MUSASHI AI

MUSASHI AI
WHITE PAPER

Understanding the Metrics for Better Evaluation of Your AI Models

Table of Contents

Brief Summary

The whitepaper highlights the mechanics, role, and practical implementation of key measures that are derived from the confusion matrix. The error matrix or confusion matrix is a uniform and effective way to understand the performance of a model. In this whitepaper, the focus of attention revolves around standard metrics used to evaluate an AI model focusing on the audience in the manufacturing industry.

Introduction

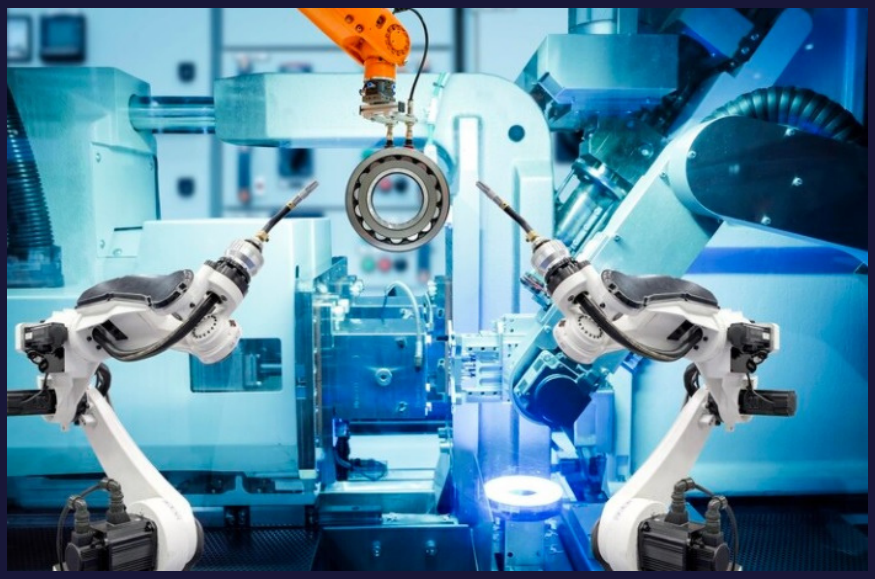
Quality Control through Visual Inspection for defects (and surface imperfections) on the finished parts is one of the most critical manufacturing processes. While manual inspection requires a trained person to be physically present to inspect each final product, most industries still bear a certain number of claims back from their customers.

The subjective nature of few imperfections, unavailability of facilities to book-keep the uncertain parts, speed in which the associates are expected to inspect, and random human errors make this challenging for the authorities. We at Musashi AI are determined in our motto as human jobs for humans. At Musashi AI, we believe that Deep Learning-based Artificial Intelligence (AI) inspection is the ultimate solution for the majority of the concerns faced by modern and contemporary manufacturing facilities.

Industry 4.0

AI-guided Automated Visual Inspection continues to prove as the most cost-effective, efficient, and reliable way to assess the quality of the manufactured parts. The most common Computer Vision tasks that are currently employed in the majority of production facilities are:

- Defect Classification (identifying whether a piece is Good or Bad).
- Anomaly Detection (finding any imperfections based on the AI's acquired knowledge about the good part).
- Defect Detection (locating the defect coordinates with a bounding box based on the past historical data).



- Defect Segmentation (finding the actual polygonal boundaries of the identified defects).

The objective of these algorithms is to evaluate with standard metrics that both researchers in academia and software developers in industries can use. But it is often misunderstood or unclear for the clients who might interact with or use such AI algorithms on daily basis.

With the increased adaption of these models in the manufacturing facilities, it has become imperative to understand and answer what does each metric measure.

NOTE: ALGORITHMS ARE REFERRED TO AS 'MODELS' THROUGHOUT THE WHITEPAPER

Consider you are the Quality Manager of a production plant that produces 1000 units of tires each day. Although the whole manufacturing process is automated, you may make about 20 tires each day that is imperfect according to your customer's standards.

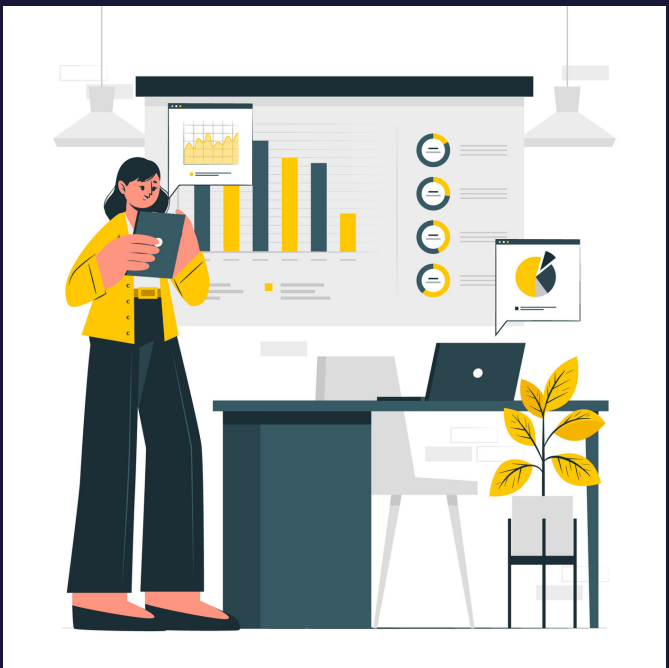
As it's crucial to catch all those defective tires before leaving your facility, you install an AI-based defect inspection model integrated into your conveyor. For example, let's say the AI model correctly identifies 15 among the 20 defective tires by missing 5 defective tires and falsely identifies additional 10 good tires to be faulty.

The Metrics

Some of the standard terms that you might be reported as the AI model's performance through their analytics tool are:

- Model Accuracy
- Precision
- Recall (or Sensitivity)
- F1 score
- Specificity

Though the terms 'accuracy' and 'precision' look similar to you, they are practically different considering an AI model's evaluation



So, before diving into their vague verbal definitions, let's first understand the model's performance. In the tire manufacturing case as mentioned earlier, the objective is to identify all the defects.

- True Positive (TP) value is 15.
  This is the count of the actual defects that the model correctly identified.

- False Positive (FP) value is 10.
  This is the number of the falsely identified good parts.

- False Negative (FN) value is 5.
  This is the count of the actual defective parts that the model missed identifying.

- True Negative (TN) value is 970.
  This is the rest of the good parts that are classified as good according to the model.

  To estimate the model's overall performance, tabulate data into an easy-to-understand 2D matrix, which we term as a Confusion Matrix or an Error Matrix.

Confusion Matrix Fundamentals

A thorough understanding of the confusion matrix makes it easier to understand the overall performance of the model and tabulate the data into the 2D matrix. Confusion Matrix or Error Matrix highlights the model accuracy. Fundamentally, the confusion matrix is used when it comes to classification problems.

n=165	Predicted: No	Predicted: Yes	
Actual: No	FN=50	FP=10	60
Actual: Yes	FN=5	TP=100	105
	55	110	

In multi-level classification issues with several potential outcome values, the confusion matrix determines the model performance. It could be something as simple as email classification [9]. Mathematically, the confusion matrix refers to a two-by-two format table that can produce four possible outcomes through a binary classifier.

Key measures such as accuracy, specificity, error rate, precision, and sensitivity stem from the hallmark confusion matrix. In addition, there are other advanced measures like precision-recall and ROC and that is based on a confusion matrix.

Instead of focusing on true positives, you can denote a confusion matrix in the form of TP, TN, FN, and FP. When there's test data in one or more classification models, the confusion matrix paints a clear picture of performance parameters. Inherently, the confusion matrix touches on the errors related to the model's performance. It is the main reason the confusion matrix form is also referred to as the error matrix.

The Essential Role of Confusion Matrix

The confusion matrix is integral when it comes to reviewing the model's performance. Additionally, the confusion matrix makes predictions based on test data and showcases the weaknesses of a classification model. Predominantly, the confusion matrix identifies the main errors in the classifiers and as well as categorizes the errors in type I and type II format.

Through a confusion matrix, you can calculate a wide range of parameters in a dedicated model. Since precision and accuracy are the hallmark measures, the confusion matrix simplifies the process to understand the performance of the model.

Key Characteristics of Confusion Matrix

- In order to make a prediction about 2 categories through a classifier, the confusion matrix needs a 2 by 2 table. Similarly, 3 categories will require a 3 by 3 table and so forth.
- The confusion matrix needs to be divided into two separate dimensions that represent actual or predicted values along with total forecasted values.
- Predicted values refer to model predictions and actual values that are true to observations.
- Once you analyze the performance metrics, plot precision-recall on suitable tools. Simultaneously, take into account the issues pertaining to ROC curves.
- Objectively, make sure use cases of precision-recall work as a better option.

Confusion Matrix Key Measures

The first two measures of the confusion matrix are vital. ACC (accuracy) and ERR (error rate) are intuitive and foundation parameters.

Error rate

You can calculate the error rate as the total number of incorrect predictions and divide it by the dataset number. The worst error rate is represented as 1.0, whereas the ideal error rate is represented as 0.0.

$$\text{ERR (Error Rate)} = \frac{FN + FP}{TN+TP+FP+FN} = \frac{FN+FP}{N+P}$$

Accuracy

You can calculate the accuracy rate as a total number of correct predictions and divide it by the total dataset number. The worst accuracy rate is 0.0, whereas the perfect accuracy rate is 1.0. You can also calculate the accuracy rate by 1-ERR.

$$\text{ACC (Accuracy Rate)} = \frac{TN+TP}{TN+TP+FP+FN} = \frac{TN+TP}{N+P}$$

Sensitivity

You can calculate sensitivity, positive, or recall rate as the total positive and correct predictions and then divide it by the total positives. In the statistics world, the sensitivity rate is also known as TPR (true positive rate) and REC (recall). The worst recall rate is 0.0, whereas the best is represented as 1.0.

$$\text{SN (Sensivity)} = \frac{TP}{FN+TP} = \frac{TP}{P}$$

Precision

You can calculate precision as the total positive and correct predictions and then divide it by total positive predictions. PREC or positive predictive value with 0.0 represents the worst precision level, whereas PPV with 1.0 represents the high precision.

$$\text{PREC (Positive Predictive Value)} = \frac{TP}{FP+TP}$$

Focus on the “Right” Data Tabulation

To measure the model's overall performance, let's tabulate these data into an easy-to-understand 2D matrix,

	Not Predicted	Predicted
Not Actual	TN = 970	FP = 10
Actual	FN = 5	TP = 15
	975	25

Now, let's break down the common metrics of the AI model in line with the confusion matrix:

Analysis: Implementation of Key Measures

Precision is the proportion of actually defective tires among all the tires that the model identified as faulty. It can be expressed as the ratio TP / (TP+FP) and is defined as "the definitive fraction of crucial instances in all retrieved instances"

Ideally, a model with lower False Positives attains a higher Precision score. However, it is essential to note that Precision doesn't consider the False Negative rate for its evaluation.

As per our example, the model has a Precision of 60% as the model has correctly identified 15 defective tires while falsely predicting 10 good tires to be faulty. Therefore, it was calculated as 15 / (15+10).

Who should care more about Precision?

- Companies that are bound by a considerable degree of False Positives.

  The recall is another metric that directly answers the question - "What proportion of defective tires were identified by our model as defective?". It is defined as "the fraction of essential retrieved instances" and can be expressed by the ratio TP / (TP+FN).

- The lower the False Negative rate, the higher the Recall score for the model. Thus, Recall is also called as Sensitivity of the model.

  Following our example, as the model missed 5 defective tires during its inspection, it has achieved a Recall of 75%, calculated as 15 / (15+5).

Who should care more about Recall?

Manufacturers that don't want even a single unit of the defective part to be shipped to their customers.

Sometimes few customers want to evaluate a model with just one score and understand its overall performance. That's when the F1 score comes in handy. It combines the aforementioned scores intuitively and is defined as the harmonic mean of Precision and Recall.

F1 score is expressed mathematically as 2 * Precision * Recall / (Precision + Recall). For our model, F1 score can be calculated as 2 * 60 * 75 / (60+75) and is equal to 66.67%.

In case if you are wondering, "Why not just take the average of Precision and Recall for a single score?" - Consider a different scenario. One fine day, you find your model not missing even a single defective part (i.e., 100% Recall) but flagging a lot of False Positives (say, only 50% Precision).

The arithmetic mean of both scores will produce a final score of 75%; while the F1 score is still only 66.67% (do the math as an exercise). F1 score naturally adds weightage over the lower rate and brings the absolute value closer to the lower score, thus alerting an unusual behavior with the model.

Specificity is the inverse of Recall and is the proportion of the good tires correctly classified by our model as good. It is calculated by the formula TN / (TN+FP).

As our model only missed 10 good parts, its Specificity score is 970 / (970+10), which is about 98.98%. It's often not widely used to evaluate a model on a manufacturing facility as it weights over the good parts rather than the defective parts.

Accuracy is the ratio of all correctly classified instances over the cumulative number of samples. It is expressed as,

Accuracy = TP+TN / (TP+FP+TN+FN)

For our example case,

= 15+970 / (15+10+970+5)

= 985 / 1000

= 98.5%

Accuracy usually depicts the model's performance adequately only if the com- position of data is balanced. If we are dealing with a significant issue in pro- duction, and almost half of the manufactured part is defective, the accuracy metric may help here.

But as for most defect inspection tasks, as the True Defective part rate is much lower than the True Good part rate, we may always end up with a score of about > 95%.

Conclusion

When it comes to key measures of the confusion matrix, each serves its own purpose and is vital than the other. In terms of use cases, measures like speci- ficity and sensitivity are highly informative than error rate and accurate rate. On the other hand, negative and positive error costs are different.

Most notably, it is imperative to cut out false negatives than false positives in the confusion matrix. If you THINK to evaluate an AI model's performance only based on its Accuracy score without looking into the underlying data distribu- tion, then THINK AGAIN!

Follow us on LinkedIn to know more about us and the values we might bring to your organization. We are looking forward to adding more blogs related to visual inspection, so do watch out for Musashi AI.