

Avoiding Object Damage in Robotic Manipulation

Erica Aduh*, Fan Wang*, Dylan Randle, Kaiwen Wang, Priyesh Shah, Chaitanya Mitash, Manikantan Nambi

Abstract—The large-scale deployment of robotic manipulation systems in warehouses has highlighted the rare but costly problem of robot-induced object damage. We present a system that uses a classification model to predict whether an object will get damaged during robotic manipulation. The model uses object attributes retrieved from warehouse information systems as well as attributes available at our robotic workcell. We evaluated different classical machine learning models, as well as a large language model (BERT) and a multimodal-transformer for our task. We show that the multi-modal transformer model that is able to leverage text and image data outperforms models that only rely on categorical and numerical data. Furthermore, our comparative analysis equips the selection of the optimal model for an application. We validate our system during an experiment in which the output of the damage prediction system is used to avoid picking objects that are likely to get damaged. In over 50k pick-and-place activities, our system reduces damage rate by 64%.

I. INTRODUCTION

Advancements in robotic hardware, grasping techniques, and perception technologies have paved the way for reliable, high-speed robotic manipulation. These breakthroughs have enabled industrial manipulation systems to handle a wide variety of objects in warehouses. These systems aim to automate processes from sorting inbound inventory to picking and packing individual orders. Amazon’s Sparrow [2], Covariant AI’s picking solutions [1], and RightHand Robotics’ RightPick 3 [3] are examples of such picking systems used for robotic manipulation in warehouses.

The large-scale deployment of robotic manipulation systems in object picking has exposed a rare, but costly issue of robot-induced object damage [34]. In addition to lost value of the object, damage leads to several adverse effects, including increased operating costs, hardware damage, and customer dissatisfaction. When an object gets damaged in a robotic workcell, an operator may need to restart the system, clear the damage and discard the object. Some damage events result in spills, which could compromise the quality of sensitive robotic hardware. Furthermore, if damage goes undetected, customers will receive a defective product which can then lead to customer attrition and mistrust.

Damage to objects in robotic manipulation is due to a number of factors such as aggressive motion trajectories, improper grasp strategies, and accidental collisions. Moreover, objects with multipart designs that are not well sealed are particularly susceptible. Manipulating objects can result in a variety of adverse outcomes: from crushing cardboard packaging, to tearing plastic wrap, and grazing a book cover.

*Authors with equal contribution. All authors are with Amazon, MA, USA. {aduheric, fanwanf, dylanran, ewakaiwe, spriyesh, cmitash, mnambi}@amazon.com

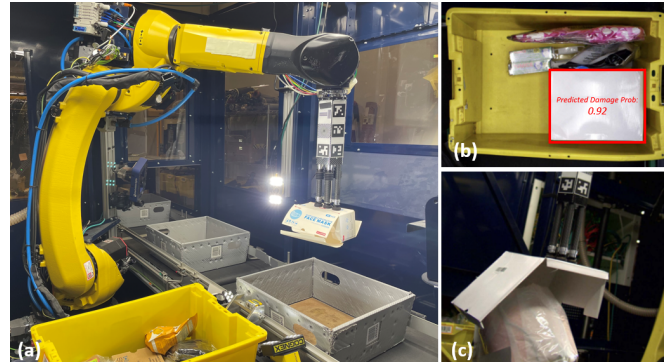


Fig. 1. (a) A robotic manipulation system operating in an Amazon warehouse. The robotic arm picks one object at a time from the yellow container and places it in a gray tray. (b) Pre-manipulation view of objects inside the yellow container. We use a system that predicts if any object will be damaged if manipulated by the robot. A white box, marked in red bounding box, is identified as having a high risk of incurring damage during the process. (c) An illustration of the consequences without damage avoidance: the object is damaged through robotic handling. Post-manipulation, a technician annotates the object’s condition, detailing whether damage occurred and categorizing the type of damage sustained.

Robot-induced object damage is rare and requires a significant number of observations to measure and understand. Therefore, despite its significance, the problem has yet to gain wide recognition within the robotics research community. We hypothesize that by identifying the factors leading to damage, we can anticipate and prevent damage events before they occur.

In this paper, we introduce the problem of object damage avoidance during robotic manipulation. We present a classification-based solution that uses descriptions of objects in the warehouse, their current state and outcomes from prior manipulations to predict the likelihood of damage and determine whether to pick the object. For this classification task, we first explore the application of various classical machine learning models and then design a multi-modal transformer that fuses encodings from a language model and a vision transformer. Thorough comparisons and an ablation study provide insights into this novel challenge and the modeling required for similar applications. Finally, we deploy this solution on a robotic manipulation workcell demonstrating damage avoidance on a large-scale, previously unseen inventory. To our knowledge, this is the first application of large language models and multi-modal transformers to the task of predicting the outcome of a robotic manipulation process.

II. RELATED WORK

A. Robotic Warehouse Automation

The rise of E-commerce has stimulated demand for automation in warehouse operations. The Amazon Robotics

Challenge (2015-2017) invigorated interest in robotic manipulation systems for warehouse tasks [39]. Designed as a real-world benchmark for warehouse operations, this event fostered the advancement of pioneering pick-and-place robot prototypes [48], [40], [49], demonstrating proficiency in recognizing and manipulating both known and new items in cluttered environments.

The industry witnessed a surge in adoption of high-speed, dependable robotic manipulation platforms in recent years. Examples include Berkshire Grey's robotic pick and pack (BG RPP) system [23], Covariant AI's pocket sorting robot [15], and XYZ robotics' piece-picking robot [47]. Another notable example of expansive robotic manipulation platform is Amazon's Sparrow, which integrates computer vision and artificial intelligence for efficient object handling [2].

B. Data-driven Predictive Analytics

Data-driven predictive analytics encompasses the systematic analysis of historical and current data through machine learning and statistical algorithms to anticipate future events. Widely applied across sectors such as manufacturing [20], [51], [44], healthcare [5], [45], [27], [4], agriculture [12], [33], [52], [35], and robotics [41], [16], [9], [31], [32], [6], its prominence has grown with the evolution of machine learning and deep learning models [25], [42], [43], [37], [21], [50], [38], [24], [30], [10], [36], [14].

In manufacturing and industrial maintenance, often coined as predictive maintenance (PdM), analytics focuses on health perception, like fault diagnosis and residual life assessment [51]. Predictive maintenance (PdM) commonly utilizes time series data, derived from sensors monitoring machinery vibration [26], temperature [46], electrical signals [22]. Time-series data typically utilize classical time series forecasting models, such as ARIMA (AutoRegressive Integrated Moving Average) [7], multi-stage prediction [51], and more recently Recurrent Neural Networks (RNNs) to recognize temporal patterns indicative of machinery anomalies [51], [29]. When visual data is incorporated, Convolutional Neural Networks (CNN) are employed to enhance prediction accuracy by processing images alongside time series data [13]. For an extensive discourse on data-driven methods in predictive maintenance, the readers are referred to [51].

Predictive analytics in robotic manipulation has been used for grasp success prediction [41], [16], [9]. For example, Calandra *et al.* [9] explored touch sensing's role in robot grasp prediction, utilizing a multimodal sensing platform that assimilates vision and tactile feedback. Using GelSight tactile sensors, the research underscored that vision and touch synergies boost grasp prediction accuracy. Cvitanic [16] also investigated closed-loop stability prediction in industrial robots, employing a laser tracker for end-effector motion measurements and suggested "Equivalent Force" for modeling robot structural dynamics. Another prevalent analytics application in robots concerns human intention or motion prediction during human-robot interactions [31], [32], [6].

Implementing predictive analytics, particularly when integrating multiple data sources, is challenging. Factors such as sensor malfunctions in manufacturing or varying imaging conditions can result in inconsistent and noisy data. Acquisition of labeled datasets can be challenging, particularly for rare events, potentially yielding non-representative datasets which are prone to overfitting by machine learning models. The interpretability of models, crucial in sectors like manufacturing and healthcare, remains a significant concern due to the black-box nature of the complex non-linear models.

III. OBJECT DAMAGE DATASET AND TAXONOMY

A. Robotic Manipulation Workcell

Data collection was performed on a robotic manipulation workcell which picks objects from a cluttered container and singulates them onto a tray [34]. The objects in the pick scene are of varying physical traits and material properties. The robot, a 6-DoF arm with a suction-based end-effector, continuously picks the items autonomously. The tray with the placed item is presented to the workcell operator who annotates the item for defects such as damage that occurred during manipulation. Once the source container is emptied, a new one is presented to the workcell and the process is repeated.

A subset of the pick-and-place activities from this testing station, consisting of over 600,000 operations, has been compiled into a dataset, which includes a small percentage of activities with damage. The data for each activity comprises: 1) a unique identifier for the object 2) properties of the object (Table II), 3) an image of the object in the container and 4) an operator annotation of whether or not the activity resulted in damage to the object.

B. Damage Taxonomy and Root Causes

We analyzed the activities that resulted in a damaged object from our robotic workcell and classified damage into nine major categories (Fig. 2):

- 1) Deconstruction: the object splits into two or more parts.
- 2) Open: the object opens but does not split into multiple parts.
- 3) Hole/Cut: object is penetrated (but not significantly deformed) by a cut or a hole.
- 4) Crush: object shape is substantially deformed (but not penetrated), including crushed items and severe dents.
- 5) Superficial: damage without penetration or deformation. These include scratches and dirt and are typically aesthetic.
- 6) Stain/Spillage: water or oil stain, spillage, etc.
- 7) Crease: a line or mark produced by folding or pressing (book cover or page, paper).
- 8) Tear: damage due to pulling apart of packaging material such as plastic wrap.
- 9) Other: any other form of damage (e.g., broken tag, chipped glass).

Figure 2 shows examples of each damage type and Table I provides a distribution for damage types observed in our workcell. We observe that deconstruction is the primary



Fig. 2. Example images for different damage types. Deconstruction, crush, and tear damage events can be catastrophic i.e., the value of object is lost. Open, superficial, stain/spillage damage events are typically less severe and the object could be recovered by a warehouse technician.

TABLE I
DISTRIBUTION OF DAMAGE TYPES IN OUR DATASET

Damage Type	Decon- struction	Open	Crush	Hole/ Cut	Super- ficial	Stain/ Spillage	Crease	Tear	Other
Proportion (%)	76.6	17.8	1.1	1.0	0.4	0.3	0.6	1.7	0.6

damage type observed in our workcell followed by open damage. These two categories combined comprise 95% of damage events observed. Deconstruction and open damage were primarily observed in four distinct types of objects: books with dust covers, telescoping boxes, tuck-lid boxes, and shoe-boxes. A suction-based end-effector, even though widely used in warehouse systems, creases or pulls off book covers. Similarly, the end-effector is not able to provide appropriate forces to maintain multiple parts of boxes together. As a result, boxes with multi-part design open and separate into multiple pieces. Example videos of deconstruction and open damage can be found in [34]. Damage such as hole, crush, or tear are caused by the interaction forces between the end-effector and the object, forces due to motion of the object, and improper trajectories when the object is placed at its destination.

IV. DAMAGE AVOIDANCE IN ROBOTIC MANIPULATION

To reduce damage to objects in our robotic workcell, we propose a system that identifies an object and predicts if the object will be damaged if it is manipulated by the workcell. We formulate the damage avoidance task as a classification problem where, given a number of attributes of the object and robotic workcell, the classifier predicts if the object will be damaged if the robotic manipulation activity would have been performed.

A. Attributes for Damage Avoidance

Through our analysis of objects damaged by our workcell during robotic manipulation, we compiled a list of attributes of objects that could be used to predict if a robotic manipulation of a particular object would result in damage. The

object attributes are classified into three categories based on the type of attribute and their source:

1) *Catalog*: A number of attributes of the objects such as weight, dimensions, shape, packaging, material, name, description, etc., are stored in a catalog by the warehouse information systems. These attributes have shown a correlation to object damage in our workcell. For instance, a heavy object packed in a plastic material is more likely to be damaged than if the same object was packaged in a cardboard box. The catalog attributes are a combination of categorical, numerical, and text data. However, the catalog data suffers from missing values as not all objects have all catalog attributes available. An example of attributes is provided in Table II.

2) *Workcell*: Top-down RGB images (Fig. 1) for each object that is picked by our system are available during each activity. We believe images of the object in the source container have relevant information embedded such as the physical and text properties as well as additional information about the object in the context of the source container. We use images of an object at the workcell in two ways: 1) predict attributes of the object such as its shape, material, and append the information to our catalog attributes, and 2) directly ingest the image in a model that can learn from multi-modal data. The attributes of the object predicted from images at the workcell reduce missing data from our catalog attributes.

3) *Experience*: Our analysis showed that if an object was damaged by a workcell, there was a high probability that another instance of the object will also be damaged if manipulated by our workcell. Based on this information, we compute a binary variable which indicates if a previous manipulation of the object at our workcell resulted in damage to the object.

B. Damage Avoidance Models

We evaluate a range of classification models with increasing model complexity and ability to leverage different sources of data.

TABLE II
ATTRIBUTES THAT ARE USED FOR PREDICTING DAMAGE IN OUR
WORKCELL.

	Attributes	Example
Catalog	Product group	[beauty, hair styling, 1000, 1200]
	Dimensions	[7.1, 4.5, 4.5] in
	Weight	0.4 lb
	Package	[container, can roll]
	Description	Keratin Silk
	Price	34.36 USD
	Brand	Colgate
Workcell	Miscellaneous	cardboard box ...
Experience	Visual Image	1

1) *Classical ML Models*: We evaluate four machine learning models: Logistic Regression [28], XGBoost (Extreme Gradient Boosting) [11], CatBoost (Categorical Boosting) [18], and Random Forests [8]. Categorical attributes are one-hot encoded and numerical variables are scaled in our dataset for training these models. With compact size and efficiency, traditional ML models have fast training and inference time and allow real-time prediction on robotic manipulation systems. In addition, these models are well-understood and widely used, making their behavior predictable and their outcomes more interpretable.

2) *BERT Sequence Classifier*: We also evaluate the BERT-based sequence classifier for the damage avoidance task. The model is built upon the BERT transformer [17] with a classification head. A dropout layer is incorporated for regularization, followed by a linear classification layer for generating binary predictions. Given BERT’s design, pre-trained on a vast text corpus, it is equipped to harness long sequences of text and the intricate attention relationships therein. A unique advantage of the BERT sequence classifier is the incorporation of unstructured text data, like object descriptions, into the model.

3) *ViT-BERT Model*: Damage to items can occur due to a number of factors, some of which may not be fully encapsulated in object-level descriptions. We hypothesize that raw images may contain additional information for predicting the occurrence of damage, such as the object’s pose, its surroundings, and proximity to other objects. To address this, we propose a multi-modal transformer model that is capable of concurrently processing both raw in-container images of an object and its attributes.

Fig. 3c shows the architecture of the model. The ViT-BERT model leverages image and text data by performing late fusion of a Vision Transformer (ViT) [19] and BERT models. Image data is processed through a ViT model and text data is processed through a BERT model which results in two distinct sets of embeddings. These embeddings consist of attention masks and tokens, which are concatenated and aligned to form a joint embedding space. To understand the relationships between the visual and text data, additional transformer layers are applied to this joint embedding space to learn the attention between the image and text embed-

dings. Regularization is achieved by applying a dropout layer to the concatenated embeddings, which is then followed by a linear classification layer for predicting the likelihood of damage.

C. Model Evaluations

To enable comparisons across various models, we’ve created an offline dataset using a selection of inductive activities captured within our robotic workcell. To manage class imbalance as damage is rare, we downsample the dataset to contain 2:1 ratio of undamaged to damaged activities. The downsampled dataset comprises data from 10782 inducts. We use a 9:1 split for train and test set. We use recall and false positive rate (FPR) as the metrics to evaluate our models. Recall indicates the proportion of activities that the model correctly predicted as resulting in object damage and FPR indicates the proportion of non-damage activities incorrectly predicted as resulting in object damage. A successful system achieves high recall and low FPR.

1) *Comparison of Classical ML Models and BERT*: We evaluate our damage avoidance models using a consistent set of attributes. Given that classical ML models are designed to handle only categorical and numerical attributes, we exclude the raw image and the text attribute “Description” from our training and testing set. As a result, we exclude ViT-BERT from this evaluation as raw images are not incorporated.

Table III shows the recall and FPR for the classical ML and BERT models. Fig. 4 shows the receiver operating characteristic (ROC) curve for the models. We observe that the BERT model outperforms other models with an average recall of 0.81 over the FPR range of 0.04 to 0.25. Amongst classical machine learning models, CatBoost showed the best performance with an average recall of 0.79 over the same FPR range. The difference in performance between BERT and CatBoost increases as the FPR increases (Fig. 4). Similarly, the classical ML models have similar performance at higher FPR (0.25) compared to performance at low FPR (0.04).

2) *Impact of text data and raw image on ViT-BERT performance*: We assess the improvement in predictive performance obtained by integrating different combinations of attributes, particularly for the raw image and text description attribute omitted in Section IV-C.1. To ensure relevance to

TABLE III
MODEL RECALL AT TARGET FALSE POSITIVE RATE LEVELS

Models	FPR						
	0.04	0.06	0.08	0.10	0.15	0.20	0.25
BERT [17]	0.73	0.75	0.78	0.80	0.84	0.87	0.90
CatBoost [18]	0.71	0.74	0.77	0.78	0.82	0.84	0.86
Logistic Regression [28]	0.68	0.73	0.75	0.77	0.83	0.85	0.87
XGBoost [11]	0.69	0.71	0.74	0.75	0.80	0.83	0.87
Random Forest [8]	0.64	0.68	0.77	0.75	0.79	0.83	0.86

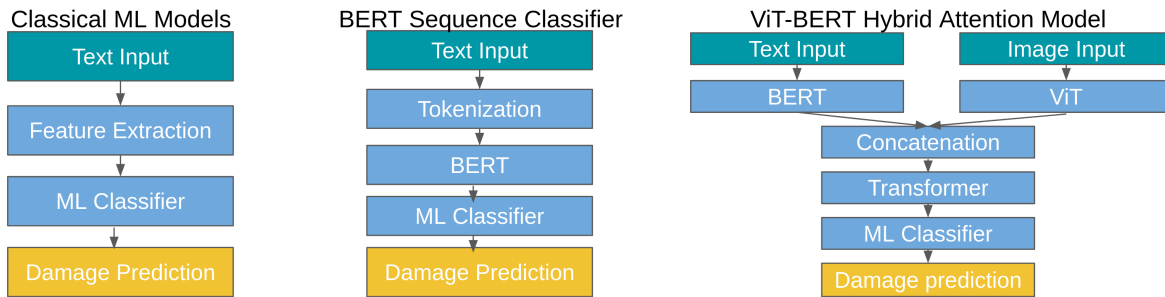


Fig. 3. The block diagram illustrates three different models used for damage avoidance. The ‘Classical ML Models’ takes text input derived from Catalog, Workcell, and Experience, and goes through a ‘Feature Extraction’ step to convert this information into an array of structured numerical data, which is then processed by the ML classifier models. The ‘BERT Sequence Classifier’ takes the same text information, concatenates it into a sentence, and then tokenizes it for processing with BERT. Finally, the ‘ViT-BERT Model’ takes text input as well as raw image as ‘Image Input’. This model uses a hybrid attention mechanism that processes both text and image data to predict damage.

other robotic manipulation systems, we’ve chosen a subset of attributes comprising ‘Product group’, ‘Dimensions’, ‘Weight’, ‘Package’ and ‘Visual’, represented as categorical and numerical values. Furthermore, we also include “Description”, described in unstructured human language, and the raw image.

Our evaluation uses the ViT-BERT model, as it is the only model capable of simultaneously processing raw images alongside other attribute types. During text data attribute analysis, the ViT head was deactivated, making the model the same as the BERT Sequence classifier. Recall rates were measured at an FPR of 0.08 by modifying the confidence threshold in binary classification.

Table IV shows the impact of adding additional attributes to the predictive power of our model. Excluding text and raw image data, the recall rate of the model is 0.75. Adding text data to the model increases the recall by 2% to 0.77. We believe that BERT Sequence classifier’s ability to process natural language present in object description enables it to outperform models that cannot process text data.

To utilize raw images in our model, we activate the ViT head of the ViT-BERT model during training. Using raw images as an input to the model increases the recall by 1.5%,

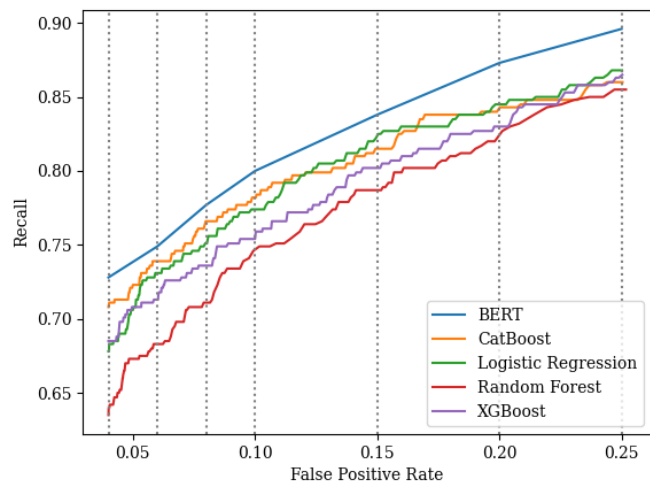


Fig. 4. Recall at False Positive Rate over the range of 0.04-0.25.

TABLE IV
INPUT FEATURE ABLATION STUDY

Feature Combination							Recall @ 0.08 FPR
product group	dimensions	weight	package	visual	item description	raw image	
✓							0.682
✓	✓						0.705
✓	✓	✓					0.718
✓	✓	✓	✓				0.733
✓	✓	✓	✓	✓			0.751
✓	✓	✓	✓	✓	✓		0.771
✓	✓	✓	✓	✓	✓	✓	0.786

validating the hypothesis that raw image adds additional value to object-level features. While the 1.5% improvement is promising, we believe there is potential to further optimize the ViT-BERT model’s performance by preparing the ViT-BERT backbone on large-scale, task relevant data, followed by fine-tuning specifically for damage avoidance to fully harness the multi-modal transformer capabilities.

3) *Feature-wise evaluation*: Finally, we breakdown the importance of single attribute groups to the damage avoidance model. Table V shows the recall for each attribute at an FPR of 0.03 as this is one of the few FPR values at which we can study the impact of all attributes individually. We observe that object description, product group, and visual attributes have high recall rates of 0.59, 0.54, and 0.51 respectively, demonstrating a strong correlation between those attributes and the likelihood of damage. Weight and package attribute have a lower recall rate in comparison (< 0.3), showing a weaker correlation.

V. ROBOT EXPERIMENT

To assess the impact of our damage avoidance system on mitigating object damage, we orchestrated a comprehensive experiment within our robotic manipulation workcell as outlined in Section III-A. This experiment involved 50,687 pick-and-place operations, categorized into control and treatment conditions for comparative analysis.

TABLE V
SINGLE FEATURE IMPORTANCE AT 0.03 FPR.

Input feature	product group	dimen- sions	weight	package	visual	item descrip- tion
Recall @ 0.03 FPR	0.543	0.459	0.132	0.279	0.510	0.589

An experiment induct is conducted as follows: Before each pick, rgb and depth images are taken of the container and an object identification system provides a unique identifier of the object. The image is used to extract workcell attributes and the identifier is used to compute the experience attributes and retrieve catalog attributes. Next, the damage avoidance model is used to determine if the object would be damaged if manipulated by the robot. We categorized our experiment into control and treatment conditions for comparative analysis. In the control group, our system estimated the likelihood of damage, but proceeded with the picking operation regardless (damage avoidance OFF). In the treatment group, the system avoided picking objects identified as having a high likelihood of damage (damage avoidance ON). The distribution between control and treatment operations was approximately even, adhering to a near 50:50 split.

We use two metrics to measure the performance of damage avoidance at the workcell: 1) **damage rate reduction**, which we define as the ratio of damaged objects to the total number of objects picked by the workcell. The damage rate reduction refers to the decrease in the damage rate between the control and treatment groups. 2) **volume loss**, identified as the increase in the proportion of objects not picked by the workcell due to damage avoidance. This is estimated by the difference in the percentage of objects picked robotically between the control and treatment groups. The damage rate reduction is mostly proportional to the recall rate of the damage avoidance model, and the volume loss is strongly correlated with the false positive rate of our model. Most objects are not picked because of a positive damage prediction or because of occlusion by objects with a positive damage prediction.

The BERT Sequence Classifier model is used in the robot experiment as the damage avoidance system due to ease of integration compared to the ViT-BERT model and better avoidance performance compared to classical ML models.

Experiment results show that the damage avoidance system reduces the damage rate by 64%. This is achieved at a volume loss of 9% between the control and treatment conditions i.e., we do not pick an additional 9% of objects when the damage avoidance system is ON.

We note variations in the damage rate across different days, which may be contributed to the diversity and change in the types of items arriving at the testing station. Moreover, due to errors in item identification, incorrect attribute data is retrieved, which reduces the performance of our damage avoidance system.

We analyze the reduction in damage rate between the control and treatment conditions for object types that are highly prone to deconstruction and open damage. The model reduces damage to books and shoe boxes by 97% and 52% respectively. The reduction in damage rate for tuck-lid and telescoping boxes are less than 50%.

VI. DISCUSSION AND FUTURE WORK

We present a taxonomy that captures the modes of damage observed during robotic manipulation of objects in a warehouse. The types of damage range from simple effects such as crease or blemish that affect the appearance of the object to catastrophic damage such as deconstruction where the object separates into multiple parts. The different types of damage are influenced by object properties and the type of end-effector used. Certain types of damage such as crush, tear, superficial could be reduced by improving manipulation policies. However, damage such as deconstruction and opening of the packages highlight a limitation of vacuum-based end-effectors. We believe that high volume of objects encountered in Amazon Fulfillment Centers and the prevalence of vacuum-based end-effectors in warehouse automation means that our damage classification could apply to other robotic manipulation processes.

We introduce a damage avoidance system designed to utilize object attributes for predicting potential damage during robotic handling in our workcell. Our evaluation spans various models, including traditional machine learning approaches, large language models, and a novel ViT-BERT model that integrates textual and visual data to assess damage risk. Our findings indicate that large language models outperform others in accuracy. Meanwhile, gradient boosting techniques like CatBoost, although slightly less effective, offer a competitive performance at a reduced computational expense. To verify our model’s efficacy, we carried out extensive testing within a robotic manipulation workcell located in an Amazon warehouse. The results demonstrate our system’s ability to mitigate object damage effectively, even amidst the inventory variations and operational errors typical of real-world conditions.

In the future, we plan to incorporate additional workcell parameters into our model. We believe that workcell parameters such as the pose of the object being picked and the parameters of the end-effector such as amount of vacuum pressure applied, number of cups engaged, etc. could further improve the performance of our damage avoidance models.

VII. ACKNOWLEDGEMENTS

We would like to thank the Sparrow [2] team members for the deployment and operation of the robotic workcell, Akshay Bhalla for the integration support on our system, and Mateusz Zwieniecki for data acquisition and analysis support. We would like to express our gratitude to Parris Wellman, Clay Flannigan, Joey Durham, Andy Marchese, Jane Shi, Kapil Katyal, and Tamir Hegazy for their valuable feedback.

REFERENCES

- [1] "AI-powered robot warehouse pickers are now ready to go to work," <http://tinyurl.com/3rhdf3>, (Accessed on 09/05/2023).
- [2] "Amazon introduces sparrow—a state-of-the-art robot that handles millions of diverse products," <https://tinyurl.com/2p8h4w7v>, (Accessed on 03/06/2023).
- [3] "Introducing next-generation rightpick 3 item-handling robot system," <http://tinyurl.com/5x856b68>, (Accessed on 09/05/2023).
- [4] M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-learning-based disease diagnosis: A comprehensive review," *Healthcare (Basel, Switzerland)*, vol. 10, no. 3, p. 541, 2022, the authors declare no conflict of interest. [Online]. Available: <https://doi.org/10.3390/healthcare10030541>
- [5] H. O. Alanazi, A. H. Abdullah, and K. N. Qureshi, "A critical review for developing accurate and dynamic predictive models using machine learning methods in medicine and health care," *Journal of Medical Systems*, vol. 41, no. 4, p. 69, 2017. [Online]. Available: <https://doi.org/10.1007/s10916-017-0715-6>
- [6] A. Antonucci, G. P. R. Papini, P. Bevilacqua, L. Palopoli, and D. Fontanelli, "Efficient prediction of human motion for real-time robotics applications with physics-inspired neural networks," *IEEE Access*, vol. 10, pp. 144–157, 2022.
- [7] M. Baptista, S. Sankararaman, I. P. de Medeiros, C. Nascimento, H. Prendinger, and E. M. Henriques, "Forecasting fault events for predictive maintenance using data-driven techniques and arma modeling," *Computers & Industrial Engineering*, vol. 115, pp. 41–53, 2018. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S036083521730520X>
- [8] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: <https://doi.org/10.1023/A:1010933404324>
- [9] R. Calandra, A. Owens, M. Upadhyaya, W. Yuan, J. Lin, E. H. Adelson, and S. Levine, "The feeling of success: Does touch sensing help predict grasp outcomes?" 2017.
- [10] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *CoRR*, vol. abs/1812.08008, 2018. [Online]. Available: <http://arxiv.org/abs/1812.08008>
- [11] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *CoRR*, vol. abs/1603.02754, 2016. [Online]. Available: <http://arxiv.org/abs/1603.02754>
- [12] A. Chlingaryan, S. Sukkarieh, and B. Whelan, "Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review," *Computers and Electronics in Agriculture*, vol. 151, pp. 61–69, 2018. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0168169917314710>
- [13] A. Choudhary, T. Mian, and S. Fatima, "Convolutional neural network based bearing fault diagnosis of rotating machine using thermal images," *Measurement*, vol. 176, p. 109196, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S02632242121002141>
- [14] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," *CoRR*, vol. abs/1604.01685, 2016. [Online]. Available: <http://arxiv.org/abs/1604.01685>
- [15] Covariant, "Robotic induction," <https://covariant.ai/robotic-induction/>, 2023.
- [16] T. Cvitanic and S. N. Melkote, "A new method for closed-loop stability prediction in industrial robots," *Robotics and Computer-Integrated Manufacturing*, vol. 73, p. 102218, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0736584521001010>
- [17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [18] A. V. Dorogush, A. Gulin, G. Gusev, N. Kazeev, L. O. Prokhorenkova, and A. Vorobev, "Fighting biases with dynamic boosting," *CoRR*, vol. abs/1706.09516, 2017. [Online]. Available: <http://arxiv.org/abs/1706.09516>
- [19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020. [Online]. Available: <https://arxiv.org/abs/2010.11929>
- [20] S. Gawde, S. Patil, S. Kumar, P. Kamat, K. Kotecha, and A. Abraham, "Multi-fault diagnosis of industrial rotating machines using data-driven approach : A review of two decades of research," *Engineering Applications of Artificial Intelligence*, vol. 123, p. 106139, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0952197623003238>
- [21] R. B. Girshick, "Fast R-CNN," *CoRR*, vol. abs/1504.08083, 2015. [Online]. Available: <http://arxiv.org/abs/1504.08083>
- [22] D. Gonzalez Jimenez, J. del Olmo Larrañaga, J. Poza, F. Garramiola, and P. Madina, "Data-driven fault diagnosis for electric drives: A review," *Sensors*, vol. 21, p. 4024, 06 2021.
- [23] B. Grey, "Pick and pack," 2023, available from: <https://www.berkshiregrey.com/portfolio/products/pick-and-pack/>. [Online]. Available: <https://www.berkshiregrey.com/portfolio/products/pick-and-pack/>
- [24] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," *CoRR*, vol. abs/1703.06870, 2017. [Online]. Available: <http://arxiv.org/abs/1703.06870>
- [25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: <http://arxiv.org/abs/1512.03385>
- [26] C. Jin, W. Zhao, Z. Liu, J. Lee, and X. He, "A vibration-based approach for diesel engine fault diagnosis," in *2014 International Conference on Prognostics and Health Management*, 2014, pp. 1–9.
- [27] D. Kiela, S. Bhooshan, H. Firooz, and D. Testuggine, "Supervised multimodal bitransformers for classifying images and text," *CoRR*, vol. abs/1909.02950, 2019. [Online]. Available: <http://arxiv.org/abs/1909.02950>
- [28] S. learn Developers. (2021) LinearRegression — scikit-learn 0.24.2 documentation. Accessed: 09-06-2023. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
- [29] Z. Li, J. Li, Y. Wang, and K. Wang, "A deep learning approach for anomaly detection based on sae and lstm in mechanical equipment," *The International Journal of Advanced Manufacturing Technology*, vol. 103, 07 2019.
- [30] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 936–944.
- [31] H. Liu and L. Wang, "Human motion prediction for human-robot collaboration," *Journal of Manufacturing Systems*, vol. 44, pp. 287–294, 08 2017.
- [32] Z. Liu, Q. Liu, W. Xu, Z. Liu, Z. Zhou, and J. Chen, "Deep learning-based human motion prediction considering context awareness for human-robot collaboration in manufacturing," *Procedia CIRP*, vol. 83, pp. 272–278, 01 2019.
- [33] Y. Mekonnen, S. Namuduri, L. Burton, A. Sarwat, and S. Bhansali, "Review—machine learning techniques in wireless sensor network based precision agriculture," *Journal of The Electrochemical Society*, vol. 167, no. 3, p. 037522, dec 2019. [Online]. Available: <https://dx.doi.org/10.1149/2.0222003JES>
- [34] C. Mitash, F. Wang, S. Lu, V. Terhija, T. Garaas, F. Polido, and M. Nambi, "ARMBench: An object-centric benchmark dataset for robotic manipulation," *arXiv preprint arXiv:2303.16382*, 2023.
- [35] K. Palanivel and C. Surianarayanan, "An approach for prediction of crop yield using machine learning and big data techniques," *International Journal of Computer Engineering and Technology*, vol. 10, 06 2019.
- [36] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3d classification and segmentation," *CoRR*, vol. abs/1612.00593, 2016. [Online]. Available: <http://arxiv.org/abs/1612.00593>
- [37] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015. [Online]. Available: <http://arxiv.org/abs/1506.01497>
- [38] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: <http://arxiv.org/abs/1505.04597>
- [39] Rutgers APC RGB-D Dataset, 2015, (Accessed Jan 28, 2019). [Online]. Available: http://pracsyslab.org/rutgers_apc_rgb_d_dataset
- [40] M. Schwarz, C. Lenz, G. M. García, S. Koo, A. S. Periyasamy, M. Schreiber, and S. Behnke, "Fast object learning and dual-arm

- coordination for cluttered stowing, picking, and packing,” *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3347–3354, 2018.
- [41] Z. Si, Z. Zhu, A. Agarwal, S. Anderson, and W. Yuan, “Grasp stability prediction with sim-to-real transfer from tactile sensing,” in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 7809–7816.
- [42] M. Tan and Q. V. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” *CoRR*, vol. abs/1905.11946, 2019. [Online]. Available: <http://arxiv.org/abs/1905.11946>
- [43] M. Tan, R. Pang, and Q. V. Le, “Efficientdet: Scalable and efficient object detection,” *ArXiv*, vol. abs/1911.09070, 2019.
- [44] N. T. Tran, H. T. Trieu, V. T. Tran, H. H. Ngo, and Q. K. Dao, “An overview of the application of machine learning in predictive maintenance,” *Petrovietnam Journal*, vol. 10, pp. 47–61, 2021.
- [45] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, “Medclip: Contrastive learning from unpaired medical images and text,” 2022.
- [46] Y. Wen, M. Fashiar Rahman, H. Xu, and T.-L. B. Tseng, “Recent advances and trends of predictive maintenance from data-driven machine prognostics perspective,” *Measurement*, vol. 187, p. 110276, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0263224121011805>
- [47] XYZ Robotics, “Piece picking robot,” <https://www.xyzrobotics.com/picking-robot/piece-picking>, 2023, available from: <https://www.xyzrobotics.com/picking-robot/piece-picking>.
- [48] K. Yu, N. Fazeli, N. C. Daffe, O. Taylor, E. Donlon, G. D. Lankenau, and A. Rodriguez, “A summary of team mit’s approach to the amazon picking challenge 2015,” *CoRR*, vol. abs/1604.03639, 2016. [Online]. Available: <http://arxiv.org/abs/1604.03639>
- [49] A. Zeng, S. Song, K. Yu, E. Donlon, F. R. Hogan, M. Bauza, D. Ma, O. Taylor, M. Liu, E. Romo, N. Fazeli, F. Alet, N. C. Daffe, R. Holladay, I. Morena, P. Qu Nair, D. Green, I. Taylor, W. Liu, T. Funkhouser, and A. Rodriguez, “Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 3750–3757.
- [50] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, Z.-L. Zhang, H. Lin, Y.-e. Sun, T. He, J. Mueller, R. Manmatha, M. Li, and A. J. Smola, “Resnest: Split-attention networks,” *ArXiv*, vol. abs/2004.08955, 2020.
- [51] W. Zhang, D. Yang, and H. Wang, “Data-driven methods for predictive maintenance of industrial equipment: A survey,” *IEEE Systems Journal*, vol. 13, no. 3, pp. 2213–2227, 2019.
- [52] M. Çelik, M. Işık, O. Yuzugullu, N. Fajraoui, and E. Erten, “Soil moisture prediction from remote sensing images coupled with climate, soil texture and topography via deep learning,” *Remote Sensing*, vol. 14, p. 5584, 11 2022.