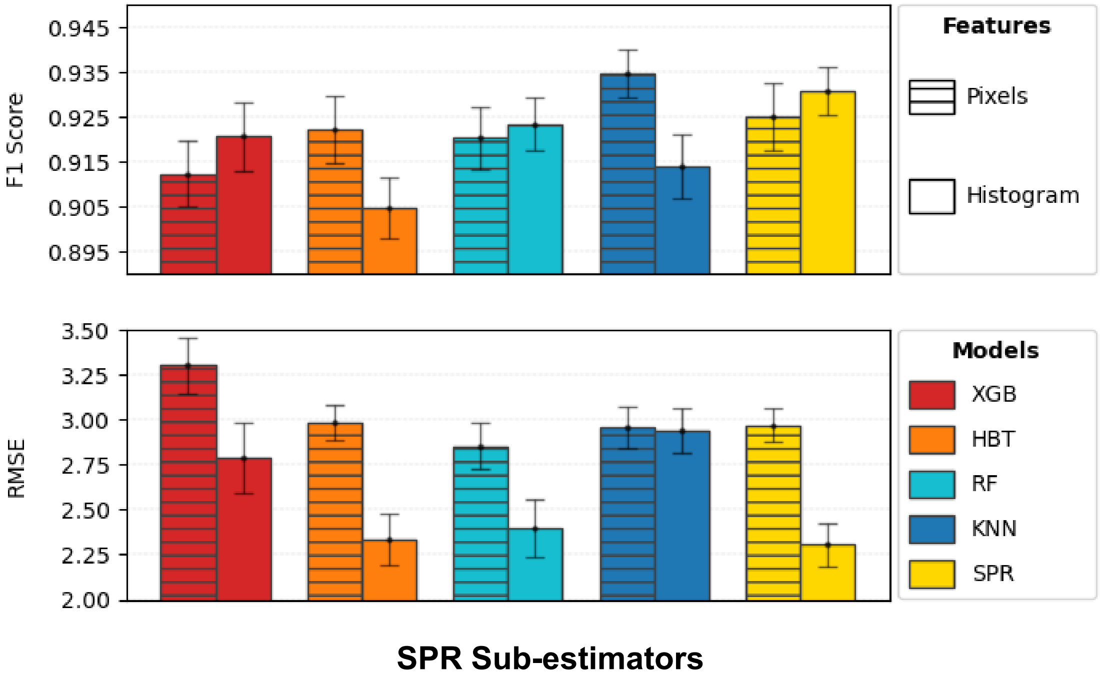
Localization
Test Strip Detection
Decode QR
Sample Metadata
Age
Sex
Ethnicity
Pre-Processing
Color Space
RGB
HSV
LAB
YCrCb
Augmentation
Mask Threshold
Crop Method
Channel Select
Normalization
Histogram
Methods
Channel-Wise
Pixel Values
Bin Config
Feature Vector
Input
Phase 1 Classification
Bin 1
Bin 2
Bin 3
...
Bin i
range 1 : [0.0, 3.9] mg/dL
Regressor 1
range 2 : [4.0, 7.5] mg/dL
Regressor 2
range 3 : [8.0, 60] mg/dL
Regressor 3
...
range i : [S_{Cr_min}, S_{Cr_max}]
Regressor i
Phase 2 Partitioned Regression
SCr
Sample Metadata
Age
Sex
Ethnicity
Phase 3 Prediction
120
60
15
0
Status

Color Space Comparison

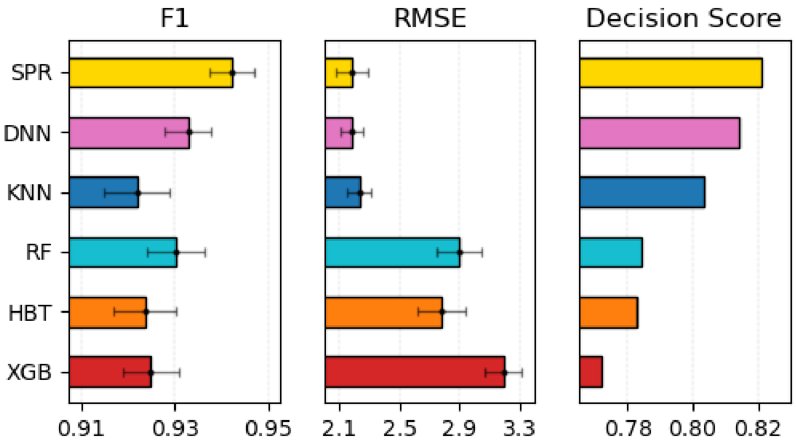




Feature Type Comparison



Comparison Against Baselines



	Method	Train (sec)	Inference (sec)
SPR	lab-histogram	5.5863	0.4111
DNN	hsv-pixels	56.1946	0.5304
KNN	hsv-histogram	0.0104	0.1202
RF	lab-histogram	1.1225	0.0504
HBT	lab-histogram	1.2412	0.0189
XGB	lab-histogram	0.5243	0.0279

Baseline Methods:
DNN: VGG-inspired CNN-based deep neural network
KNN: k-Nearest Neighbor classifier/regressor
RF: Random Forest classifier/regressor
HBT: Histogram Gradient Boosting Decision Tree classifier/regressor
XGB: Extreme Gradient Boosting Tree classifier/regressor
DT: Decision Tree classifier/regressor (*not shown in figure*)
SVM: Support Vector Machines classifier/regressor (*not shown in figure*)

Other Current Projects

[FIM’22] Identification of Distinct Characteristics of Antibiofilm Peptides and Prospection of Diverse Sources for Efficacious Sequences

Antibiofilm and Antithrombotic Peptide Prediction

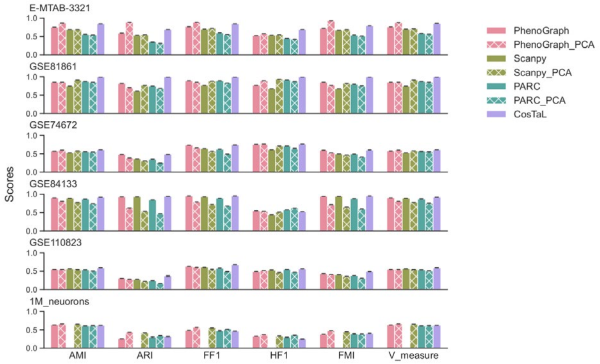
w/ Anand K. Ramasubramanian, SJSU

- Given a peptide’s amino acid sequence
    - Determine its ability to prevent biofilm production or the clotting of the blood.
    - Determine whether the compound that is derived from the peptide is likely to have unintended side effects for the patient.

Mass-Cytometry Screening

w/ Edgar A. Arriaga, University of Minnesota

- Analyze large multidimensional single-cell datasets
    - Developed a graph-based clustering algorithm to identify related compounds
    - CosTaL transforms high-dimensional cells into a weighted k-NN graph
    - Weights are refined via Tanimoto
    - Community detection via Leiden’s algorithm



[Bioinformatics’23] Yijia Li, Jonathan Nguyen, David C. Anastasiu, Edgar A. Arriaga. CosTaL: an accurate and scalable graph-based clustering algorithm for high-dimensional single-cell data analysis. Briefings in Bioinformatics, 2023.

Open Modification Spectral Library Search

SANTA CLARA UNIVERSITY

[Grant NSF 1850557] CRII: III: RUI: Effective Protein Characterization via Fast Exact Open Modification Searching

w/ William Stafford Noble, Genome Sciences, UW

- Methods for characterizing the protein composition of biological samples
  - Mass spectrometers output relative abundance histograms (spectra)
  - Massive databases exist for protein-associated spectra (spectral libraries)
  - Task is to match unknown spectra against nearest neighbor in library

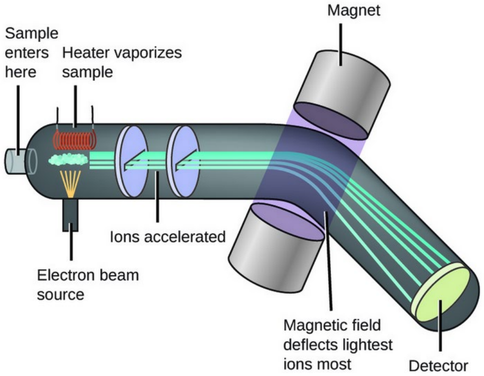
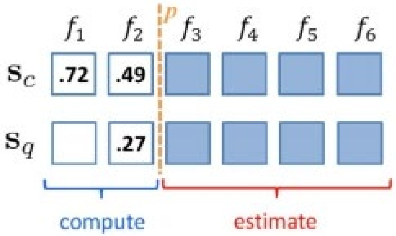
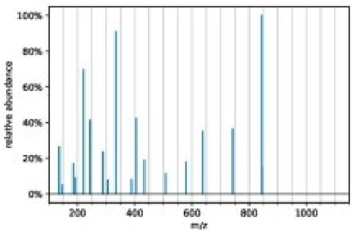


Image: https://i.stack.imgur.com/iVYVY.png

- Challenges
  - Imperfect ionization/spectrometry
  - Size of databases (10’s to 100’s or million)

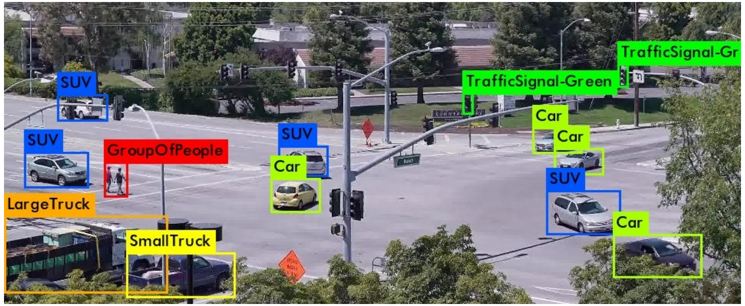




[IEEE CVPR’19] CityFlow: A City-Scale Benchmark for Multi-Target Multi-Camera Vehicle Tracking and Re-Identification
[IEEE CVPRW’23, ‘22, ‘21, ‘20, ‘19, IEEE SOSE’19, IEEE SOSE’19, IEEE MC’19, IEEE CVPRW’18, IEEE SmartWorld’17]

w/ NVIDIA, Toyota, Johns Hopkins, Iowa State, Boston Univ., Univ. of Albany – SUNY, IIT Kanpur, Australian National Univ.

- Organizing member and Evaluation Chair for the AI City Challenge.
- Address challenges in traffic analysis from video, including:
  - Multi-camera vehicle tracking and multi-movement counting
  - Speed estimation from video
  - Anomaly detection
  - Accident description
  - Driver activity recognition



Questions?

References

- [BigData'23] Yanhong Li, Jack Xi & David C. Anastasiu. SEED: An Effective Model for Highly-Skewed Streamflow Time Series Data Forecasting. In the 2023 IEEE International Conference on Big Data (IEEE BigData 2023), Dec 15-18, 2023, Sorrento, Italy.
- [DataBrief'23] Mohammed Shaiqur Rahman, Archana Venkatachalapathy, Anuj Sharma, Jiyang Wang, Senem Velipasalar Gursoy, David Anastasiu, Shuo Wang. Synthetic distracted driving (SynDD1) dataset for analyzing distracted behaviors and various gaze zones of a driver. Data in Brief, 46:108793, 2023.
- [Bioinformatics'23] Yijia Li, Jonathan Nguyen, David C. Anastasiu, Edgar A. Arriaga. CosTaL: an accurate and scalable graph-based clustering algorithm for high-dimensional single-cell data analysis. Briefings in Bioinformatics, 2023.
- [AAAI'23] Yanhong Li, Jack Xi & David C. Anastasiu. An Extreme-Adaptive Time Series Prediction Model Based on Probability-Enhanced LSTM Neural Networks. In Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, AAAI Press, 2023.
- [CVPRW'23] Arpita Vats, David C. Anastasiu. Enhancing Retail Checkout through Video Inpainting, YOLOv8 Detection, and DeepSort Tracking. In IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 5530-5537, 2023.
- [GHTC'22] Alex Whelan, Soham Phadke & David C. Anastasiu. On-Device Prediction for Chronic Kidney Disease. In 2022 IEEE Global Humanitarian Technology Conference (GHTC) (GHTC 2022), 2022.
- [ECTEL'22] Arpita Vats, Gheorghi Guzun & David C. Anastasiu. CLP: A Platform for Competitive Learning. In Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption (EC-TEL 2022), pages 615-622, Springer International Publishing, 2022.
- [Microbio'22] Bipasa Bose, Taylor Downey, Anand K. Ramasubramanian & David C. Anastasiu. Identification of Distinct Characteristics of Antibiofilm Peptides and Prospection of Diverse Sources for Efficacious Sequences. Frontiers in Microbiology, 12, 2022.
- [CVPRW'22] Arpita Vats & David C. Anastasiu. Key Point-Based Driver Activity Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pages 3274-3281, 2022.
- [AIC'21] Milind Naphade, Shuo Wang, David C. Anastasiu, Zheng Tang, Ming Ching Chang, Xiaodong Yang, Yue Yao, Liang Zheng, Pranamesh Chakraborty, Christian E. Lopez, Anuj Sharma, Qi Feng, Vitaly Ablavsky & Stan Sclaroff. The 5th AI City Challenge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (CVPRW'21), pages 4263-4273, 2021.
- [AIC'20] Milind Naphade, Shuo Wang, David C. Anastasiu, Zheng Tang, Ming Ching Chang, Xiaodong Yang, Liang Zheng, Anuj Sharma, Rama Chellappa & Pranamesh Chakraborty. The 4th AI City Challenge. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (CVPRW'20), 1:2665-2674, 2020.
- [BDAT'20] David C. Anastasiu, Jack Gaul, Maria Vazhaeparambil, Meha Gaba & Prajval Sharma. Efficient City-Wide Multi-Class Multi-Movement Vehicle Counting: A Survey. Journal of Big Data Analytics in Transportation, 2(3):235-250, 2020.
- [CVPR'19] Zheng Tang, Milind Naphade, Ming Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David C. Anastasiu & Jenq Neng Hwang. CityFlow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In CVPR 2019: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019), 2019.

References

SANTA CLARA UNIVERSITY

- [AIC'19] Milind Naphade, Zheng Tang, Ming Ching Chang, David C. Anastasiu, Anuj Sharma, Rama Chellappa, Shuo Wang, Pranamesh Chakraborty, Tingting Huang, Jenq Neng Hwang & Siwei Lyu. The 2019 AI City Challenge. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (CVPRW'19), 1:452-460, 2019.
- [iDSC'19] Manika Kapoor & David C. Anastasiu. A Data-Driven Approach for Detecting Autism Spectrum Disorders. In Data Science -- Analytics and Applications (iDSC 2019), Springer Fachmedien Wiesbaden, 2019.
- [MC'19] Anupama Upadhayula, Avinash Ravilla, Ishwarya Varadarajan, Sowmya Viswanathan & David C. Anastasiu. Study Area Recommendation via Network Log Analytics. In The Seventh IEEE International Conference on Mobile Cloud Computing, Services, and Engineering, IEEE, 2019.
- [SOSE'19] Shuai Hua & David C. Anastasiu. Effective Vehicle Tracking Algorithm for Smart Traffic Networks. In Thirteenth IEEE International Conference on Service-Oriented System Engineering (SOSE), IEEE, 2019.
- [CIKM '18] Swapnil Gaikwad, Melody Moh & David C. Anastasiu. Data Structure for Efficient Line of Sight Queries. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM '18), ACM, 2018.
- [AIC'18] Milind Naphade, Ming Ching Chang, Anuj Sharma, David C. Anastasiu, Vamsi Jagarlamudi, Pranamesh Chakraborty, Tingting Huang, Shuo Wang, Ming Yu Liu, Rama Chellappa, Jenq Neng Hwang & Siwei Lyu. The 2018 NVIDIA AI City Challenge. In 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'18), 1:53-60, 2018.
- [CVPRW'18] Shuai Hua, Manika Kapoor & David C. Anastasiu. Vehicle Tracking and Speed Estimation from Traffic Videos. In 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'18), IEEE, 2018.
- [JDSA'17] David C. Anastasiu & George Karypis. Efficient identification of Tanimoto nearest neighbors; All Pairs Similarity Search Using the Extended Jaccard Coefficient. Springer International Journal of Data Science and Analytics, 4(3):153-172, 2017.
- [StatsRef'17]David C. Anastasiu & Andrea Tagarelli. Document Clustering. Wiley StatsRef: Statistics Reference Online, pages 1-11, American Cancer Society, 2017.
- [JPDC'17] David C. Anastasiu & George Karypis. Parallel cosine nearest neighbor graph construction. Elsevier Journal of Parallel and Distributed Computing, 2017.
- [SCI'17] Swapnil Gaikwad & David C. Anastasiu. Optimal Constrained Wireless Emergency Network Antenna Placement. In Proceedings of the IEEE Smart City Innovations 2017 Conference (IEEE SCI 2017), 2017.
- [iDSC'17] David C. Anastasiu. Cosine Approximate Nearest Neighbors. In Data Science -- Analytics and Applications (iDSC 2017), pages 45-50, Springer Fachmedien Wiesbaden, 2017.
- [SmartWorld'17] Niveditha Bhandary, Charles MacKay, Alex Richards, Ji Tong & David C. Anastasiu. Robust Classification of City Roadway Objects for Traffic Related Applications. In 2017 IEEE Smart World NVIDIA AI City Challenge (SmartWorld'17), IEEE, 2017.

References

- [DSAA'16] David C. Anastasiu & George Karypis. Efficient Identification of Tanimoto Nearest Neighbors. In 2016 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2016, pages 156-165, 2016.
- [IA3'16] David C. Anastasiu & George Karypis. Fast Parallel Cosine K-Nearest Neighbor Graph Construction. In 2016 6th Workshop on Irregular Applications: Architecture and Algorithms (IA3) (IA3 2016), pages 50-53, 2016.
- [IA3'15] David C. Anastasiu and George Karypis. Pl2ap: Fast parallel cosine similarity search. IA3 2015. In conjunction with SC'15, IA3 2015, 2015.
- [CIKM'15] David C. Anastasiu and George Karypis. L2knng: Fast exact k-nearest neighbor graph construction with l2-norm pruning. CIKM '15, pages 791-800, New York, NY, USA, 2015. ACM.
- [ICDE'15] David C. Anastasiu, Al M. Rashid, Andrea Tagarelli, and George Karypis. Understanding computer usage evolution. ICDE 2015, pages 1549-1560, 2015.
- [ICDE'14] David C. Anastasiu and George Karypis. L2ap: Fast cosine similarity search with prefix l-2 norm bounds. ICDE 2014, pages 784-795, 2014.
- [SI'14] David C. Anastasiu, Jeremy Iverson, Shaden Smith, and George Karypis. Big data frequent pattern mining. In Frequent Pattern Mining, pages 225-260. Springer International Publishing, Switzerland, 2014.
- [CRC'13] David C. Anastasiu, Andrea Tagarelli, and George Karypis. Document clustering: The next frontier. In Data Clustering: Algorithms and Applications, pages 305-338. CRC Press, Boca Raton, FL, USA, 2013.
- [WWW'13] David C. Anastasiu, Byron J. Gao, Xing Jiang, and George Karypis. A novel two-box search paradigm for query disambiguation. World Wide Web, 16(1):1-29, 2013.
- [IC'13] Byron J. Gao, David Buttler, David C. Anastasiu, Shuaiqiang Wang, Peng Zhang, and Joey Jan. User-centric organization of search results. IEEE Internet Computing, 17(3):52-59, May 2013.
- [CIKM'11] David C. Anastasiu, Byron J. Gao, and David Buttler. A framework for personalized and collaborative clustering of search results. CIKM '11, pages 573-582, New York, NY, USA, 2011. ACM.
- [SIGIR'11] David C. Anastasiu, Byron J. Gao, and David Buttler. Clusteringwiki: personalized and collaborative clustering of search results. SIGIR 2011, pages 1263-1264, 2011.
- [COLING'10] Byron J. Gao, David C. Anastasiu, and Xing Jiang. Utilizing user-input contextual terms for query disambiguation. COLING '10, pages 329-337, Stroudsburg, PA, USA, 2010.
- [CIKM'09] Byron J. Gao, Mingji Xia, Walter Cai, and David C. Anastasiu. The gardener's problem for web information monitoring. CIKM '09, pages 1525-1528, New York, NY, USA, 2009. ACM.
- [DMIN'09] Walter Cai, David C. Anastasiu, Mingji Xia, and Byron J. Gao. Olap for multicriteria maintenance scheduling. DMIN '09, pages 35-41. CSREA Press, 2009.