CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

The influence of data annotation process requirements on performance criteria of ML models

Bachelor of Science Thesis in Software Engineering and Management

Maab Mohammedali

Muntasir Adam

Department of Computer Science and Engineering
UNIVERSITY OF GOTHENBURG
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2023

Demonstrating the influence of the data annotation process on the final performance of machine learning models.

Supervisor: Hans-Martin Heyn
Examiner: Richard Berntsson Svensson

University of Gothenburg
Chalmers University of Technology
Department of Computer Science and Engineering
SE-412 96 Göteborg
Sweden
Telephone + 46 (0)31-772 1000

Department of Computer Science and Engineering
University of Gothenburg
Chalmers University of Technology
Gothenburg, Sweden 2023

The influence of data annotation process requirements on performance criteria of ML models

1st Maab Mohammedali
dept. of Computer Science and Engineering
University of Gothenburg
Gothenburg, Sweden
gusmaamo@student.gu.se

2nd Muntasir Adam
dept. of Computer Science and Engineering
University of Gothenburg
Gothenburg, Sweden
gusadammu@student.gu.se

***Abstract*—The data annotation process is a critical step in the development of machine learning (ML) models, as it entails labeling data to help supervised learning. This study investigates the impact of data annotation process requirements on the performance of ML models. Employing an experimental approach, the study compares the performance of ML models using different annotated datasets and various process requirements. Performance metrics, including average precision, precision, recall, and F1 score, are used to compare the outcomes.**

The study reveals that the requirements imposed on the data annotation process have a substantial influence on the performance criteria of ML models. These findings shed light on the crucial role that the data labeling process plays in the creation of ML models, providing valuable insights for both academic researchers and industry professionals. keyword-Data annotation, performance metrics, precision, recall, true positive, False negative, false positive, Intersection over Union, mean average precision, experiment methods.

I. Introduction

A. Background and Topic

Artificial intelligence is "concerned with the development of computers able to engage in human-like thought processes such as learning, reasoning, and self-correction" [1]. A distinguishing feature of the fourth industrial revolution is the use and expansion of AI (Artificial Intelligence) in various applications [2] such as Natural language processing (NLP), optical character recognition (OCR), image and video recognition. One of the main reasons for this expansion is the availability of massive amount of data that can be used to train AI models [3], However the correctness and accuracy of this data is crucial to the performance of the resulting machine learning model.

Supervised machine learning requires a large amount of annotated data [5]. The process of adding tags or annotations to the data to enhance the significance is referred to as data annotation. The purpose of the data annotation is labelling the sensor data with meaningful classes. Supervised learning and modeling techniques become available with labeled data [6].

For data annotation open-source datasets were used such as ImageNet and Microsoft Common Objects in Context (COCO) [7]. However, these datasets have a limited number of object classes, which may not be adequate or appropriate for specific tasks [5]. Requirements on the data annotation process can have a significant impact on the performance criteria of ML models.

In order to ensure high-quality annotated data, various standard methods are used as part of the machine learning development cycle. Some examples of these practices are:

Manual annotation humans manually tag or label data to ensure its accuracy and relevance of annotations. Although (to a certain extent) manual annotation is more accurate, it is very labor-intensive and commonly used to train a machine to perform automatic annotation. [8]. Manual annotation is a commonly utilized technique across several fields, including natural language processing, image recognition, and speech recognition, among others. Annotators who are involved in this process may possess specialized knowledge and expertise in a particular domain or task, or they may have no prior experience in the field.

Automated annotation involves automated techniques like machine learning algorithms to annotate data. Automated annotation may not be as accurate, but it has the capability to process a significantly larger number of documents than what humans can handle. [8]. Automated annotation is a widely used technique in various machine learning applications such as natural language processing and image recognition. For example, image recognition algorithms can be trained to automatically classify objects within images. In natural language processing, automated annotation is used for labeling text data with pertinent tags or to identify named entities.

Crowdsourcing involves the use of many non-expert annotators to annotate data, enabling cost-effective data annotation. However, the quality of annotations may vary based on the expertise and skill of the annotators. Crowdsourcing does present several difficulties, though, including controlling participant incentives and motives as well as assuring the quality and consistency of contributions. Crowdsourcing has been employed successfully in several industries, including business, healthcare, and scientific research, despite these obstacles. For instance, platforms like CrowdFlower and Amazon Mechanical Turk have grown in popularity for crowdsourcing work in industries like picture recognition and data annotation.

The identification of requirements for the annotated data is crucial in the data annotation process, and these requirements may include data types and formats, annotation types, annotation guidelines, times and cost constraints.

Time restrictions consider the primary requirements in data annotation. It is crucial to set a strict deadline for the annotation process to guarantee timely completion of annotations. This minimizes potential development delays by ensuring that machine learning models are trained on the most recent data. Time limitations often restrict the duration annotators can dedicate to each activity within annotation processes. Nonetheless, the quality of annotations may deteriorate when annotators operate under predetermined time budgets, such as a maximum of 20 seconds per image. This is due to the reduced time available to focus on the intricate details of each frame, potentially resulting in lower-quality annotations [18].

Precision requirements are yet another crucial necessity. In terms of precision, we mean how accurate the annotations are. The quality of the annotated data will depend on how precisely the annotations must be made; hence this need must be clearly stated. The clarity of the annotation standards and the annotators' comprehension of the necessary degree of precision should also be prioritized. Additionally, a crucial requirement for data annotation is consistency checks. The annotations must be made with the utmost care, making sure they follow the set annotation standards and are uniform among all annotators. Implementing a review process that examines the consistency of the annotations and resolves any differences is one approach to accomplish this.

Finding the best balance between the time and cost required for data annotation and the performance requirements of ML models are crucial because data annotation may be both time and cost consuming. The accuracy and quality of the annotations, and consequently, the performance of the resulting ML models, can be affected by a variety of variables, including the level of information in the annotations, the consistency of labeling, and the number of annotators engaged.

B. Review of current knowledge and related work

The importance of the ML (Machine Learning) models is becoming more and more a common element of software systems. Therefore, as one of many aspects of integrating ML models in software systems, several studies aim to explore the consequences of data annotation process requirements on the performance of machine learning models. These studies can help understanding the requirements that are used in data annotation process, because the annotation process is an essential element of the machine learning development cycle and should be treated clearly defined through process requirements.

The effect of data annotation process requirements on the performance criteria of machine learning models is receiving increased attention in the industry. Recent years have seen a significant focus on scrutinizing dataset annotation practices [9]. Research has been conducted to investigate the relationship between data annotation quality and ML model performance. Studies indicate that high-quality and accurate data annotations can greatly enhance the performance of ML models, particularly in tasks like image classification and object detection [5].

Alhazmi et al. [5] examined how the quality of data annotations influenced the performance of machine learning models. The researchers evaluated the performance of various computer vision models, such as object recognition and picture classification models, using image datasets with differing levels of annotator quality. The researcher employed three different datasets to assess how the models' performance was influenced by the quality of annotations. The annotations were of high quality in the first dataset and become medium quality in the second dataset, and the third dataset had annotations of low quality. The models were trained and evaluated on each dataset, and the outcomes were compared.

The observed results show a clear relationship between the quality of annotations and the performance of the models. The model's accuracy shows an increase when accompanied by high-quality annotations and low error rates .The experience of the annotators, the chosen annotation method, and the specific annotation tool used are considered main factors that affect the quality of the annotations. Furthermore, the study indicates that the effect of annotation quality on model performance changes according to the difficulty of the task. The effect of annotation quality was more significant for easier activities like image categorization and less significant for harder tasks like object detection.

The research emphasizes the need for stringent quality control procedures in the annotation process and emphasizes the significance of high-quality annotations for machine learning models. The authors propose that to enhance the overall effectiveness of machine learning models, future study should concentrate on developing techniques for automatically evaluating and enhancing annotation quality.

Nazari et al. [11] show how class noise affects the performance of machine learning algorithms. Class noise refers to instances or data points that are misclassified into a wrong class. The performance of various machine learning algorithms was assessed through experiments on several datasets with differing degrees of class noise. Many factors control the class noise impact such as model complexity, training dataset size and the amount of noise.

Additionally, it has been found that some machine learning techniques—including decision trees, random forests, and support vector machines—are more impervious to class noise than others. The study highlights the importance of reducing class noise in training data if you want machine learning algorithms to work more effectively. The researcher suggests that future research should focus on developing more resilient algorithms that can manage class noise and improve noise reduction techniques in order to raise the precision of machine learning models.

Taran et al. [13] highlights the significance of achieving accurate and efficient semantic image segmentation in traffic situations by underscoring the requirement for high-quality ground truth annotations. The research evaluated for semantic image segmentation three deep learning models on a publicly available traffic dataset, using annotations of varying quality. The results indicate that the ground truth annotations' quality affects the performance of segmentation models. Low-quality annotations resulting in decreased accuracy and greater computational expense. The research highlights the significance of achieving accurate and efficient semantic image segmentation in traffic situations by underscoring the requirement for high-quality ground truth annotations. The findings stress the need for more research to improve the quality of ground truth annotations in traffic datasets. This is critical for professionals and academics involved in computer vision and traffic analysis.

Hu et al. [14] suggest a unique method for formulating specifications for artificial intelligence systems that learn perceptual tasks from human performance. The authors contend that due to the complexity and ambiguity of machine-learned perception systems, conventional methods for defining requirements for perception systems, such as input-output connections and performance measurements, are insufficient.

The researcher suggested, instead, utilizing human performance as a standard for assessing how well machine-learned perception systems work. They contend that this strategy is more understandable and intuitive than more conventional ones. The authors establish requirements for the system, such as false positive rates and detection rates, using data on human performance. The paper presents a novel way for developing requirements for AI (Artificial Intelligence) perception systems based on human performance. The technique has the potential to improve the accuracy and dependability of machine-learned perception systems and can be used to a wide range of applications across a number of areas.

According to Hauptmann et al. [15] although video retrieval systems have made notable advancements in recent years, there are still several difficulties that machines alone cannot resolve. Therefore, the researchers suggested a method to enhance video retrieval by combining human and computer performance. Instances where human perception and comprehension of videos outperform machines include recognizing intricate events, identifying emotions, or extracting semantic significance. The authors suggested a collaborative optimization structure that maximizes both human and computer performance in video retrieval to tackle these constraints. The approach involves presenting videos to human annotators, who provide feedback on the relevance and quality of the results returned by the computer system. This feedback is used later to improve the performance of the computer system in subsequent searches.

The researchers also introduce several techniques to enhance the interaction between human and computer performance, such as selecting diverse subsets of videos for annotation, incorporating user preferences and feedback into the retrieval process, and adapting the system to the specific needs of different users. The suggested method is tested on several benchmark datasets, and the findings demonstrate that it beats both systems that purely rely on either human or computer performance as well as state-of-the-art video retrieval systems. Overall, the research highlights the need of combining human and artificial intelligence to improve video retrieval results and provides a solid framework for future research in this area.

Hao et al. [16] discuss the impact of label inaccuracies on the classification performance of weakly-supervised models. It highlights the importance of addressing label noise and inaccuracies, as they can significantly degrade the model's ability to generalize and make accurate predictions. The researcher aims to automatically identify and correct inaccurate labels to improve the performance of classification models.

The paper addresses the challenge of inaccurate labels in weakly-supervised deep learning scenarios. In this type of dataset, where labels are partially or noisily annotated, the paper introduces strategies to estimate the accuracy of each label. This estimation is achieved by comparing the predictability of a model trained with enhanced labels to that of a model trained with the original labels. The experimental results demonstrate that the automatic identification and correction of inaccurate labels can significantly improve the classification performance of weakly supervised deep learning models. The corrected models achieve higher accuracy and F1-score compared to models trained without label correction.

C. Gap in knowledge

The existing literature emphasizes the importance of considering data annotation process requirements and their impact on the performance of machine learning (ML) models. Studies have explored various aspects such as the quality of annotations, dataset noise, class noise, and the combination of human and computer performance, shedding light on the relationship between these factors and ML model performance. However, there is still a gap in knowledge regarding the optimal annotation requirements that can lead to improved ML model performance.

Conducting a study on the influence of data annotation process requirements on performance criteria of ML models will reduce uncertainty in the role of process requirements for annotations as well as boost the ability of software engineer to produce with model with high performance and to decide on process requirement for the annotation cost vs. precision. The software engineer will provide detailed information regarding data annotations process requirements. When a user does not know the requirements, they can look at this study to gain as insight into how the data annotations process requirements affect the performance of ML model. Our related work explored the impact of annotation errors and dataset noise on the performance of machine learning models [5], [11].

They have not considered the significance of annotation process requirements in causing these errors, which results in a gap in knowledge. Although they briefly mentioned that human errors, intentional or unintentional, are the root causes of such errors, the impact of annotation process requirements on model performance has not been thoroughly examined.

Currently, our understanding of the relationship between data annotation process requirements and the performance criteria of ML models is not clearly defined, which is why it is necessary to investigate the precise influence of data annotation process requirements on the efficiency of ML(Machine Learning) models.

This thesis covers the selected annotation process and aims to help create a model with high performance.

D. *Statement of the problem*

The problem is that the performance criteria of ML models can be significantly impacted by the requirements on the data annotation process. To learn and produce precise predictions, ML models need a lot of labeled data, but the accuracy and consistency of the annotations can have a a significant impact on how well the models perform. Performance of ML models, typically measured as precision, average precision, average IoU, recall and F1 score can be impacted by different annotation requirements, such as the level of detail, labeling consistency, and labeling quality. Therefore, it is crucial to carefully consider the annotation requirements and their potential impact on the performance criteria of ML models before conducting the annotation process.

E. *Purpose of the study*

The purpose of this study is to investigate the influence of selected process requirements on typical performance metrics of ML models. The reason this study is required is that although ML model is used in a lot in many applications as mentioned above there is not enough research made on how the data annotation process requirement affects the performance of the model. This research is needed because it helps to lead to better ML model performance once we determine the requirement of data annotation process. Additionally, involving students in this field of study can enhance their ability to create high-performance ML models. Moreover, given the increasing demand for ML models, this research can have wide-ranging applications, including in computer vision and speech recognition.

F. *Aim of your research and how it fits into the gap*

The aim of this study is to demonstrate the influence of the data annotation process on the final performance of machine learning models. This obviously fits the gap of understanding the role of data annotation in ML model performance and how data annotation process requirements can be designed to perfect model performance. This is a crucial area that has to be filled because reliable data labeling is a necessary step in the creation of powerful ML models.

II. Research Methodology

This study aims to determine how individual aspects of the data annotation process affect the performance standards of machine learning models. These aspects include time to annotate and the data set size. To ensure transparency and reproducibility, the entire codebase, including the scripts for creating the dataset and model implementation, is made publicly available on GitHub[1]. Researchers interested in duplicating our work or further exploring the data annotation process can refer to the repository for detailed instructions and access to the necessary resources.

A. *Research questions and/or hypotheses*

To achieve the aims of the study, we answer the following research questions.

- RQ1: How do individual annotation process requirements impact ML model performance?
- RQ2: What recommendation can be made for the integration of data annotation process requirements in the machine learning development cycle based on the experiment's result?

We also include hypotheses based on a review of the existing literature and the research questions that we formulated and answered. The following hypotheses were set:

- $\mathcal{H}_1$: It cannot be argued that the specifications for data annotation significantly affect the performance metrics of the machine learning model.
- $\mathcal{H}_0$: We cannot argue for or against that the specifications for data annotation affect the performance metrics of the machine learning model.

B. *Method used*

In this study entails an experiment which compares the performance of machine learning models using different annotated datasets under various process requirements. Scientific studies often use experimental techniques, which involve the use of controlled environments to investigate and track the effects of specific variables on a particular occurrence. The main objectives of an experimental study are to establish causal links between variables and test a hypothesis.

Overall, the use of experimental methods in this study provides a rigorous and systematic approach for evaluating the influence of data annotation process requirements on the performance criteria of machine learning models.

By controlling various variables and systematically manipulating the annotated datasets, researchers can establish causal relationships between the process requirements and the model's performance. This allows for more precise conclusions and better generalizability of the results. The chosen research strategy facilitates the examination of cause-effect relationships between variables. The experiment aims to test the hypothesis that changes to the data annotation process requirements will lead to alterations in the performance criteria of the machine learning models. Furthermore, the experiment enables researchers to control and manipulate the data annotation process requirements and assess their impact on the performance of the machine learning models.

[1] GitHub repository: https://github.com/adammuntasir/Data-Annotation-Process-and-ML-Model-Performance

In order to conduct the experiment, two real-life scenarios have been identified for emulation. Both scenarios involve a company that provides data labelling services where we assume that a fixed time and monetary budget exist.

Scenario 1: In this scenario, we assume a requirement that each frame had to be labeled within a short amount of time, resulting in each annotator having only a fixed, small amount of time to decide on a label out of a larger set of possible labels. This was expected to result in a higher level of wrong labelled data.

Scenario 2: In this scenario, we assume that the process requirement had changed such that correctness of the labels was prioritized. To achieve this, the company allowed annotators unlimited time to carefully select the correct label and employed double annotation, where two annotators independently labeled the same dataset to reduce the risk of mislabeling. This resulted in a longer time allocated to each data frame, which was expected to lead to higher accuracy. However, due to the same fixed time and money budget as before, significantly smaller amounts of labeled data were obtained.

Building upon these scenarios, our experiment consists of three setups:

Experiment 1: This experiment examined the impact of dataset noise on the machine learning models by varying the percentage of labeling error, while keeping the dataset size constant. Specifically, the labeling error was incrementally increased from 0% to 20% in fixed intervals (e.g., 5%). TABLE I provides further details regarding the amount of data noise and corresponding model name for Experiment 1.

Experiment 2: This experiment examined the impact of dataset size on the resulting machine learning models by varying the dataset size, while keeping the percentage of labeling error constant. the dataset size was incrementally decreased from 900 to 300 in fixed intervals (e.g., 100 images). Table II provides details regarding the amount of dataset size and corresponding model name for Experiment 2.

Experiment 3: This experiment examined the impact of both the dataset size and the percentage of labeling errors on the performance of the resulting machine learning models. the data set size was incrementally increased from 100 to 1000 images in fixed intervals, simultaneously the labeling error was increased from 0% to 50% in 5 fixed intervals (e.g., 5%). TABLE III provides further details regarding the amount of dataset size, the percentage of labelling error and corresponding model name for Experiment 3.

TABLE I
EXPERIMENT 1:VARY PERCENTAGE OF LABELING ERRORS WHILE KEEPING THE SAME DATASET SIZE

Model Name	data size	Label Error Percentage
E1L0	1000	0%
E1L5	1000	5%
E1L10	1000	10%
E1L15	1000	15%
E1L20	1000	20%

TABLE II
EXPERIMENT 2:VARY DATA SET SIZE WHILE KEEPING PERCENTAGE OF LABELING ERRORS

Model Name	data size	Label Error Percentage
E2D90	900	0%
E2D80	800	0%
E2D70	700	0%
E2D50	500	0%
E2D30	300	0%

TABLE III
EXPERIMENT 3:VARY BOTH OF DATA SET SIZE AND PERCENTAGE OF LABELING ERRORS

Model Name	dataset size	Label Error Percentage
E3D10L0	100	0%
E3D15L5	150	5%
E3D20L10	200	10%
E3D30L15	300	15%
E3D40L20	400	20%
E3D50L25	500	25%
E3D60L30	600	30%
E3D70L35	700	35%
E3D80L40	800	40%
E3D90L45	900	45%
E3D100L50	1000	50%

In this study, a naming convention has been adopted for the machine learning models resulting from the experiments. The convention involves assigning names in the form of $E(x)D(y)L(z)$, where (x) represents a numerical value ranging from 0 to 100. Specifically, ”Ex” denotes the experiment number, ”Dy” indicates the size of the dataset used to train the model, and ”Lz” represents the percentage of labeling error added to the dataset. This naming convention has been adopted to facilitate the readers’ understanding of the various combinations of factors that were utilized to train the machine learning models, and to aid in distinguishing between different models in a straightforward manner.

C. Terminology

To better understand the relationship between time to annotate, annotation error rate, and the performance of a machine learning model on unknown data, we developed a causal diagram Fig.1. The diagram shows that time to annotate has a direct causal effect on annotation error rate, which in turn has a causal effect on the performance of the ML model on unknown data. Time to annotate is a causal factor, while annotation error rate is a control variable, and the performance of the ML model is an effect. By examining this causal relationship, we can better understand the impact of time to annotate and annotation error rate on the performance of an ML model, and potentially identify strategies to improve model performance through perfecting the annotation process. Another probable

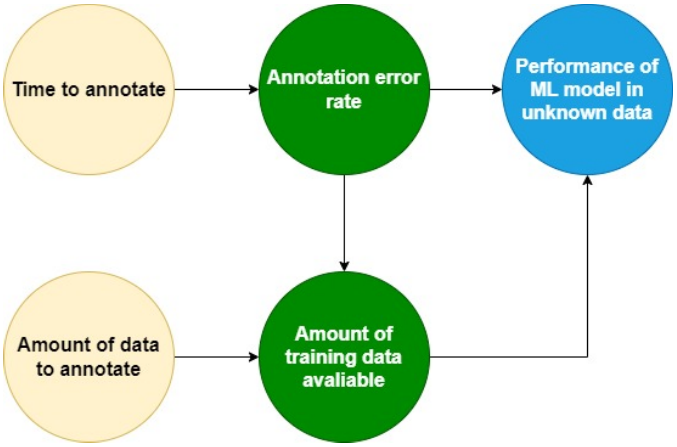
Time to annotate
Annotation error rate
Performance of ML model in unknown data
Amount of data to annotate
Amount of training data avaliable

Fig. 1. Annotating Time, Error Rate, and Model Performance Diagram

cause that we can show in the annotation process is amount of data to annotate, and the control variables could be amount of training data available and that this can have influence on the performance of the machine learning model.

1) Data collection: The process of examining the impact of data annotation process requirements on performance criteria of ML models involves passing data, which is a crucial stage. The overall performance of the resulting ML models can be significantly affected by the quality and quantity of data that is passed. To ensure that the data obtained is indicative of the real-world settings in which the ML models will be used, it is crucial to organize and carry out the data passing procedure. The data for this study is passed from the Microsoft COCO (Common Objects in Context) dataset, which consists of 328,000 images and 2.5 million labeled instances of 91 object types [7].

The COCO (Common Objects in Context) dataset passes a broad number of photos across a wide range of item categories, including animals like cats, which is one benefit of using it for data collecting. This can ensure that the data passes representative of real-world situations and can enhance the accuracy of the ML models that are produced. Additionally, The COCO dataset's pre-labeled and annotated nature passes it simpler to begin model training and evaluation, which is another benefit of using it for data collecting. The availability of an application programming interface (API) and detailed documentation further facilitates its usage and integration into the study. Furthermore, the Microsoft COCO dataset recommends the use of a third-party open-source application called FiftyOne, which makes it even easier to download the entire dataset or subsets of it and reprocess it in an efficient manner. Moreover, the application offers various tools to visualize and export the COCO dataset into different formats such as YOLO providing greater flexibility in data processing and model training.

Set of tools are used to conduct the experiments:

• Google colaboratory notebook and Jupyter Notebook

• VOXEL51 Fiftyone

• ALVIS cluster, National Academic Infrastructure for Supercomputing in Sweden (NAISS) designed for Artificial intelligence and Machine Learning research with powerful Graphical processing Units (GPUs) accelerator cards.

• Open OnDemand web portal for accessing Jupyter Notebook that connected to compute node on ALVIS cluster.

• AlexeyAB Darknet YOLOV4, object detection model AlexeyAB Darknet is an open-source neural network framework that is primarily used for object detection, image classification,and other computer vision tasks.

YOLOv4 was chosen for its high accuracy, fast processing speed, and ability to detect small objects and multiple categories [12].

2) Data analysis: To assess the impact of the process requirements of data annotations on the performance of machine learning models, a disturbance was introduced into the data annotations by intentionally randomly relabeling number of the annotations of cats to other classes. This was done to simulate the errors that annotators may make and is known as class noise [10] or labeling error [11]. This type of noise occurs for two reasons: either contradictory instances, where the same instance appears in the dataset with two class labels, or misclassification, where some instances are labeled incorrectly.

The dataset was partitioned into three subsets: training, validation, and testing, consisting of a total of 4000 images. Among these, 3000 images were allocated to the test dataset, while the remaining 1000 images were assigned to the training dataset. Notably, the training data underwent specific modifications for each model, involving either a reduction in size or the introduction of label errors. Following the resizing process, a designated portion constituting 20% of the training data was set as the validation set. The test dataset exclusively depicting cats, and it is noteworthy that this dataset is free from any labeling errors, thereby ensuring the accuracy of the ground truth annotations. The considerable size of the test dataset has been deliberately chosen to facilitate thorough evaluation of the model's performance. Moreover, this specific test dataset has been utilized consistently across all 21 models under examination. Consequently, the results obtained from this standardized testing procedure provide a reliable basis for comparative analysis, as the models have been evaluated under identical conditions, yielding clear distinctions between their respective performances.

For evaluating the resulting model, a combination of performance criteria is being used to obtain a comprehensive evaluation of the models performance. These criteria are widely used in industry for measuring machine learning performance, and include:

- **Intersection over Union (IoU)** "measure gives the similarity between the predicted region and the ground-truth region for an object present in the image, and is defined as the size of the intersection divided by the union of the two regions" [17] .

$$IoU = \frac{\text{Area of Intersection}}{\text{Area of Union}} \tag{1}$$

- **Average IoU** a metric used to evaluate the performance of object detection or segmentation models. It measures the average overlap between predicted bounding boxes or segmentation masks and the ground truth annotations.

$$\text{Average IoU} = \frac{1}{\text{Total Objects}} \sum_{i=1}^{\text{Total Objects}} \text{IoU}_i \tag{2}$$

- **Precision** represents the ratio of true positive predictions made by the model to all positive predictions. A high precision value indicates that the classifier has a low rate of false positive predictions, while a low precision value suggests that the classifier makes a considerable number of false positive predictions.
- **Average Precision (AP)** measures the average precision of correctly identified relevant items among the retrieved results.
- **Mean Average Precision (mAP)**is a metric commonly used to evaluate the performance of object detection models. It provides an overall measure of how well the model performs across multiple object classes.

$$\text{mAP} = \frac{\sum_{i=1}^{C} AP_i}{C} \tag{3}$$

Where:

C represents the total number of object classes.

AP_i denotes the Average Precision for class i.

- **Recall** is the proportion of true positive prediction out of all positive instances. High recall indicates that the model can detect most of the positive instances and low recall indicates that the model is missing many positive instances.
- **F1- score** is the harmonic mean of precision and recall.

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$F_1\text{-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

Where TP is the true positive, TN is the true negative, FP is the false positive and FN is the false Negative. [19].

III. RESULTS AND DISCUSSION

In the context of the Darknet framework, weight files are saved at regular intervals during the training process. Specifically, the weight file is saved every 1000 iterations, aligning with the training configuration. Given that the models in this study were trained on 6000 batches, a total of six weight files are saved, each corresponding to a set of 1000 iterations.

Furthermore, the Darknet framework employs a selection process to identify the weight file with the highest mean average precision. This selected weight file is considered the best weight among the saved checkpoints. Additionally, the framework also saves the weight file obtained after the 6000th iteration as the final weight.

To facilitate result comparison in this study, the evaluation is performed using the final weight on the testing dataset. By utilizing the final weight displayed in Table. IV, which represents the model's learned parameters after the completion of the training process, a fair and consistent basis for evaluating the model's performance on unseen data is established.

Results of the first experiment:

This experiment showed the impact of changing labeling error percentages on the performance of machine learning models. Performance metrics such as precision, recall, F1-score, true positives (TP), false positives (FP), false negatives (FN), mean average precision (mAP), average IoU, and average precision (AP) were utilized to demonstrate the effect of this change. All models tested in this experiment had a data size of 1000 images in total.

We started with model E1L0 (Experiment1 Label error 0) with 0% labeling errors. This model achieved a high precision of 0.92, showing a high rate of true positives, and a recall of 0.89. The F1-score, which balanced precision and recall, reached 0.91, indicating a strong overall performance. With a high mAP of 30.15% and an average IoU of 79.11%, this model proved accurate and consistent predictions across the dataset. Overall, this served as a benchmark for comparing the impact of increasing labeling errors on model performance.

The precision, recall, and F1-score decreased in model E1L5 (Experiment1 Label error 5%) when compared to the first model as shown in Fig. 3. The number of false negatives (FN) increased slightly. Despite the decrease in precision, recall, and F1-score all these led to reduce the Ap to 90.62

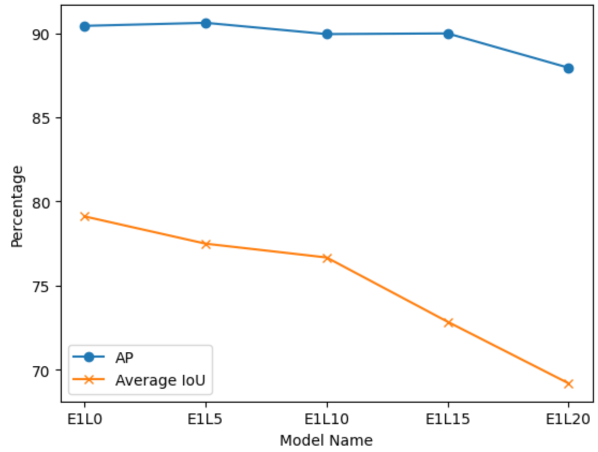

Fig. 2. Comparison of average precision and average IoU for E1.

In model E1L10 (Experiment1 Label error 10%), with a 10% labeling error, we observed further declines in precision, recall, and F1-score. This suggested that the increase in labeling errors negatively affected the model's ability to correctly classify instances. The number of FP and FN increased noticeably compared to the baseline model. The decrease in mAP, average IoU, and AP reinforced the notion that the model's performance was being affected by the labeling errors.

As we progressed to model E1L15 (Experiment1 Label error 15%), which had a 15% labeling error, we observed a more significant impact on the model's performance. The precision, recall, and F1-score showed further decreases, indicating an increased number of misclassifications. The decrease in mAP, average IoU, and AP was more noticeable, emphasizing the deteriorating performance of the model as depicted in Fig. 2.

Finally, in model E1L20 (Experiment1 Label error 20%), which experienced a 20% labeling error, we saw a substantial decrease in precision, recall, and F1-score. The model's ability to correctly classify instances was significantly compromised, as shown by the larger numbers of FN (False Negatives). The decrease in mAP, average IoU, and AP indicated a considerable decline in the model's overall performance.

In conclusion for experiment 1, the analysis of the examined machine learning models revealed a clear correlation between the presence of labeling errors and a decline in model performance as depicted in Fig. 2 and Fig. 3. As the percentage of labeling errors increased, the models' precision, recall, and F1-score decreased, indicating reduced accuracy. The number of false negatives increased, leading to a deterioration in the model's overall performance.

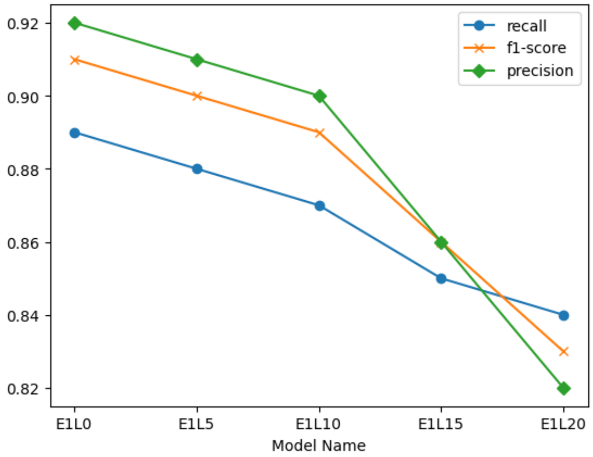

Fig. 3. Comparison of recall, f1-score and precision for E1.

Results of the second experiment:

The results obtained from examining the impact of dataset size on machine learning models provided insights into how varying the dataset size affected model performance while keeping the labeling error constant. We used the same performance metrics that we used in the first experiment to show the effects of dataset size on model performance. All models tested in this experiment had a 0 % labelling error in total.

Starting with discussing the result for model E2D90 (Experiment 2 data size 900), which consisted of a dataset size of 900 and had no labeling errors. High precision, recall, and F1-score in this model showed that it accurately classified occurrences with few false positives and false negatives. The TP value was high, and the FP and FN values were low. The mAP of 30.24% and average IoU of 79.15% indicated accurate and consistent predictions across the dataset. Additionally, the high average precision (AP) of 90.73% confirmed the high performance of the model.

Moving to model E2D80 (Experiment 2 data size 800), with a dataset size of 800, we observed similar performance to the baseline model (E2D90). The precision, recall, and F1-score remained high, indicating that the model maintained its ability to accurately classify instances. The slight decrease in TP and FN values suggested a small reduction in overall performance, but it did not significantly affect the model's accuracy. The mAP, average IoU, and AP scores also remained consistently high, indicating that the model could still provide reliable predictions despite the reduction in dataset size.

In model E2D70 (Experiment 2 data size 700), with a dataset size of 700, we started to observe more noticeable changes in model performance. The precision, recall, and F1-score remained high but showed slight decreases compared to the previous models. The decrease in TP and increase in FP and FN values indicated a higher rate of misclassification. However, the model still maintained a reasonably high level of performance. The decline in mAP, average IoU, and AP suggested a slight decrease in overall performance, reflecting the impact of the reduced dataset size.

In model E2D50 (Experiment 2 data size 500), with a dataset size of 500, we observed further decreases in precision, recall, and F1-score. The model's ability to accurately classify instances was affected as the TP value decreased and the FP and FN values increased. The decline in mAP, average IoU, and AP indicated a noticeable reduction in performance, emphasizing the impact of the reduced dataset size on the model's predictive capabilities as shown in Fig. 4.

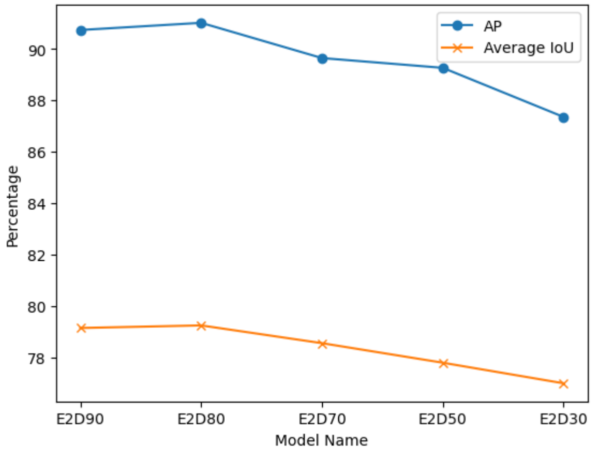

Fig. 4. Comparison of average precision and average IoU for E2.

Finally, in model E2D30 (Experiment 2 data size 300), with a dataset size of 300, we observed significant decreases in precision, recall, and F1-score. The model's ability to accurately classify instances was considerably compromised, as indicated by the lower TP value and higher FP and FN values. The decrease in mAP, average IoU, and AP underscored the substantial decline in the model's overall performance due to the significantly reduced dataset, examine the graphic displayed in Fig. 4 and Fig. 5. The results suggested that a smaller dataset size adversely affected the model's performance.

In conclusion for experiment 2, the analysis of the examined machine learning models revealed a clear correlation between the dataset size and model performance. As the dataset size decreased, the precision, recall, and F1-score of the models also decreased, indicating a deterioration in the overall performance of the models.

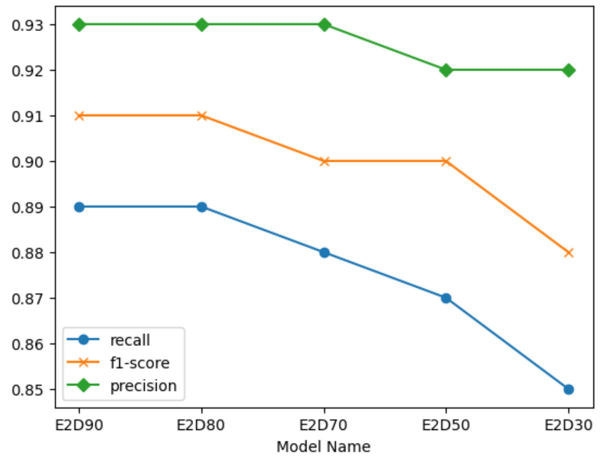

Fig. 5. Comparison of recall, f1-score and precision for E2.

Results of the third experiment:

The findings showed that as the dataset size and labeling error increased (from D10L0 to D30L15), there was an improvement in the values of mAP, AP, and IoU, indicating enhanced overall performance see Fig. 6. The best performance was observed at the E3D30L15 model, while the lowest values were attained at E3D100. These results suggest that, for relatively small dataset sizes, the impact of size supersedes the influence of labeling errors. However, after reaching the E3D30L15 model, the effect of labeling errors becomes more prominent, resulting in decreasing values even with an increased dataset size.

For instance, the E3D10L0 model, with a dataset size of 100 images and a labeling error of 0%, achieved an mAP@0.50 of 27.54%. Conversely, the E3D30L15 model, with a dataset size of 300 images, achieved an mAP@0.50 of 28.82%. These findings showed that a larger amount of accurately labeled data positively influenced the model's ability to accurately detect objects.

Similarly, as the labeling error rate increased (from L20 to L50), the performance metrics consistently decreased, signifying a reduction in performance attributable to erroneous annotations.

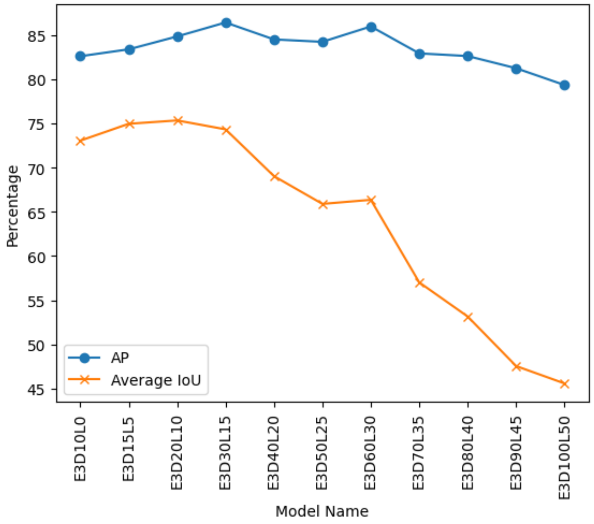

Fig. 6. Comparison of average precision and average IoU for E3.

In conclusion, the results of this experiment reflected the effect of both data size and labeling errors on model performance. Increasing the dataset size generally resulted in improved performance; however, the presence of labeling errors could diminish these gains, refer to Fig. 6 and Fig. 11. These findings underscore the importance of carefully considering both dataset size and the quality of annotations when training machine learning models.

Comparing the three experiments:

From the results, it is shown that when the data size remained constant and the label error percentage increased, the average precision and the number of true positives (TP) decreased. This effect is particularly evident when comparing the E1L0 model, characterized by a label error rate of 0%, with the E3D100L50 model, where the label error rate reaches 50% this indicated by Fig. 7. Since the dataset size remains consistent across these experiments with total of 1000 images, it allows for a valid comparison, highlighting the highest values of label errors. Specifically, the average precision declined from 90.44% to 79.37%, TP count decreased from 3098 to 2091, F1-score and recall decreased as indicated in Fig. 8. These trends were consistently observed across all models within the E1 series, as well as in the E3D100L50 model.

On the other hand, when maintaining a constant label error percentage while increasing the data size, both average precision and TP count exhibited an upward trend. This pattern becomes evident when transitioning from a dataset size of 1000 to 100 images. The average precision decreased from 90.44% to 82.63% , accompanied by a TP count rise from 2787 to 3098. Average precision,average IoU,score and recall were slightly decreased as shown in Fig. 9 and Fig. 10. This effect was consistently observed across all models in the E2 series, as well as in the E1L0 and E3D10L0 models,

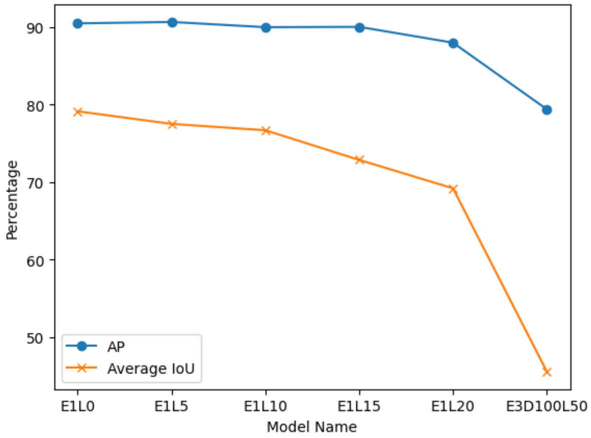

Fig. 7. Comparison of average precision and average IoU for E1 and E3D100L50.

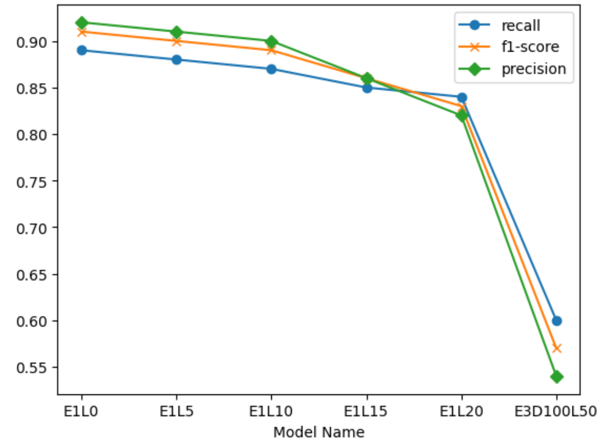

Fig. 8. Comparison of recall, f1-score and precision for E1 and E3D100L50.

which represent full-sized datasets without label errors and the smallest dataset size without label errors, respectively.

Answers to the research questions:

Based on the information compiled and explained in the previous section, along with the results obtained from the experiments, we will present our analysis and draw conclusions to address the research questions in two areas: the requirements of the individual annotation process (RQ1) and the integration of the data annotation process (RQ2).

Starting with **RQ1: How do individual annotation process requirements impact ML model performance?**

To answer this question, we will refer to the two scenario that we had:

Scenario 1: Here the requirement was to label each frame within a short amount of time. This setup was expected to result in a higher level of noise and a larger amount of data. From the results we chose some models e.g., E1L0, E1L5, E1L10, E1L15, E1L20): These models were trained with limited time for annotation. As expected, the performance metrics show a relatively average IoU compared to other

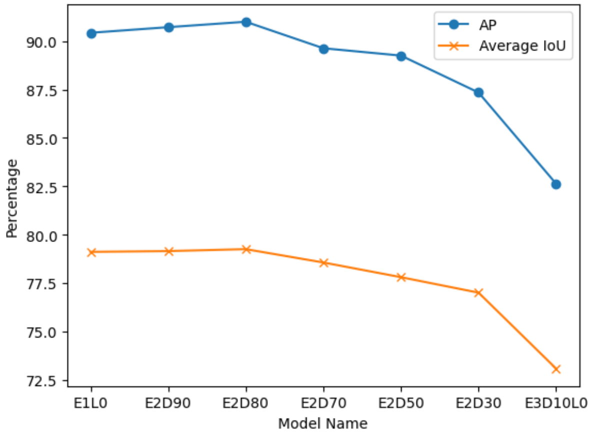

Fig. 9. Comparison of average precision and average IoU for E1L0, E2 and E3D10L0

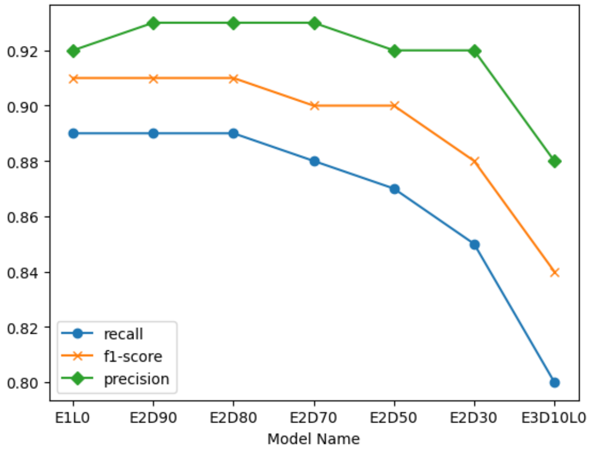

Fig. 10. Comparison of recall, f1-score and precision for E1L0, E2 and E3D10L0.

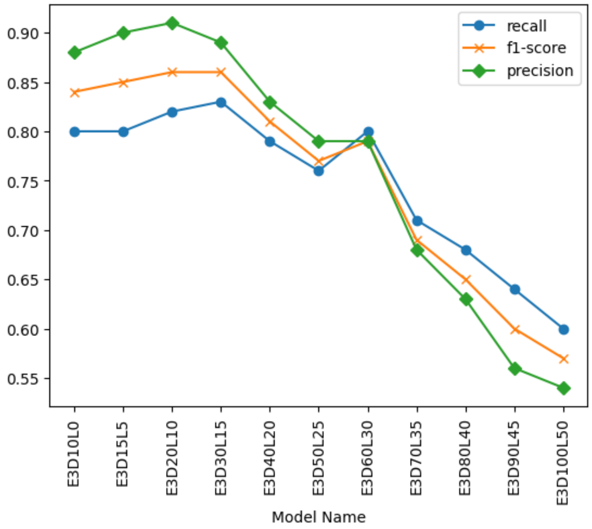

Fig. 11. Comparison of recall, f1-score and precision for E3.

TABLE IV
MODEL TESTING RESULTS(FINAL WEIGHTS)

Model Name	mAP	AP	TP	FP	FN	Precision	Recall	F1-score	Average IoU	Detections Count	Unique Truth Count	Detection Time
E1L0	30.15%	90.44%	3098	254	283	0.92	0.89	0.91	79.11%	4600	3481	56
E1L5	30.21%	90.62%	3076	232	405	0.91	0.88	0.90	77.49%	4854	3481	24
E1L10	29.98%	89.95%	3036	238	445	0.90	0.87	0.89	76.67%	5215	3481	24
E1L15	29.66%	89.99%	2967	234	514	0.86	0.85	0.86	72.85%	5961	3481	23
E1L20	29.32%	87.95%	2917	250	564	0.82	0.84	0.83	69.19%	6338	3481	55
E2D90	30.24%	90.73%	3088	239	393	0.93	0.89	0.91	79.15%	4614	3481	53
E2D80	30.34%	91.01%	3105	237	376	0.93	0.89	0.91	79.25%	4657	3481	53
E2D70	29.88%	89.64%	3071	247	410	0.93	0.88	0.90	78.56%	4469	3481	52
E2D50	29.75%	89.26%	3099	254	442	0.92	0.87	0.90	77.80%	4485	3481	23
E2D30	29.12%	87.35%	2960	265	541	0.92	0.85	0.88	77.00%	4249	3481	23
E3D10L0	27.54%	82.63%	2787	366	694	0.88	0.80	0.84	73.07%	4218	3481	25
E3D15L5	27.80%	83.41%	2785	284	696	0.90	0.80	0.85	74.98%	4265	3481	24
E3D20L10	28.29%	84.87%	2839	256	642	0.91	0.82	0.86	75.36%	4422	3481	25
E3D30L15	28.82%	86.45%	2901	250	580	0.89	0.83	0.86	74.35%	4828	3481	26
E3D40L20	28.18%	84.53%	2743	233	738	0.83	0.79	0.81	69.05%	5731	3481	26
E3D50L25	28.09%	84.26%	2658	220	823	0.79	0.76	0.77	65.91%	6072	3481	25
E3D60L30	28.67%	86.00%	2779	225	702	0.79	0.80	0.79	66.37%	6570	3481	26
E3D70L35	27.65%	82.95%	2467	159	1014	0.68	0.71	0.69	57.02%	7552	3481	24
E3D80L40	27.55%	82.65%	2353	149	1128	0.63	0.68	0.65	53.16%	7819	3481	24
E3D90L45	27.09%	81.27%	2218	168	1263	0.56	0.64	0.60	47.55%	8411	3481	24
E3D100L50	26.46%	79.37%	2091	107	1390	0.54	0.60	0.57	45.58%	8678	3481	24

scenarios.

Scenario 2: In this scenario, annotators were given unlimited time to carefully select the correct label. We selected some models (e.g., E2D90, E2D80, E2D70, E2D50, E2D30): the performance metrics showed higher average IoU compared to Scenario 1.

From the two scenarios can conclude that individual annotation process requirements have a notable impact on ML model performance. Scenario 2, with a focus on correctness of the labels, generally leads to higher performance. However, it also results in fewer labeled data samples. On the other hand, Scenario 1, with limited time for annotation, provides a larger amount of data but at the cost of lower performance.

For **RQ2: What recommendation can be made for the integration of data annotation process requirements in the machine learning development cycle based on the experiment's result?**

Based on the experiment's results, several recommendations can be made for integrating data annotation process requirements in the machine learning development cycle:

1. Implementation of consistent annotation methods: It is advisable to establish a set of standardized annotation methods that ensure consistency across the annotated data. By employing such methods, the annotation process can be expedited, thereby reducing the time required for annotation. Moreover, the utilization of consistent annotation methods can effectively mitigate the occurrence of errors made by annotators. This serves to enhance the overall quality and reliability of the annotated data.

2. Determination of best data quantity and label error rate: An important consideration in the data annotation process is the choice of a suitable combination of the dataset size and the allowable label error rate. To achieve this, it is recommended to carefully analyze and find the optimal interval of time for each frame during the annotation process. This interval should strike a balance between obtaining a sufficiently large dataset and ensuring that the label error rate remains within acceptable limits.

The study done by Nazari et al. [11] illustrates how class noise affects the performance of machine learning algorithms. In order to increase the efficiency of machine learning models, it needs to reduce class noise in training data. The paper recommends that future research concentrates on developing more resilient algorithms and improving noise reduction techniques. In relation to this work, our research findings align with Nazari et al.'s study by highlighting the importance of individual annotation process requirements on ML model performance. Our study shows that factors like annotation method and experience of annotators, labelling error strongly affect the performance of ML models, much as class noise impacts the performance of machine learning algorithms.

Similarly, Taran et al. [13] emphasize the significance of high-quality ground truth annotations for correct and efficient semantic image segmentation. They found that variations in annotation quality can lead to decreased performance and increased computational expense. Our research echoes these findings by showing that the quality of annotations, influenced by individual annotation process requirements, affects the performance of ML models.

The study used data from the Microsoft COCO (Common Objects in Context) dataset. By employing diverse datasets that employ similar metrics like precision, recall, and F1-score, it is possible to balance false positives (FPs) and false negatives (FNs). To achieve that we could use ensemble approaches, which combine the predictions of various models by training several models with various setups or techniques and combining their predictions [23].

In data annotation and model creation, the trade-off between using expert annotations, crowd-sourcing tactics, unsupervised learning, or automated annotation approaches is crucial to consider. Expert annotations by subject matter experts or skilled annotators provide high-quality annotations with precise labels, making them useful for complex datasets. However, they could be expensive, time-consuming, and resource intensive [20]. Crowdsourcing annotation tasks allow for cost efficiency, and quicker response times, but may cause an annotation's quality to vary because of the variety of annotators. For dependable findings, effective quality control procedures are crucial [21]. Without human intervention, unsupervised learning automates labeling using clustering and machine learning. It is useful for huge datasets; however, biases and errors could develop, needing adjustment for better accuracy [22]. The choice of annotation approaches depends on dataset characteristics, resources, timelines, and quality requirements. Balancing factors like cost, time and quality is crucial. The decision should consider the strengths and limitations of each approach, based on the project's specific requirements and constraints.

Future research

- Instead of using the YOLOv4 object detection model, we could employ different machine learning models such as Faster R-CNN (Region-based Convolutional Neural Network) and SSD (Single Shot MultiBox Detector) to compare their performance under various data annotation scenarios.
- Investigate different process requirements to provide a better understanding of how these requirements could affect model performance.
- Using different dataset instead of Microsoft COCO dataset for the experiments.
- Add another scenario to run more experiments which lead to a more correct result.

IV. Limitations and Validity threats

We considered three validity issues in our study namely external , internal and construct validity.

External Validity: In our study, external validity was associated with the sample we collected from the Microsoft COCO (Common Objects in Context) dataset, which included specific types of images and labeled instances. The representativeness of this dataset to other domains or applications may have varied, thereby affecting the generalizability of the findings. Furthermore, the study focused on two scenarios related to data annotation requirements within a particular company, which may not directly apply to other industries, organizations, or annotation processes with distinct characteristics.

Internal Validity: The study's objective was to enhance internal validity by investigating the influence of controlled variables, specifically dataset noise and dataset size, on the performance of the model. This was achieved through meticulously designed scenarios and experiments conducted in controlled environments. By concentrating on these particular variables, the study aimed to establish a clear relationship between the independent variables and the dependent variable, thus meeting the criteria for claiming internal validity, as outlined in the provided definition.

Construct Validity: In this study, we aimed to ensure construct validity by accurately standing for the theoretical concept of data annotation requirements and their impact on the performance of machine learning models. To achieve this, we conducted an extensive literature review and formulated research questions and hypotheses that directly aligned with our study goals. To evaluate the machine's performance, we employed industry-standard performance criteria, including intersection over Union (IoU), average IoU, precision, recall, and F1-score. By adhering to these widely recognized metrics, we made sure there was a strong match between the variables being examined and the theoretical ideas we wanted to investigate. Recognizing the crucial role of the data annotation process in advancing machine learning models, we investigated various aspects of this process and their overall influence.

V. Conclusions

The study focused on examining the influence of data annotation process requirements on the performance criteria of machine learning (ML) models. Performance metrics such as precision, average precision, average IoU, F1-score and recall that the requirements for the data annotation process significantly affected the performance criteria of machine learning models. Therefore, it is important to have a well-defined data annotation process with clear requirements to ensure the quality of the annotated data. Three experiments with different value of data size and labelling error conducted in this study. The first experiment proved the impact of label errors on ML models. It was observed that the percentage of labeling errors increased, precision, recall, and F1-score decreased, indicating a decline in performance. Moreover, an increase in false positives and false negatives had a negative effect on the overall performance of the model.

The second experiment demonstrated the impact of dataset size on ML models. It was observed that as decreasing the dataset size had noticeable effects on precision, recall, and F1-score and that led to a decrease in overall model performance.

The third experiment explored the performance of ML models with varying dataset sizes and labeling error percentages. The results indicated that larger dataset sizes generally improved object detection performance, as reflected in higher mAP values. However, the presence of labeling errors negatively affected model performance, even with a larger dataset.

The main findings of the study indicate that the performance of ML models is strongly influenced by the data annotation process requirements. By addressing this knowledge gap, it paves the way for future research and the development of high-performance models.

Future work includes comparing different machine learning models, using alternative datasets, and conducting more scenarios and experiments to enhance our understanding of data annotation's impact on model performance. These efforts will contribute to advancing the field and guiding future research.

VI. Acknowledgment

We would like to express our sincere gratitude to our academic supervisor, Hans-Martin Heyn, for his direction and assistance throughout the research process. His insightful comments and helpful suggestions greatly influenced this research.

We would also like to express our gratitude to our examiner, Beatriz Cabrero-Daniel, for her invaluable advice and feedback that greatly assisted us in crafting this thesis.

The computations, data handling and storage were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at www.naiss.se partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

We thank our academic colleagues for their insightful criticism and recommendations on this study. Their feedback has improved and strengthened our work.

Finally, we want to applaud our university for giving us the tools such as computers and previous research that we needed to carry out this study. Their help has been extremely helpful.

Thank you to everyone who helped with this thesis. Your help and direction were invaluable and made this study possible.

References

[1] Kok, Joost N., et al. ”Artificial intelligence: definition, trends, techniques, and cases.” Artificial intelligence 1 (2009): 270-299.

[2] Xu, Min, Jeanne M. David, and Suk Hi Kim. ”The fourth industrial revolution: Opportunities and challenges.” International journal of financial research 9.2 (2018): 90-95.

[3] Collins, Christopher, et al. ”Artificial intelligence in information systems research: A systematic literature review and research agenda.” International Journal of Information Management 60 (2021): 102383.

[4] Mahesh, Batta. ”Machine learning algorithms-a review.” International Journal of Science and Research (IJSR).[Internet] 9 (2020): 381-386.

[5] Alhazmi, Khaled, et al. ”Effects of annotation quality on model performance.” 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC). IEEE, 2021.

[6] Schreiner, Christopher, et al. ”Using machine learning techniques to reduce data annotation time.” Proceedings of the human factors and ergonomics society annual meeting. Vol. 50. No. 22. Sage CA: Los Angeles, CA: SAGE Publications, 2006.

[7] Lin, Tsung-Yi, et al. ”Microsoft coco: Common objects in context.” Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014.

[8] Petrillo, Matthew, and Jessica Baycroft. ”Introduction to manual annotation.” Fairview research (2010): 1-7.

[9] Paullada, Amandalynne, et al. ”Data and its (dis) contents: A survey of dataset development and use in machine learning research.” Patterns 2.11 (2021): 100336.

[10] Zhu, Xingquan, and Xindong Wu. ”Class noise vs. attribute noise: A quantitative study.” The Artificial Intelligence Review 22.3 (2004): 177.

[11] Nazari, Zahra, et al. ”Evaluation of class noise impact on performance of machine learning algorithms.” IJCSNS Int. J. Comput. Sci. Netw. Secur 18 (2018): 149.

[12] Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao. ”Yolov4: Optimal speed and accuracy of object detection.” arXiv preprint arXiv:2004.10934 (2020).

[13] Taran, Vlad, et al. ”Impact of ground truth annotation quality on performance of semantic image segmentation of traffic conditions.” Advances in Computer Science for Engineering and Education II. Springer International Publishing, 2020.

[14] Hu, Boyue Caroline, et al. ”Towards requirements specification for machine-learned perception based on human performance.” 2020 IEEE Seventh International Workshop on Artificial Intelligence for Requirements Engineering (AIRE). IEEE, 2020.

[15] Hauptmann, Alexander G., et al. ”Extreme video retrieval: joint maximization of human and computer performance.” Proceedings of the 14th ACM international conference on Multimedia. 2006.

[16] Hao, Degan, et al. ”Inaccurate labels in weakly-supervised deep learning: Automatic identification and correction and their impact on classification performance.” IEEE journal of biomedical and health informatics 24.9 (2020): 2701-2710.

[17] Nowozin, Sebastian. ”Optimal decisions from probabilistic models: the intersection-over-union case.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[18] Heyn, Hans-Martin, et al. ”Automotive Perception Software Development: An Empirical Investigation into Data, Annotation, and Ecosystem Challenges.” arXiv preprint arXiv:2303.05947 (2023).

[19] Davis, Jesse, and Mark Goadrich. ”The relationship between Precision-Recall and ROC curves.” Proceedings of the 23rd international conference on Machine learning. 2006.

[20] Snow, Rion, et al. ”Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks.” Proceedings of the 2008 conference on empirical methods in natural language processing. 2008.

[21] Su, Hao, Jia Deng, and Li Fei-Fei. ”Crowdsourcing annotations for visual object detection.” Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence. 2012.

[22] Barlow, Horace B. ”Unsupervised learning.” Neural computation 1.3 (1989): 295-311.

[23] Liu, Lijue, et al. ”Solving the class imbalance problem using ensemble algorithm: application of screening for aortic dissection.” BMC Medical Informatics and Decision Making 22.1 (2022): 1-16.

APPENDIX A
MODEL TESTING RESULTS(BEST WEIGHTS)

TABLE V
MODEL TESTING RESULTS(BEST WEIGHTS)

Model Name	**mAP**	**AP**	**TP**	**FP**	**FN**	**Precision**	**Recall**	**F1-score**	**Average IoU**	**Detections Count**	**Unique Truth Count**	**Detection Time**
E1L0	31.10%	93.31%	3155	341	326	0.90	0.91	0.90	69.19%	13356	3481	52
E1L5	31.05%	93.15%	3090	505	391	0.86	0.89	0.87	68.93%	15980	3481	26
E1L10	30.56%	91.67%	3200	702	281	0.82	0.92	0.87	62.10%	24505	3481	25
E1L15	31.08%	93.24%	3187	617	294	0.84	0.93	0.87	65.60%	18189	3481	26
E1L20	30.87%	92.62%	3076	981	405	0.76	0.88	0.82	58.20%	16306	3481	56
E2D90	31.16%	93.47%	3216	395	265	0.89	0.92	0.91	72.54%	7825	3481	53
E2D80	30.71%	91.12%	3161	306	320	0.91	0.91	0.91	76.46%	5185	3481	53
E2D70	30.99%	92.97%	3170	287	311	0.92	0.91	0.91	72.78%	8687	3481	51
E2D50	30.66%	91.97%	3171	635	310	0.83	0.91	0.87	61.20%	12911	3481	26
E2D30	30.68%	92.03%	3126	309	355	0.91	0.90	0.90	73.44%	8190	3481	26
E3D10L0	28.62%	85.87%	2903	349	578	0.89	0.83	0.86	72.73%	5744	3481	26
E3D15L5	29.32%	87.96%	2939	616	542	0.83	0.84	0.84	62.52%	11294	3481	26
E3D20L10	29.00%	87.01%	2934	406	547	0.88	0.84	0.86	71.93%	5825	3481	26
E3D30L15	28.36%	85.07%	2786	495	695	0.85	0.80	0.82	68.74%	5961	3481	25
E3D40L20	28.97%	86.91%	2836	794	645	0.78	0.81	0.80	62.33%	2836	3481	26
E3D50L25	28.69%	86.07%	2611	938	870	0.74	0.75	0.74	58.70%	8608	3481	25
E3D60L30	27.64%	82.91%	2663	1824	818	0.59	0.77	0.67	44.71%	17717	3481	25
E3D70L35	28.91%	86.74%	2406	1924	1075	0.56	0.69	0.62	43.08%	15235	3481	24
E3D80L40	29.28%	87.84%	2759	3654	722	0.43	0.79	0.56	32.33%	32440	3481	24
E3D90L45	30.41%	91.24%	2814	3545	667	0.44	0.81	0.57	34.49%	20915	3481	25
E3D100L50	30.18%	90.55%	2648	2854	833	0.48	0.76	0.59	38.07%	16786	3481	24

Appendix B
Main performance indicators by models

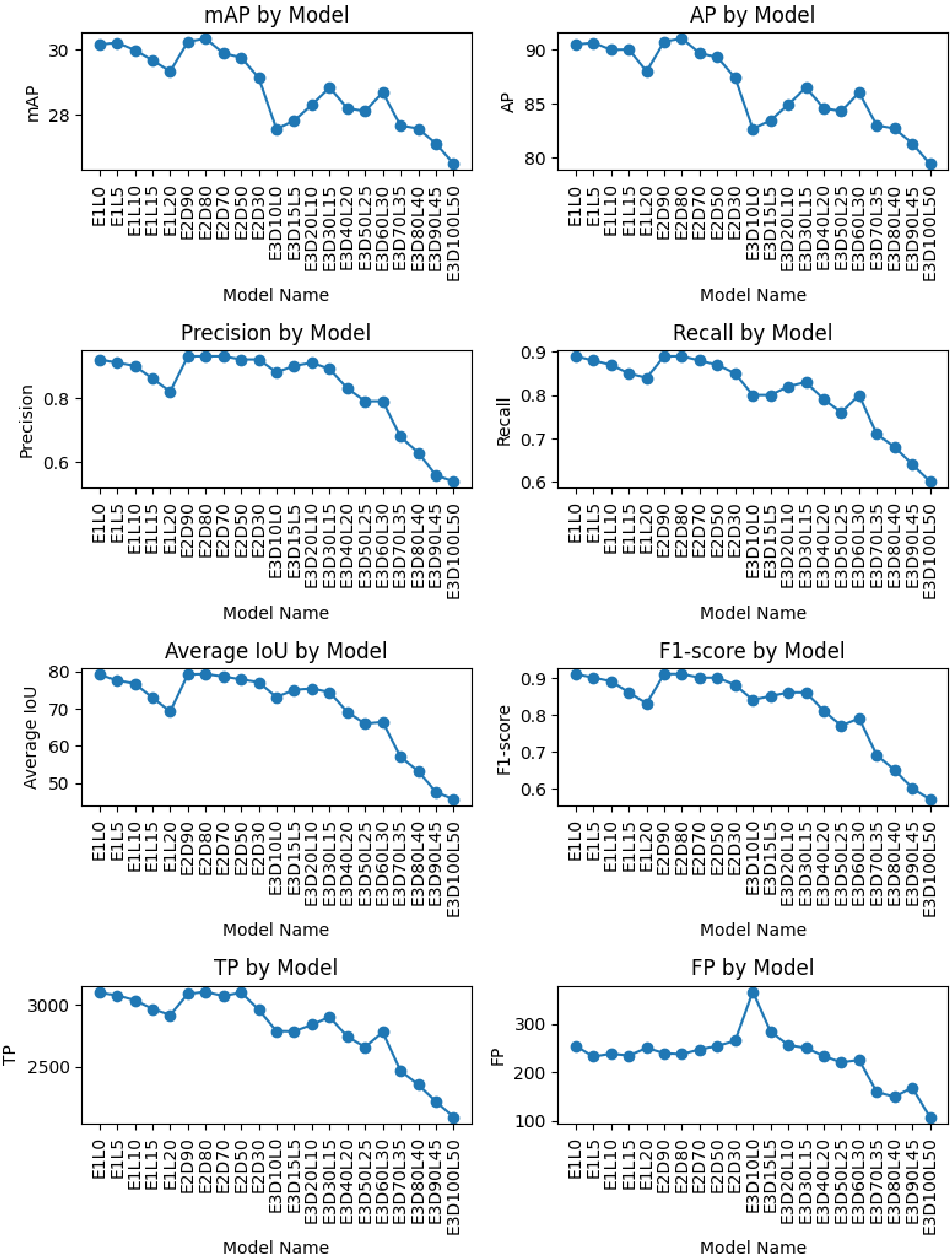

Fig. 12. Comparison of main performance indicators by models.