Running Out of Gains: Understanding the Relation between Dataset Size and Model Performance for Detecting Race Bib Numbers using YOLO

Finn Alberts (8526855751)
Research Methods for Artificial Intelligence

June 23, 2026

Abstract

This study investigates the relationship between training dataset size and the performance of YOLO-based race bib detection models. Models were trained on the RBNR and TGCRBNW datasets using varying amounts of training data and evaluated using precision, recall, mAP50, and mAP50-95. Linear, logarithmic, and power-law regression models were fitted to characterize the relationship between dataset size and performance.

The results show a consistent positive relationship between training data size and model performance across both datasets and all evaluation metrics. However, no regression model consistently outperformed the others, and no statistically significant differences were found after Bonferroni correction. These findings suggest that increasing dataset size generally improves race bib detection performance, but do not provide conclusive evidence for a specific functional relationship between dataset size and performance.

1 Introduction

During running events, organizers often have photographers present to capture runners and the atmosphere. In larger events, which may involve hundreds of participants, these photographs are often published online, for instance on the event website. As a result, participants may find it difficult to locate images of themselves among the large volume of available photos. Nonetheless, runners can in principle be identified through their race bib, which is typically pinned to the chest (Wong et al., 2019).

Computer vision, more specifically the You Only Look Once (YOLO) model[1], could allow for the automatic detection of these race bibs and, in combination with optical character recognition (OCR), therefore allow easier retrieval of photographs of a specific participant (Wong et al., 2019; Castrillón-Santana et al., 2024). It is important to note that OCR alone is not sufficient as images may contain other text than race bib numbers, such as text on

[1] https://docs.ultralytics.com/

T-shirts or license plates of cars parked on the side of the parcours. Therefore, detection is required before OCR. In this paper, we will be focusing on the detection stage of the pipeline.

In practice, race bib detection has several real-world challenges complicating accurate detection. First, occlusion is a major challenge as often race bibs are partially hidden behind other runners or by the runner's own arms. Moreover, motion blur is often present in the images due to the nature of running which reduces detection accuracy. Furthermore, the position of the camera relative to the runner causes different perspectives which further complicates detection (Wong et al., 2019; Castrillón-Santana et al., 2024).

Although prior research into race bib detection with YOLO exists (which we will discuss in Section 2), little attention has been given to the relation between the amount of training data and model performance in this specific context. As building datasets is a time-intensive task, this study will investigate this relationship and aims to provide practical insights for the required amount of data when building a race bib detection tool.

2 Related work

2.1 Race bib detection and recognition

Research on race bib number recognition dates back as early as 2012, with Ben-ami et al. (2012) framing it as a combination of text detection and recognition with some unique challenges due to the nature of running (e.g. occlusion, motion blur). In their paper, they proposed one of the earliest approaches in solving this problem by using domain priors, such as first detecting the face, then torso and using this to find the race bib.

This approach, however, was limited in that it made strong assumptions about the position and visibility of the race bib. With the emergence of deep learning, research shifted towards detection using Convolutional Neural Networks (CNNs) and sequence recognition (Wong et al., 2019; Castrillón-Santana et al., 2024). A commonly adopted pipeline consists of YOLO (or similar) for detecting the bib number and using a Convolutional Recurrent Neural Network (CRNN) to recognize the digits.

Although an improvement over earlier approaches, the solution was limited due to possible error propagation between detection and recognition stages. Alternative approaches such as using a YOLO-only pipeline as proposed by Rayhan and Lhaksmana (2020) allow for simpler pipelines, but may require dataset-specific tuning to achieve robust performance across different race conditions.

At its core, a key issue can be defined as the lack of a publicly available annotated dataset with a diverse set of images with varying circumstances such as lighting conditions, weather, or background noise. Two commonly used datasets are the RBNR dataset from Ben-ami et al. (2012) and the TGCRBNW dataset from Hernández-Carrascosa et al. (2021).

The RBNR dataset consists of 217 high-quality images by professional and amateur photographers. The photographs come from different races, meaning the race bibs have different fonts and sizes. As this dataset was created at the time when the approach was still using domain priors such as the face of a runner, the dataset contains images with not all race bibs being labeled (Castrillón-Santana et al., 2024).

RBNR	TGCRBNW
217 images	2530 images
290 annotated bibs	3232 annotated bibs
Various different races	Single race
Photographs	Frames extracted from five different videos
Day-time lighting conditions	Low-light conditions

Table 1: Comparison of the RBNR and TGCRBNW datasets

The second dataset, the TGCRBNW dataset (Hernández-Carrascosa et al., 2021) contains frames extracted from videos of ultra-distance trail runners and consists of 2530 images and 3232 annotated bibs extracted from five different video recordings of a single race. In contrast to RBNR, TGCRBNW includes images captured in low-light conditions. The authors of the TGCRBNW dataset claim that their dataset is more extensive and more challenging than the RBNR dataset (Hernández-Carrascosa et al., 2021; Castrillón-Santana et al., 2024). Although the TGCRBNW dataset is significantly larger than the RBNR dataset, it is still relatively small. Furthermore, as it only contains images from a single race, it limits its ability to generalize across different race settings. A comparison of the RBNR and TGCRBNW dataset can be seen in Table 1.

As evaluated by Castrillón-Santana et al. (2024), performance varies significantly between racing bib datasets. For example, performance drops markedly on more challenging in-the-wild datasets such as TGCRBNW. Furthermore, if real-time constraints appear, e.g. for live identification of runners, trade-offs arise between speed and accuracy (Martínez et al., 2024).

2.2 Relation between dataset size and deep learning model performance

Previous research has shown that performance on a wide range of computer vision tasks improves with increasing dataset size, often following a logarithmic trend (Sun et al., 2017). Using the JFT-300M dataset containing approximately 300 million images, Sun et al. (2017) demonstrated that larger training datasets can substantially improve performance across tasks such as image classification, object detection, semantic segmentation, and human pose estimation. Furthermore, they found that larger models benefit more from additional data than smaller models, highlighting the interaction between model capacity and data scale.

Gholinavaz et al. (2025) investigated the effect of dataset size on object detection performance using the DAWN dataset, which consists of 1000 annotated traffic images captured under adverse weather conditions such as fog, rain, snow, and sandstorms. By training on 25%, 50%, 75%, and 100% of the dataset, they demonstrated a clear positive relationship between dataset size and performance.

Their results further suggest a diminishing returns effect, as performance gains between 75% and 100% of the data are relatively small, aligning with the logarithmic trend observed in prior work. Additionally, they show that recall is more sensitive to dataset size than precision, indicating that smaller datasets primarily reduce the model's ability to detect all relevant objects in an image (recall) rather than increasing the number of incorrect detections

(precision) (Gholinavaz et al., 2025).

However, collecting and labeling data is time-intensive and costly, leading to a trade-off between dataset size and model performance. To address this, Moskovskaya et al. (2023) propose methods for estimating the dataset size required to achieve a target accuracy for object detection models such as Fast R-CNN and YOLOv8. Their approach constructs learning curves by training models on datasets of varying sizes and fitting Neural Scaling Laws to the observed performance. This enables extrapolation of the relationship between dataset size and model performance, allowing the minimum dataset size required for a desired accuracy to be estimated. Their study, conducted across 168 datasets, further highlights the substantial impact of dataset size on object detection performance.

3 Research questions

In this research, we will bridge the gap between the current knowledge about race bib detection and the current knowledge about dataset size and model performance. More specifically, this research aims to answer the following questions:

1. What is the relationship between model performance and dataset size in the context of race bib detection using YOLO?
2. Is the relationship consistent for mAP50, mAP50-95, precision, and recall?
3. Is the relationship consistent across different datasets?

4 Research methodology

In this study, several YOLO26m models will be trained. YOLO26 is the most recent YOLO models, with its medium variant providing a good balance between performance and inference speed[2]. All training will be conducted with `image size` = 640 (default for YOLO), `batch size` = 16, and `50 epochs`.

Training will be done using both the RBNR dataset as well as the TGCRBNW dataset separately. As discussed in Section 2, the TGCRBNW dataset is expected to be more challenging than the RBNR dataset, which is also illustrated in Figure 1. Additionally, the RBNR dataset contains images with incomplete annotations. Given its relatively small size (217 images), missing race bibs will be manually annotated in this study using Label Studio[3].

For the RBNR dataset, a 5-fold cross-validation procedure is employed. For the TGCRBNW dataset, the five recording perspectives present in the dataset are treated as predefined folds to prevent temporal and viewpoint-related data leakage. Since frames originating from the same recording perspective are highly similar, assigning entire perspectives to individual folds ensures that visually near-identical samples cannot appear in both training and test data. A consequence of this approach is that the TGCRBNW folds are not equally sized. However, preventing data leakage is considered more important than creating equal sized folds. The

[2]https://docs.ultralytics.com/models/yolo26
[3]https://labelstud.io/

(a) RBNR dataset

(b) TGCRBNW dataset

Figure 1: Examples of both datasets

RBNR and TGCRBNW datasets are analyzed independently and are never combined during training or evaluation.

For each fold, the data is divided into 70% training, 15% validation, and 15% test data. Subsequently, the YOLO model is trained using nested subsets comprising 10%, 25%, 50%, 70%, 85%, 90%, and 100% of the training partition, such that each larger subset contains all samples from the smaller subsets. This results in 2 `datasets` $\times$ 5 `folds` $\times$ 7 `training percentages` = 70 `experiments`. For each experiment, mAP50, mAP50-95, precision, and recall are recorded. Mean performance across folds is reported and visualized.

To investigate the relationship between training dataset size and model performance, three regression models are considered: (1) a linear model of the form $y = ax + b$, (2) a logarithmic model of the form $y = a\log(x) + b$, and (3) a power-law model of the form $y = ax^b$, where $x =$ `percentage of training data` and $y =$ `performance measure`. For visualization purposes, regression curves are fitted to the mean performance values across folds for each dataset and performance metric. RMSE and R^2 scores for these curves are reported.

For statistical analysis, the three regression models are fitted separately for each fold. For each dataset, fold, performance metric, and regression model, the Root Mean Squared Error (RMSE) is computed. To compare the regression models, paired t-tests are performed on the fold-level RMSE values for each pair of regression models at 95% confidence ($\alpha = 0.05$). Since multiple pairwise comparisons are conducted, Bonferroni correction is applied. In addition to statistical significance, t-values and p-values, Cohen's d is reported as a measure of effect size.

The consistency of the observed relationship across performance metrics and datasets is assessed by comparing the fitted regression curves, R^2 values, and RMSE scores. Consistency is interpreted based on whether similar performance trends are observed and whether the same regression models provide the best fit across metrics and datasets.

5 Results

5.1 Performance across train percentages

Table 2 presents the mean and standard deviation of precision, recall, mAP50, and mAP50-95 across the five folds for each training data percentage and dataset. Overall, performance generally tends to improve with increasing training data, although the magnitude of this effect differs between datasets and metrics. The results of the 70 individual experiments can be found in Appendix A. The code for running all experiments and subsequent visualizations and analysis can be found at `https://github.com/FinnAlberts/Running-out-of-gains`.

Dataset	**Training data percentage**	**Precision**	**Recall**	**mAP50**	**mAP50-95**
RBNR	10	0.947±0.054	0.829±0.061	0.918±0.035	0.581±0.039
RBNR	25	0.916±0.065	0.928±0.057	0.948±0.027	0.608±0.050
RBNR	50	0.937±0.012	0.929±0.053	0.960±0.027	0.642±0.056
RBNR	70	0.949±0.040	0.957±0.019	0.969±0.013	0.654±0.031
RBNR	85	0.951±0.018	0.948±0.043	0.972±0.020	0.677±0.048
RBNR	90	0.962±0.040	0.970±0.022	0.981±0.011	0.663±0.052
RBNR	100	0.968±0.023	0.943±0.029	0.972±0.018	0.674±0.048
TGCRBNW	10	0.857±0.042	0.797±0.073	0.847±0.048	0.341±0.057
TGCRBNW	25	0.878±0.047	0.786±0.065	0.851±0.076	0.376±0.065
TGCRBNW	50	0.874±0.030	0.796±0.062	0.857±0.055	0.381±0.080
TGCRBNW	70	0.865±0.041	0.805±0.092	0.854±0.078	0.359±0.075
TGCRBNW	85	0.879±0.068	0.818±0.081	0.865±0.096	0.391±0.115
TGCRBNW	90	0.865±0.048	0.823±0.071	0.867±0.066	0.360±0.080
TGCRBNW	100	0.869±0.078	0.843±0.107	0.871±0.096	0.384±0.078

Table 2: Summary of the evaluation metrics across different datasets and training percentages. The mean and standard deviation of precision, recall, mAP50, and mAP50-95 are reported for each combination of dataset and training percentage.

5.2 Relationship analysis

To investigate the relationship between training data size and model performance, linear, logarithmic, and power-law regression models were fitted for each performance metric and dataset. The fitted curves for the RBNR and TGCRBNW datasets are shown in Figures 2 and 3, respectively. The logarithmic and power-law models produce highly similar fits and therefore largely overlap in the figures, causing the logarithmic curve to be partially obscured.

An average RMSE and R^2 is computed for all fitted models and can be found in Table 3. No clear best-performing regression model emerged across all datasets and performance metrics. In many cases, the differences in R^2 and RMSE between the fitted models were

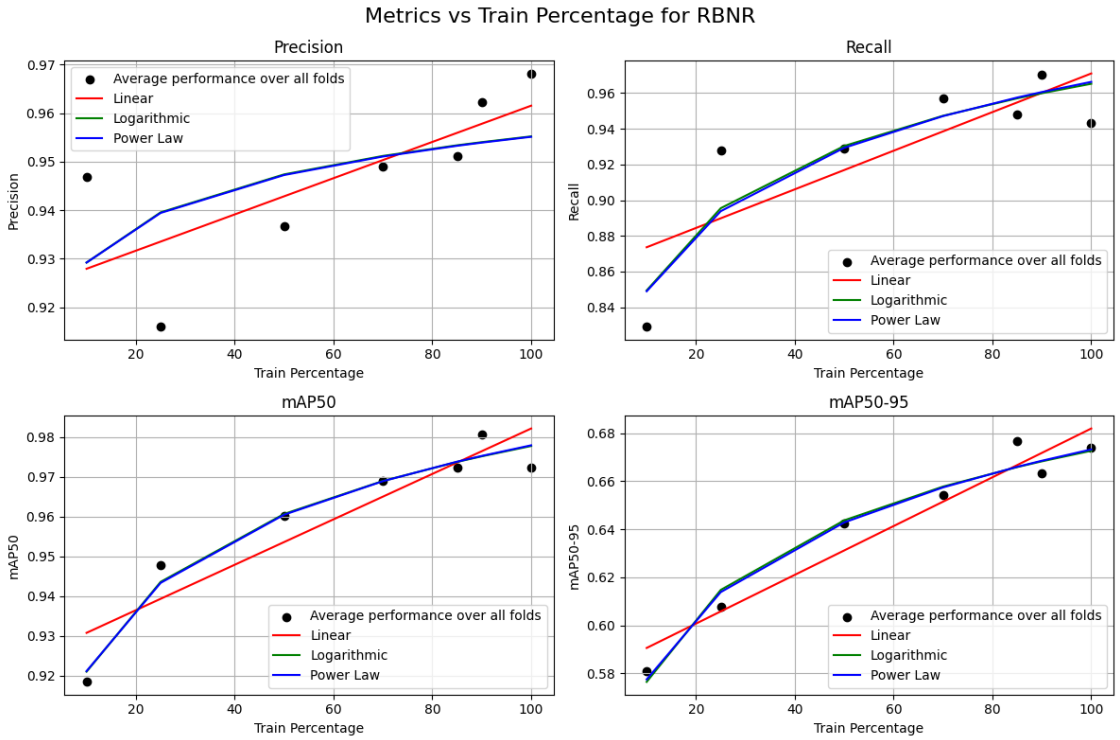

Figure 2: Trends for the RBNR dataset. Note that the logarithmic and power law fit are almost fully overlapping.

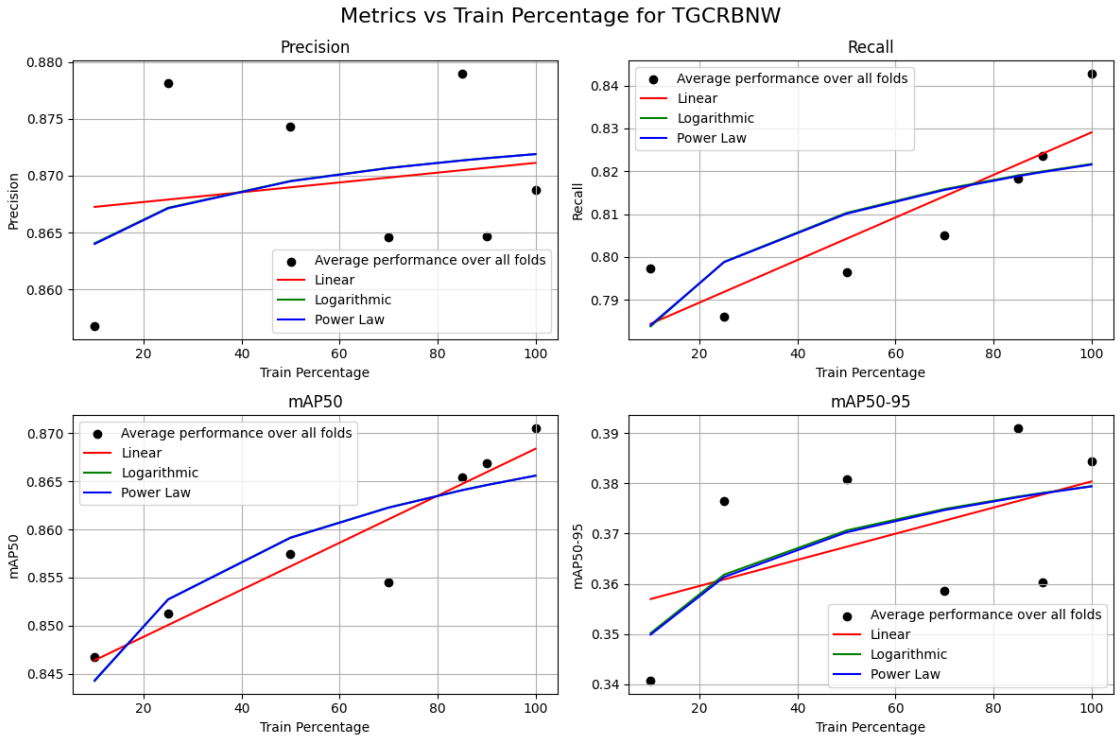

Figure 3: Trends for the TGCRBNW dataset. Note that the logarithmic and power law fit are almost fully overlapping.

small. This was particularly evident for the logarithmic and power-law models, which often produced nearly identical fits.

		Linear		**Logarithmic**		**Power law**	
Dataset	**Metric**	R^2	RMSE	R^2	RMSE	R^2	RMSE
RBNR	Precision	**0.5538**	**0.0106**	0.3086	0.0132	0.3117	0.0132
RBNR	Recall	0.6303	0.0263	**0.8301**	**0.0178**	0.8196	0.0183
RBNR	mAP50	0.8522	0.0075	**0.9677**	**0.0035**	0.9663	0.0036
RBNR	mAP50-95	0.9398	0.0081	0.9706	0.0057	**0.9738**	**0.0054**
TGCRBNW	Precision	0.0327	0.0074	**0.1247**	**0.0071**	0.1241	**0.0071**
TGCRBNW	Recall	**0.7607**	**0.0088**	0.5102	0.0126	0.5157	0.0125
TGCRBNW	mAP50	**0.8897**	**0.0027**	0.7806	0.0038	0.7823	0.0038
TGCRBNW	mAP50-95	0.2510	0.0143	**0.3651**	**0.0131**	0.3602	0.0132

Table 3: Regression fit statistics for the three candidate models. Best-performing models according to R^2 and RMSE are highlighted in bold.

5.3 Statistical testing

To evaluate whether differences in regression model fits are statistically significant, paired t-tests were performed. Unlike the analyses presented in Figures 2 and 3 and Table 3, which are based on mean performance across folds, the statistical analysis was conducted at the fold level. More specifically, each regression model was fitted separately for each fold, yielding five RMSE values per combination of regression model and performance metric for pairwise comparison.

Since three pairwise comparisons are performed for each dataset–metric combination, Bonferroni correction was applied by multiplying the obtained p-values by three. The results are reported in Appendix B. Consistent with the small differences in goodness-of-fit observed in Table 3, no statistically significant differences were found. All Bonferroni-corrected p-values exceeded the significance threshold of $\alpha = 0.05$.

Nevertheless, several comparisons yielded moderate to large effect sizes. For recall specifically, effect sizes for comparisons between the linear model and the logarithmic and power-law models ranged from -1.08 to 1.34 across datasets. However, none of these comparisons remained statistically significant after Bonferroni correction.

6 Discussion and conclusion

The goal of this research was to investigate the relationship between model performance and dataset size for race bib detection using YOLO, and to assess whether this relationship is consistent across datasets and performance metrics. To this end, YOLO models were trained and evaluated on two datasets using four performance metrics.

The results show that increasing the amount of training data generally leads to improved performance. However, no clear evidence was found in favor of a linear, logarithmic, or

power-law relationship. Although several comparisons yielded large effect sizes (Cohen's d), these differences did not reach statistical significance after Bonferroni correction.

One possible explanation for the absence of statistically significant differences is the limited number of observations available for fitting the regression models. Since only seven training percentages were evaluated, the fitted curves were often highly similar, resulting in only minor differences in goodness-of-fit.

With respect to the performance metrics, the positive relationship between dataset size and performance was observed consistently across precision, recall, mAP50, and mAP50-95. However, the regression model providing the best fit varied between metrics, preventing the identification of a single consistently superior model.

A similar pattern was observed when comparing the RBNR and TGCRBNW datasets. While both datasets showed an overall positive relationship between training data size and performance, no regression model consistently outperformed the others across datasets. Future research using a larger number of training percentages may provide a clearer picture of the underlying relationship and whether a particular functional form consistently offers the best fit.

Additionally, future research could include a broader range of race bib datasets to investigate whether the observed trends generalize across different race conditions, image qualities, lighting conditions, and levels of occlusion. The present study was limited to two available datasets, which may not fully represent the diversity encountered in real-world running events.

Furthermore, repeating the analysis with a larger number of folds or independent train-test splits could increase statistical power and provide more robust comparisons between regression models.

Finally, future studies could investigate whether the observed relationship depends on the object detection architecture by comparing YOLO with alternative detectors. Such research would help determine whether the positive relationship between dataset size and performance observed in this study generalizes beyond race bib detection and YOLO specifically.

References

Idan Ben-ami, Tali Basha, and Shai Avidan. Racing bib numbers recognition. In *Proceedings of the British Machine Vision Conference*, pages 19.1–19.10, 2012.

Modesto Castrillón-Santana, David Freire-Obregón, Daniel Hernández-Sosa, Oliverio Santana, Francisco Ortega-Zamorano, José Isern-González, and Javier Lorenzo-Navarro. An evaluation of general-purpose optical character recognizers and digit detectors for race bib number recognition. In *Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods - ICPRAM*, pages 910–917. INSTICC, SciTePress, 2024.

Sana Gholinavaz, Nima Saeedi, and Sina Samadi Gharehveran. Robustness analysis of yolo and faster r-cnn for object detection in realistic weather scenarios with noise augmentation. *Scientific Reports*, 15(1):44888, 2025.

Pablo Hernández-Carrascosa, Adrian Penate-Sanchez, Javier Lorenzo-Navarro, David Freire-Obregón, and Modesto Castrillón-Santana. Tgcrbnw: A dataset for runner bib number detection (and recognition) in the wild. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9445–9451, 2021.

Rafael Martínez, Álvaro Llorente, Alberto del Rio, Javier Serrano, and David Jimenez. Performance evaluation of yolov8-based bib number detection in media streaming race. *IEEE Transactions on Broadcasting*, 70(3):1126–1138, 2024.

Elizaveta Moskovskaya, Olesya Chebotareva, Valeria Efimova, and Sergey Muravyov. Predicting dataset size for neural network fine-tuning with a given quality in object detection task. *Procedia Computer Science*, 229:158–167, 2023. 12th International Young Scientists Conference in Computational Science, YSC2023.

Muhammad Aditya Rayhan and Kemas Muslim Lhaksmana. Racing bib number recognition method using deep learning. *Jurnal Media Informika Budidarma*, 4(3):815–824, 2020.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era, 2017. URL `https://arxiv.org/abs/1707.02968`.

Yan Chiew Wong, Li Jia Choi, Ranjit Singh Sarban Singh, Haoyu Zhang, et al. Deep learning-based racing bib number detection and recognition. *Jordanian Journal of Computers and Information Technology*, 5(3), 2019.

A Raw experiment results

Training data percentage	Fold idx	Precision	Recall	mAP50	mAP50-95
100	1	0.941	0.925	0.947	0.631
100	2	0.973	0.950	0.990	0.674
100	3	0.953	0.978	0.989	0.687
100	4	0.975	0.904	0.971	0.631
100	5	1.000	0.961	0.965	0.748
10	1	0.926	0.755	0.899	0.548
10	2	0.873	0.861	0.891	0.573
10	3	1.000	0.795	0.950	0.613
10	4	0.934	0.821	0.890	0.542
10	5	1.000	0.914	0.962	0.629
25	1	0.864	0.943	0.962	0.626
25	2	0.878	0.950	0.948	0.656
25	3	0.866	0.956	0.962	0.628
25	4	0.972	0.827	0.901	0.524
25	5	1.000	0.963	0.965	0.605
50	1	0.920	0.906	0.950	0.622
50	2	0.949	0.975	0.989	0.664
50	3	0.933	0.956	0.980	0.655
50	4	0.932	0.846	0.920	0.560
50	5	0.948	0.963	0.962	0.711
70	1	0.898	0.943	0.967	0.647
70	2	0.928	0.975	0.967	0.642
70	3	0.978	0.975	0.991	0.665
70	4	0.942	0.932	0.955	0.617
70	5	1.000	0.960	0.965	0.701
85	1	0.954	0.906	0.969	0.642
85	2	0.970	1.000	0.993	0.674
85	3	0.956	0.970	0.992	0.681
85	4	0.921	0.901	0.947	0.632
85	5	0.954	0.963	0.962	0.755
90	1	0.963	0.978	0.982	0.655
90	2	0.998	1.000	0.995	0.666
90	3	0.897	0.971	0.981	0.680
90	4	0.961	0.939	0.980	0.585
90	5	0.993	0.963	0.965	0.730

Table 4: Raw experiment results for the RBNR dataset experiments

Training data percentage	Fold idx	Precision	Recall	mAP50	mAP50-95
100	1	0.881	0.796	0.868	0.394
100	2	0.955	0.921	0.968	0.485
100	3	0.745	0.678	0.712	0.271
100	4	0.858	0.935	0.907	0.362
100	5	0.904	0.884	0.896	0.410
10	1	0.825	0.710	0.797	0.346
10	2	0.908	0.877	0.903	0.428
10	3	0.802	0.770	0.807	0.274
10	4	0.873	0.871	0.890	0.314
10	5	0.875	0.758	0.835	0.341
25	1	0.829	0.707	0.798	0.352
25	2	0.919	0.851	0.931	0.468
25	3	0.825	0.728	0.747	0.301
25	4	0.917	0.804	0.904	0.413
25	5	0.899	0.840	0.876	0.348
50	1	0.895	0.762	0.854	0.393
50	2	0.915	0.895	0.934	0.465
50	3	0.847	0.728	0.782	0.270
50	4	0.869	0.799	0.877	0.442
50	5	0.846	0.799	0.840	0.334
70	1	0.861	0.724	0.817	0.377
70	2	0.916	0.903	0.947	0.477
70	3	0.821	0.769	0.798	0.295
70	4	0.896	0.904	0.930	0.348
70	5	0.829	0.725	0.781	0.296
85	1	0.880	0.774	0.851	0.380
85	2	0.918	0.937	0.969	0.504
85	3	0.769	0.728	0.722	0.223
85	4	0.946	0.855	0.935	0.494
85	5	0.881	0.798	0.850	0.354
90	1	0.857	0.807	0.858	0.387
90	2	0.902	0.873	0.934	0.474
90	3	0.817	0.711	0.778	0.259
90	4	0.925	0.894	0.929	0.360
90	5	0.822	0.832	0.836	0.322

Table 5: Raw experiment results for the TGCRBNW dataset experiments

B Results of statistical tests

Dataset	**Metric**	**Comparison**	**t**	**Bonf.** p	**Cohen's** d
RBNR	Precision	Linear vs Logarithmic	-1.15	0.938	-0.52
RBNR	Precision	Linear vs Power Law	-1.11	0.988	-0.50
RBNR	Precision	Logarithmic vs Power Law	1.22	0.871	0.54
RBNR	Recall	Linear vs Logarithmic	2.34	0.238	1.05
RBNR	Recall	Linear vs Power Law	2.41	0.221	1.08
RBNR	Recall	Logarithmic vs Power Law	-1.71	0.487	-0.76
RBNR	mAP50	Linear vs Logarithmic	0.89	1.000	0.40
RBNR	mAP50	Linear vs Power Law	0.92	1.000	0.41
RBNR	mAP50	Logarithmic vs Power Law	-0.33	1.000	-0.15
RBNR	mAP50-95	Linear vs Logarithmic	0.04	1.000	0.02
RBNR	mAP50-95	Linear vs Power Law	0.12	1.000	0.05
RBNR	mAP50-95	Logarithmic vs Power Law	0.66	1.000	0.29
TGCRBNW	Precision	Linear vs Logarithmic	-0.18	1.000	-0.08
TGCRBNW	Precision	Linear vs Power Law	-0.20	1.000	-0.09
TGCRBNW	Precision	Logarithmic vs Power Law	-1.79	0.443	-0.80
TGCRBNW	Recall	Linear vs Logarithmic	-2.68	0.166	-1.20
TGCRBNW	Recall	Linear vs Power Law	-2.71	0.160	-1.21
TGCRBNW	Recall	Logarithmic vs Power Law	1.48	0.638	0.66
TGCRBNW	mAP50	Linear vs Logarithmic	-1.44	0.670	-0.64
TGCRBNW	mAP50	Linear vs Power Law	-1.43	0.677	-0.64
TGCRBNW	mAP50	Logarithmic vs Power Law	0.95	1.000	0.42
TGCRBNW	mAP50-95	Linear vs Logarithmic	1.07	1.000	0.48
TGCRBNW	mAP50-95	Linear vs Power Law	0.95	1.000	0.42
TGCRBNW	mAP50-95	Logarithmic vs Power Law	-1.91	0.386	-0.85

Table 6: Results of the paired t-tests comparing fold-level RMSE values between regression models. Bonferroni-corrected p-values are reported. No statistically significant differences were observed after correction for multiple comparisons.