A green AI model selection strategy for computer-aided mpox detection.

EZENKWU, C.P., STEPHEN, B.U.-A., AFFIAH, I. and DANIEL, B.

2023

A Green AI Model Selection Strategy for Computer-Aided Mpox Detection

Chinedu Pascal Ezenkwu[1], Bliss Utibe-Abasi Stephen[2], Iniabasi Affiah[2], Betabasi Daniel[2]
School of Creative and Cultural Business, Robert Gordon University, UK[1]
Department of Computer Engineering, University of Uyo, Nigeria[2]
chineduezenkwu@ieee.org,blissstephen@uniuyo.edu.ng, iniabasi.bernard@gmail.com, betabasi.daniel@gmail.com

***Abstract*—With the recent global surge in Mpox (formerly Monkeypox) cases, researchers have proposed deep learning technologies for early detection of the disease from skin lesion images. However, many of these researchers follow the current Red AI trend of seeking to improve the performance accuracies of classifiers with no consideration given to the efficiency and environmental-friendliness of their models. This paper proposes a Green AI model selection strategy based on a multi-criteria decision technique, incorporating computational time in identifying the optimal model for final deployment. We have experimented with end-to-end ResNet50, VGG19, and InceptionV3 networks and the transfer-learning of their pre-trained versions with SVMs. Using our proposed Green AI strategy, we have identified the optimal models based on efficiency and performance. The results have been assessed using expert-level validation. We demonstrate that our proposed method can select the best model. The outcomes of our model selection strategy are similar to experts' choices of the optimal model when presented with both model error and computation time. This paper's contributions are significant as they support the ongoing call for Green AI, especially within the healthcare sector.**

***Index Terms*—Mpox, machine learning, Green AI, monkeypox, deep learning, cost-aware AI, model selection**

I. Introduction

Health informatics has proven helpful in managing diseases, especially when health resources are overwhelmed or strict surveillance is needed. For example, Surveillance Outbreak Response Management and Analysis System (SORMAS), a mobile digital system, was deployed with positive outcomes during the 2017 to 2019 Mpox outbreak in Nigeria [1]. SORMAS addressed challenges ranging from information delay and difficulties with updating and verifying case data to integrating laboratory tests and managing contact tracing, except diagnosis. Artificial Intelligence (AI), especially machine and deep learning technologies, have been applied in different aspects of public health and epidemiology, including monitoring and control of virus spread [2], early disease detection [3], [4], medical imaging [5], and drug discovery [6], [7]. AI can help overcome some of the known challenges of medical approaches to epidemic control. For example, using a Mpox detection mobile app can help democratise virus detection and reporting processes and eliminate patient-to-caregiver transmission and the need for only specialised trained personnel to control the spread of the virus.

In a similar trend to COVID-19 [8], an increasing number of research papers have recommended the use of deep learning, especially convolutional neural networks (CNN), in Mpox detection from images [9]. While deep learning models have been successful in this task, there is an enormous environmental concern associated with large machine and deep learning models, which the authors of these papers have not factored into their evaluation of AI models – i.e., the computational cost and CO_2 emissions. For example, training a deep learning model known as the transformer has been estimated to produce more CO_2 than a car would emit in its lifetime [10]. Similarly, Schwartz et al report that between 2012 and 2017, the computations needed for deep learning research increased by 300,000X and doubled every two months, resulting in significant and challenging contributions to the carbon footprint [11]. Despite the preceding concerns, over 90% of ACL papers, 80% of NeurIPS papers, and 75% of CVPR papers target improvements in accuracy [11]. Schwartz et al refer to the current practice of paying attention to accuracy alone within the AI community as Red AI [11].

A review of the current literature reveals that most published research papers involving AI applications to Mpox epidemiology follow the Red AI approach, making the techniques inefficient and environmentally unsafe. To illustrate, Sahin et al compared the accuracies of convolutional neural works (CNN) such as ResNet18, GoogleNet, EfficientNetb0, NasnetMobile, ShuffleNet, and MobileNetv2 in skin lesion image classifications into Mpox and the non-Mpox images [12]. Ali et al developed an online tool powered by an ensemble of three deep CNN models to improve Mpox image detection accuracy; these models are VGG16, ResNet50, and InceptionV3 [13]. Similar to Ali et al, Sitaula and Shahi used an ensemble of Xception and DenseNet-169 to improve performance [14]. We argue that increasing performance by expanding a model's complexity contradicts the required commitments to Net-Zero emissions by 2050 [1].

Although Green and Cost-aware AI fields are relatively new, a few researchers have committed to pursuing efficient AI models in healthcare applications. To increase the efficiency of an AI model, researchers have optimised the number of features used in training and predictions by eliminating irrelevant ones. Erion et al follow a model-agnostic feature selection approach using a cost-aware framework for developing AI models for healthcare applications [15]. The feature

[1]https://www.un.org/en/climatechange/net-zero-coalition

selection methodology incorporates features' predictive power and expert annotations of feature cost. Other authors have also employed this approach of incorporating feature annotation costs in the feature selection process [16], [17]. Outside healthcare, Kim et al have proposed a cost-efficient diagnostic framework for fault detection and isolation (FDI) for solenoid pumps, employing extraction and selection of discriminative features [18]. This paper seeks to contribute to the ongoing calls to transition from Red AI to Green AI through a model selection strategy using Mpox detection as a case study.

The significant contribution of this paper is a novel Green AI strategy for model comparison and selection using a strategy based on a multi-criteria decision-making model. Although the proposed model has the potential to incorporate as many variables that contribute to the inefficiency of machine learning models with performance in our model inefficiency equation, we have only experimented with computational time. Other variables that the equation can be extended to include the number of hyperparameters, the number of epochs, the amount of CO_2 emission, and data size. This paper provides insights that can drive the development of more efficient AI models for applications within and outside medicine. To validate our Green AI strategy, we experimented with six different learning algorithms, involving deep learning and support vector machines with a pre-trained feature extractor, on the Mpox detection task, demonstrating that the proposed strategy can identify an optimal model based on the experimenter's choice of a parameter(s) that conveys the desirability of each decision variable.

The remainder of this paper is as follows: Section II presents relevant background information on Green AI. The research methodology has been provided in Section III, while the experiments, including the data descriptions and results, are illustrated in Section IV. Section V concludes the paper.

II. Green AI

Despite the successes recorded by computationally intensive deep learning models, generally referred to as Red AI approaches by Schwartz et al, these models are both environmentally unfriendly and expensive. For example, GPT-3 has been ranked the best in NLP [19], with 175 billion parameters and 96 total layers. Similarly, EfficientNet-L2 ranks amongst the best-performing image classifier, with 480 million parameters trained on 130 million images [20]. These requirements for many parameters and a large amount of data to solve a task using deep learning necessitate using GPUs and TPUs to expedite model training. These trends generate concerns about the cost of hardware and power and the carbon footprint caused by using energy-intensive hardware [10], [21]. Schwartz et al proposed the Red AI equation presented in equation 1, capturing different variables contributing to the cost of an AI (R)esult. The cost grows linearly with the cost of processing a single (E)xample, the size of the training (D)ataset and the number of (H)yperparameters [11].

$$Cost(R) \propto E.D.H \tag{1}$$

Green AI aims to develop efficient AI technologies to reduce the contributions of AI research to the carbon footprint and make this research more inclusive through practical AI algorithms that are trainable with minimal hardware and data requirements [11]. One way to achieve Green AI results is by optimising one or more variables in the Red AI equation. Schwartz et al recommend including measures of efficiency such as CO_2 emissions, electricity usage, elapsed real time, and the number of parameters associated with a model in AI research papers to drive research towards Green AI. Research on topics such as cost-aware AI pursues a similar goal of achieving efficient AI models [18], [22].

III. Methodology

In response to calls to decarbonise the energy system and a large amount of CO_2 accompanying digital transformations, this research proposes a Green AI strategy for model selection in machine learning, using Mpox detection as a case study. In this paper, we investigate the performance and efficiency benefits of employing the transfer learning of a pre-trained convolutional neural network for feature extraction and support vector machine (SVM) for the final classification of Mpox, using our proposed cost-aware strategy to identify the optimal model. The pre-trained CNN models considered in this paper include ResNet50, VGG19, and InceptionV3. To evaluate the performance of the transfer learning strategies with SVM, the CNN models (i.e., ResNet50, VGG19, and InceptionV3) were implemented and applied end-to-end on the same task. Our cost-aware model selection strategy proposed in this paper incorporates each model's performance and computational cost for model selection.

Figure 1 presents the pipeline for the model's training, evaluation and selection. The first stage of the workflow is the pre-processing stage. Augmentation techniques such as rotation, translation, reflection, shear, hue, saturation, contrast and brightness jitter, noise, and scaling, were applied to the images to increase the data's size, improve the model's robustness, and minimise overfitting. Either in the end-to-end CNN approach or convolutional SVM, the data is split into training and test sets at the ratio of 80:20, respectively. The model's training is executed through 10-fold cross-validation and grid search for model selection and hyperparameter optimisation. During the train-test split and the k-fold cross-validation, stratified sampling is employed to maintain the original distribution among the different splits.

We propose the model inefficiency equation, which we use as a Green AI strategy for selecting the final model for deployment as shown in equations 2 and 3. The model inefficiency equation is based on the weighted product model. It incorporates both model error and the computational time for each model to help select a model based on performance and efficiency for final deployment. This concept can be generalised to other factors contributing to the cost of machine learning models, including the number of hyperparameters, epochs, and CO_2 emission, and so on.

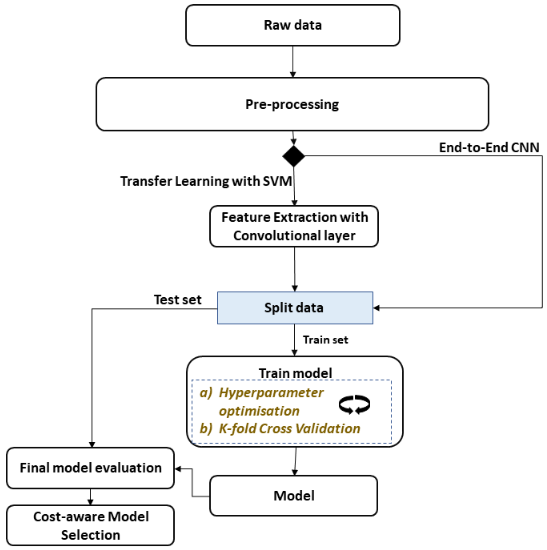

Fig. 1. Workflow for model training and evaluation

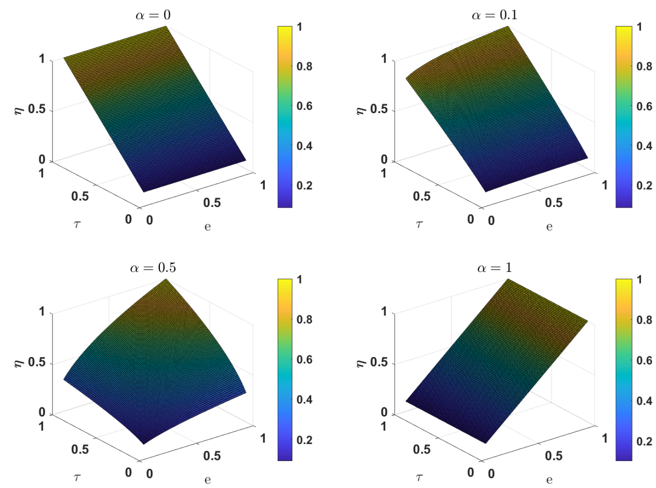

Fig. 2. Surface plots for the model inefficiency equation

$$\eta^{<m>} = \left(\frac{e^{<m>}}{||\{e^{<m>}\}_{m=1}^{M}||_{\infty}}\right)^{\alpha} \cdot \left(\frac{\tau^{<m>}}{||\{\tau^{<m>}\}_{m=1}^{M}||_{\infty}}\right)^{(1-\alpha)} \tag{2}$$

Where $\eta^{<m>}$ is the inefficiency of model m $\forall$ m $\in$ $\{1, ..., M\}$; $e^{<m>}$ and $\tau^{<m>}$ are the error and the computational time for model m respectively. α is an inclusive number between 0 and 1, and $||.||_{\infty}$ is the infinity norm. α describes how much attention is paid to error relative to computational time. If α is 1 then the models are selected entirely based on error, and if α is 0, the computational time becomes the only factor for model selection. An α of 0.5 provides equal attention to the two variables. Figure 2 presents surface plots for the model inefficiency equation for different values of α. As described above, a decrease in α shifts the focus of the model selection process from performance error to computational time of the models.

The optimal model m^* is the model that minimises the inefficiency equation as follows:

$$m^* = argmin_m(\eta^{<m>}) \tag{3}$$

IV. Experiment

A. Data Description

The Mpox Skin Lesion Dataset (MSLD) is used in this experiment. MSLD was prepared by Ali et al. through the collation of images from online case reports, news portals, and websites [13]. After eliminating images with poor resolution and quality, the authors compiled 228 images, including 102 cases of Mpox and 126 of other skin conditions like measles and chickenpox. Each image has the following dimension $224 \times 224 \times 3$. Following the augmentation process, the 228 images were expanded to 3192, consisting of 1428 images of Mpox and 1764 other skin conditions. We treated the task as a binary problem detecting whether a skin lesion image was Mpox.

B. Results and Discussion

All experiments used a Windows operating system with Intel Core i5 @ 2.30GHz with 8GB RAM. The Google Colaboratory tool was used to train the CNN models. Google Colaboratory provides access to a single 12GB NVIDIA Tesla K80 GPU for training our models. As well as using Scikit-learn for the SVM implementation, our model development involves using Keras, an easy-to-use open-source library with a high-level interface to TensorFlow for pre-trained models and end-to-end implementation of the different deep learning architectures.

Table I summarises the hyperparameters derived for the experiments in this paper. The images were cropped to 150×150 pixels. Each end-to-end CNN implementation utilised the Adam optimiser, a learning rate of 0.1, and a batch size of 32, using a dropout probability of 0.2 to prevent overfitting. While the binary cross-entropy loss function was used in training the end-to-end CNN models, the sum of the squared hinge function and the L2 penalty function is optimised for the SVM component of the transfer learning approaches. As shown in Table I, the regularisation term C varied with the convolutional layer used with SVM.

Table II illustrates the performance of the models in terms of accuracy, precision, recall, and F1 score. From the results, the end-to-end InceptionV3 model and its SVM variant performed better than other models on most metrics, with the former yielding the overall best result. The end-to-end ResNet50 performed the worst on this task, scoring less than its SVM variant on all metrics except recall. VGG19 and VGG19+SVM performed relatively better than ResNet50 and ResNet50+SVM, with VGG19 producing the overall best

TABLE I
HYPERPARAMETERS FOR THE LEARNING ALGORITHMS

	ResNet50+SVM	**VGG19 + SVM**	**InceptionV3 + SVM**	**ResNet50, VGG19 , InceptionV3**
Optimizer	-	-	-	Adam
Image size	150×150	150×150	150×150	150×150
Learning rate	-	-	-	0.01
Batch size	-	-	-	32
Activation Function	-	-	-	ReLU
Loss	squared hinge	squared hinge	squared hinge	binary cross-entropy
Dropout	-	-	-	0.2
Epoch	-	-	-	30
Penalty	L2	L2	L2	-
Regularization, C	0.1	0.01	0.0001	-

recall. While most research in the literature follows the Red AI strategy of favouring the best-performing model, the end-to-end InceptionV3 model, we developed a novel Green AI approach to model selection.

This study utilises our proposed Green AI strategy for the selection of an optimal model for Mpox detection. We consider two factors in the model selection: model performance (% error i.e., 100 - % accuracy) and computational time. Table III reports each model's computation time (training duration). The results demonstrate that the end-to-end CNN models have higher computational time than their SVM variants, with InceptionV3 having the highest computation time and VGG19+SVM having the least by combining model error and computational time for each model using equations 2 and 3, we can select a performing but a cost-aware final model for the Mpox classification task. This contradicts the Red AI approach, which focuses on performance only, favouring InceptionV3 with the highest computation cost. [!htbp]

Figure 7 demonstrates the selection of the optimal model m^* for different values of α. We have plotted model error, computation time and the model inefficiency on the same axes to convey an understanding of the choices of m^* for different values of α. As can be seen in that figure, in an extreme situation where α = 0, the model selection is based entirely on computational time, making VGG19+SVM to become the most desirable model in that case. Similarly, when α = 1, the model selection is 100% based on error, similar to the Red AI selection approach, hence selecting InceptionV3 as the optimal model despite its undesirable computational time. To identify a model that will accommodate both objectives based on user-defined desirability, we considered α = 0.5 and α = 0.6. At α = 0.5, the model selection process pays equal attention to performance and computational time. In this case, VGG19+SVM is considered optimal. In the case of α = 0.6, the computational time is slightly traded off for accuracy, resulting in InceptionV3+SVM as the optimal model. From our results, the best model can be decided between VGG19+SVM and InceptionV3+SVM depending on the specifics of the use case. InceptionV3, identified as the optimal model using the Red AI model selection approach, is not favoured by our Green AI model selection strategy.

Using outcomes from expert-level evaluations conducted among 56 experts, we carried out a statistical hypothesis test to determine if the introduction of computation time significantly influences the model selection by experts. After eliminating responses that indicate apparent misunderstanding following the reasons provided by the expert, a statistical hypothesis test was conducted using the remaining N = 49 responses. The experts were asked to choose from the six models using error as the only criterion. Afterwards, the experts made their choices using both error and computation time. In the first case, all 49 responses favoured InceptionV3. On introducing computational time as a second criterion, all the experts' choices changed from InceptionV3 to either VGG19+SVM or InceptionV3, with 31 choosing VGG19+SVM and 18 choosing InceptionV3+SVM. The result aligns with the choice we can make using the model selection strategy we propose in this paper. We used the marginal homogeneity test to establish the statistical significance of this change in proportion from InceptionV3 to either VGG19+SVM or InceptionV3+SVM. The hypotheses are defined as follows:

H0: There is no significant difference between the model choices experts make due to error alone and the choices they make due to a combination of error and computation time.

H1: There is a significant difference between the model choices experts make due to error alone and the choices they make due to a combination of error and computation time.

From the survey sample, the observed Chi-square statistic, $\mathcal{X}^2$(df=1, N=49) = 98, p-value $<$ 0.00001. Since p-value$<$ 0.05 or $\mathcal{X}^2 >$ the critical value, $\mathcal{X}^{2*}$ = 3.841, for a significance level and degrees of freedom, df of 0.05 and 1 respectively, we reject H_0 with 95% confidence, indicating that the presentation of computation time strongly influenced experts' choice of the optimal model. Although manual model selection can be easy for a bivariate selection case, as presented in this paper, it will not be the case for more than two variables. We argue that the strategy presented in this paper will be more effective than relying on an expert's discretion for situations involving more than two variables. Our future work will explore the proposed strategy, factoring in the number of model parameters, electricity usage, size of data, CO_2 emissions and other important variables that can represent the cost of a machine learning

TABLE II
MODEL PERFORMANCES (MEAN ± SD)

	ResNet50+SVM	VGG19 + SVM	InceptionV3 + SVM	ResNet50	VGG19	InceptionV3
Accuracy (%)	81.05 ± 3.86	89.18 ± 1.69	92.21 ± 1.41	66.18 ± 2.50	88.11 ± 4.49	**94.19 ± 2.43**
Precision	81.06 ± 5.53	88.05 ± 1.59	90.58 ± 1.66	63.24 ± 2.03	84.18 ± 5.95	**93.51 ± 3.89**
Recall	86.87 ± 9.56	93.05 ± 2.37	95.90 ± 1.98	92.60 ± 4.93	**97.43 ± 2.62**	96.39 ± 2.02
F1	83.29 ± 4.35	90.46 ± 1.53	93.15 ± 1.25	75.07 ± 1.95	90.17 ± 3.28	**94.86 ± 2.00**

TABLE III
COMPUATION TIME OF THE MODELS (MEAN ± SD)

Model	Computation time (mins)
ResNet50 + SVM	28.13 ± 5.62
VGG19 + SVM	20.03 ± 11.07
InceptionV3 + SVM	29.72 ± 10.76
ResNet50	84.89 ± 14.62
VGG19	112.90 ± 7.94
InceptionV3	100.83 ± 9.66

model.

V. CONCLUSIONS

Research involving the application of AI in medical image analysis primarily follows the Red AI approach, focusing on improving the percentage accuracies of learning algorithms. While this approach has resulted in some success within health informatics, it contradicts the commitments of many governments and organisations towards decarbonising energy systems within digital transformations. This paper presents an approach based on Green AI in the selection of an efficient model for Mpox detection from skin lesion images. In line with the preceding, our research has produced an inefficiency equation which leverages the concept of multi-criteria decision-making in optimising a model based on performance and factors that can contribute to the efficiency of that model. Even though the proposed model can generalise to unlimited variables, we considered only performance and computation time as the decision variables for selecting the optimal model. Our expert-level validation reveals that the proposed model can realise a similar choice of model as an expert would, considering the performances and computation times of the learning algorithms. The current issue with the proposed model selection strategy is that it relies on the manual selection of α. Our future research will address this issue using a multi-objective optimisation algorithm like Pareto optimisation.

REFERENCES

[1] B. C. Silenou, D. Tom-Aba, O. Adeoye, C. C. Arinze, F. Oyiri, A. K. Suleman, A. Yinka-Ogunleye, J. Dörrbecker, C. Ihekweazu, and G. Krause, "Use of surveillance outbreak response management and analysis system for human monkeypox outbreak, nigeria, 2017–2019," *Emerging infectious diseases*, vol. 26, no. 2, p. 345, 2020.

[2] R. Vaishya, M. Javaid, I. H. Khan, and A. Haleem, "Artificial intelligence (ai) applications for covid-19 pandemic," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, no. 4, pp. 337–339, 2020.

[3] A. Colubri, M.-A. Hartley, M. Siakor, V. Wolfman, A. Felix, T. Sesay, J. G. Shaffer, R. F. Garry, D. S. Grant, A. C. Levine *et al.*, "Machine-learning prognostic models from the 2014–16 ebola outbreak: data-harmonization challenges, validation strategies, and mhealth applications," *EClinicalMedicine*, vol. 11, pp. 54–64, 2019.

[4] T. Shaik, X. Tao, N. Higgins, L. Li, R. Gururajan, X. Zhou, and U. R. Acharya, "Remote patient monitoring using artificial intelligence: Current state, applications, and challenges," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, p. e1485, 2023.

[5] N. Benameur, R. Mahmoudi, S. Zaid, Y. Arous, B. Hmida, and M. H. Bedoui, "Sars-cov-2 diagnosis using medical imaging techniques and artificial intelligence: A review," *Clinical Imaging*, vol. 76, pp. 6–14, 2021.

[6] D. Paul, G. Sanap, S. Shenoy, D. Kalyane, K. Kalia, and R. K. Tekade, "Artificial intelligence in drug discovery and development," *Drug discovery today*, vol. 26, no. 1, p. 80, 2021.

[7] C. Cerchia and A. Lavecchia, "New avenues in artificial-intelligence-assisted drug discovery," *Drug Discovery Today*, p. 103516, 2023.

[8] Z. A. A. Alyasseri, M. A. Al-Betar, I. A. Doush, M. A. Awadallah, A. K. Abasi, S. N. Makhadmeh, O. A. Alomari, K. H. Abdulkareem, A. Adam, R. Damasevicius *et al.*, "Review on covid-19 diagnosis models based on machine learning and deep learning approaches," *Expert systems*, vol. 39, no. 3, p. e12759, 2022.

[9] K. K. Mohbey, G. Meena, S. Kumar, and K. Lokesh, "A cnn-lstm-based hybrid deep learning approach to detect sentiment polarities on monkeypox tweets," *arXiv preprint arXiv:2208.12019*, 2022.

[10] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in nlp," *arXiv preprint arXiv:1906.02243*, 2019.

[11] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green ai," *Communications of the ACM*, vol. 63, no. 12, pp. 54–63, 2020.

[12] V. H. Sahin, I. Oztel, and G. Yolcu Oztel, "Human monkeypox classification from skin lesion images with deep pre-trained network using mobile application," *Journal of Medical Systems*, vol. 46, no. 11, pp. 1–10, 2022.

[13] S. N. Ali, M. Ahmed, J. Paul, T. Jahan, S. Sani, N. Noor, T. Hasan *et al.*, "Monkeypox skin lesion detection using deep learning models: A feasibility study," *arXiv preprint arXiv:2207.03342*, 2022.

[14] C. Sitaula and T. B. Shahi, "Monkeypox virus detection using pre-trained deep learning-based approaches," *Journal of Medical Systems*, vol. 46, no. 11, pp. 1–9, 2022.

[15] G. Erion, J. D. Janizek, C. Hudelson, R. B. Utarnachitt, A. M. McCoy, M. R. Sayre, N. J. White, and S.-I. Lee, "A cost-aware framework for the development of ai models for healthcare applications," *Nature Biomedical Engineering*, pp. 1–15, 2022.

[16] Q. Wei, Y. Chen, M. Salimi, J. C. Denny, Q. Mei, T. A. Lasko, Q. Chen, S. Wu, A. Franklin, T. Cohen *et al.*, "Cost-aware active learning for named entity recognition in clinical text," *Journal of the American Medical Informatics Association*, vol. 26, no. 11, pp. 1314–1322, 2019.

[17] S. Das, R. Iyer, and S. Natarajan, "A clustering-based selection framework for cost-aware and test-time feature elicitation," in *8th ACM IKDD CODS and 26th COMAD*, 2021, pp. 20–28.

[18] S. Kim, U. E. Akpudo, and J.-W. Hur, "A cost-aware dnn-based fdi technology for solenoid pumps," *Electronics*, vol. 10, no. 19, p. 2323, 2021.

[19] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[20] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 687–10 698.

[21] L. F. W. Anthony, B. Kanding, and R. Selvan, "Carbontracker: Tracking and predicting the carbon footprint of training deep learning models," *arXiv preprint arXiv:2007.03051*, 2020.

[22] G. Erion, J. D. Janizek, C. Hudelson, R. B. Utarnachitt, A. M. McCoy, M. R. Sayre, N. J. White, and S.-I. Lee, "Coai: Cost-aware artificial intelligence for health care," *medRxiv*, 2021.

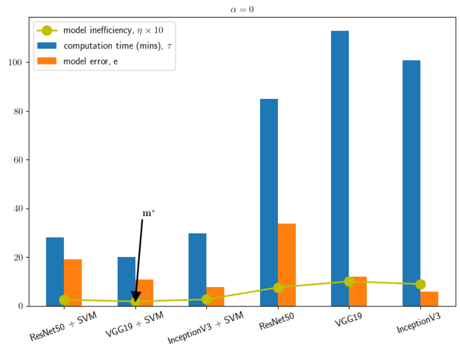

Fig. 3. $\alpha = 0$

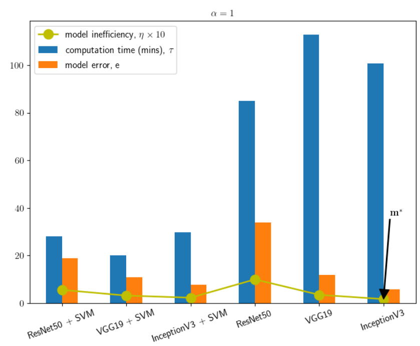

Fig. 4. $\alpha = 1$

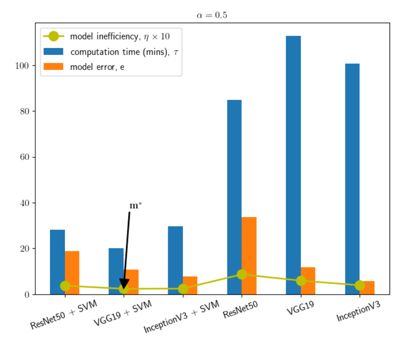

Fig. 5. $\alpha = 0.5$

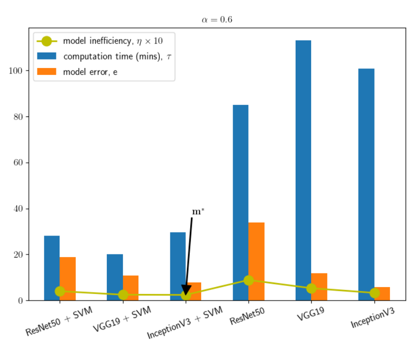

Fig. 6. $\alpha = 0.6$

Fig. 7. Model selection using the model inefficiency equation for different values of α