Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

HYBRID MODEL WITH DISCRETE-EVENT SIMULATION AND REPEATED MACHINE LEARNING PREDICTION-BASED QUALITY INSPECTION OF INBOUND DISTRIBUTION CENTER DELIVERIES

Joost R. Remmmelts
Alexander Hübl

University of Groningen
Faculty of Science and Engineering
Engineering Systems and Design Group
Nijenborgh 4
Groningen, 9747 AG, THE NETHERLANDS

ABSTRACT

Business-to-business distributors deem it necessary to inspect the quality of inbound deliveries to their distribution centers. This paper observes a company that experiences an inefficient quality inspection and wishes to improve the process. The broader product-receiving process is under-researched in warehousing literature but possesses similarities with manufacturing quality control. Moreover, the paper aims to extend prediction-based quality inspection to the warehousing field. It applies a hybrid model, combining discrete-event simulation and machine learning multi-label classification to decrease the required inspection volume and evaluate its effects on the ability of the inspection and the workload and costs of distribution center operations. The results show that the inspection volume can drastically be decreased, reducing the workload and costs at the expense of the inspection capability of infrequently occurring delivery quality flaws in training data. The configuration of the classification model determines the degree of inspection volume reduction and wrongly predicted delivery quality flaws.

1 INTRODUCTION

Gan et al. (2023) observe a manufacturing shift to high-mix, low-volume (HMLV) production as a response to the changing market trend of increased demand for product differentiation and personalization. The effects of the HMLV industry can also be seen in downstream processes such as in distribution centers (DCs) and customer orders (Zhang and Tseng 2009). A business-to-business distribution company in fast-moving consumer goods operates multiple DCs in the Netherlands and is observed during this research. The company has encountered the HMLV effect on top of an already diverse cosmetics branch and, to minimize inventory, purchase order deliveries are increasing in HMLV nature, causing the workload to significantly increase. The company strives for a Lean Six Sigma working environment and wants to improve the product-receiving phase to increase customer satisfaction and decrease costs. In line with the Lean Six Sigma principles, waste, and non-value-adding streams must be reduced. A quality check can be seen as non-value adding if no issue is identified. However, if products do not meet the standard or if the quantity does not match what is ordered, necessary additional steps are taken. In 2022, around 6% of inbound product batches did not meet the quality standards, indicating the need for a quality inspection.

The product-receiving stage of warehousing is less researched than storage and order-picking related topics (Davarzani and Norrman 2015). However, an inefficient receiving process is not a new phenomenon and different approaches have been proposed. The topic also shows similarities with the widely explored

quality control field in manufacturing. The improvement of the quality inspection of the receiving stage can roughly be divided into two themes: *Process Optimization* and *Inspection Volume Reduction*.

The current inspection strategy is invented by managers, conservatively based on intuition. Strategies depend on company rules which should be adjusted to the relevant situation and can range from random or periodic inspections to full-scale control (Ten Hompel and Schmidt 2008). The inspection volume can be decreased by adopting a less intensive strategy, yet inspection reliability must remain acceptable. The enhancement of the WMS has greatly increased the available data on various elements of the process, which the company is increasingly using to control its operations. The data can be used to formulate a data-driven inspection strategy adapted to the real-world and up-to-date situation (McAfee and Brynolfsson 2012). Results from Chavez et al. (2017) show that data-driven supply chains are positively associated with manufacturing capabilities that lead to customer satisfaction. Furthermore, Agrawal et al. (2023) show promising data-driven quality management cases in supply chains that have gained interest in the last ten years. The usage of data as a support for decision-making in supply chains can thus be advantageous over conventional speculative strategies. The current lack of data usage in the quality inspection of the company indicates a great possibility for improvement.

Likewise, warehouse processes can be optimized and automated to improve the product-receiving stage. Related to the processes are optimization techniques such as layout optimization, workstation balancing, and routing optimization (Ten Hompel and Schmidt 2008). Scheduling optimization can also offer a solution to bottlenecks in the supply chain and contribute to a constant flow of work (Ambroziak and Lewczuk 2008). Liu et al. (2019) show that the conventional barcode can be replaced by RFID technology and automated visual inspection can replace the manual visual checks (Abd Al Rahman and Mousavi 2020).

Due to the high variability in products, automation of product-receiving processes is fairly complex. Moreover, the company has already invested in the automation of its processes, whereas the usage of data is exploited to a lesser degree. The available data can immediately be used to improve the process by decreasing the required inspection volume. It is therefore determined that decreasing the inspection volume is more viable and should be the main focus. This paper offers a data-driven solution to significantly decrease inspection volume at the expense of a slight decrease in quality inspection reliability.

Several methods to decrease inspection volume are explored in Section 2. Thereafter, the case study is described and a hybrid model, consisting of a simulation model and a classification model, is presented. Finally, experimental results and a conclusion are given.

2 LITERATURE REVIEW

An often applied quality control method is statistical process control (SPC). SPC uses statistical methods such as univariate control charts for multivariate quality control by using quality data (Kourti and MacGregor 1995). Important variables are observed and located in a specification range, outside which a fault is reported. However, both Kim et al. (2012) and He and Wang (2007) indicate limitations of univariate SPC that are relevant to the company's case, and propose machine learning algorithms as an alternative. Even though multivariate SPC methods such as Principal Component Analysis can manage multiple variables, much of the original information can be lost and the methods assume that data is linear and unimodal, which is often not the case. On the contrary, machine learning algorithms methods have been successfully implemented in such multi-model and nonlinear cases (Markou and Singh 2003). Moreover, when important quality variables are unknown, SPC does not offer a method to identify the important variables and investigate variable interactivity.

Since the need for a quality inspection is based on the state of the delivery by suppliers and transporters, a second option is Supplier Performance Management. A set of performance-deciding aspects can be drafted, each with a relevancy weight (Gunasekaran et al. 2004). The dataset can be acquired to give suppliers a score based on their performance. Shokoohyar (2018) investigates the application of supplier scorecards and shows its positive effect on delivery quality. The applied method in our paper can be helpful in achieving the long-term goal to abolish the need for a quality inspection altogether to save capacity.

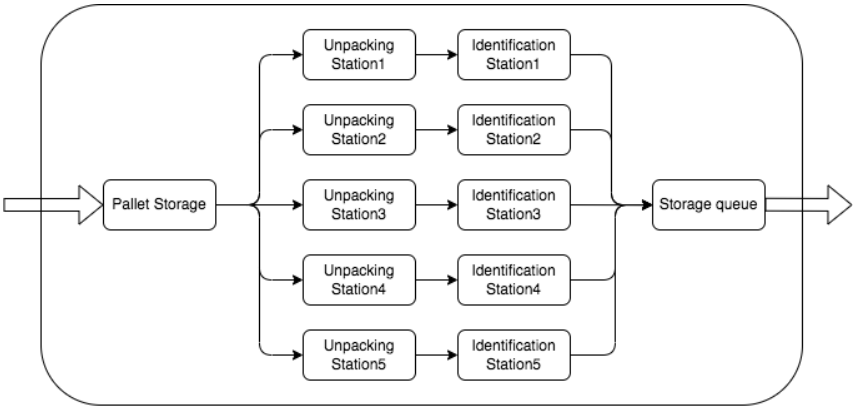


Figure 1: A simplified representation of the observed company's receiving area.

In that regard, predictive analytics is introduced. Machine-learning algorithms can be used to find patterns in historical data, which can then be used to predict a specific outcome (Kumar and Garg 2018). The method is common in manufacturing quality control, where positive effects are obtained (Arif et al. 2013; Kim et al. 2012). Furthermore, Schmitt et al. (2020) show that a predictive model-based quality inspection with machine-learning classification can reduce the required inspection volume up to 29%. There are multiple similarities between the case studies in the literature and the company's case study. A prediction-based inspection is therefore deemed to be a viable method and will be explored in Section 4.

Hybrid simulation models, where discrete event simulation is combined with an optimizing technique is used applied in recent years to gain more realistic insights in stochastic models by testing different possible designs (Ooms and Hübl 2022). An optimization model, such as a machine-learning-dependent predictive classification model, can be combined with a discrete-event simulation model to create a hybrid simulation approach, which is argued to lead to greater solving capabilities of complex systems than a non-hybrid approach (Brailsford et al. 2019). A generic simulation framework can be designed where certain inputs and decisions of a simulation depend on the outcome of an external optimization model (Hübl et al. 2011). However, to achieve a short-term solution to reduce the increasing quality inspection costs, it must be determined if a prediction-based quality inspection can be applied in such a hybrid model approach and be extended to the domain of DC order deliveries while obtaining favorable results. This paper investigates if the same throughput on inspected items can be achieved by less manpower.

3 CASE STUDY

At the observed DC, $3,503$ purchase orders were delivered in 2022. The condition and content of the shipments are systematically inspected to check if the deliveries satisfy the order expectation. The inbound products are stored in boxes on pallets, of which $20,953$ of the latter were required. The pallets contained a total of $121,209$ individual product batches, amounting to a total of $27,539,658$ products. Of the $121,209$ product batches, $9,374$ batches' quantity did not match the order, $7,332$ did not meet quality standards, and 758 contained damaged products. These numbers are not acceptable and thus a check is necessary.

The inspection is divided into 5 parallel sequences, visualized in Figure 1, of an unpacking stage where the products are counted and protective wrapping is removed, and an identification stage where products are visually inspected in detail. It is infeasible to check each individual product and thus, the company has devised a strategy where one product of each inbound product batch is subjected to the identification inspection. The current inspection process is time-consuming and 15 to 20 employees are needed to manage the workload. The warehouse management system (WMS) was recently enhanced to streamline the process which significantly increased efficiency. However, manual labor still takes up the majority of the effort. Moreover, meticulous tasks are performed that rely on the full attention of the operator, and, especially under a time constraint due to a high workload, details are often overlooked.

Table 1: The acquired product batch data of the inbound deliveries of 2022, with the possible data elements.

Batch Data	Elements	Batch Data	Elements	Label	Elements
Delivery Type	Complete / Partial	Product	27,657	Quantity	Flaw / Non-flaw
Batch Size	$1-105,600$	Supplier	370	Quality	Flaw / Non-flaw
Batches per Pallet	$1-239$	Customs Status	39	Damage	Flaw / Non-flaw
Batches per Delivery	$1-1,013$	Packaging	1,082		
Pallets per Delivery	$1-294$	Origin Country	31		

The company data of the inbound deliveries of 2022 is collected. Since the current inspection strategy is to check one product of each product batch, the data is collected at a product batch level to ensure that 100% of the data has been tested. The result of the quality inspection is labeled to the corresponding product batches, as well as ten other data aspects, which can be seen in Table 1. The elements in the second column, apart from that of the delivery type, are described by the range of values in the data. Of these, only the pallets per delivery can take on a decimal number, and the average of a delivery is taken and assigned to each product batch of that delivery. The fourth column displays the totality of different entries present in the corresponding batch data, where the customs status describes the status that is given to ordered goods related to Dutch and European customs regulations. The fourth column entries are transformed into numerical data via the label-encoding method. Finally, the delivery type data and the labels are transformed into binary numbers to end up with a dataset of 121,209 batches with 1,575,717 numbers.

4 HYBRID MODEL

Hybridization in simulation integrates simulation with another method, such as, machine learning or optimization, whereby different designs (iterations) can be tested with the hybrid system (Ooms and Hübl 2022). Such a hybrid model will be designed in our paper for the quality inspection system of the introduced case study. In our paper, the simulation model is executed and a classification model (machine learning) is called each time – **repeated** – when a delivery arrives. The product batch data, seen in columns one and three of Table 1, is passed to the classification model, which returns the label prediction output for each product batch of the delivery. If, for example, 50 deliveries arrive, the classification model will be repeatedly called 50 times during the single simulation run.

The DC receiving phase is applied in a discrete-event environment and a classification model is designed to predict delivery quality. The software AnyLogic 8.8.1 is used for the discrete-event simulations and Python 3.11 is used for the classification model. The Python machine learning library *scikit-learn* is used for its convenient applicability (Pedregosa et al. 2011). Finally, the hybridization is realized by using the Pypeline 1.4 extension for AnyLogic. The connector calls a Python script from a running AnyLogic model, allowing for the transfer of information between the two methods through a created workspace environment. The transformation of information itself happens relatively quickly, but some processing time is required each time Pypeline calls a Python script. It is, therefore, beneficial to minimize the required repeated classification executions.

4.1 Classification Model

In the case mentioned above of Schmitt et al. (2020), labeled data is used to classify a product either to the class *Defective* or *Defect-free*. The observed company in this paper also possesses labeled data, thus classification algorithms can be applied. The three delivery flaw categories, quality, quantity, and damage are documented. This enables the prediction of three separate outcomes, i.e., quantity, quality, and damage, as a *Flaw* or *Non-flaw*. The three subjects are not mutually exclusive and thus, Multi-Label Classification (MLC) algorithms can be used to generate detailed predictions (Zhang and Zhou 2007). The optimal MLC algorithm can generally not be determined a priori and the performance depends on case-specific

aspects (Kotthoff 2016). Zhang and Zhou (2013) introduce seven MLC algorithms, divided into problem transformation and algorithm adaption methods. According to Pushpa and Karpagavalli (2017), problem transformation methods have good flexibility, and therefore, the focus will be on the Binary Relevance (BR), Classifier Chains (CC), and Label Powerset (LP) algorithms. The methods will briefly be described. For a more elaborate description, the reader is referred to (Zhang and Zhou 2013).

Binary Relevance is a first-order approach where the multi-label problem is transformed into an *independent* classification problem for each label. Three separate binary classifiers are generated with a base classifier, where the correlation between labels is not considered. BR is a versatile and often used method but can be relatively slow to train. **Classifier Chains** is a higher-order method that consists of separate binary classifiers. It can be viewed as a chain of *dependent* classifiers, where the next is built upon the predictions of the previous, and so on. Again, three classifiers are generated where label correlation is taken into account. The sequence in which classifiers are generated will affect the performance of the algorithm. **Label Powerset** is a method that transforms a multi-label problem into a multi-class problem, where each combination of label values is considered a class. The LP approach benefits from label correlation as the base classifier considers each combination of labels that occurs in training data. A drawback of the method is the low performance of infrequent classes and the complexity of high-order data.

The three methods make use of a base classifier. There is a vast selection of base classifiers to choose from and new and better algorithms are often being discovered. The extensive research by Tercan and Meisen (2022) on machine learning for predictive quality in manufacturing is used as a guideline to find a suitable base classifier. The three machine learning models that appear most often in publications in 2020 and 2021 are Multilayer Perceptron, Convolutional Neural Network, and Random Forest. Of these, Random Forest (RF) performs well with categorical data (Yu et al. 2019). Furthermore, RF can be applied in a wide range of prediction problems and is generally recognized for its accuracy (Biau and Scornet 2016). In addition, RF returns measures of variable importance, which is useful to detect processes that affect quality flaws (Cerrada et al. 2016). Azab et al. (2021) conducted a similar study, machine-learning-assisted simulation, of a predictive maintenance system and concluded that RF was the best-performing algorithm for some of the regression prediction problems. Our pre-model training studies indeed proved, of the three most frequently published machine learning models, RF generated the most promising results and will therefore be chosen as the base classifier.

First, a Python application is used to randomly split the gathered dataset into a training set of 75% of the data and a test set of 25% that will be used for the preliminary studies of the MLC methods. Next, the RF base classifier is generated. Three design choices for the algorithm are the minimum number of nodes for each tree, the number of trees, and the number of randomly chosen variables to generate the trees. The minimum number of nodes is set to 1 and all the aforementioned variables are considered for the maximum usage of information. The optimal number of trees is determined via a grid-search and set to 300. The three MLC algorithms are then constructed with the base classifier. The algorithms produce a probability value between 0 and 1 for all three labels per product batch. By default, all values greater than 0.5 are classified as a *Flaw*. Preliminary studies showed that this configuration resulted in a careless classifier that failed to detect many flaws. The so-called threshold value was decreased from 0.5 to a conservative value of 0.02, meaning that only product batches with at least a 98% certainty of not containing a flaw are classified as *Non-Flaw*. Moreover, the probability outcome of the individual MLCs can be combined before applying the threshold value. By doing so, the benefits of the three methods are combined and more flaws are detected, at the cost of a smaller decrease in inspection volume.

Preliminary studies concluded that random partitions into train and test sets produced similar results. Furthermore, it was concluded that decreasing the percentage of the train set below 75% decreases the performance of the algorithm, whereas increasing it beyond 75% does not significantly improve performance but does increase the model training time. For this reason, a train set of 75% is chosen as a suitable rate.

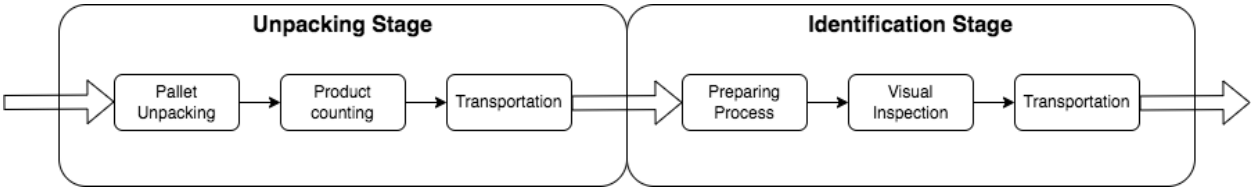


Figure 2: The Unpacking and Identification stages, where the Visual Inspection step is a stochastic process.

Observation 1: The minimum training set percentage should be 75% and increasing beyond this does not significantly increase performance but does increase model training time.

4.2 Simulation Model

The company's simplified receiving area of Figure 1 is applied in a discrete-event simulation whereby only the shop-floor activities are considered and long- and medium-term levels are excluded (Hübl 2015). Due to the high variability of activities and a high degree of human interaction, designing a detailed model where each possible action is integrated proved unachievable. Therefore, a baseline model is designed that resembles the outcome of the actual system as much as possible. This baseline model is used for validating the simulation model to see if the KPIs of the current process of the case study can be met (Kleijnen 1995). Because the data used for machine learning is at a product batch level and to keep the simulation size manageable, product batches are considered as agents for the simulation instead of single products. An aggregated model is hereby designed with the additional advantage that 100% of the agents in the simulations, i.e., all the product batches, have been inspected. The batches are first grouped onto pallets to better represent the processes of the shop-floor and secondly grouped into deliveries for which the Python scripts will be called, since running the Python scripts each time a delivery arrives instead of for each product batch drastically reduces simulation time. The deliveries can later be separated into pallets and product batches once more. Furthermore, the work on the shop-floor is operated $8:00-16:00$ from Monday to Friday. To recreate this, all product batch agents are seized at $16:00$, where the agents wait in place of the procedure until $8:00$, when the agents are discharged once more.

To compare the envisioned prediction-based quality inspection system to the baseline model, a new design is modeled where arriving product batches are inspected based on the prediction outcome of the classification model. If both the labels *Quality* and *Damage* are predicted to be a *Non-flaw*, the products bypass the identification stage and move directly to the storage queue. Data regarding the start and finish time of the visual inspection process of the identification stage is available and is used in the simulation. The processing time of the visual inspection step is set as an exponentially distributed value, which is lower bound by the recorded minimum, to incorporate a form of the stochastic behavior of the system. As can be seen in Figure 2, the stochastic process is applied in a simplified sequential discrete-event simulation model, amongst the five remaining deterministic processing time steps. The reality-based data-driven identification time allows for a realistic evaluation of the effect of the decrease in inspection volume. The unpacking stage, where the product batches are counted, is still essential for preparing products for storage and can thus not easily be bypassed. Furthermore, the company wishes to make absolutely sure that all paid-for products have arrived and therefore, the prediction of the *Quantity* label will not affect the simulation. The result will still produce a useful insight into the causes of quantity issues via the feature importance capabilities of the RF algorithm.

With the available historical data of the entering and leaving times of products in the system and interviews with company staff responsible for the quality inspection, the suitability of the model is determined (Correa Espinal et al. 2012). The five deterministic processes of Figure 2 are designed, partially based on visual shop-floor time observations, to achieve satisfactory model behavior conforming to the historical data. The two designed simulation models can then each be subjected to (a period of) the historical arrival times and shipment content of 2022. A comparison can be made between the current

Table 2: The Confusion matrix for binary classification.

	Pred. *Flaw*	Pred. *Non-Flaw*
True *Flaw*	True Positive (TP)	False Negative (FN)
True *Non-Flaw*	False Positive (FP)	True Negative (TN)

quality inspection model simulation and the hybrid simulation with a prediction-based model with repeated classification. Moreover, the theoretical result of different classification models can be directly tested on historical data to evaluate its effect on the quality inspection system.

4.3 Key Performance Indicators

Both the results of the classification model and the simulation have to be evaluated. Therefore, key performance indicators (KPIs) must be determined. For the evaluation of the classification model, the confusion matrix for binary classification, as shown in Table 2, is considered according to (Schmitt et al. 2020). When dealing with an imbalanced set, the usual *accuracy* metric can give a false indication of the actual classification performance. In our case, for a total of $363,927$ checks, $17,464$ were a flaw. If the classifier predicts all non-flaws the accuracy would therefore be high but a non-desirable result is obtained. The focus is therefore placed on the rate of flaws that are detected (*recall*), the rate of accurate flaw predictions (*precision*), and the harmonic mean of the two, giving an indication of overall performance (F_1-*score*). The three metrics are computed as follows: $precision = \frac{TP}{TP+FP}$, $recall = \frac{TP}{TP+FN}$ and $F_1\text{-}score = 2 \times \frac{precision \times recall}{precision+recall}$. To tackle the multi-label case, both the *micro*- and *macro*-average are used which are two different approaches to represent the same classification result. The *micro*-average presents the mean performance of each prediction for each instance (product batch) and the *macro*-average takes the average performance of the classification of each label (Zhang and Zhou 2013). The *micro-precision* can thus be calculated by taking the sum of all True Positives (TPs) and dividing it by the sum of all TPs plus the sum of False Positives (FPs). Differently, the *macro-precision* can be found by taking the average of the individual precision scores of the three labels.

The KPIs for the simulation model are chosen to evaluate the effect of decreasing the inspection volume. First of all the degree to which the inspection volume is decreased is considered. Moreover, the product batch leadtime and the identification station utilization are considered. The leadtime is taken as the time between a batch entering the system at the receiving area and a batch entering the storage queue, which is de facto the end of the quality inspection stage. The utilization is calculated by dividing the time the stations were active by the total time the stations could have been active. Finally, the system's work-in-progress (WIP) is tracked. All simulation KPIs indicate the systems business under a set of configurations.

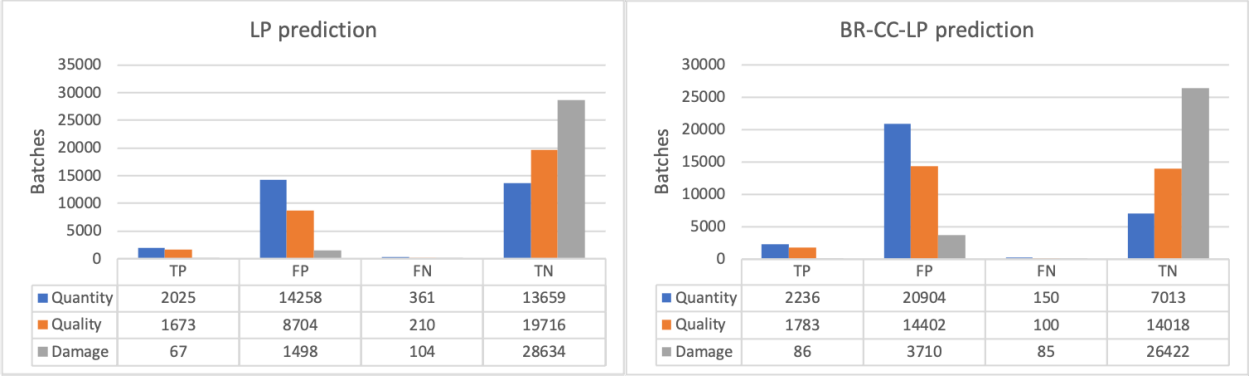
5 EXPERIMENTAL RESULTS

5.1 Quality Prediction

The labels of the test dataset are predicted and compared to the actual data. Seven different MLC models are considered, i.e., the three methods BR, CC, and LP and all possible combinations of the three. The six possible orders in which the Classifier Chains method can be generated were also tested, of which the order *Damage*>*Quality*>*Quantity* produced slightly better results. The resulting performance metrics of the seven methods are displayed in Table 3, where for instance, the combination of Binary Relevance and Classifier Chains is denoted as BR-CC. Not surprisingly, the single methods outperform the combined methods when it comes to overall performance (F_1) as these are directly trained on the dataset whereas the combined methods are an accumulation of optimized algorithms that need not be optimal themselves. However, because of the differences in the MLC methods, different flaws are found with different methods. After applying the low threshold value, this results in a much higher recall of the combined methods than that of the single methods. Unfortunately, this comes at the price of lower precision, as a larger

Table 3: The performance of the respective multi-label classification (MLC) methods and combination of methods. The methods marked with an asterisk will be used in the simulations.

	BR	CC	LP*	BR-CC	BR-LP	CC-LP	BR-CC-LP*
***micro**-Precision*	13.4%	13.6%	13.3%	10.9%	10.8%	10.8%	9.7%
***micro**-Recall*	85.3%	85.2%	84.8%	90.2%	90.3%	90.3%	92.5%
***micro**-F_1score*	23.2%	23.4%	23.5%	19.4%	19.3%	19.3%	17.6%
***macro**-Precision*	11.0%	11.1%	10.9%	8.8%	8.7%	8.7%	7.8%
***macro**-Recall*	71.7%	71.6%	71.0%	76.6%	77.0%	77.1%	79.6%
***macro**-F_1score*	19.0%	19.2%	18.9%	15.7%	15.5%	15.6%	14.1%



(a) The LP prediction result of the test set.

(b) The BR-CC-LP prediction result of the test set.

Figure 3: Bar charts of the confusion matrix data of the classification models prediction of the test set.

inspection volume is required to obtain the high recall. The threshold value can be tuned to find an optimal precision-to-recall rate for the company's wishes but for this research, the value is kept at 0.02.

> **Observation 2:** The combinations of Multi-Label Classification methods are more conservative than the single Multi-Label Classification methods.

Furthermore, Table 3 clearly shows that the *macro*-average is generally worse than the *micro*-average for the same experiment. Keep in mind, that the experiment data for micro and macro is identically, only the presentation of data is different. This can be explained by the difference in label classification performance. The MLC methods perform much better in the classification of the *quantity* and *quality* labels than that of the *damage* label. The cause of this is the low occurrence of damage cases in the dataset. The dataset contains more than 10 times fewer damage cases than quantity flaws and thus, the models are better trained to find quantity and quality flaws than damage cases. The difference in performance over each label for the single LP method can be seen in Figure 3a. The method with the highest F_1 score correctly predicts 85% of quantity and 89% of quality flaws but only 39% of damage flaws. However, LP, which is the least cautious method with the highest precision but lowest recall, still performs well at predicting quality cases, while reducing the overall checks by 83%, or 66% for just the quality label. Comparing the LP result to that of the BR-CC-LP method, the method with the highest recall, seen in Figure 3b, we indeed see a much lower False Negative total for the combined method. In this case, around 50% of all the damage cases are predicted for which only 12.5% of the total instances need to be checked. Notably, 95% of all quality flaws are correctly predicted while only 53% of the instances have to be inspected. The quantity cases are also predicted significantly well but at a much higher cost than the quality label, resulting in a

Table 4: The result of simulations of the two different hybrid models and the model of the current system.

	Insp. Volume (batches)	avg. Lead-time (hrs)	Utilization (%)	avg. WIP (batches)	Simulation Runtime
Base Model	25,501	34.7	57%	401.5	5 min
LP Model	9,991 (−61%)	19.5 (−44%)	22% (−61%)	237.3 (−41%)	30 min
BR-CC-LP Model	16,670 (−35%)	23.5 (−32%)	37% (−35%)	285.6 (−29%)	5.5 hrs

total of 20,904 False Positive cases. A higher threshold value would likely improve the performance of the prediction of this particular label.

> **Observation 3:** The prediction of the *Damage* label is considerably worse than the other two as it occurs less frequently in training data.

Finally, the ten product batch data aspects of Table 1 used to generate the MLC models were not of equal importance and the respective importance of each data type out of 100% is recorded. The feature of most importance is unsurprisingly the product at 35%. After that, the type of packaging and the batch size have the most impact on the classification, both at around 15.5%. The remaining data instances have importance from $3\% - 9\%$, with the exception of the delivery type, which has close to zero importance and can thus be neglected in future work.

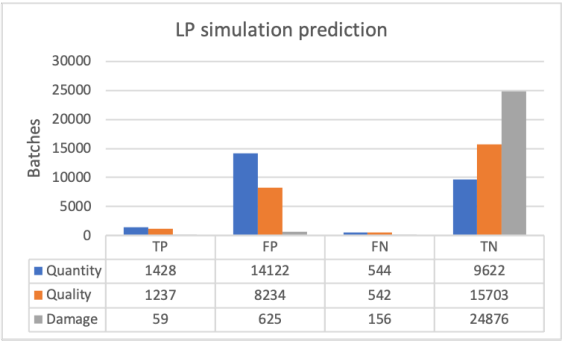
> **Observation 4:** Product type, packaging type, and batch size data are most relevant to the quality of deliveries.

5.2 Inspection Reduction

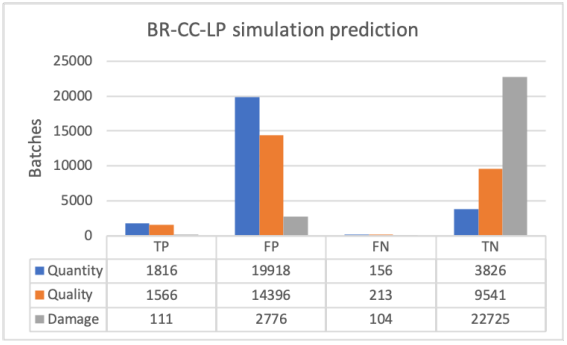
The best-performing overall method LP and the method with the highest recall, the combination of all methods (BR-CC-LP), will be tested in the simulations. These two methods have the largest and lowest theoretical inspection volume reduction, respectively. Ten random simulation replications with the historical data of January up until March of 2022 are executed with the base- and LP model and three with the BR-CC-LP model, due to the long simulation time, which entail a total of 25,716 product batches each iteration. The resulting average KPIs are recorded and can be seen in Table 4. LP, the least cautious method, achieves a significant inspection volume reduction of 61% with an all-around positive impact on the KPIs and can thus significantly reduce costs. Most notable is the tremendous reduction in identification station utilization. A shift to a station utilization of 22% suggests that the number of stations could likely be reduced, freeing up much of the product receiving budget. However, in line with the findings in Section 5.1, many flaws are missed. Figure 4a displays the confusion matrix data of the classification of the LP simulation. The damage label has a significantly low flaw detection rate at 27% as opposed to the quantity and quality label where a more acceptable rate of around 70% was observed.

> **Observation 5:** A significant inspection volume reduction is realized by a prediction-based inspection strategy at the cost of a degree of missed cases. The volume reduction can lead to reduced warehouse operation costs.

The most cautious combined method BR-CC-LP realizes a lower decrease in the workload than the LP method. Nonetheless, an inspection volume reduction of 35% is significant and the system is considerably less busy than the base system, which allows for a significant cost reduction. Furthermore, the confusion matrix data of the BR-CC-LP simulation, seen in Figure 4b, demonstrates the high recall of the method presented in Section 5.1. Here, 52% of actual damage cases and around 90% of actual quantity and quality flaws are predicted. The consideration can be made that for a 43% lower inspection reduction, 29% more

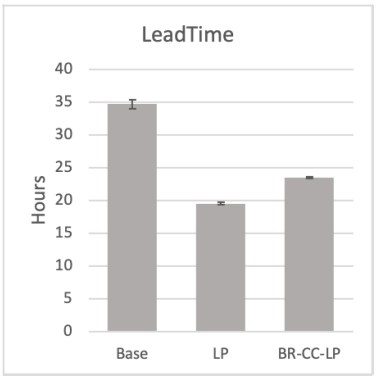


(a) The LP prediction result of the simulation.

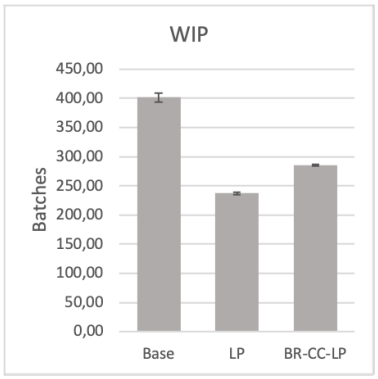


(b) The BR-CC-LP prediction result of the simulation.

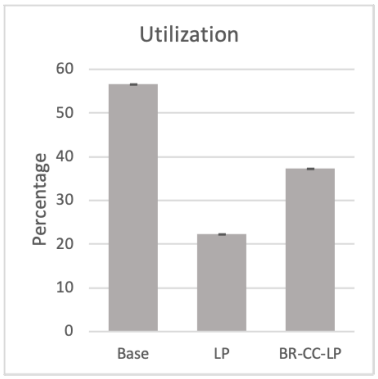
Figure 4: Bar charts of the confusion matrix data of the classification models prediction of the simulation.



(a) The Leadime comparison.



(b) The WIP comparison



(c) The Utilization comparison.

Figure 5: A bar chart of the stochastic KPIs with the 95% confidence interval. The confidence interval of the Base and LP model was determined with 10 replications and that of the BR-CC-LP method with three.

quality and quantity cases and 126% more damage cases are predicted. These results are all obtained with the threshold value set at 0.02. By tuning this value and applying it to the various methods, a favorable result can be sought. It should be kept in mind that training the models for the combined method takes much longer than that of a single method and the same is true for the prediction time.

> **Observation 6:** The Multi-Label Classification configuration can be modified to obtain a favorable inspection volume reduction for a degree of delivery quality flaws that are missed in the quality inspection phase.

The stochastic behavior of the system is evaluated by determining the confidence interval of the 10 replications of the base- and LP models and three replications of the BR-CC-LP model and can be seen in Figure 5. Because there was only one process step of Figure 2 that could be determined from data, this was taken as the only random processing time as it allowed for a realistic stochastic distribution. This results in a relatively small confidence interval.

6 CONCLUSION

This research proposes a hybrid simulation model to evaluate the effect of a machine learning prediction-based quality inspection of inbound deliveries. Every time a delivery arrives, a classification model predicts

the quality, which determines the routing of the products in a discrete-event simulation model. The results showed that the inspection volume can be greatly decreased, reducing product-receiving costs, at the cost of a certain rate of missed flaws. The research extends prediction-based quality control in manufacturing to the warehousing domain and shows promising applications in the data-driven machine learning field for this sector. Furthermore, it showed that combining simulation with machine learning allowed for a real-time interpretation of the implementation of a data-driven management method.

This research considers the Random Forest base learning classifier. Future work in similar case studies can investigate the performance of different base classifiers and Multi-Label Classification methods. Furthermore, a method to optimize the threshold value can be analyzed to improve the overall result. The simulation model used in the research gives a good indication of the effect of an inspection volume decrease but was heavily simplified. Hence, a more precise simulation model, containing more realistic randomness, could give a better representation of the system and therefore also of the effect of an inspection volume reduction. Ultimately, machine learning prediction-based quality inspection should not be the standalone goal to optimize the product-receiving process. Optimizing and automating processes and implementing methods such as Supplier Performance Management can additionally improve overall efficiency.

One of the main findings of this paper concludes that the minimum training set percentage of the classification model should be 75% and increasing beyond this does not significantly increase performance but does increase model training time. The trained models can then be combined to detect more flaws than single methods, at the price of a much lower reduction in inspection volume. If a delivery quality flaw does not occur sufficiently often in the training set, like the *Damage* label in this paper, the model is not trained adequately to predict the cases. Regarding the data in the training set, the type of product, type of packaging, and batch size contribute the most to the prediction of the delivery quality. It was concluded that a prediction-based inspection strategy greatly reduces inspection volume but comes at the cost of a degree of missed delivery quality flaws. Thus, warehouse operation costs can be cut down as a result of the volume reduction. Finally, the configuration of the classification model determines the degree of inspection volume reduction and wrongly predicted delivery quality flaws and can be altered according to preference.

REFERENCES

Abd Al Rahman, M., and A. Mousavi. 2020. "A Review and Analysis of Automatic Optical Inspection and Quality Monitoring Methods in Electronics Industry". *Ieee Access* 8:183192–183271.

Agrawal, R., V. A. Wankhede, A. Kumar, and S. Luthra. 2023. "A Systematic and Network-Based Analysis of Data-Driven Quality Management in Supply Chains and Proposed Future Research Directions". *The TQM Journal* 35(1):73–101.

Ambroziak, T., and K. Lewczuk. 2008. "A Method for Scheduling the Goods Receiving Process in Warehouse Facilities". *Total Logistic Management* 5(1):7–14.

Arif, F., N. Suryana, and B. Hussin. 2013, 05. "A Data Mining Approach for Developing Quality Prediction Model in Multi-Stage Manufacturing". *International Journal of Computer Applications* 69:40.

Azab, E., M. Nafea, L. A. Shihata, and M. Mashaly. 2021. "A Machine-Learning-Assisted Simulation Approach for Incorporating Predictive Maintenance in Dynamic Flow-Shop Scheduling". *Applied Sciences* 11(24):11725.

Biau, G., and E. Scornet. 2016. "A Random Forest Guided Tour". *Test* 25:197–227.

Brailsford, S. C., T. Eldabi, M. Kunc, N. Mustafee, and A. F. Osorio. 2019. "Hybrid Simulation Modelling in Operational Research: A State-of-the-Art Review". *European Journal of Operational Research* 278(3):721–737.

Cerrada, M., G. Zurita, D. Cabrera, R.-V. Sánchez, M. Artés, and C. Li. 2016. "Fault Diagnosis in Spur Gears based on Genetic Algorithm and Random Forest". *Mechanical Systems and Signal Processing* 70:87–103.

Chavez, R., W. Yu, M. A. Jacobs, and M. Feng. 2017. "Data-Driven Supply Chains, Manufacturing Capability and Customer Satisfaction". *Production Planning & Control* 28(11-12):906–918.

Correa Espinal, A. A., R. A. Gómez Montoya, and J. A. Sánchez Alzate. 2012. "Improvement of Operations of Picking and Dispatch for a Business in the Mattress Industry, supported by Discrete Simulation". *Dyna* 79(173):104–112.

Davarzani, H., and A. Norrman. 2015. "Toward a Relevant Agenda for Warehousing Research: Literature Review and Practitioners' Input". *Logistics Research* 8:1–18.

Gan, Z. L., S. N. Musa, and H. J. Yap. 2023. "A Review of the High-Mix, Low-Volume Manufacturing Industry". *Applied Sciences* 13(3):1687.

Gunasekaran, A., C. Patel, and R. E. McGaughey. 2004. "A Framework for Supply Chain Performance Measurement". *International Journal of Production Economics* 87(3):333–347.

He, Q. P., and J. Wang. 2007. "Fault Detection using the K-Nearest Neighbor Rule for Semiconductor Manufacturing Processes". *IEEE Transactions on Semiconductor Manufacturing* 20(4):345–354.

Hübl, A. 2015. *Stochastic Modelling in Production Planning*. Springer.

Hübl, A., K. Altendorfer, H. Jodlbauer, M. Gansterer, and R. F. Hartl. 2011. "Flexible Model for Analyzing Production Systems with Discrete Event Simulation". In *Proceedings of the 2011 Winter Simulation Conference*, edited by S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu, 1554–1565. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Kim, D., P. Kang, S. Cho, H.-j. Lee, and S. Doh. 2012. "Machine Learning-Based Novelty Detection for Faulty Wafer Detection in Semiconductor Manufacturing". *Expert Systems with Applications* 39(4):4075–4083.

Kleijnen, J. 1995. "Verification and Validation of Simulation Models". *European Journal of Operational Research* 82(1):145–162.

Kotthoff, L. 2016. "Algorithm Selection for Combinatorial Search Problems: A Survey". *Data Mining and Constraint Programming: Foundations of a Cross-Disciplinary Approach*:149–190.

Kourti, T., and J. F. MacGregor. 1995. "Process Analysis, Monitoring and Diagnosis, using Multivariate Projection Methods". *Chemometrics and Intelligent Laboratory Systems* 28(1):3–21.

Kumar, V., and M. Garg. 2018. "Predictive Analytics: a Review of Trends and Techniques". *International Journal of Computer Applications* 182(1):31–37.

Liu, H., Z. Yao, L. Zeng, and J. Luan. 2019. "An RFID and Sensor Technology-Based Warehouse Center: Assessment of New Model on a Superstore in China". *Assembly Automation* 39(1):86–100.

Markou, M., and S. Singh. 2003. "Novelty Detection: a Review—Part 2:: Neural Network Based Approaches". *Signal Processing* 83(12):2499–2521.

McAfee, A., and E. Brynolfsson. 2012. "Big Data: the Management Revolution". *Harvard Business Review* 90(10):60–68.

Ooms, J., and A. Hübl. 2022. "Applying a Hybrid Model to Solve the Job-Shop Scheduling Problem with Preventive Maintenance, Sequence-Dependent Setup Times and Unknown Processing Times". In *Proceedings of the 2022 Winter Simulation Conference*, edited by B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C. Corlu, L. Lee, E. Chew, T. Roeder, and P. Lendermann, 1750–1761. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. "Scikit-learn: Machine Learning in Python". *Journal of Machine Learning Research* 12:2825–2830.

Pushpa, M., and S. Karpagavalli. 2017. "Multi-Label Classification: Problem Transformation Methods in Tamil Phoneme Classification". *Procedia Computer Science* 115:572–579.

Schmitt, J., J. Bönig, T. Borggräfe, G. Beitinger, and J. Deuse. 2020. "Predictive Model-Based Quality Inspection using Machine Learning and Edge Cloud Computing". *Advanced Engineering Informatics* 45:101101.

Shokoohyar, S. 2018. *Supplier Performance Management: a Behavioral Study*. Ph. D. thesis.

Ten Hompel, M., and T. Schmidt. 2008. *Warehouse Management*. Springer.

Tercan, H., and T. Meisen. 2022. "Machine Learning and Deep Learning Based Predictive Quality in Manufacturing: a Systematic Review". *Journal of Intelligent Manufacturing* 33(7):1879–1905.

Yu, B., S. Bao, F. Feng, and J. Sayer. 2019. "Examination and Prediction of Drivers' Reaction when provided with V2I Communication-Based Intersection Maneuver Strategies". *Transportation Research Part C* 106:17–28.

Zhang, M.-L., and Z.-H. Zhou. 2007. "ML-KNN: a Lazy Learning Approach to Multi-Label Learning". *Pattern Recognition* 40(7):2038–2048.

Zhang, M.-L., and Z.-H. Zhou. 2013. "A Review on Multi-Label Learning Algorithms". *IEEE Transactions on Knowledge and Data Engineering* 26(8):1819–1837.

Zhang, Q., and M. Tseng. 2009. "Modelling and Integration of Customer Flexibility in the Order Commitment Process for High Mix Low Volume Production". *International Journal of Production Research* 47(22):6397–6416.

AUTHOR BIOGRAPHIES

JOOST R. REMMELTS is a graduate master student in Industrial Engineering and Management at the University of Groningen. He is specializing in the track of Production Technology and Logistics. His e-mail address is j.r.remmelts@student.rug.nl.

ALEXANDER HÜBL is a member of the Engineering Systems and Design Group at the University of Groningen (Netherlands). He holds a PhD in logistics and operations management from the University of Vienna, Austria. His research interests include discrete-event simulation, agent-based simulation, queuing theory, hybrid modeling and their applications in logistics and operations management. His email address is a.hubl@rug.nl.