Scalability Analysis of Mini-Cluster Jetson TX2 for Training DNN Applied to Healthcare

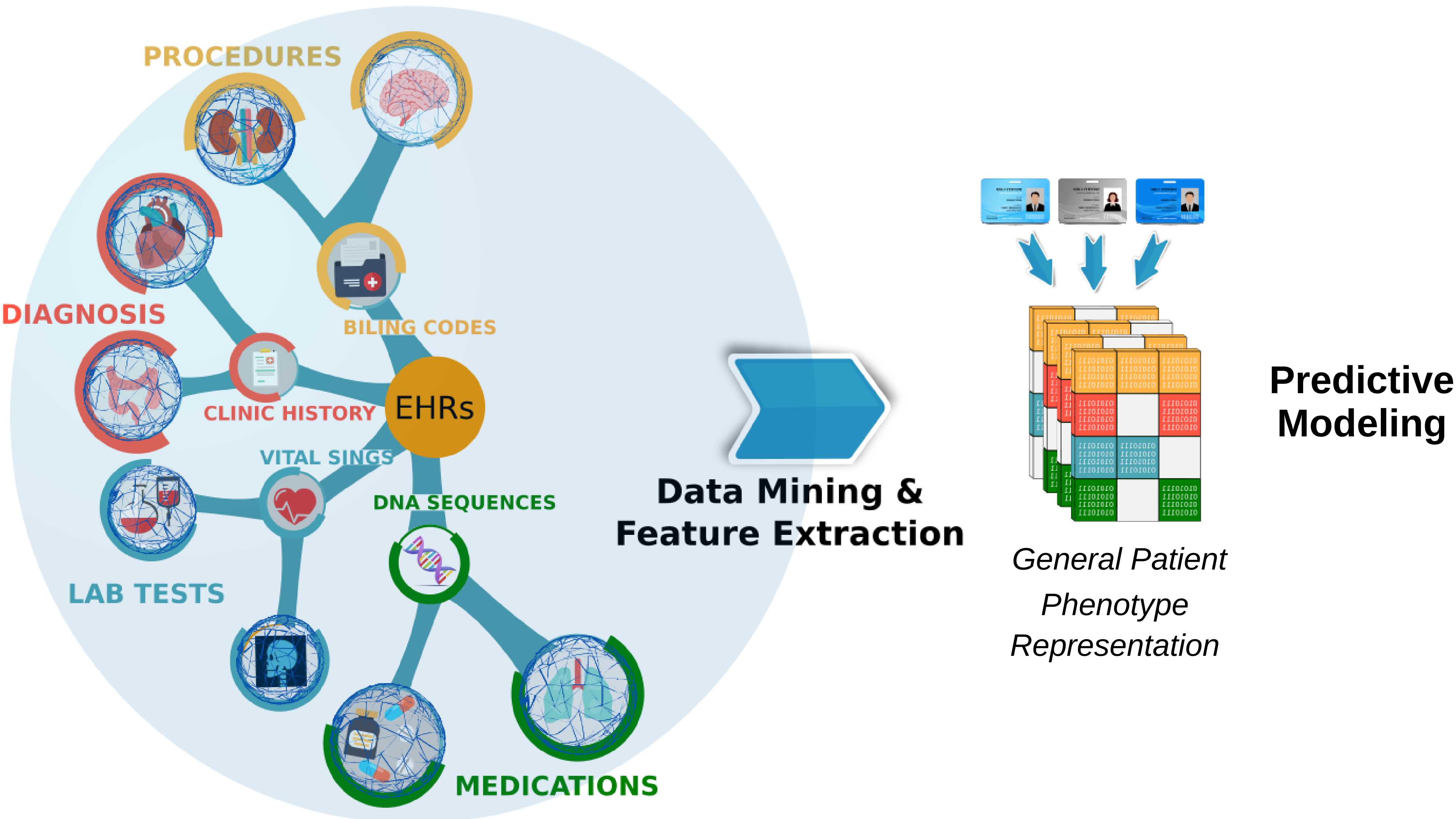
John A. GARCÍA. H.

Frédéric PRECIOSO, Pascal STACCINI, Michel RIVEILL

Université Côte d'Azur, CNRS

Laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis – I3S
Département d'Informations Médicales DIM – CHU Nice

Motivation: Health Care Decision-Making



Challenges

- *A common challenge in healthcare today is that physicians have access to massive amounts of data on patients, but have short time to analyze all of them.*

- *One limitation is that hospitals without robust computational systems for processing, storing and drawing conclusions requires to outsource the clinical tasks and that is a risk for privacy clinical data.*

Developing a Green Intelligence Medical System to derivate a patient representation for predict general medical targets and improving the computational resources usage.

DiagnoseNET

Green Intelligence Medical System

Provides three high-level features:

1) A framework to build full Deep Neural Networks (DNN) workflow;

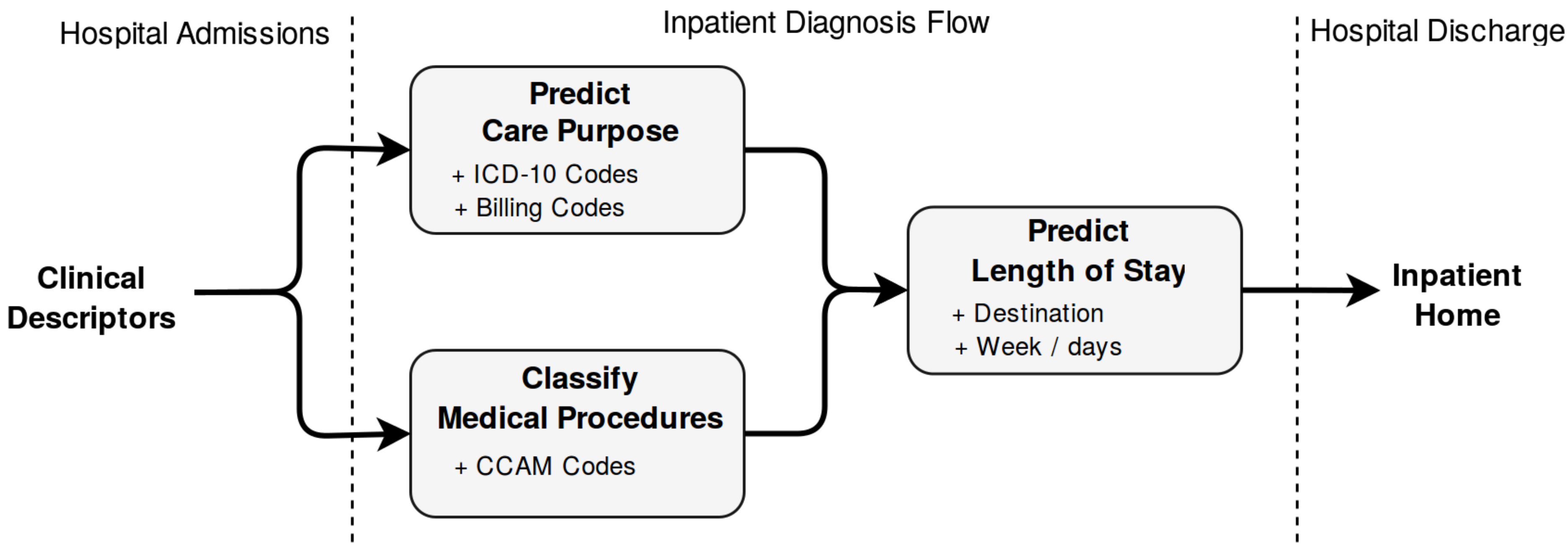
2) An energy-monitoring tool for workload characterization;

3) A distributed processing for training DNN on Jetson TX2 Mini-Clusters.

It is part of the **IADB Project**
Integration and Analysis of Biomedical Data: http://www.i3s.unice.fr/~riveill/IADB/

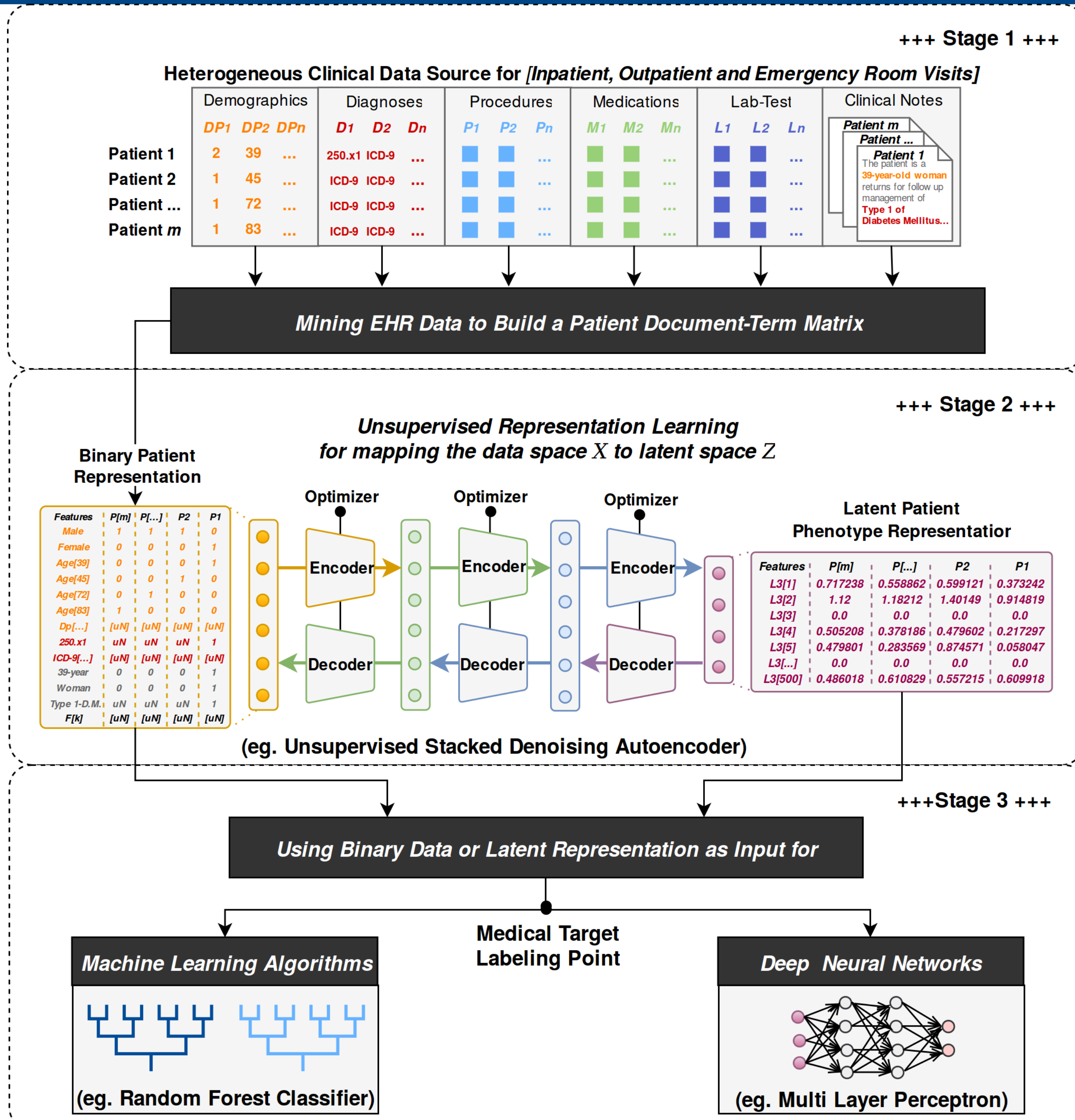
Case Study: Predict the Medical Future of Hospitalized Patients

Medical Target Pipeline



	Diagnosis-related Group	ICD-10 Codes	Definition
Patient 1	Morbidity Principal	R402	Unspecified coma
	Etiology	I619	Nontraumatic intracerebral hemorrhage, unspecified
Medical Target	Care Purpose	Z515	Encounter for palliative care
Label used	Clinical Major Category	20	Palliative care
Patient 2	Morbidity Principal	R530	Neoplastic (malignant) relate fatigue
	Etiology	C20	Malignant neoplasm of rectum
Medical Target	Care Purpose	Z518	Encounter for other specified aftercare
Label used	Clinical Major Category	60	Other disorders

DiagnoseNET: Framework to automize the Patient Phenotype Representation

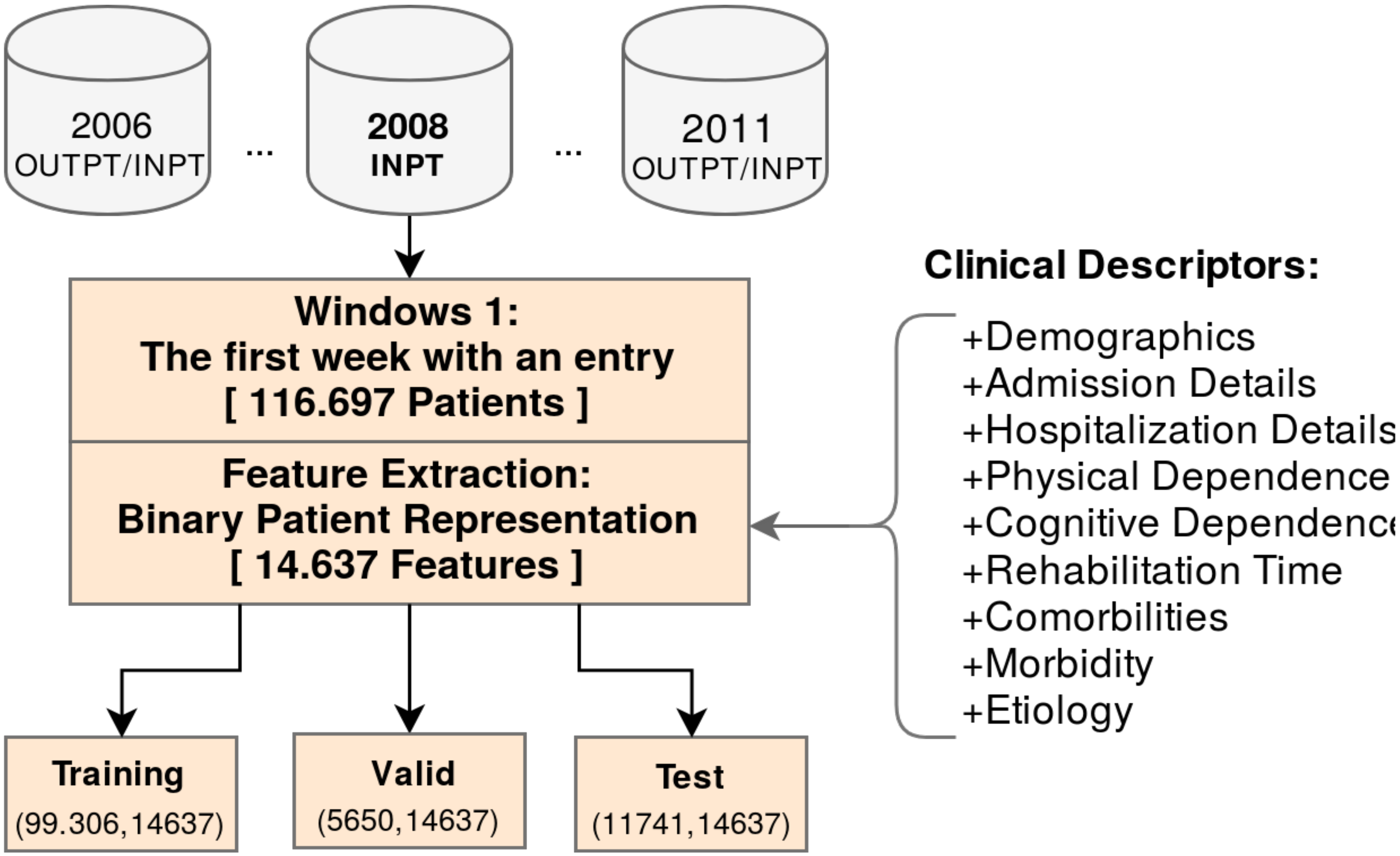


3. Adapted from: **Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records**. By Riccardo Miotto et al. SCIENTIFIC REPORTS, 2016.

Mining Electronic Health Records
To Build A Patient Entity-Term Matrix

PMSI-PACA Clinical Dataset

*As Input we are using **result features** that describe the patient clinical descriptors to predict the medical targets.*



Data-mining: Feature Extraction From Electronic Health Records

Serialized a Patient Record in a Clinical Document Architecture Schema

Patients	x1_demographics			x4_physical_dependance			x7_related_diagnoses		
	gender	...	age	feeding	...	displacement	Das1	...	Das 3
Patient 1	2	...	61	4	...	2	Z431	...	Z501
Patient 2	2	..	65	4	...	2	J459	...	F322
....	...	...	...	...	...	...	...	...	...
Patient m	1	...	95	1	...	2	C259	...	F322

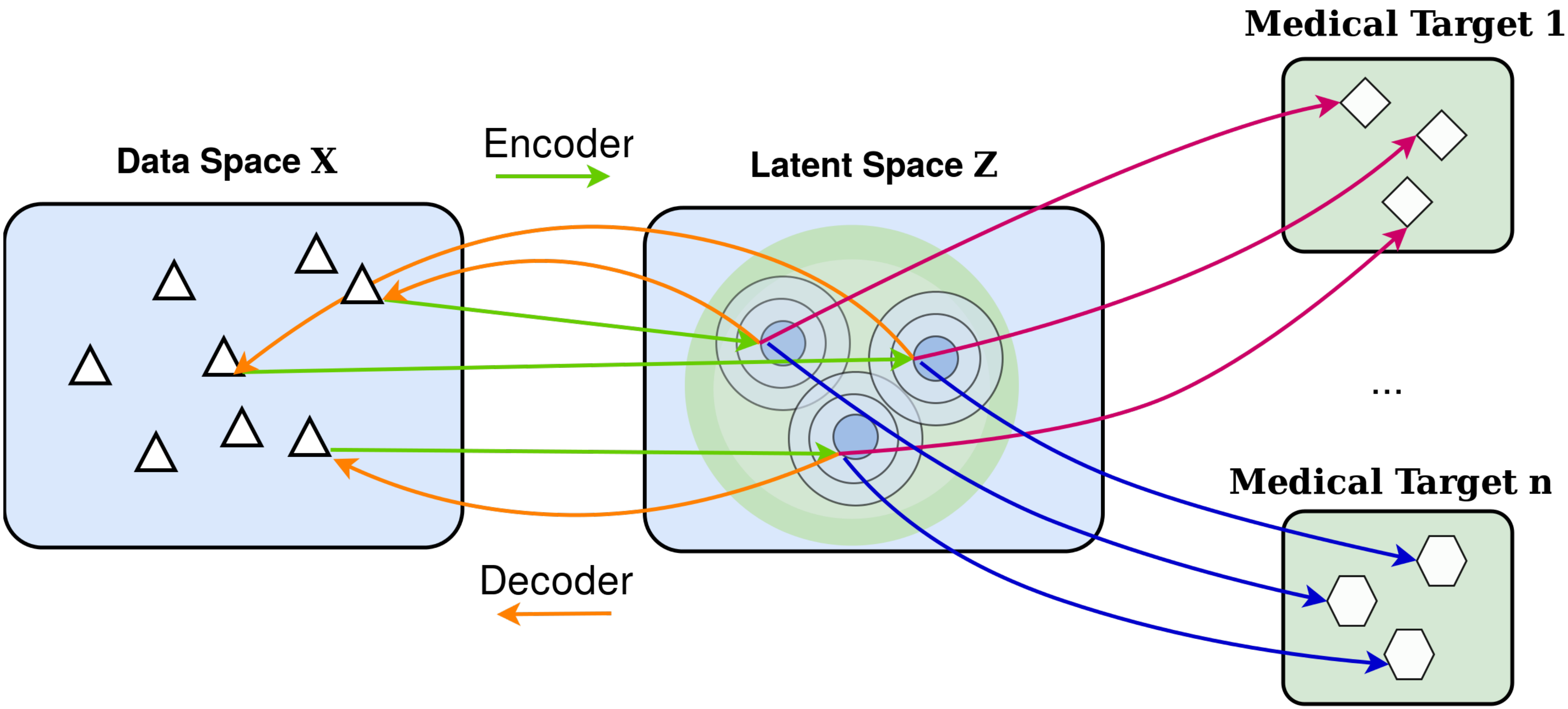
Build a Term-Document Matrix or Binary Patient Representation

Patients	x1 :demographics			x4 :physical_dependance			x7 :related_diagnoses		
	[1 :male]	[2 :female]	60-74	[4 :Assistance]		[2 :normal_transfer]	Z431	...	F322
Patient 1	0	1	1	1	...	1	1	...	0
Patient 2	0	1	1	1	...	1	0	...	1
....	...	...	...	...	...	...	...	...	...
Patient m	1	0	0	0	...	1	0	...	1

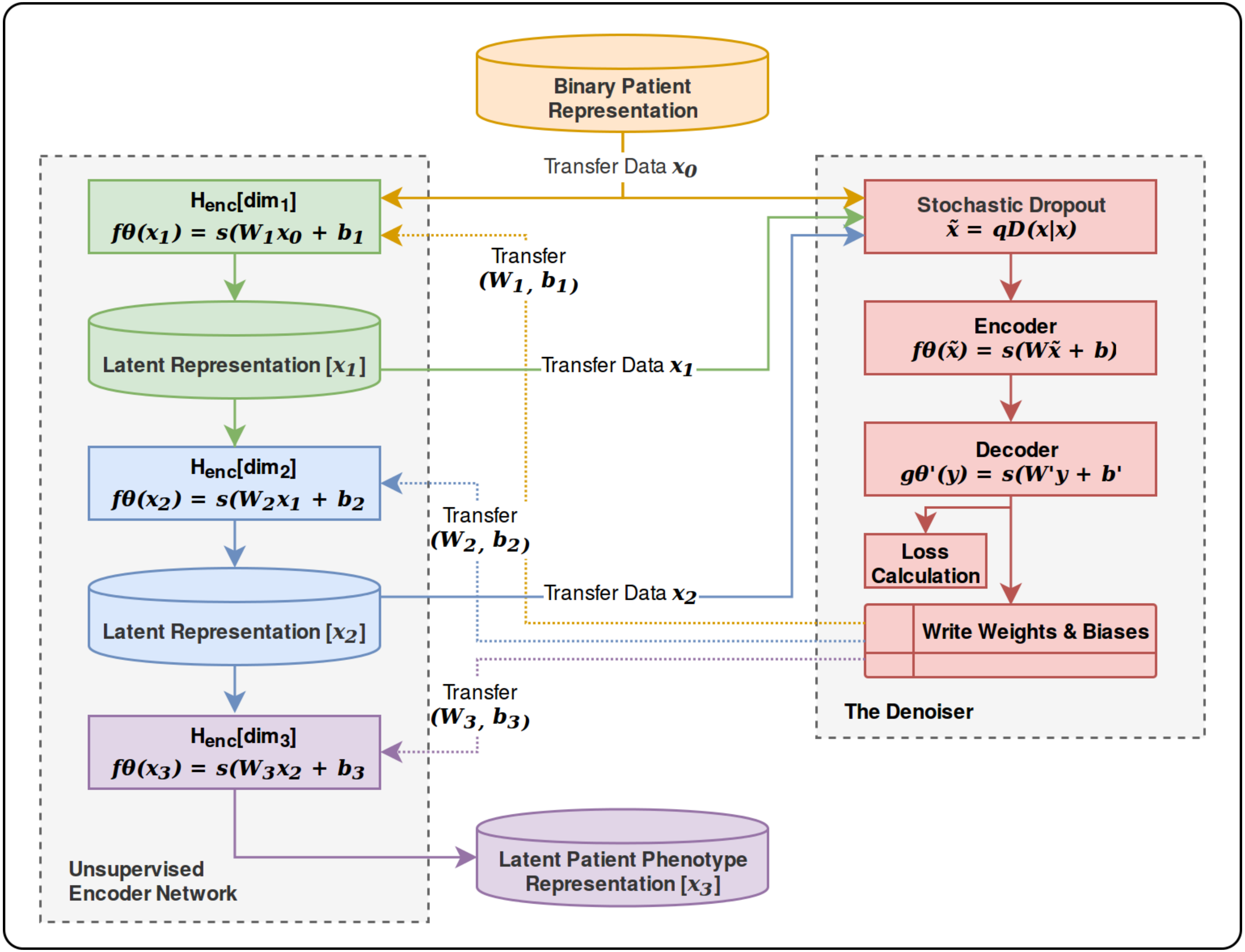
Unsupervised Patient Phenotype Representation

Unsupervised Patient Phenotype Representation

+ From a binary patient representation **{X}** derive a latent patient representation **{Z}.**
+ Once the general representation is obtained mapping it for the different medical targets.



Unsupervised Stacked Denoising Autoencoder Network



4. Source **Parallel and Distributed Processing for Unsupervised Patient Phenotype Representation.** By GARCÍA H. John et al. 2018

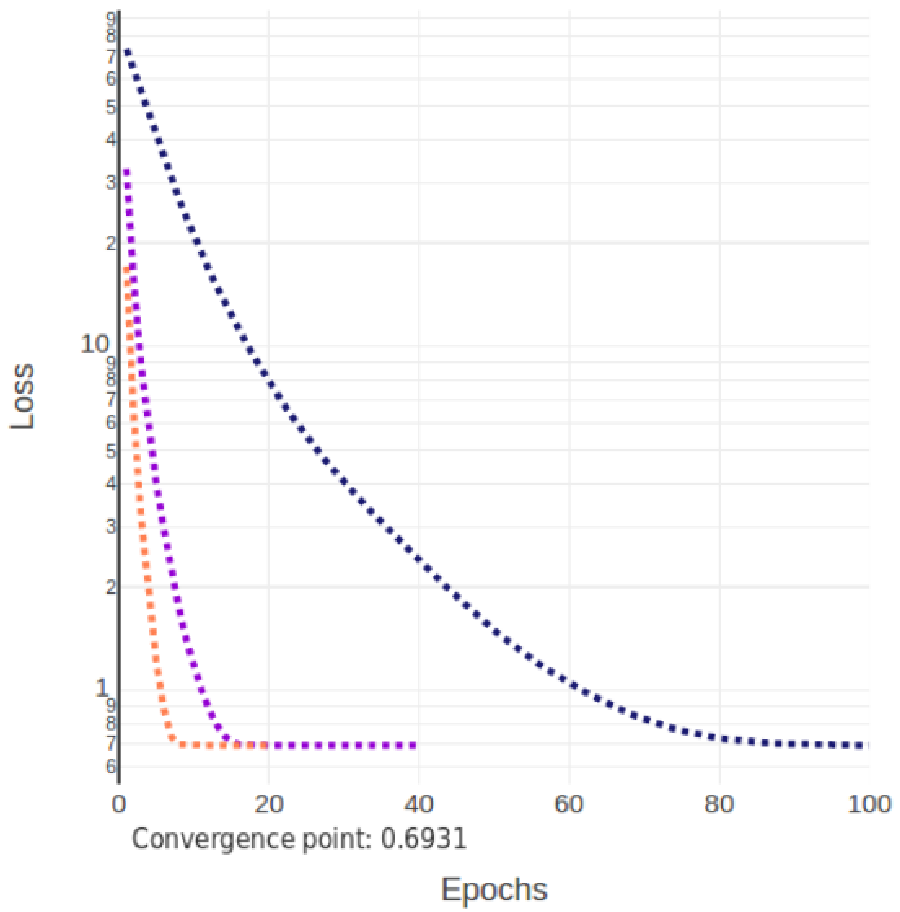
Experiment Analysis

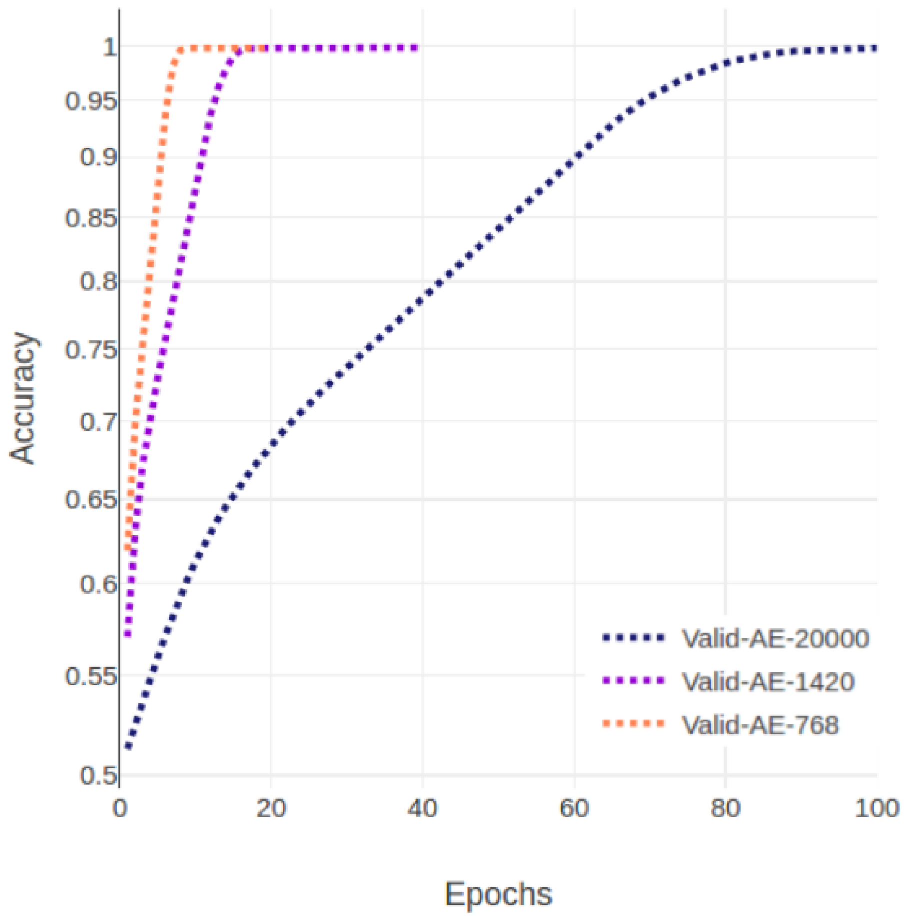
1) Number of Gradient Updates as Factor to Early Model Convergence.

1) Number of Gradient Updates as Factor to Early Model Convergence

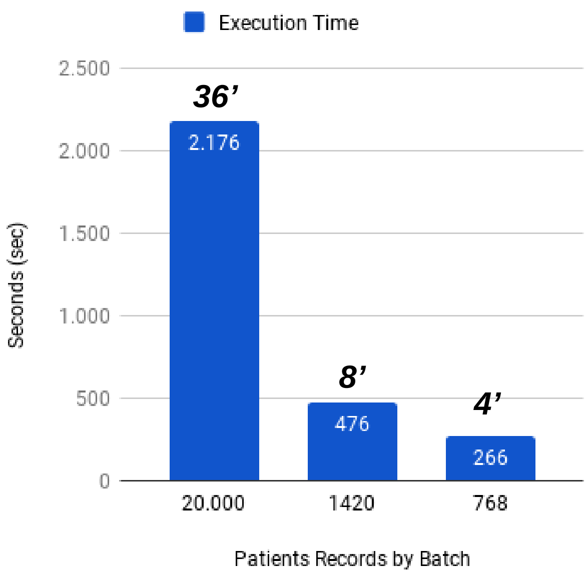
	1-Layer1	2-Layer2	3-Layer3	4-Activation_funct	5-GD_Optimizer	6-Learning_rate	7-Dropout-rate
0	2048	2048	768	relu	adam	0.0001	0.5

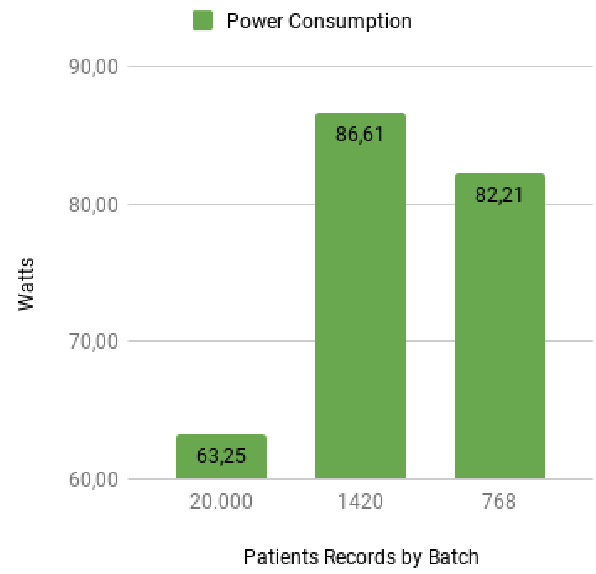
Network convergence using batch partitions of [20000, 1420, 768] records to generate [4, 59, 110] gradient updates by epoch respectively.

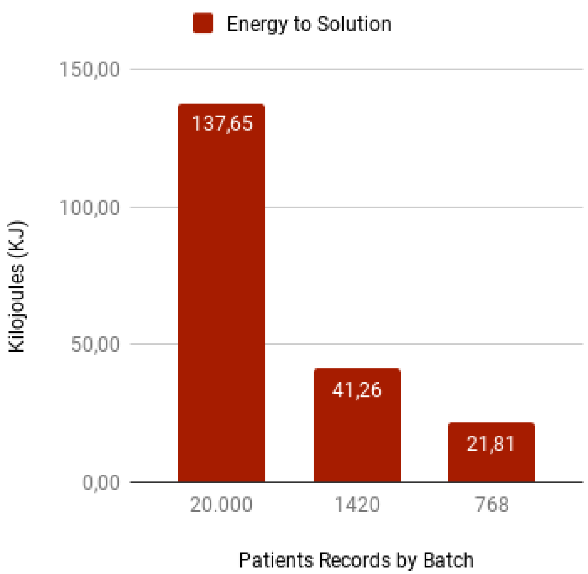




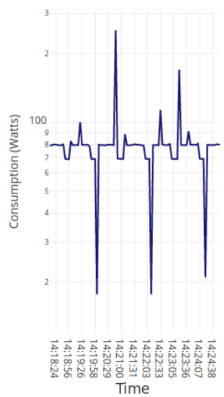
Number of Gradient Updates Impact the Energy to Solution

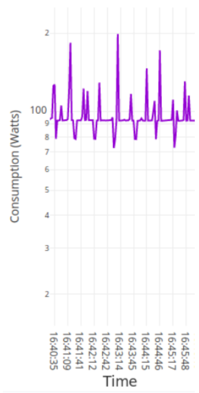


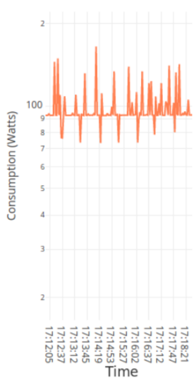




Power consumption in a window of 6 minutes







BF: 20.000	***BF: 1420***	***BF: 768***
17 Epochs	***15 Epochs***	***14 Epochs***
68 GRAD UPD	***885 GRAD UPD***	***1540 GRAD UPD***

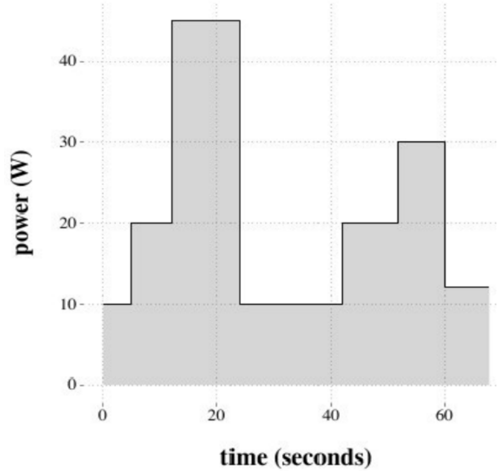
Energy Monitoring tool for Workload Characterization

Energy Consumption Metrics

1) Discretization of Energy Consumption per Time-Unit:

$$Energy = \int Power(t)\ dt$$

$$Energy = \sum_{i=1}^{N} Power(i) * \Delta t(i)$$



2) Energy to Solution (E2S):

$$E2S = \sum_{i=1}^{N} Power_{Node}(i) * max\ (\Delta t_{CPU}(i), \Delta t_{GPU}(i))$$

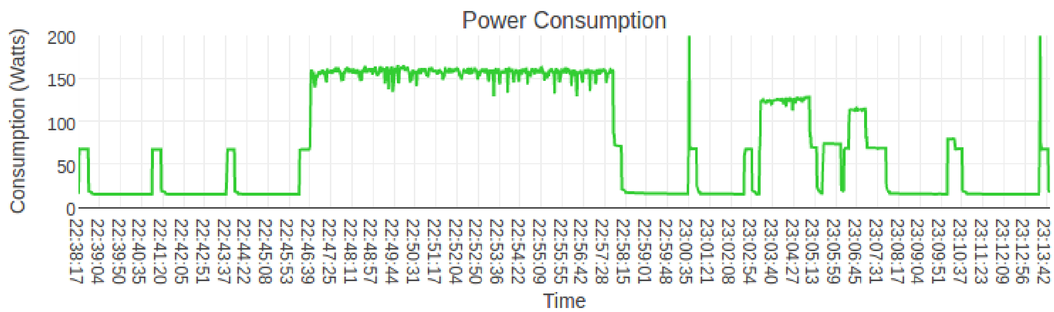
3) Instantaneous Power Consumption per Node:

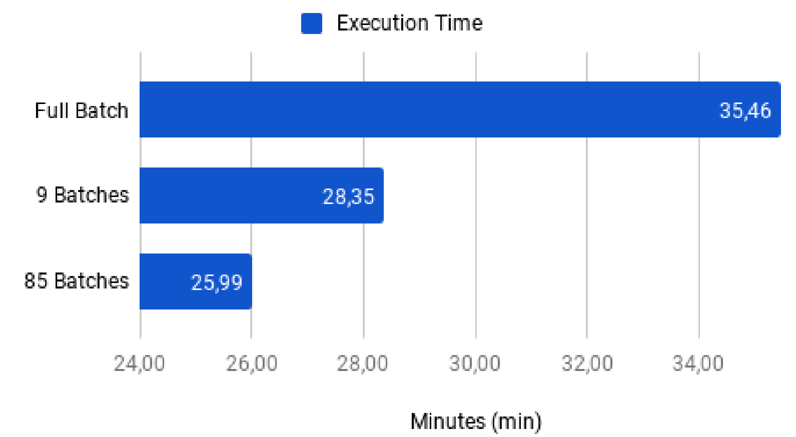
$$Power_{Node}(i) = \sum_{j=1}^{nc} P^{j}_{CPU}(i) + \sum_{j=1}^{ng} P^{j}_{GPU}(i) + \sum_{j=1}^{nm} P^{j}_{RAM}(i)$$

5. Source **enerGyPU and enerGyPhi Monitor for Power Consumption and Performance Evaluation on Nvidia Tesla GPU and Intel Xeon Phi.** By GARCÍA H. John et al. 2016

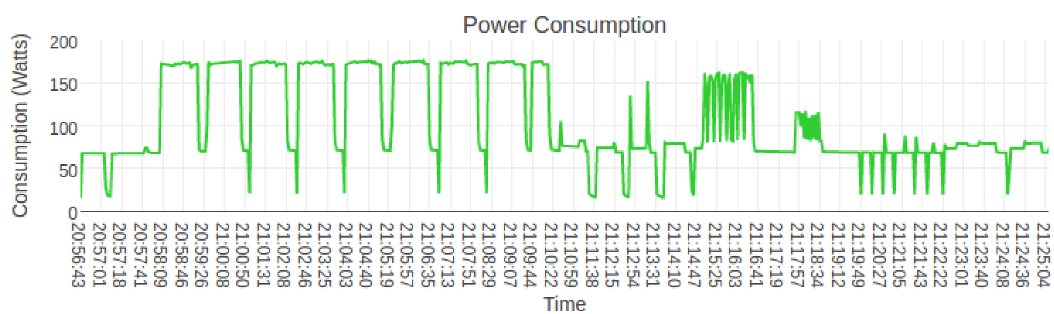
Analysis of Energy Efficiency for Training the USDA on 1-GPU

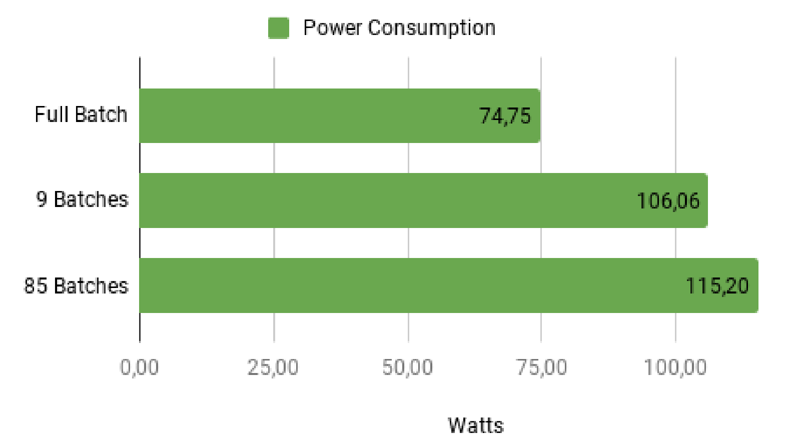
S-1: Full Batch with 84.999 Patients Records



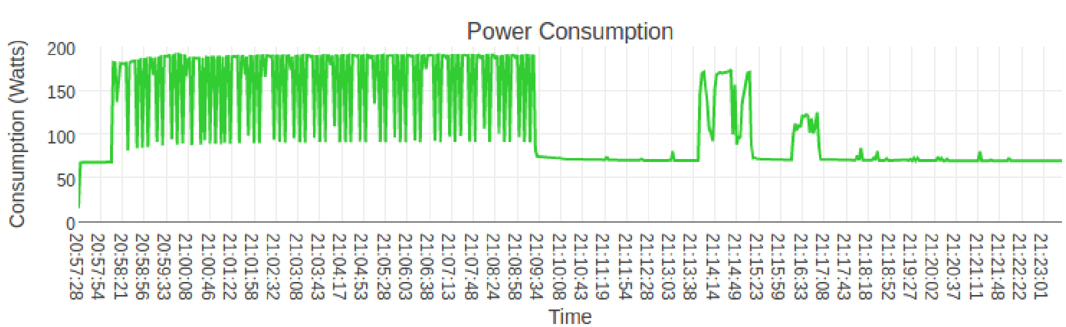


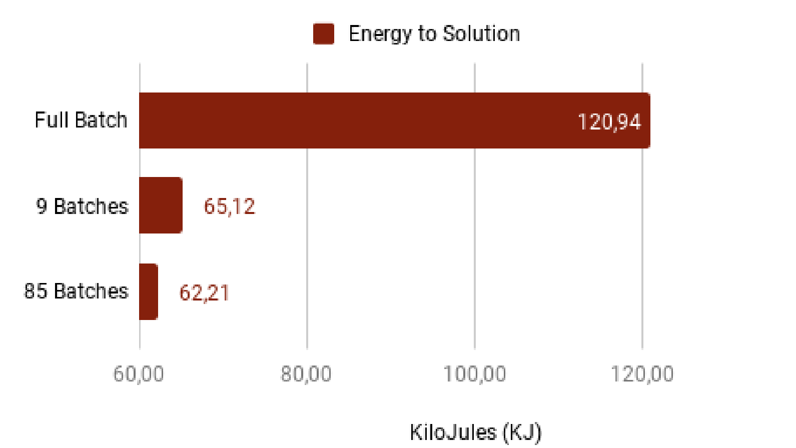
S-2: 9 Batches that Contain 10.000 Patients Records





S-3: 85 Batches that Contain 1000 Patients Records

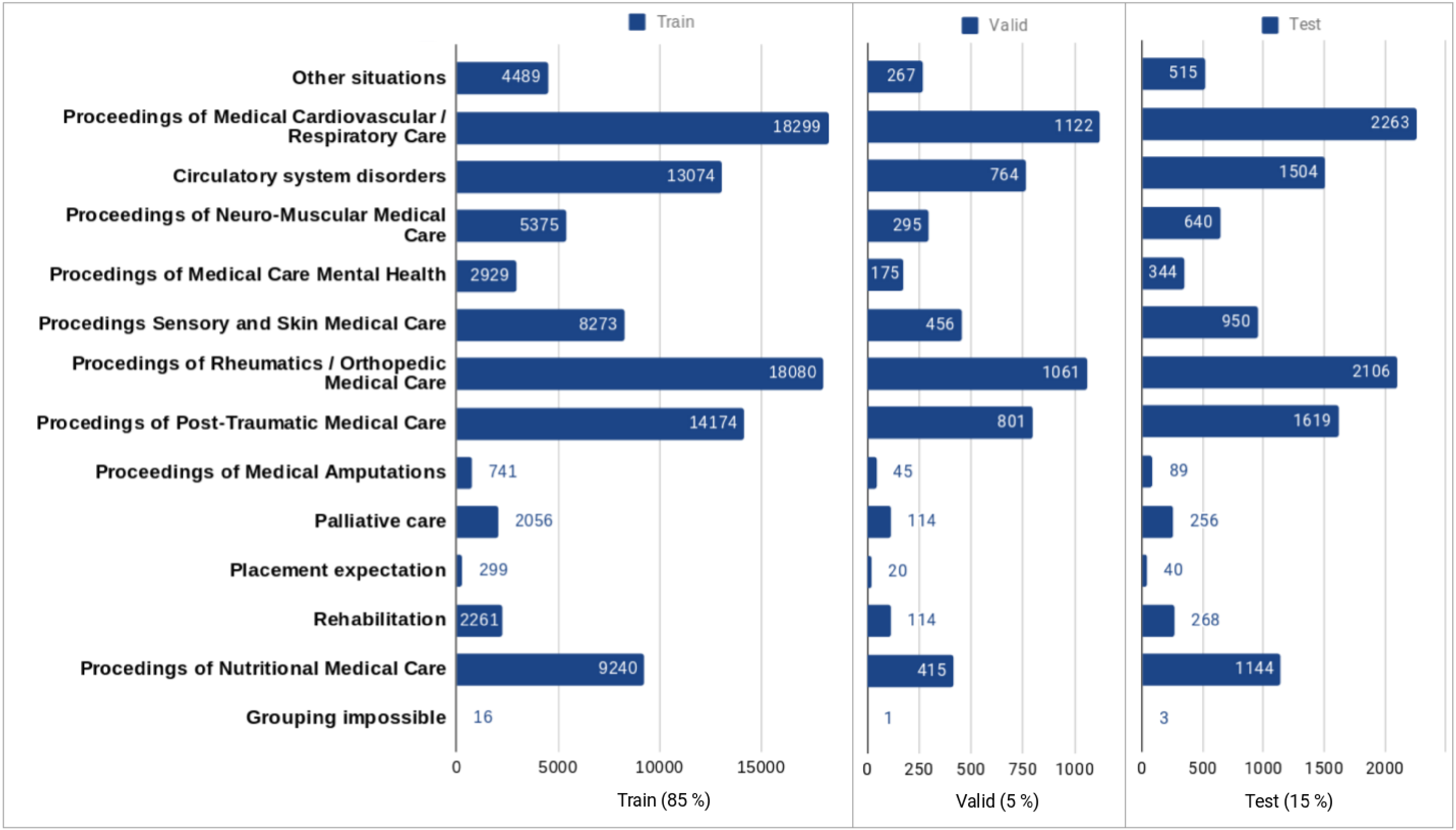




Supervised Learning

Medical Target 1

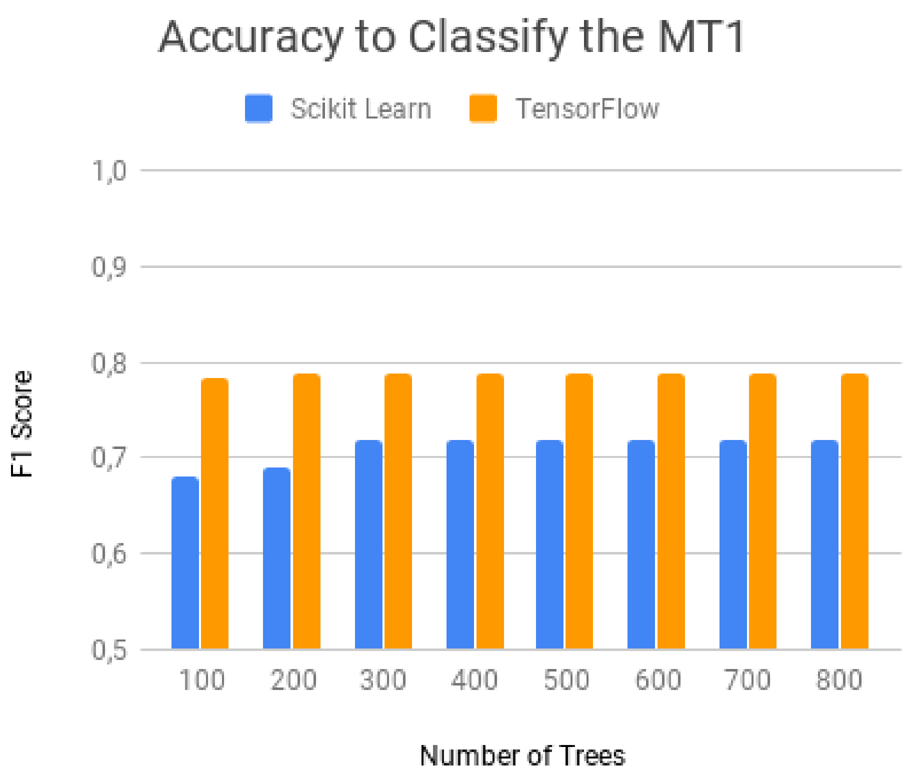
Medical Target 1: Care Purpose Description Labels at ICU

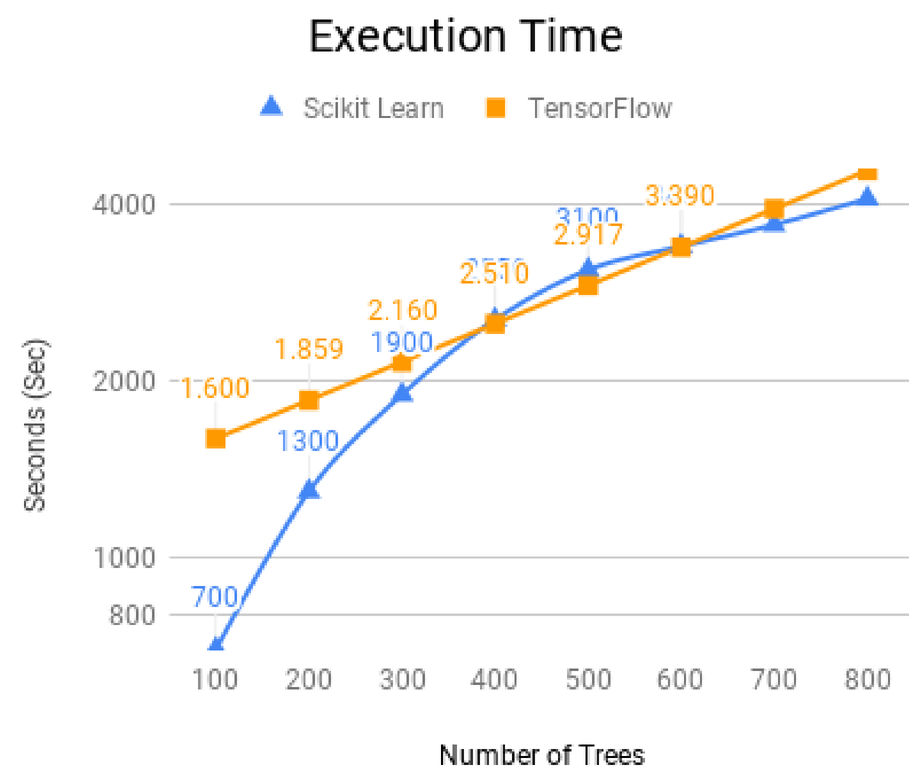


Supervised Learning

Random Forest Classifier

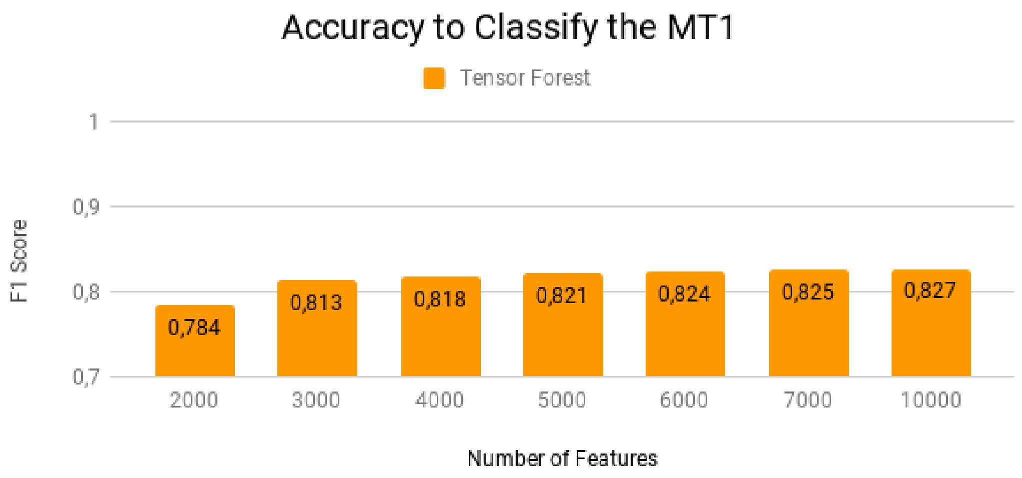
Experiments: Comparison of Random Forest off-Line Vs on-Line

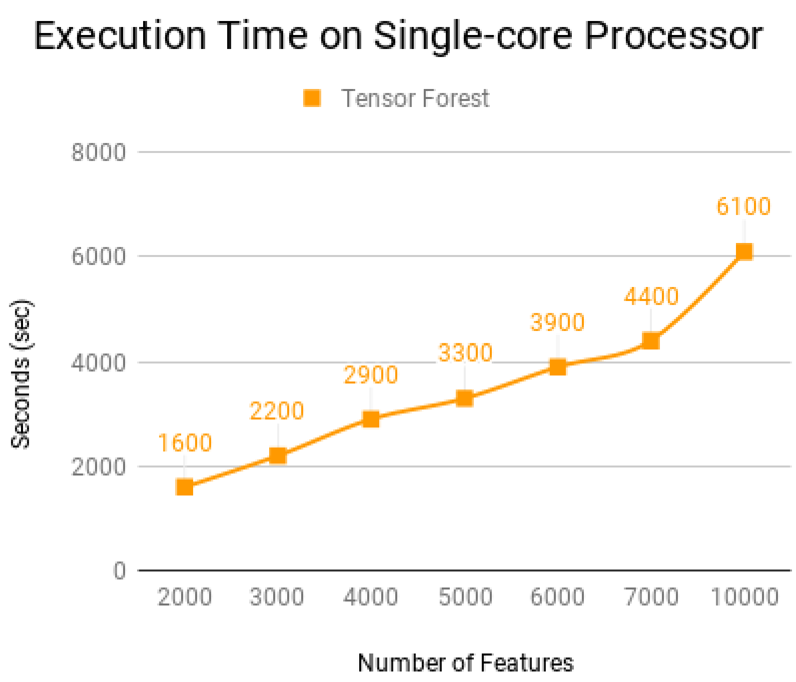




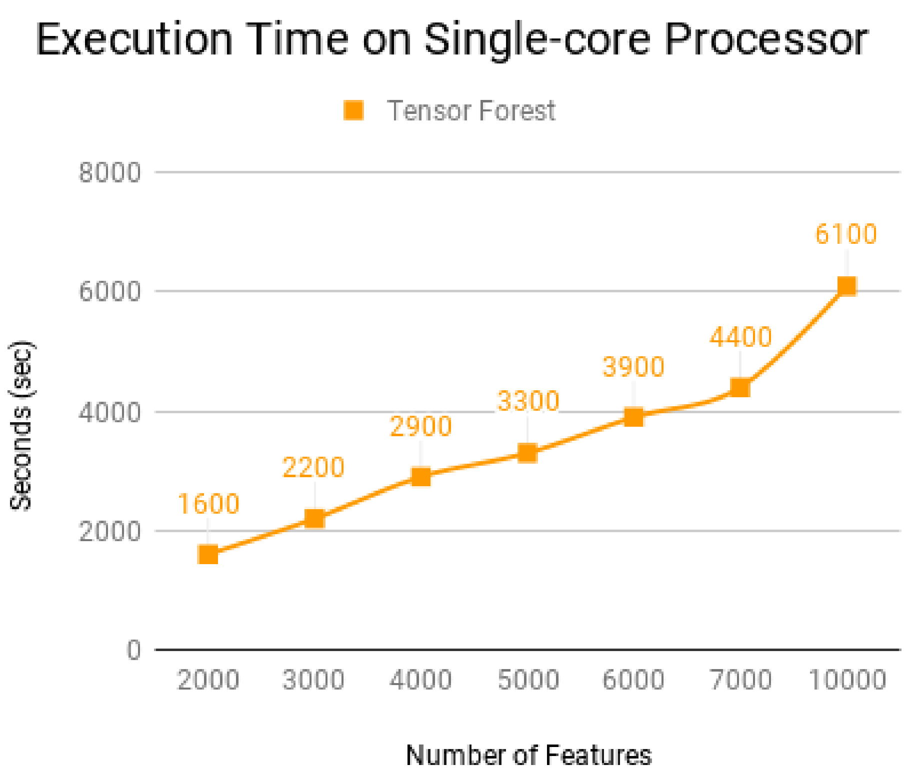
6. Source **TensorFores: Scalable Random Forests on TensorFlow.** By Colthurst, Tomas et al. 2016

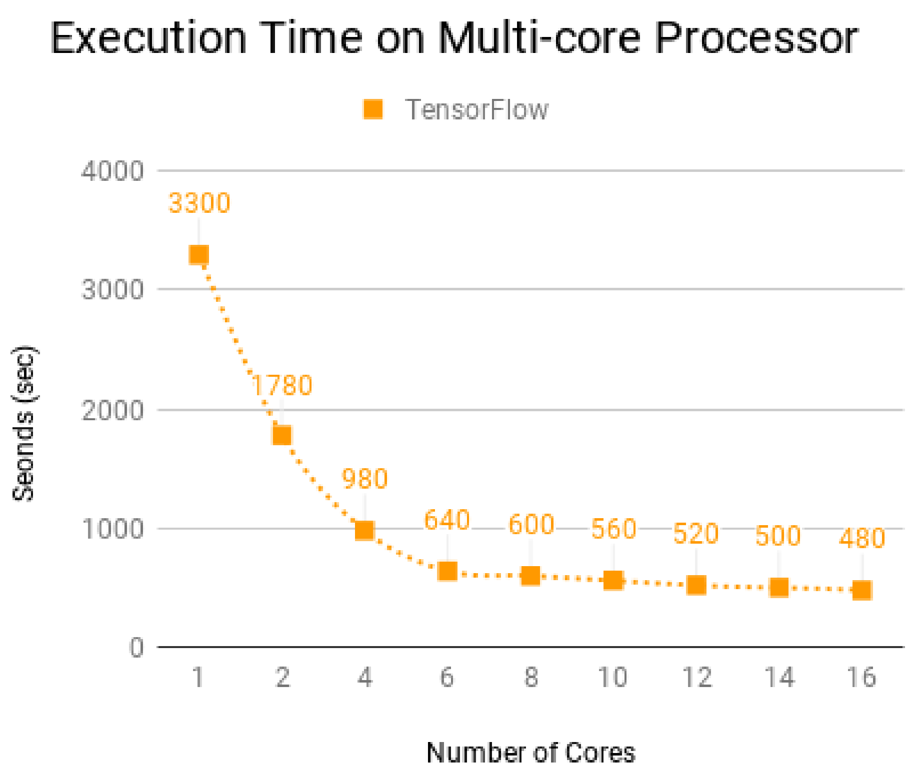
Experiments: Tensor Forest Analysis on Different Number of Features





Experiments: Tensor Forest Analysis on Multi-Core Processor

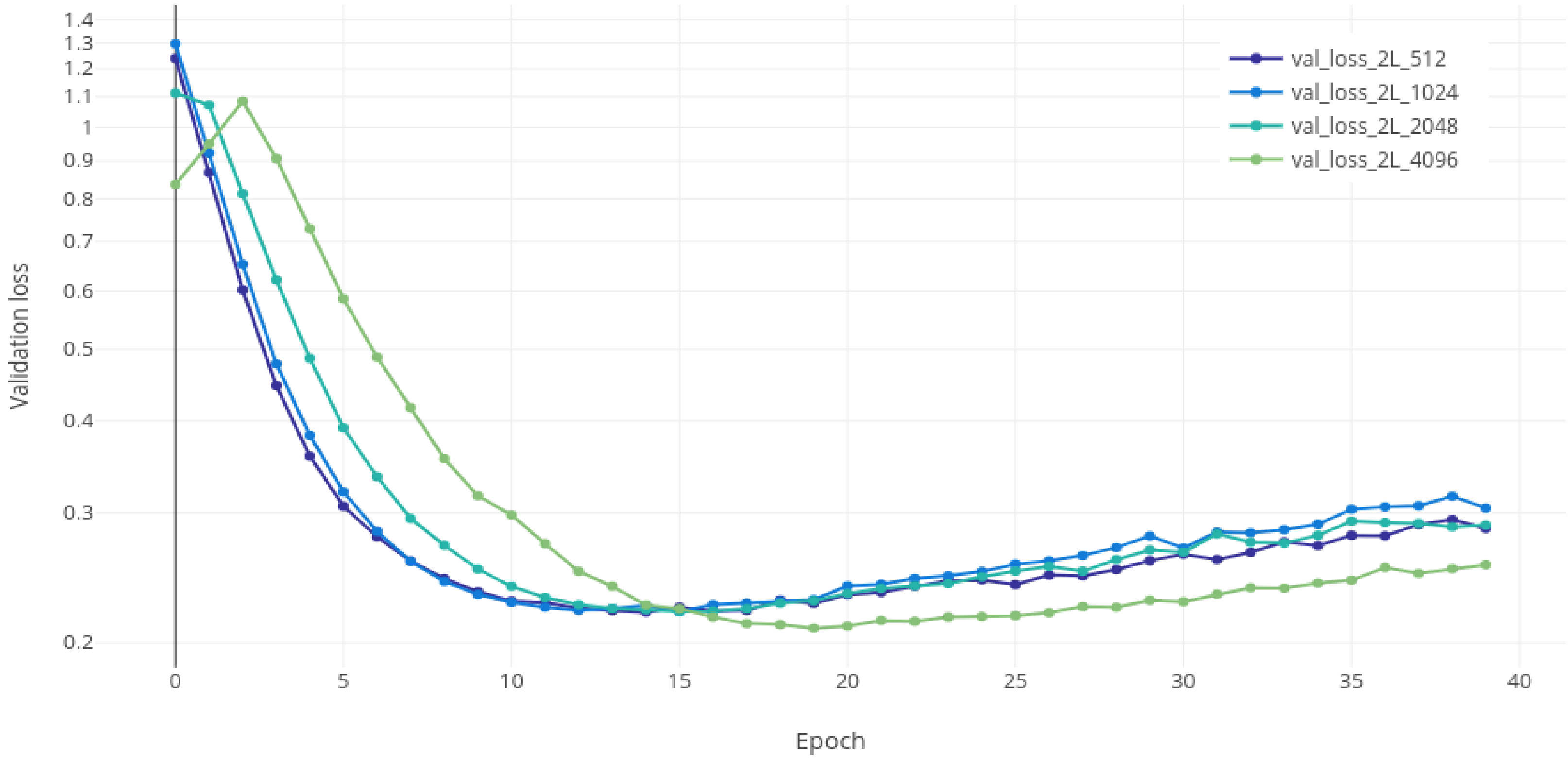




Supervised Learning

Feed-forward Multilayer Perceptron

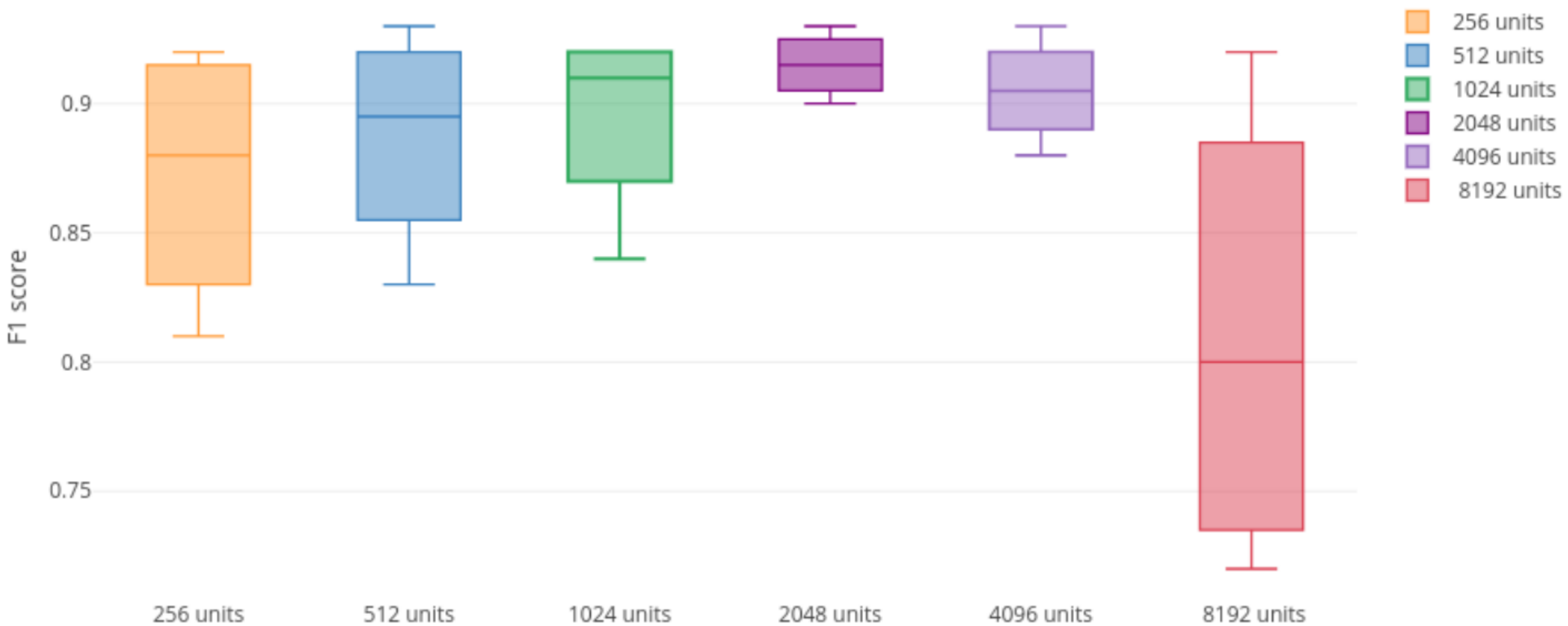
Experiments: Feed-forward Multilayer Perceptron



Stretching the 4096 Neurones over Deep Architectures

Number of units = 4096	F1 score	execution time (hours)	power consumption (watts)	energy consumption Mj
2 layers - 4096 units	0.92	4.85	214.86	3.75
4 layers - 2048 units	0.85	4.8	173.17	32.99
8 layers - 1024 units	0.72	4.76	153.43	2.68
16 layers - 512 units	0.75	4.72	130.95	2.23

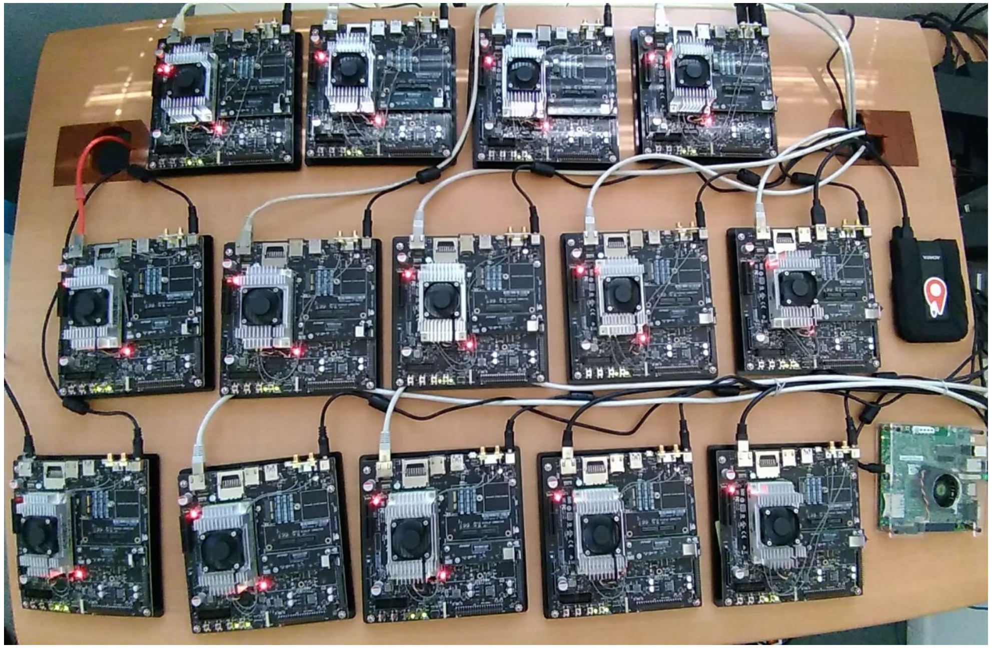
Preliminary Results: F1-score for Different Stretching Configurations



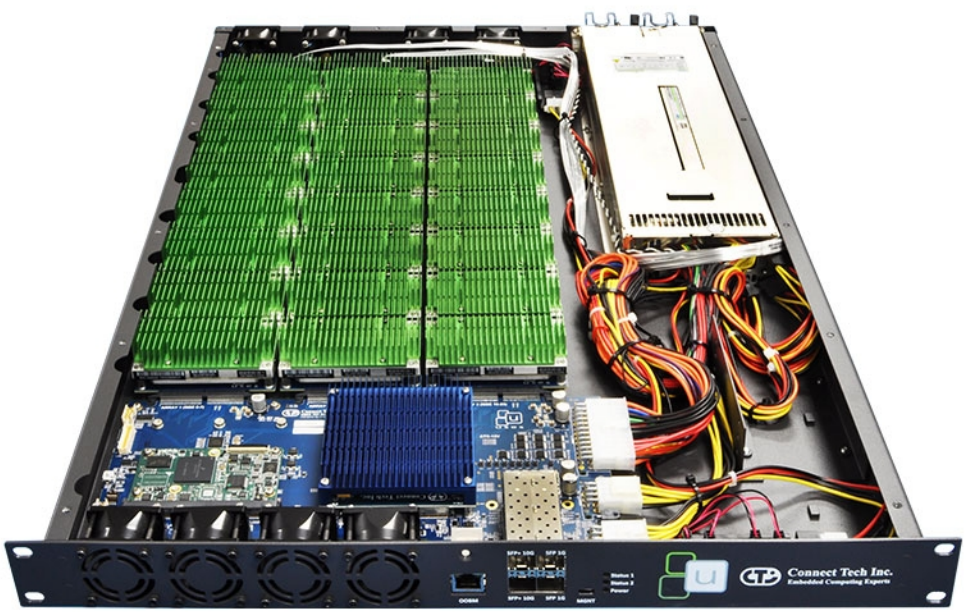
Architecture	F1 score	execution time (hours)	power consumption (watts)	energy consumption Mj
256 units - 2 layers - 128 units	0.91	4.71	92.27	1.57
2048 units - 8 layers - 256 units	0.91	4.73	101.68	1.73
8192 - 2 layers - 4096 units	0.92	4.85	214.86	3.75

Distributed Processing for Training DNN on Jetson TX2 Mini-Clusters

Computational Resources



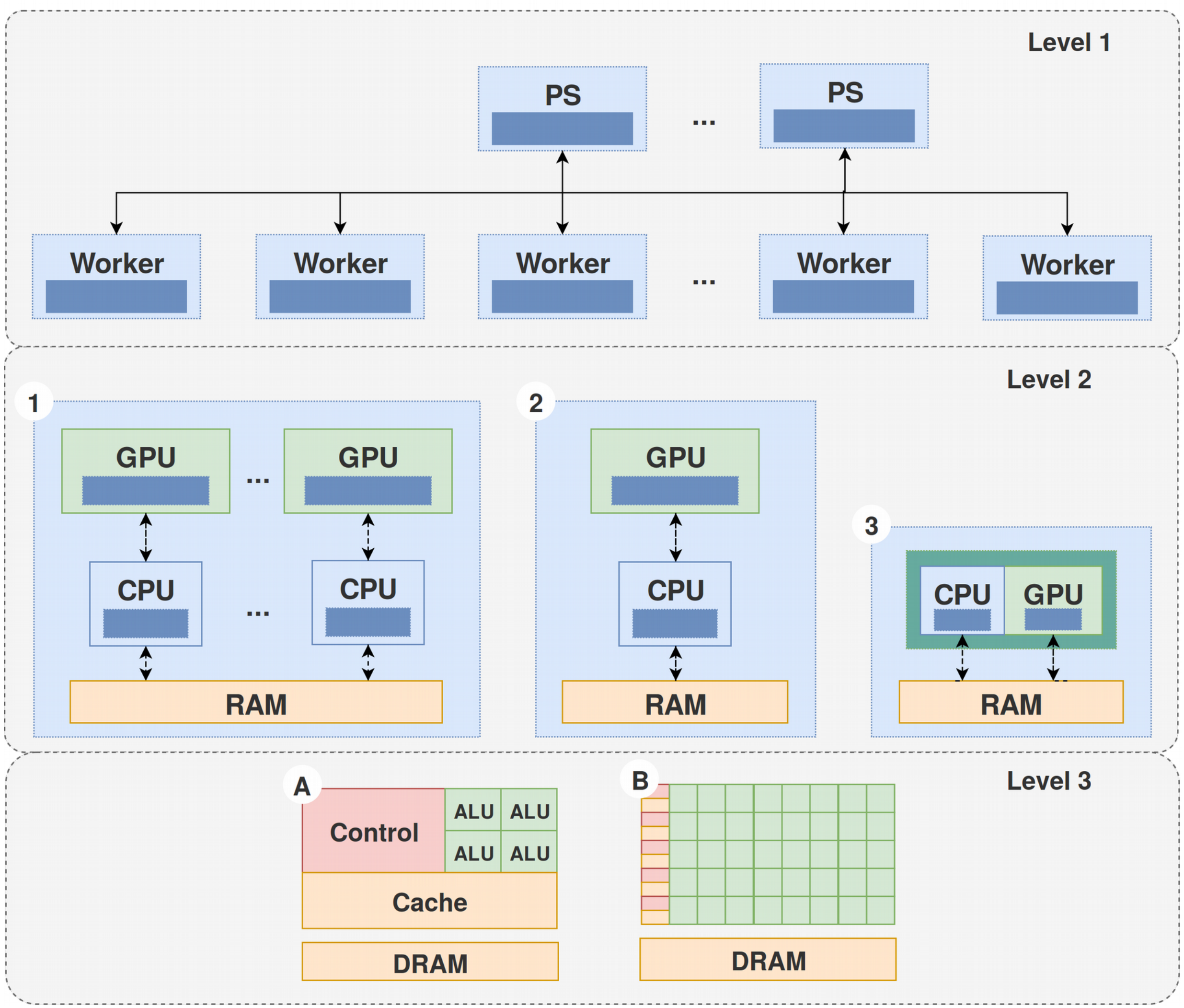
Mini-Cluster Jetson TX2



Array Node with 24 Jetson TX2

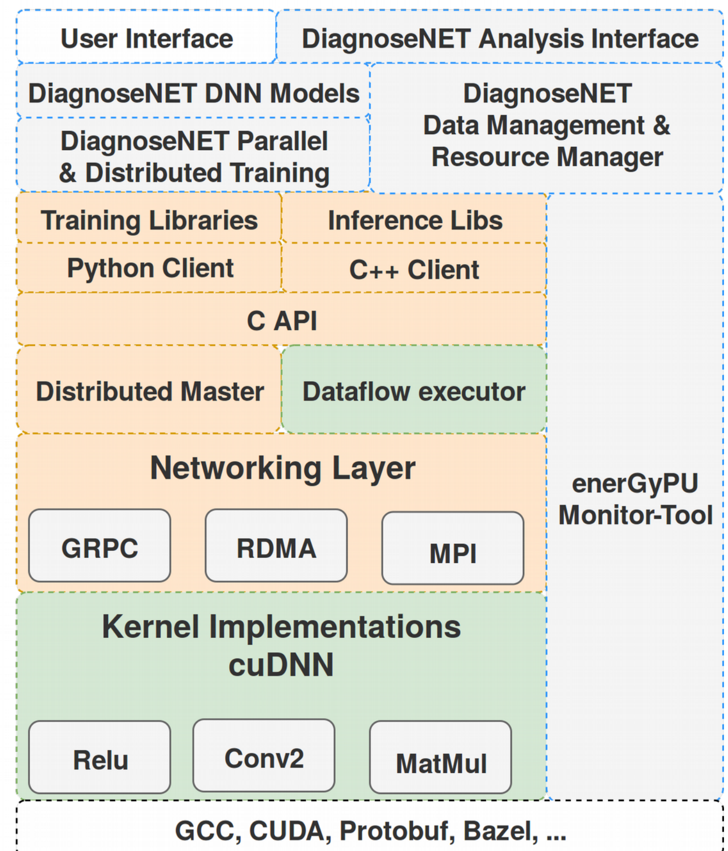
Develop DiagnoseNET for Training Large-Scale DNN on Distributed Systems

Levels of Parallel and Distributed Processing



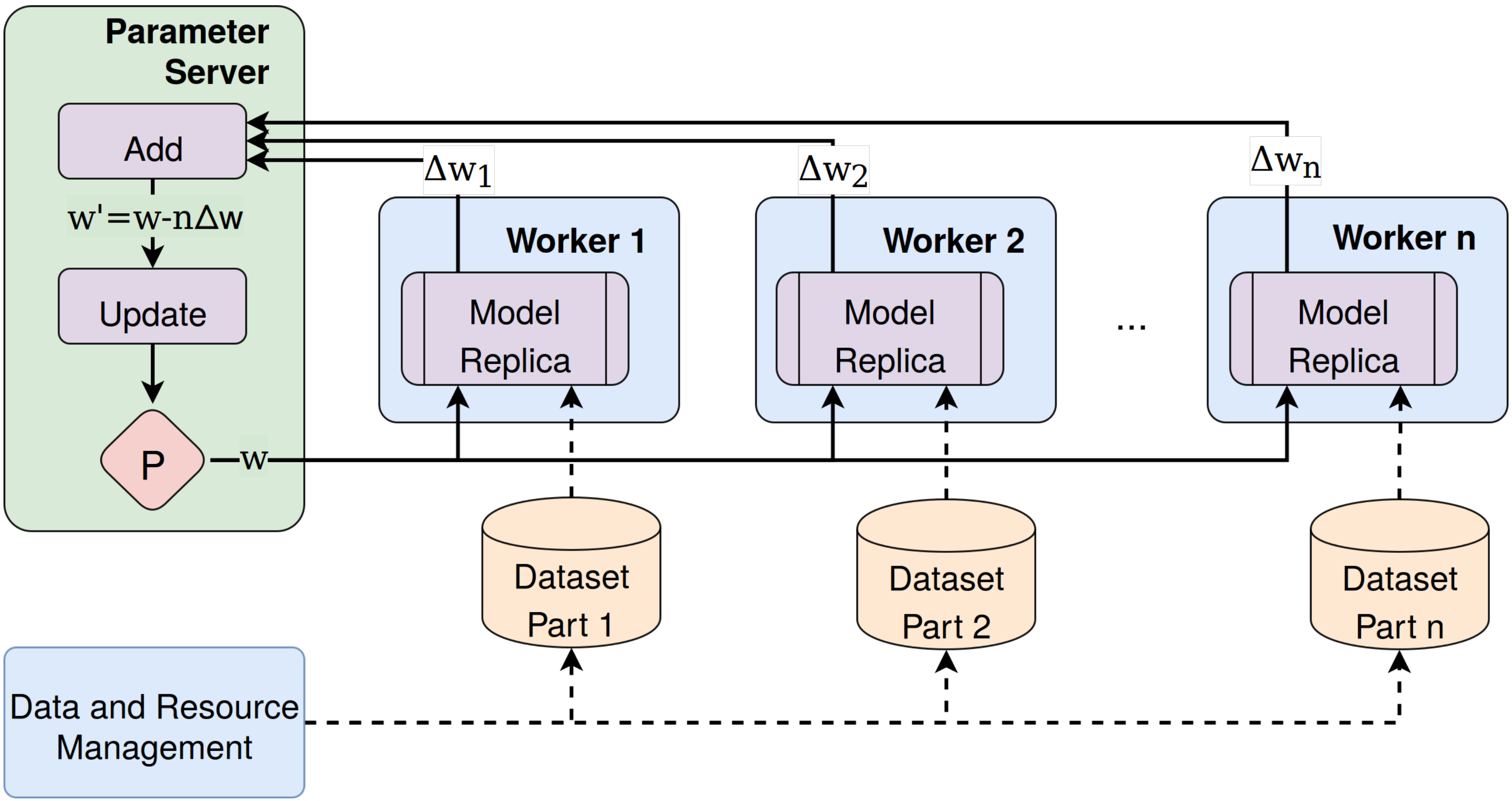
DiagnoseNET

Cross-platform Library



7. Adapted from Snap Machine Learning. By IBM Research et al. 2018
8. Adapted from TensorFlow Architecture. By Google Research.

Task-Based Data Parallelism: Synchronous



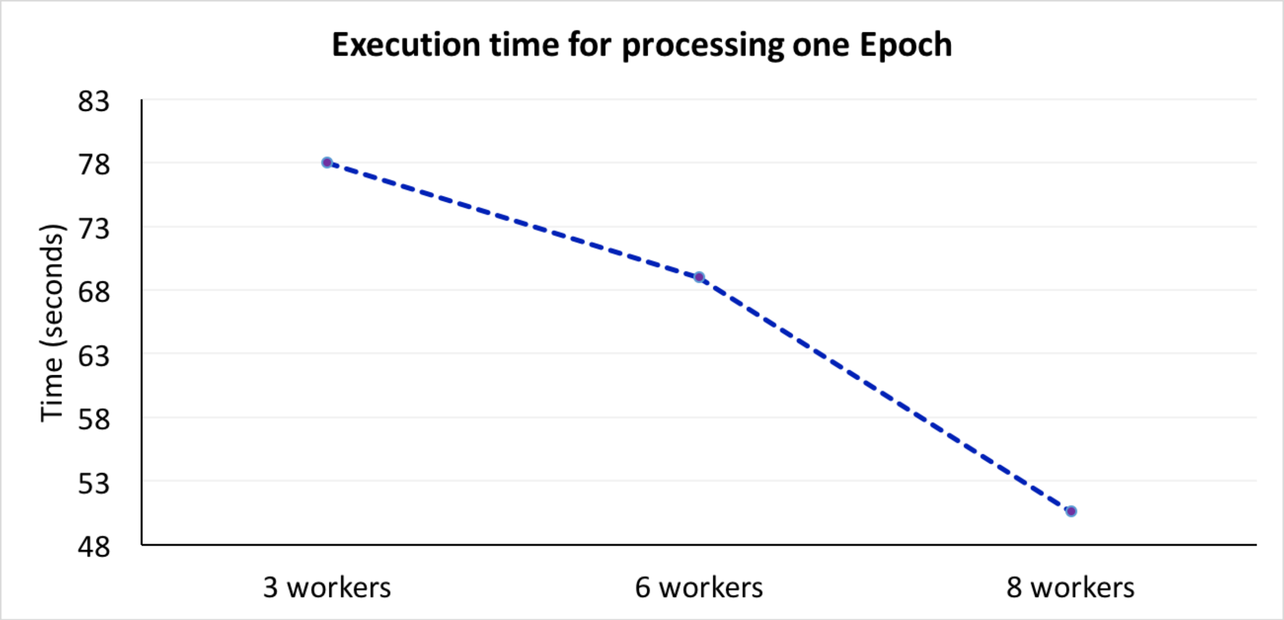
(Data Size):

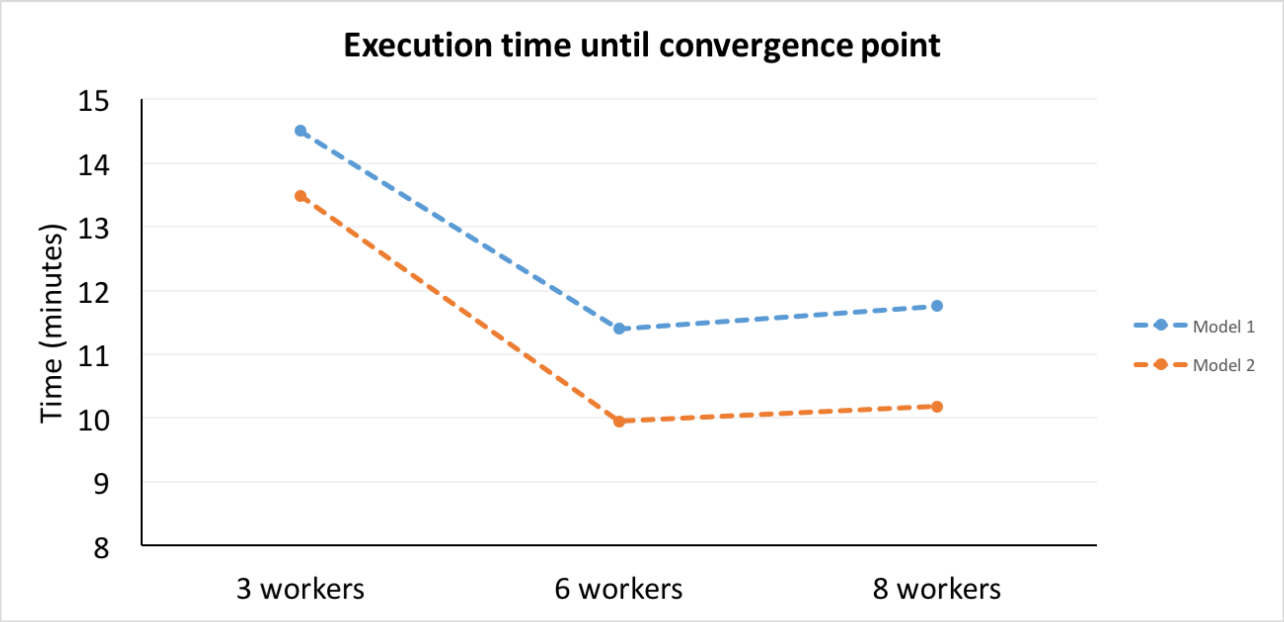
+ Setting the number of workers and micro batch.
+ Fine-tuning DNN hyperparameters.
+ Speeds up the training.
+ I/O Intensive.

9. Adapted from TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. By Google Research. 2015.

Preliminary Results to Scale the Unsupervised Representation Learning

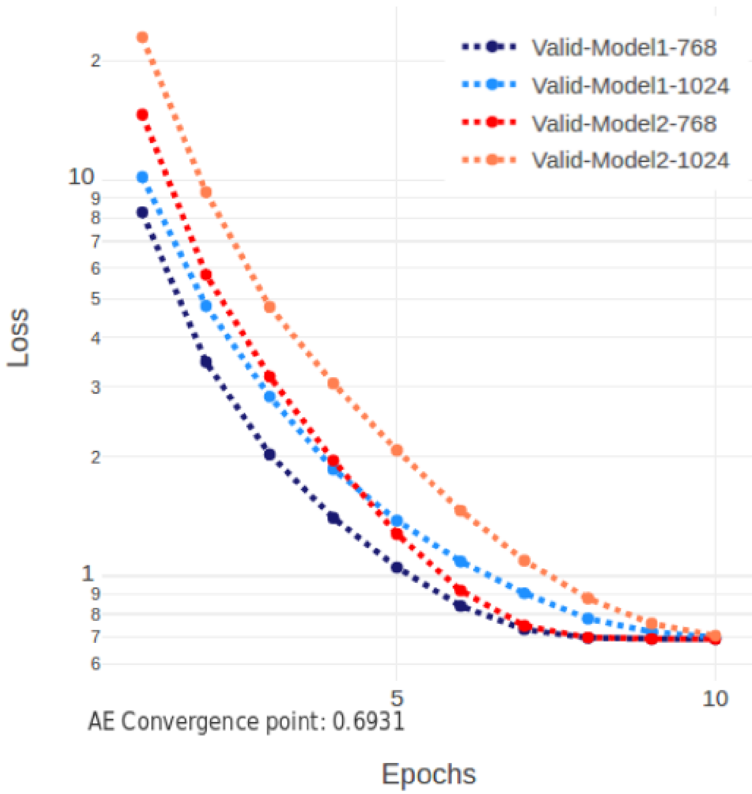
Preliminary results using: 10.000 records and 11.466 features.



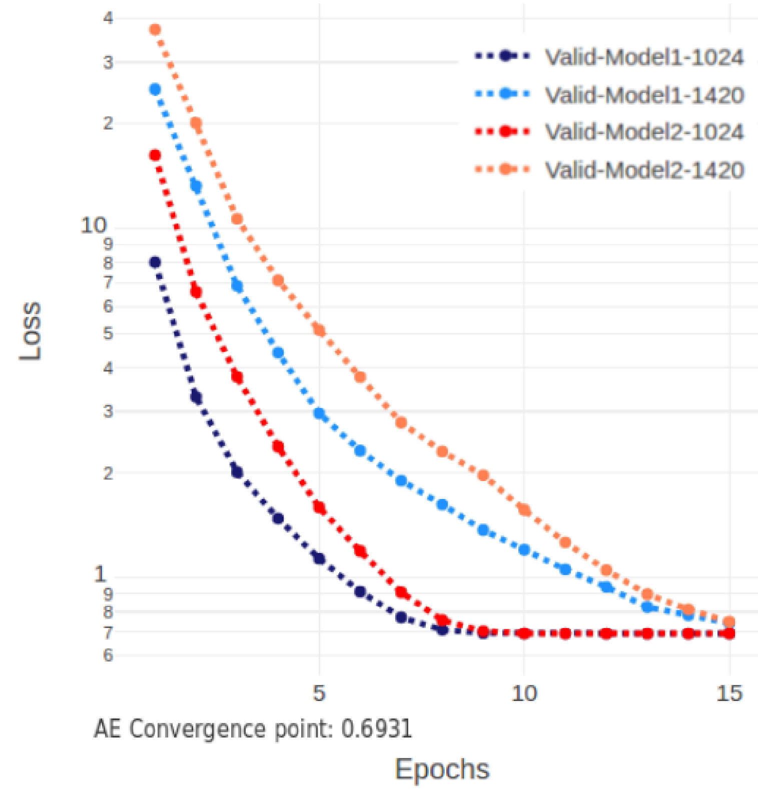


Number of Workers and Task Granularity as Factor to Early Model Convergence

- *Early convergence comparison between different groups of workers and task granularity for distributed training with 10.000 records and 11.466 features.*



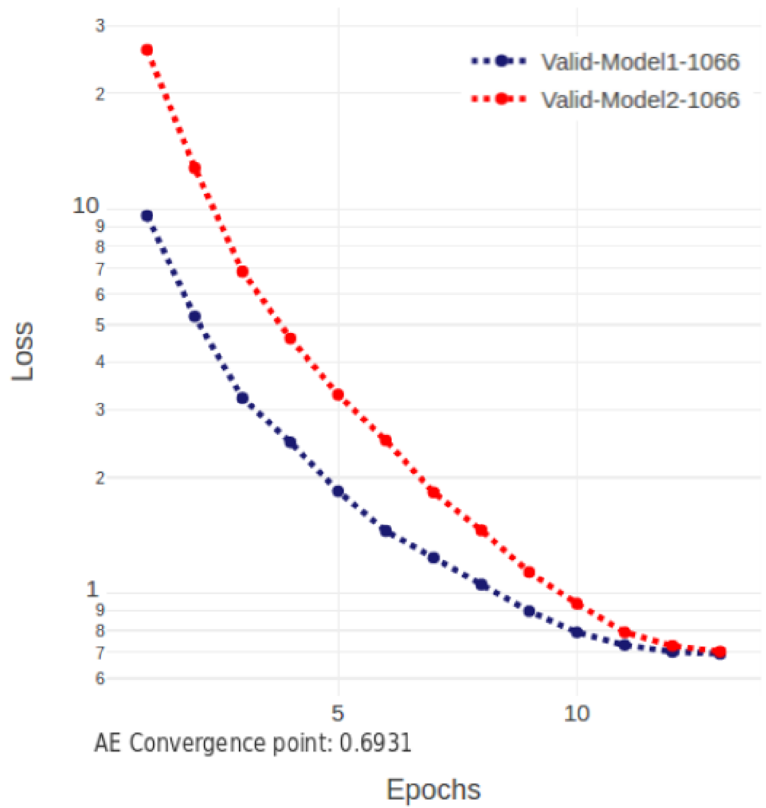
1.30 mins in avergange for processing one epoch on 1 PS 3 workers.



1 min in avergange for processing one epoch on 1 PS 6 workers.

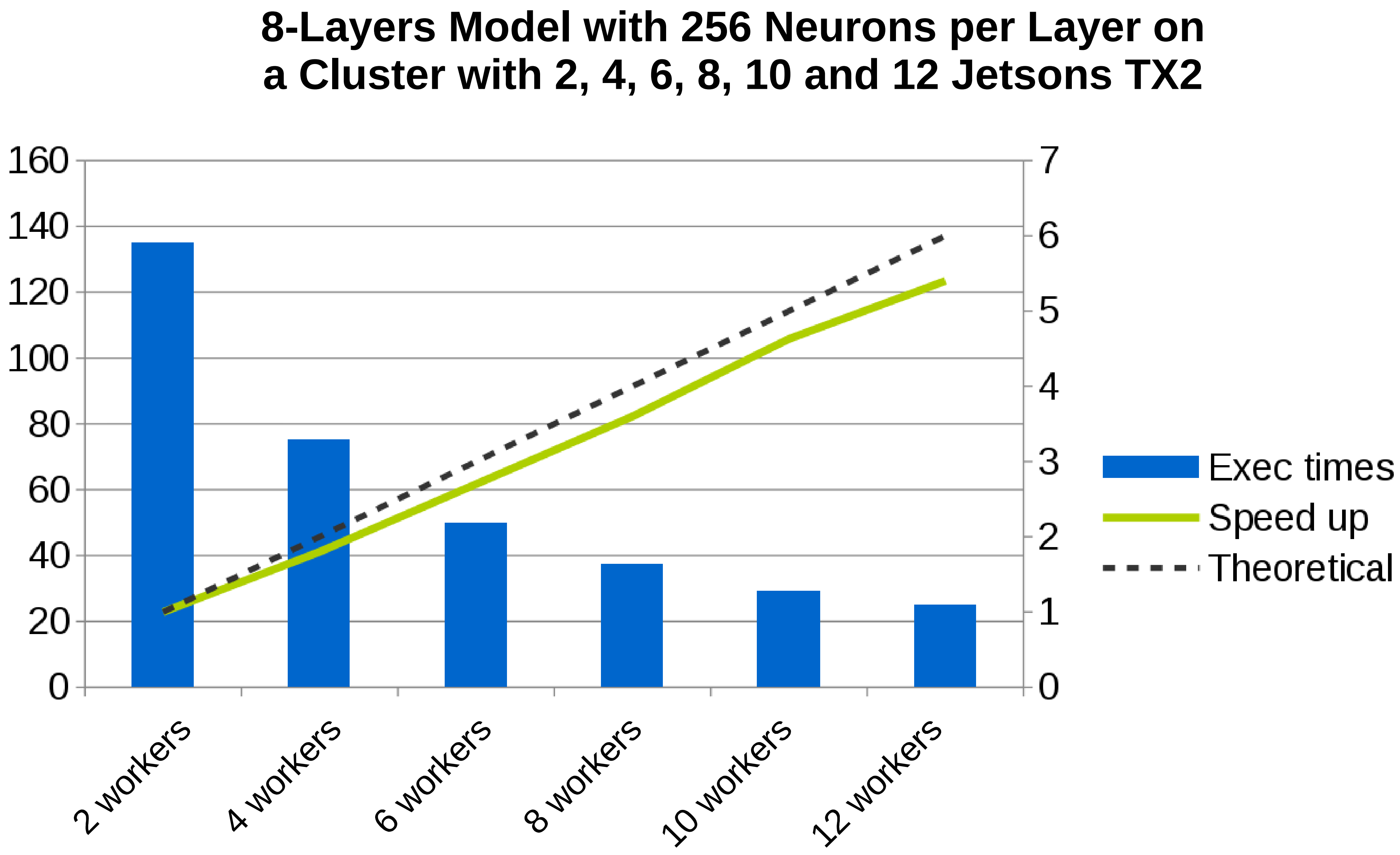
Number of Workers and Task Granularity as Factor to Early Model Convergence

- *Early convergence comparison between different groups of workers and task granularity for distributed training with 10.000 records and 11.466 features.*



50.6 secs in avergange for processing one epoch on 1 PS 8 workers.

Preliminary Results to Scale the Feed-forward Multilayer Perceptron



Conclusions

Using hundreds of **Gradient Updates by Epochs** with synchronous data parallelism offer an efficient distributed DNN training to early convergence.

Adapting the **Number of Records by Batch** or the **Model dimensionality** to minimize the bottleneck of data transfer from host memory to device memory reducing the GPU idle status.

Use a mini-Cluster of Jetson TX2 presents good results in distributed training using synchronous data parallelism.

Therefore, this can be used as a learning center with minimal infrastructure requirements and low power consumption and brings the opportunity to use a dataset with more patient’s features.

Latent Representation:

- Reduces the number of sparse features without loss of precision in future classification.
- Reduces training time (41 %) to classify the first medical target.

Next Work on DiagnoseNET

- Distributed others kind of DNN architectures Like (CNN, RNN) and Tensor Forest.

- Compare several architecture:
  - multi-GPU (share memory)
  - vs Cluster (distributed memory)
  - vs Array (hybrid memory)

- For several medical task:

  MT-1: Predict the '***Care Purpose or Major Clinical Category'*** of patients in (coarse grain CMC / fine grain GHJ) from inpatients features recorded at the admission time.

  MT-2: Predict the ***'Clinical Procedures'*** from Inpatients features recorded at the admission time and the Primary Morbidity.

  MT-3: Predict the ***'Inpatient Destination'*** (home, transfer, death) and length of hospitalization stay from inpatients features recorded at the admission time and Primary Morbidity and Clinical Procedures.

Scalability Analysis of Mini-Cluster Jetson TX2 for Training DNN Applied to Healthcare

John A. GARCÍA. H.
Frédéric PRECIOSO, Pascal STACCINI, Michel RIVEILL

Université Côte d'Azur, CNRS
Laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis – I3S
Département d'Informations Médicales DIM – CHU Nice