DEVELOPPEMENT DE MODELE PREDICTIFS DANS LE CADRE DE L'INDUSTRIE MANUFACTURIERE A GROS VOLUMES

Eva Jabbar [1] Philippe Besse [2]Jean-Michel Loubes[3] Christophe Merle[4]

[1] *IMT - Institut de Mathématiques, Université de Toulouse, CNRS UMR5219, 118 Route de Narbonne, 31400 Toulouse*

[1] *Continental Powertrain France SAS,1 avenue Paul Ourliac, 31036 Toulouse, Eva.jabbar1@gmail.com*

[2] *IMT - Institut de Mathématiques, Université de Toulouse, CNRS UMR5219, 118 Route de Narbonne, 31400 Toulouse, philippe.besse@insa-toulouse.fr*

[3] *IMT - Institut de Mathématiques, Université de Toulouse, CNRS UMR5219, 118 Route de Narbonne, 31400 Toulouse, loubes@math.univ-toulouse.fr*

[4] *Continental Powertrain France SAS, 1 avenue Paul Ourliac, 31036 Toulouse, Christophe.merle@continental.com*

Résumé.

La qualité de la production des cartes électroniques atteint un excellent niveau de performance, avec un taux de rejet extrêmement faible. De ce fait, il devient de plus en plus complexe de déterminer les paramètres/facteurs sur lesquels il faut agir pour améliorer encore davantage la production. Nous avons donc eu recours aux solutions dites data driven à travers l'application des techniques de machine learning comme un moyen d'investigation pour améliorer la qualité et réduire les couts.

Nous nous focalisons dans cette étude sur deux problématiques industrielles. La première concerne la présence de taux élevé de faux défauts à la suite d'inspection optique impliquant une pénibilité de tache des opérateurs et une perte de temps importante. La deuxième concerne l'amélioration de la détection d'anomalies actuellement qui se base sur des définitions univariées des alarmes complexes et dépendante d'une référence de produit et des composants associés.

L'étude de la première problématique a requiert le développement d'un modèle de classification qui soit un outil d'aide à la décision pour les opérateurs. Nous avons comparé les performances des techniques de machine learning à base d'arbres : CART, Random Forest, Adaboost, XGBoost. Le résultat de cette étude positionne XGBoost comme meilleur algorithme avec un score de précision de 99.4% et un rappel de 98.6%. Cette étude est présentée par [Jabbar et al., 2018]. Une étape d'interprétabilité est ensuite étudiée à travers l'analyse des valeurs 'SHapley Additive exPlanations' présenté par [Lundberg and Lee, 2017].

Pour la deuxième problématique, nous proposons l'application de techniques de détection d'anomalies conditionnelles. La méthodologie adaptée est d'identifier d'abord les sous-populations homogènes caractérisant ainsi une même condition. Puis appliquer sur

chaque sous-population des méthodes de détection d'anomalie. Nous avons analysé les performances des Variational Autoencoders [Kingma and Welling, 2013](VAEs) entièrement connectés qui utilisent l'erreur de reconstruction comme score d'anomalie. Les VAEs donnent les meilleurs résultats avec une précision de 97.5%, un rappel de 98.44% et un temps de calcul respectant les contraintes de production (temps réel). Les résultats obtenus ont aussi permis de détecter des anomalies dont certaines ont réellement été identifiées comme des défauts plus tard dans le processus.

Mots-clés. Industrie 4.0, Intelligence artificielle, Outil d'aide à la décision, Production de cartes électroniques, Apprentissage supervisé, Détection d'anomalie conditionnelle, Apprentissage profond, Interprétabilité et Explicabilité.

Abstract.

The production quality of electronic devices has reached an excellent level of performance, with an extremely low rejection rate. As a result, it becomes very complex to determine which parameters/factors need to be addressed for further improvement. We therefore made use of so-called data driven solutions through the application of machine learning techniques as an investigation tool to improve quality and reduce costs.

In this study, we focus on two industrial challenges. The first concerns the presence of high rates of false defects as a result of optical inspection involving a laborious task for operators and a significant loss of time. The second concerns the improvement of the detection of anomalies currently based on univariate definitions of complex alarms that depend on a product reference and related components.

The study of the first problem requires the development of a classification model that will help operators to make decisions. We compared the performance of tree based machine learning techniques: CART, Random Forest, Adaboost, XGBoost. The result of this study positions XGBoost as the best algorithm with an accuracy score of 99.4% and a recall of 98.6%. This study is presented by [Jabbar et al., 2018]. An interpretability step is then studied through the analysis of the SHapley Additive exPlanations values presented by [Lundberg and Lee, 2017].

For the second problem, we propose the application of conditional anomaly detection techniques. The methodology adapted is to first identify the homogeneous subpopulations that characterize the same condition. Then apply anomaly detection methods to each sub-population. We analyzed the performance of fully connected Variational Autoencoders [Kingma and Welling, 2013] (VAEs) that use reconstruction error as an anomaly score. The VAEs give the best results with an accuracy of 97.5%, a recall of 98.44% and a calculation time respecting production constraints (real time). The results achieved allowed to detect anomalies, among them, some were identified as defects later in the process.

Keywords. Industry 4.0, Artificial intelligence, Decision-making tool, Electronic circuit board production, Supervised learning, Conditional anomaly detection, Deep Learning, interpretability and Explainability.

1 Introduction

La qualité de la production des cartes électroniques atteint un excellent niveau de performance, avec un taux de rejet extrêmement faible. De ce fait, il devient de plus en plus complexe de déterminer les paramètres/facteurs sur lesquels il faut agir pour améliorer encore davantage la production. Nous avons donc eu recours aux solutions dites *data driven* à travers l'application des techniques de machine learning comme un moyen d'investigation pour améliorer la qualité et réduire les coûts.

Cette étude concerne le domaine de production des cartes électroniques ou circuit imprimé *printed circuit board* (PCB). Chaque PCB comporte des *pads* où la pâte est déposé et des composants. Les joints de soudure prennent ensuite forme par la *refusion* de la pâte à braser. Dans la ligne de production, nous distinguons deux principales stations d'inspection : Inspection de la pâte à braser *Solder Paste Inspection* (SPI) et l'Inspection optique automatisée *Automated Optical inspection* (AOI). Ces inspections se font sur un panel de PCB qui sont ensuite découpés à la fin de ligne de production.

Une des approches classiques de détection de défaut est la mise en place d'une station d'inspection/contrôle AOI en aval dans le processus de production supporté par une vérification humaine afin de valider les produits en vrais défauts et remettre en production ceux avec un faux défaut. Cette approche est caractérisée par un taux de faux défauts très élevé sur les produits écartés, autour de 95%, ce qui implique une pénibilité de tâche des opérateurs et une perte de temps importante.

Une autre approche utilisée dans la production est la mise en place d'un programme d'inspection SPI basé sur des analyses univariées à travers la définition de limites pour chaque paramètre d'un composant donné et pour chaque référence de produit. Imposer le codage d'alarme complexe limité à chaque paramètre peut engendrer l'envoi d'un grand nombre de fausses alarmes et en même temps ne pas verrouiller les situations à risque combiné.

C'est dans cette perspective que nous nous intéressons au développement de modèles prédictifs en utilisant les données historiques disponibles dans le cadre de l'industrie manufacturière à gros volumes. La chaîne de production est en effet un milieu très contrôlé, suivi par un grand nombre de capteurs divers mélangeant plusieurs types d'informations (capteurs fonctionnels, qualitatifs, booléens, etc.), générant une volumétrie de données très importante.

2 Problématique 1

2.1 Méthodologie

L'objectif de ce travail est de proposer une nouvelle approche basée sur l'apprentissage supervisé afin de réduire le temps nécessaire à la détection des faux défauts en station d'inspection AOI.

Nous nous focalisons sur six types de PCB $P_1, \cdots, P_6$. Chaque produit P_i est associé à une matrice X_i de taille $n_i \times f$ avec n_i mesures et $f = 187$ nombre de *features* comme suit:

$$X_i = \begin{pmatrix} m_{11}^{(i)} & m_{12}^{(i)} & \cdots & m_{1f}^{(i)} \\ m_{21}^{(i)} & m_{22}^{(i)} & \cdots & m_{2f}^{(i)} \\ \vdots & \vdots & \ddots & \vdots \\ m_{n_i1}^{(i)} & m_{n_i2}^{(i)} & \cdots & m_{n_if}^{(i)} \end{pmatrix}, \tag{1}$$

où le nombre de *pads* n_i varie selon le type de PCB. Chaque vecteur de *features* $m_{k.}^{(i)}$ est constitué de mesures de soudure SPI et de mesures de refusion. Il est associé à une étiquette Y définie par AOI vérification qui appartient à l'ensemble discret :

$$\{good, false\ call, real\ defect\}.$$

Comme mentionné précédemment, parmi les algorithmes existants, nous nous concentrons sur une méthode d'apprentissage basée sur l'arbre CART [Breiman et al., 1984], et plusieurs améliorations de cette méthode par agrégation : Random Forest [Breiman, 2001], en utilisant la méthodologie de boosting AdaBoost[Yoav and Robert, 1997] et enfin l'un des derniers algorithmes d'apprentissage XGBoost [Tianqi and G., 2016].

Les résultats de cette étude place XGBoost en première position avec un score de précision de 99.4% et un rappel de 98.6%.

Une étape d'interprétabilité est ensuite étudiée à travers l'analyse des valeurs 'SHapley Additive exPlanations' présenté par [Lundberg and Lee, 2017]. L'idée de cette technique est de calculer la répartition des *features* sur la base des démonstrations de la théorie des jeux. Nous discutons les résultats d'interprétation générale à travers le classement des paramters les plus influents sur la prédiction ainsi qu'une interprétation générale pour une meilleures compréhension des valeurs des paramètres impactant la prévision des classes.

3 Problématique 2

3.1 Méthodologie

Dans cette approche, nous nous focalisons sur une source de données qui correspond à la première station de mesure dans la ligne de production. Nous avons choisi de développer une approche nonsupervisée de détection d'anomalie à ce stage en amont afin d'identifier le plus tôt possible des éventuelles anomalies. Cette solution apporte aussi une amélioration importante au systeme actuel d'inspection qui se base sur le codage de seuil paramètre par paramètre. Ce système ne prend donc pas en compte les éventuelles interactions entre ces différents paramètres.

Nous supposons que nous avons un échantillon $X = (x_1, \cdots, x_{d_m}, r_1, \cdots, r_{d_c}) = (x, r)$ qui ont deux ensembles de variables (x, r) où x représente l' ensemble des paramètres de mesure et r représente les paramètres des condition. Dans notre cas, les r *features* ont été sélectionnées selon expertise des ingénieurs de production , à savoir la référence du produit et le package du composant.

Chaque ligne des données représente un vecteur de mesures SPI au niveau *pad* d'un composant d'une catégorie k dans un panel d'ID de produit P_i. Notez que k dépend de chaque P_i. Nous désignons par $C_i^k = (P_i, k_i)$ la condition homogène dans notre cas.

L'idée principale est d'apprendre d'abord un modèle de détection d'anomalies robuste et stable basé sur des données de qualité. Pour cela, les données historiques massives sont extraites, prétraitées et ensuite divisées en ensembles de données d'apprentissage, de validation et de test. Comme le nombre de dimensions a un effet significatif sur le fonctionnement de l'apprentissage machine (concept de *fléau de la dimension*), nous apprenons d'abord ce qu'on appelle *representation* en effectuant des techniques de réduction non linéaire de la dimensionnalité pour nous concentrer sur un sous-espace compacté conditionnellement à C_i^k. Le modèle normal conditionnel est alors formé. L'objectif est de trouver la région où la plupart des points sont situés sur l'espace du *manifold* de faible dimension. À l'aide des données de validation, le score d'anomalie est calculé pour chaque C_i^k. Pour une identification rapide et globale de l'anomalie, il est nécessaire d'agréger tous ces scores conditionnels au niveau du panel pour permettre à l'expert process d'identifier d'abord le produit dévié et ensuite d'apporter les actions correctives appropriées. Le seuil le plus approprié est ensuite choisi en fonction des exigences de qualité. La configuration optimale consiste à détecter le plus grand nombre d'anomalies (critiques et/ou potentielles) avec un taux de fausses alarmes le plus faible. Cette étape est itérative car elle nécessite une évaluation et une analyse approfondies des résultats. Après avoir approuvée une première version du modèle, nous les testons à nouveau sur différents ensembles de données de test. La performance est évaluée en termes de précision, de rappel et de temps de calcul.

Dans ce travail, nous avons présenté, testé et évalué la performance d'une approche de détection d'anomalie conditionnelle profonde basée sur des Variational AutoEncoders (VAEs) [Kingma and Welling, 2013]. Les VAEs donnent les meilleurs résultats avec une précision de 97.5%, un rappel de 98.44% et un temps de calcul respectant les contraintes de production (temps réel). Les résultats obtenus ont aussi permis de détecter des anomalies dont certaines ont réellement été identifiées comme des défauts plus tard dans le processus.

References

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

[Breiman et al., 1984] Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA.

[Jabbar et al., 2018] Jabbar, E., Besse, P. C., Loubes, J.-M., Roa, N. B., Merle, C., and Dettai, R. (2018). Supervised learning approach for surface-mount device production. In Nicosia, G., Pardalos, P. M., Giuffrida, G., Umeton, R., and Sciacca, V., editors, *LOD*, volume 11331 of *Lecture Notes in Computer Science*, pages 254–263. Springer.

[Kingma and Welling, 2013] Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[Lundberg and Lee, 2017] Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *NIPS*, pages 4768–4777.

[Tianqi and G., 2016] Tianqi, C. and G., C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

[Yoav and Robert, 1997] Yoav, F. and Robert, S. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139.