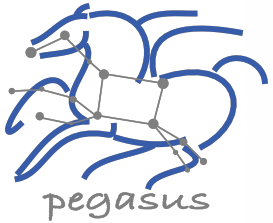


Using Gantt Charts and CNNs to Study end-to-end Scientific Workflows and their Anomalies through Visual Analysis

Patrycja Krawczuk, George Papadimitriou, Shubham Nagarkar,
Mariam Kiran, Anirban Mandal, Ewa Deelman

This work was funded by the US Department of Energy under Grant DE-SC0012636M

OUTLINE

Motivation

High Level Approach Overview

Dataset Creation

Dataset Summary

Machine Learning Methods

Model Training Workflow

Experimental Results

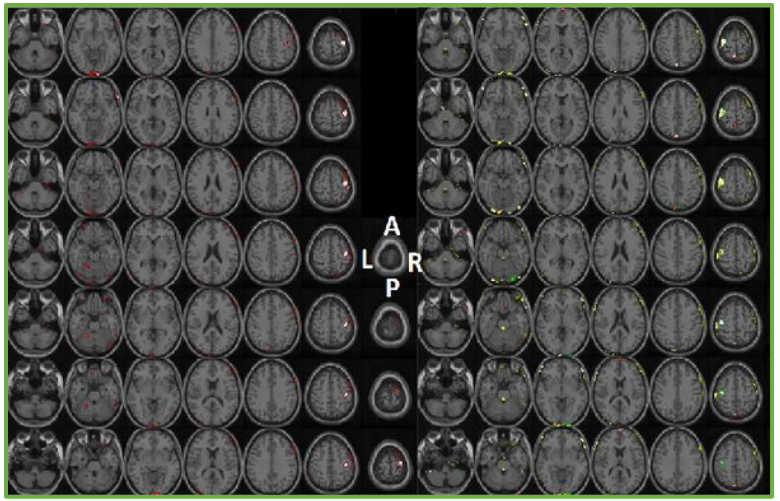
Limitations and Future Work

Patrycja Krawczuk

George Papadimitriou

Motivation

Large-scale population-level cancer surveillance study run on Oak Ridge National Lab's Summit computer.

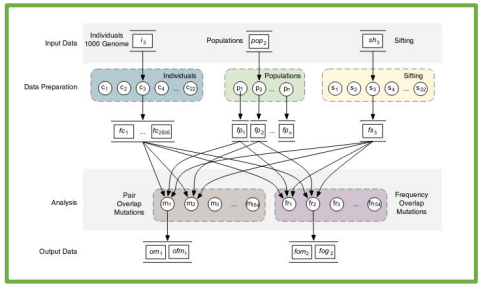


Classification of fMRI volumes with 3D convolutional neural networks (Vu H. et al.)

Question: *Can we employ ideas from the field of computer vision to identify anomalies on workflow execution traces?*

High Level Overview of the Approach

Workflow

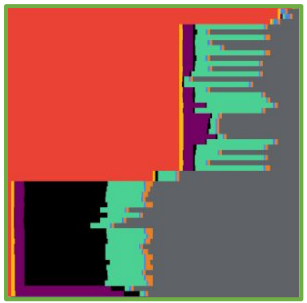


Monitoring Data

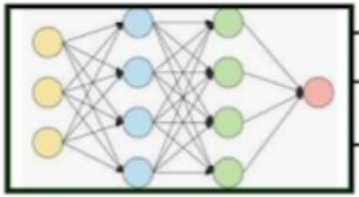
	ready	submit	execute_start	execute_end	post_script_start	post_script_end
0	1591766744	1591766748	1591766750	1591766754	1591766754	1591766759
1	1591766759	1591766766	1591767046	1591767093	1591767093	1591767098
2	1591766759	1591766766	1591767006	1591767047	1591767047	1591767052
3	1591766759	1591766766	1591766770	1591767077	1591767077	1591767093
4	1591766759	1591766766	1591766770	1591767083	1591767083	1591767098

panorama360

Processed Data



Anomaly Detection and Classification via CNN



Network Issues

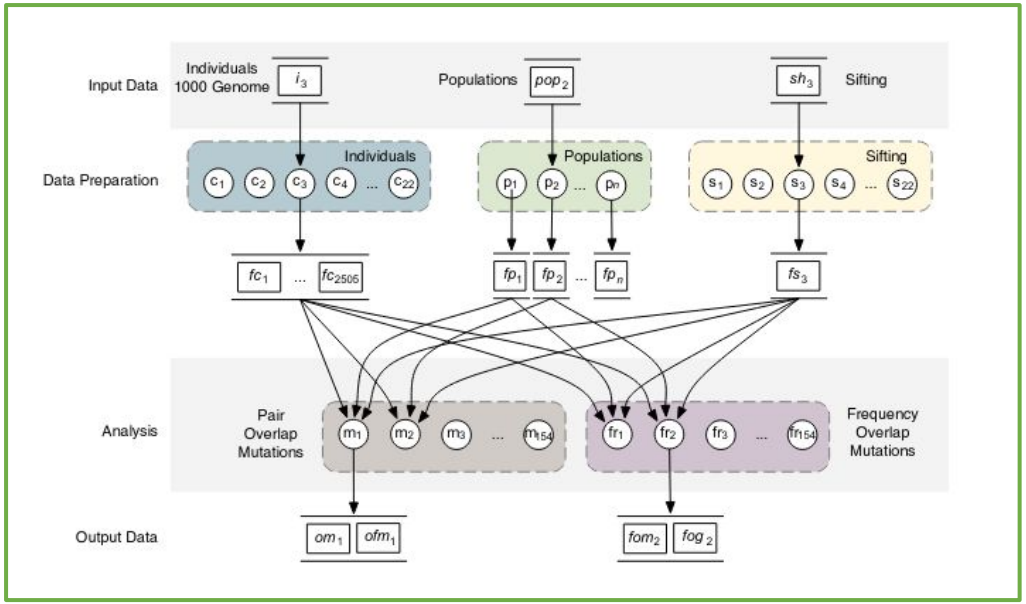
CPU Problem

....

Healthy Execution

Human in the Loop:
Troubleshooting

Dataset Creation: Workflow Executions



Pegasus 1000 Genome Workflow

Executed on an HTCondor pool deployed on the ExoGENI testbed using

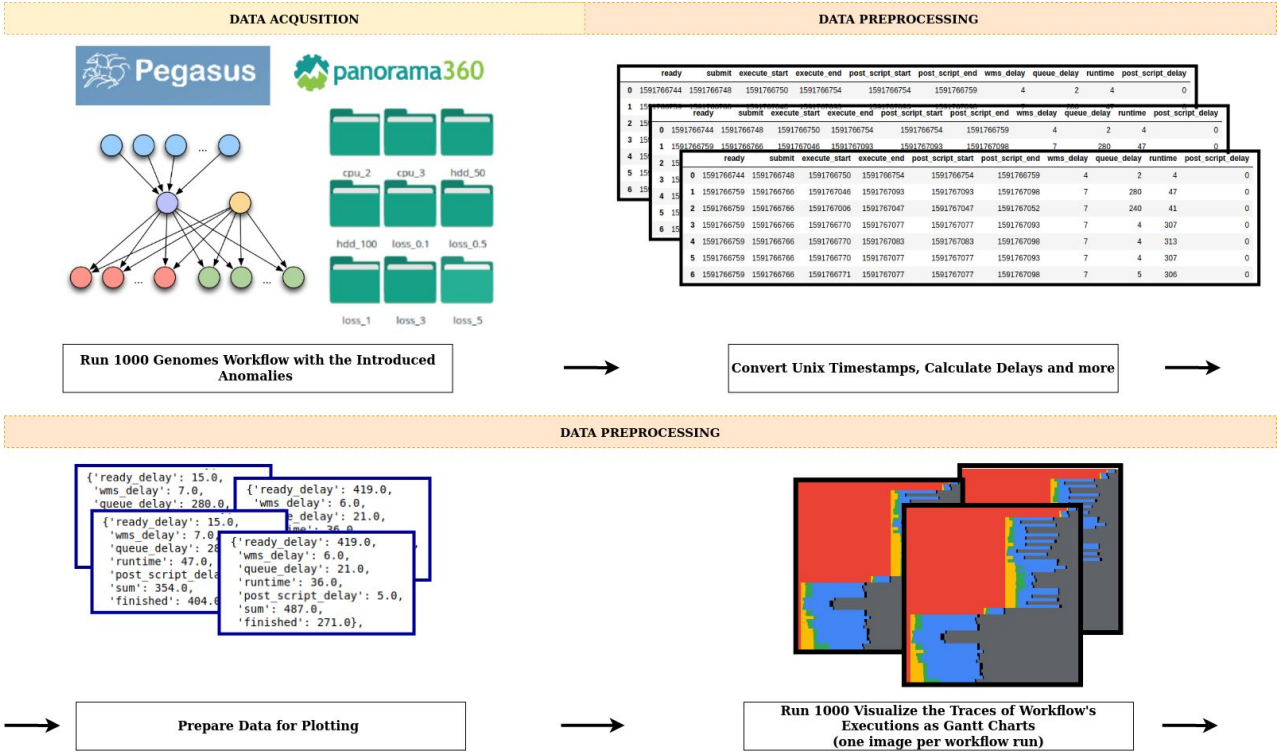
- 1 submit node
- 1 data node
- 5 worker nodes

The version of the executed workflow

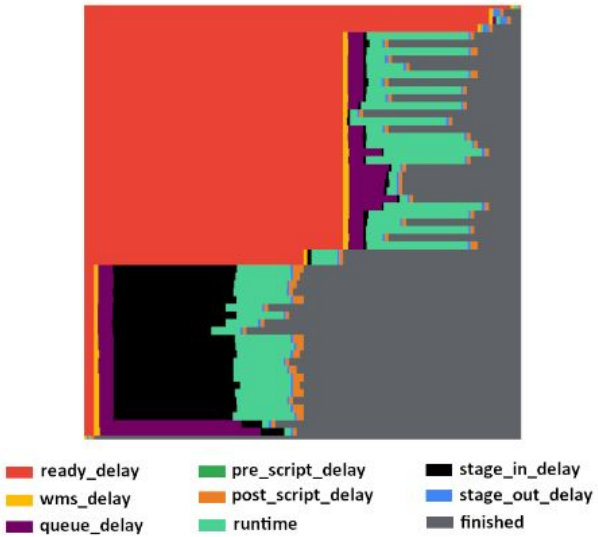
- Had 52 compute tasks
- and transferred over 22 GBs

During the execution we introduced synthetic anomalies affecting **compute** and **network resources.**

Dataset Creation: Data Collection and Preprocessing

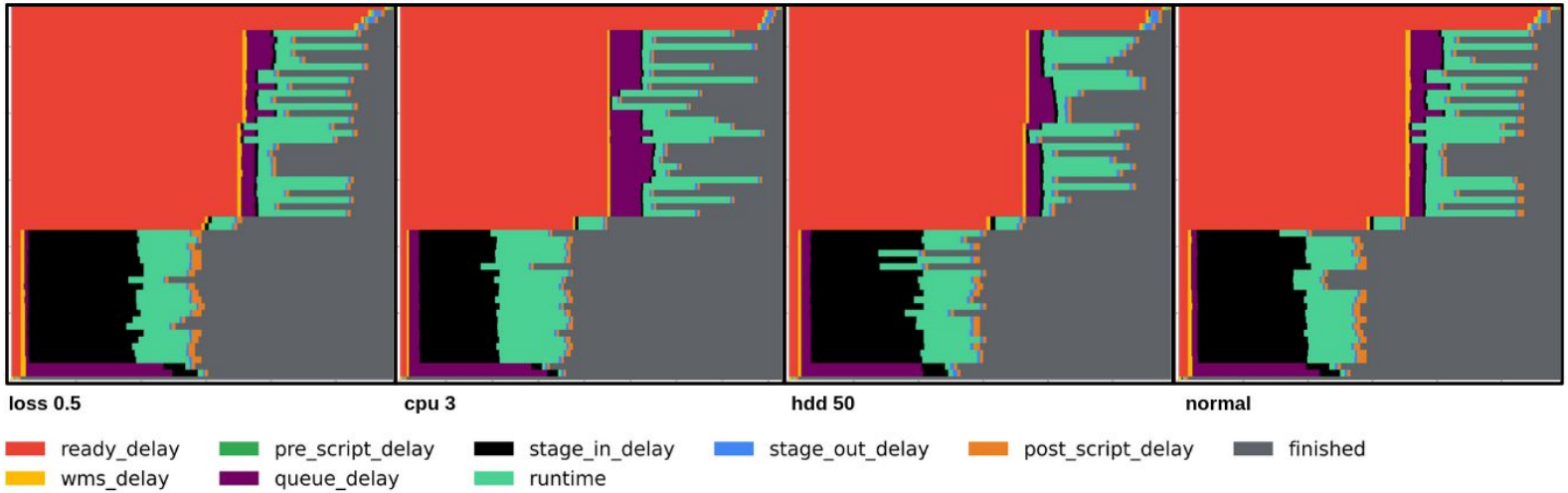


Dataset Creation: High Resolution Gantt Charts



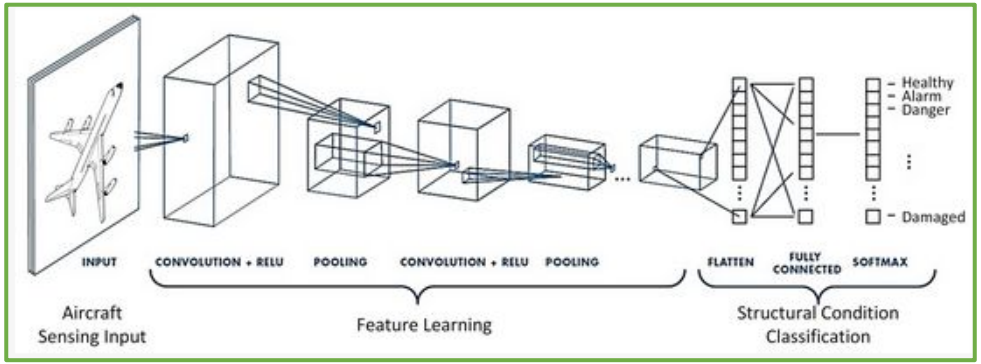
- **Ready time:** Timestamp since the beginning of the workflow, where all dependencies have been met and job can be dispatched.
- **Pre script delay:** Time spent on a script that is executed before job submission (if exists).
- **WMS delay:** Time spent by the workflow management system to prepare and submit the job.
- **Queue delay:** Time spent in the queue waiting for resources.
- **Stage in delay:** Time spent transferring input data.
- **Runtime:** Time spent during computation.
- **Stage out delay:** Time spent transferring data to the intermediate scratch directory or final output directory.
- **Post script delay:** Time spent on a script executed after job exits (e.g., wms parses stdout and exit code).
- **Completion time:** Timestamp marking job completion, since beginning of workflow.

Dataset Summary

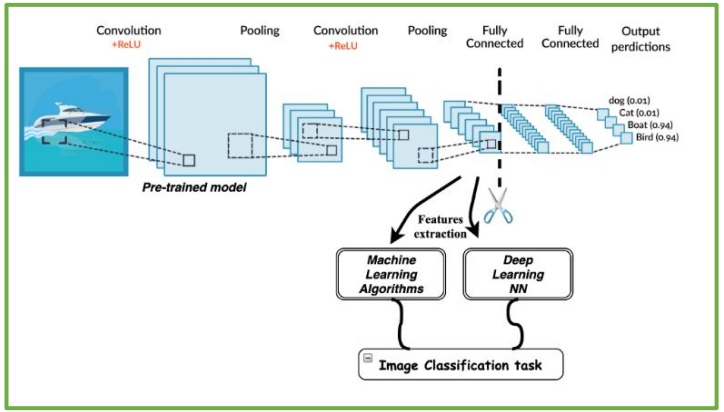


Labels	Normal	CPU		HDD						Loss				
		2	3	50	60	70	80	90	100	0.1%	0.5%	1.0%	3.0%	5.0%
# Traces	250	125	125	67	37	35	31	31	49	50	50	50	50	50

Machine Learning Methods: CNN and Transfer Learning

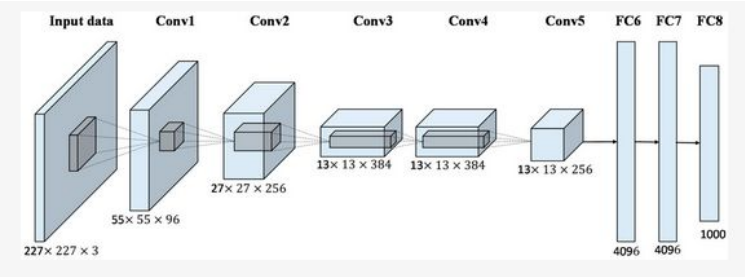


Convolutional Neural Networks (CNNs)

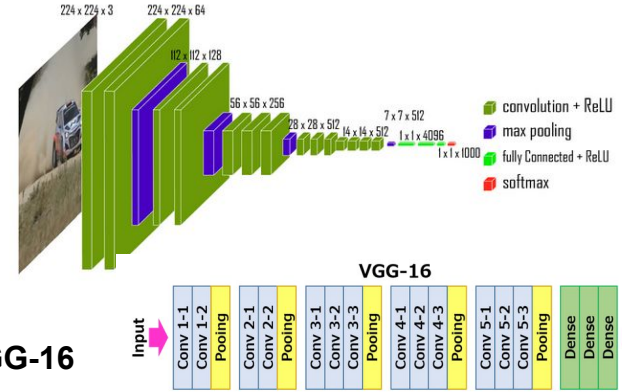


Transfer Learning Methodology

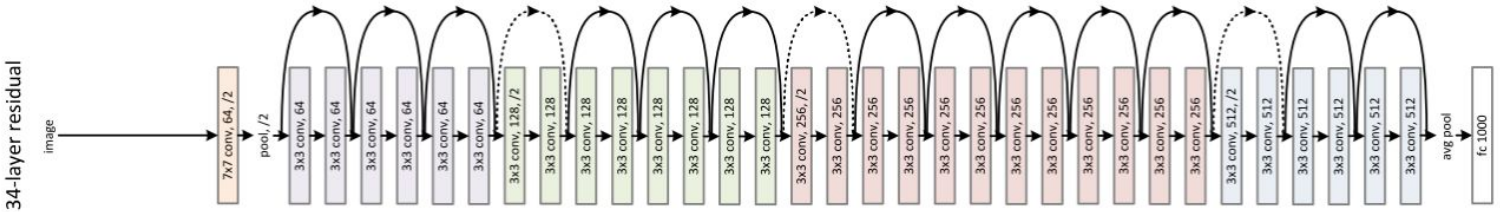
Machine Learning Methods: CNN Pre-Trained Architectures



AlexNet

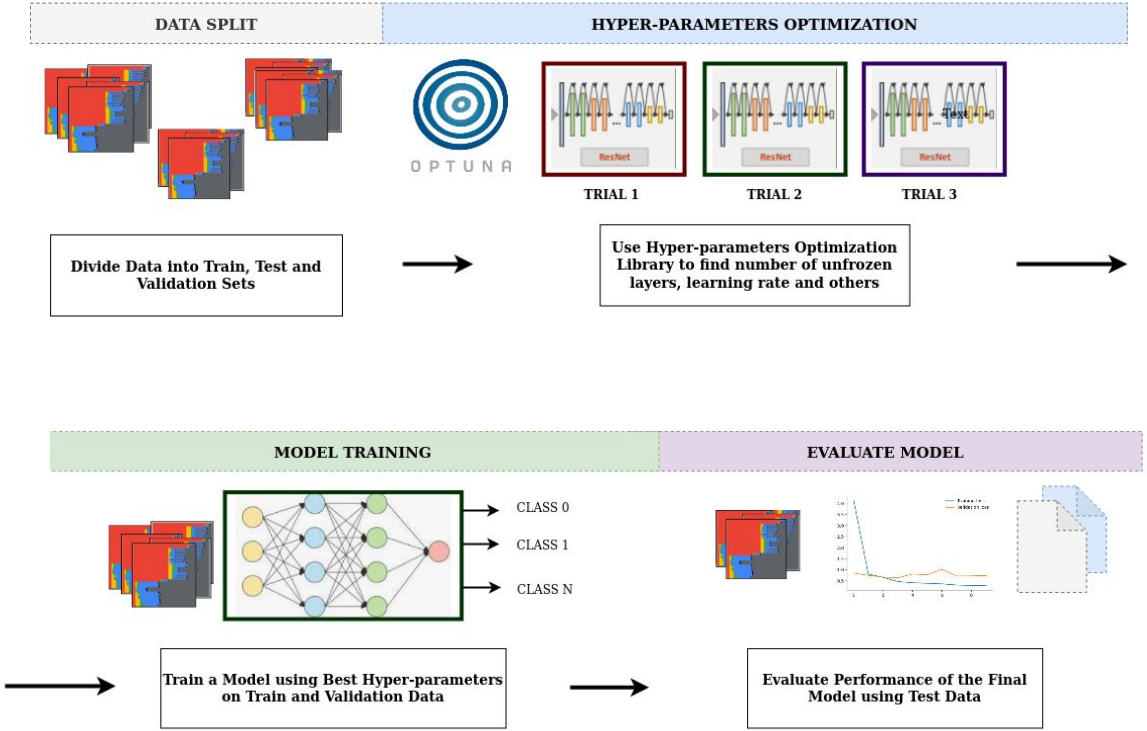


VGG-16

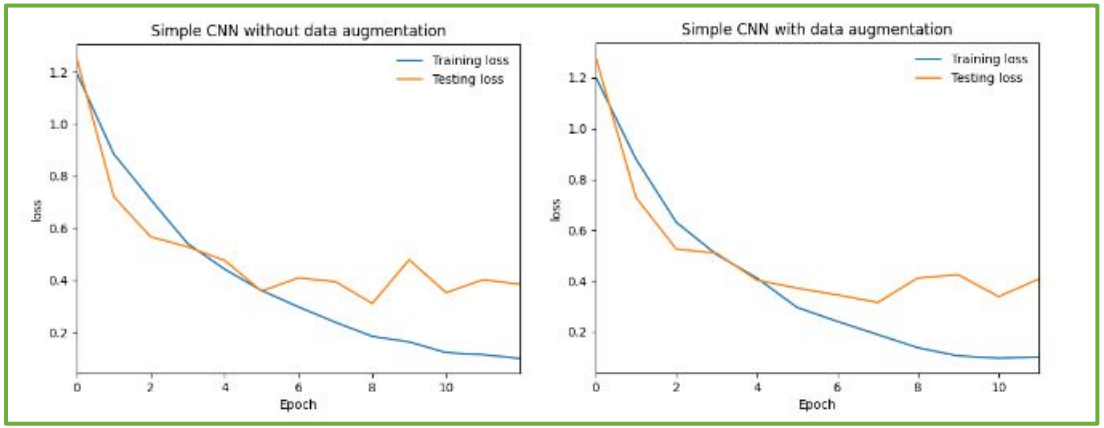


ResNet

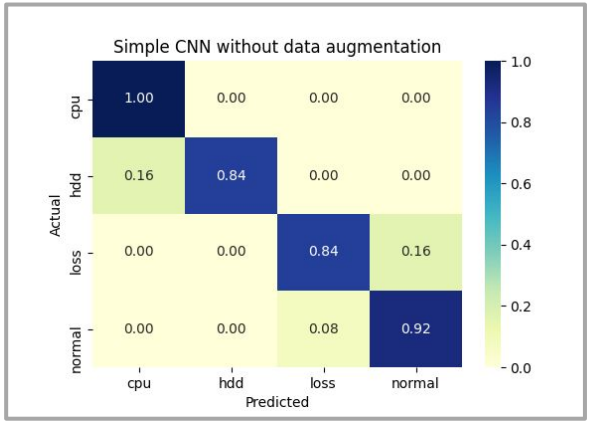
Model Training Workflow



Experimental Results: Training from Scratch



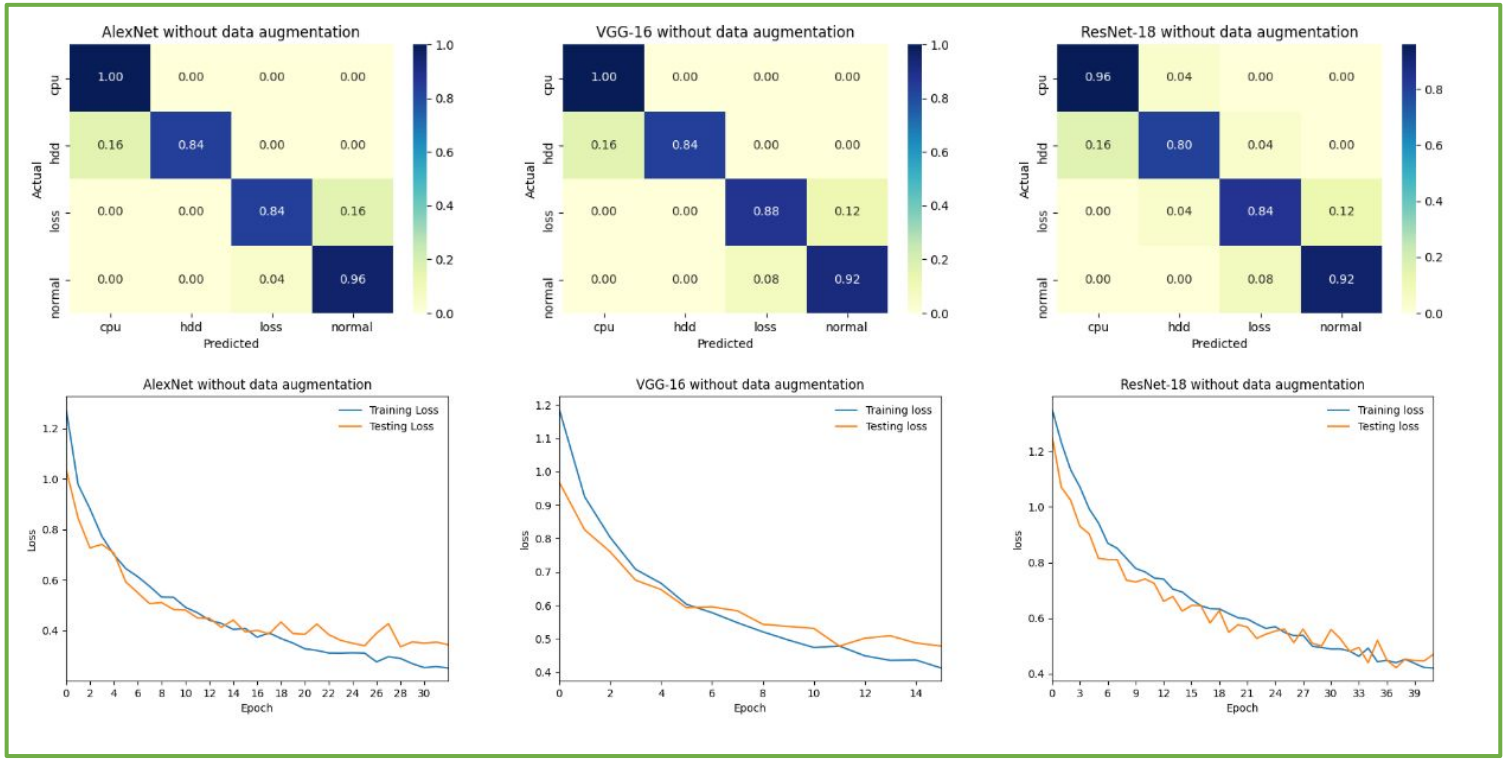
Training curves for our model trained from scratch without and with data augmentation.



Confusion matrix with the results for the model trained without data augmentation.

Model	Acc.	Recall	Prec.	F-score	Time (s)
our CNN	**0.900**	**0.900**	**0.907**	**0.900**	77.09
our CNN+aug	0.870	0.870	0.885	0.869	77.85

Experimental Results: Transfer Learning



Confusion matrices and training curves for the pre-trained models.

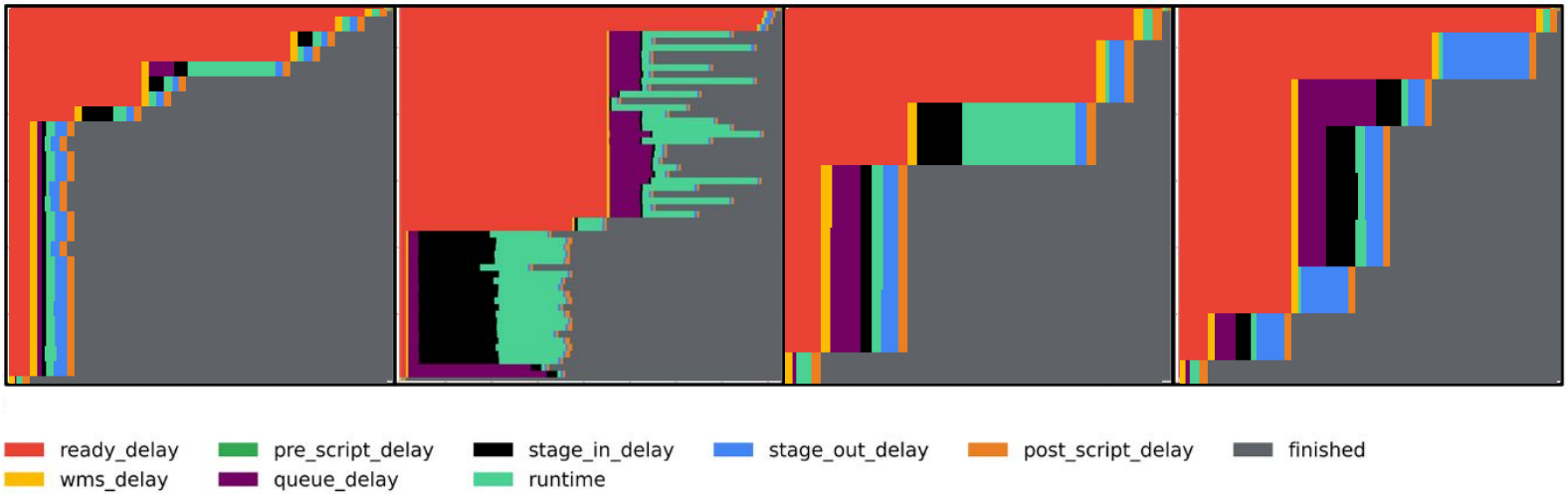
Experimental Results: Impact of Data Augmentation

Model	Acc.	Recall	Prec.	F-score	Time (s)
Without Data Augmentation					
AlexNet	**0.910**	**0.910**	**0.918**	**0.910**	222.75
VGG-16	**0.910**	**0.910**	**0.916**	**0.910**	283.42
ResNet-18	0.880	0.880	0.881	0.879	320.43
With Data Augmentation					
AlexNet	0.910	0.910	0.918	0.910	268.83
VGG-16	**0.930**	**0.930**	**0.916**	**0.930**	288.41
ResNet-18	0.890	0.890	0.900	0.890	325.52

Summary of the results for the pre-trained models trained without and with data augmentation.

Limitations and Future Work

GOAL: Collect diverse dataset of workflows' executions that allows for training of a robust CNN model.



Thank you