Quality prediction/classification of a production system under uncertainty based on Support Vector Machine

Wahb Zouhri

HESAM UNIVERSITÉ

Arts et Métiers Sciences et Technologies

ÉCOLE DOCTORALE SCIENCES DES MÉTIERS DE L'INGÉNIEUR
[Laboratoire de Conception, Fabrication et Commande – Campus de Metz]

THÈSE

présentée par : **Wahb ZOUHRI**

soutenue le : **10 décembre 2020**

pour obtenir le grade de : **Docteur d'HESAM Université**

préparée à : **École Nationale Supérieure d'Arts et Métiers**

Spécialité : **Génie industriel**

Prédiction/classification de la qualité des systèmes de production sous incertitudes par la méthode des machines à vecteurs supports (SVM)

THÈSE dirigée par :

Pr. DANTAN Jean-Yves

et co-encadrée par :

MCf. HOMRI Lazhar

Jury

M. Jean-Marc LINARES, Professeur, IUT d'Aix Marseille, Aix Marseille Université	Président
M. Nabil ANWAR, Professeur, LURPA, Université de Paris-Saclay	Rapporteur
M. George LIBEROPOULOS, Professeur, Dpt. Mech. Eng., University of Thessaly	Rapporteur
M. Rikard SÖDERBERG, Professeur, Wingquist Laboratory, Chalmers University	Rapporteur
Mme. Hind BRILEL-HAOUZI, Professeure, ENSTIB, Université de Lorraine	Examinatrice
M. Xavier GENDRE, Maître assistant, ISAE-SUPAERO, Université de Toulouse	Examinateur
M. Lazhar HOMRI, Maître de Conférences, LCFC, Arts et Métiers Metz	Encadrant
M. Jean-Yves DANTAN, Professeur, LCFC, Arts et Métiers Metz	Encadrant

THÈSE

Acknowledgements

I would like to begin by thanking my thesis supervisors, Prof. Jean-Yves DANTAN, and Dr. Lazhar HOMRI for their help, their remarks, their confidence and their commitment during the three years of my thesis. I would also like to take this opportunity to thank them for guiding me through the ups and downs and the different phases of my thesis work, as well as for making the last few years fun and entertaining.

I would like to thank the members of my thesis committee: Prof. Nabil ANWAR, Prof. George LIBEROPOULOS, Prof. Rikard SÖDERBERG, Prof. Hind BRILEL-HAOUZI, Dr. Xavier GENDRE and Prof. Jean-Marc LINARES, for devoting their time and expertise to the evaluation of my research work.

Many thanks to my friends and colleagues for making the lab a kind of second home, and without them, these years would have been different. Also many thanks to my friends from outside the lab, for their support and for the good memories over the last three years.

One last thought goes to my family, especially my parents, brother and little sister for trusting me from the beginning and for helping me to be the person I am today.

The Director General of Arts et Métiers authorized the writing of this thesis in English in order to apply for the European label. Therefore, the manuscript includes an extended summary in French (without tables and figures) and a detailed description of the research in English (with tables and figures).

Résumé étendu en Français

Avant-propos

Les travaux de recherche de cette thèse de doctorat ont été menés au sein de l'école d'Arts et Métiers au campus de Metz, au LCFC (Laboratoire de Conception, Fabrication et Contrôle), où les activités consistent à développer les futurs systèmes de production dans les secteurs des services et de l'industrie.

La thèse de doctorat est réalisée dans le cadre d'une chaire de recherche industrielle intitulée "Systèmes de production reconfigurables - sûrs - performants" en partenariat avec Thyssenkrupp, le Fonds européen de développement régional "Programme opérationnel FEDER-FSE Lorraine et Massif de Vosges 2014- 2020", l'UIMM F2I, et l'UIMM Lorraine. L'objectif de cette chaire est de fournir des solutions industrielles qui améliorent la flexibilité et la réactivité des systèmes de production face aux fluctuations de la demande, qui améliorent leur efficacité, qui facilitent leur maintenance, qui gèrent les risques opérationnels et qui améliorent leurs performances en prenant en compte tous les facteurs, y compris les facteurs humains. L'un des défis de cette chaire de recherche est :

Caractérisation des performances des systèmes de production complexes et modulaires : comment permettre une meilleure compréhension des performances des systèmes de production par des méthodes d'apprentissage automatique ?

La thèse se focalise sur ce défi, et donc sur la manière dont l'apprentissage automatique peut être utilisé pour améliorer la qualité des systèmes de production, en tenant compte des incertitudes rencontrées dans ces systèmes. Il est également important de préciser que les travaux de recherche ont été réalisés en collaboration avec des partenaires industriels, en utilisant leur expertise et leurs données industrielles pour le développement et l'évaluation de nouvelles approches robustes.

De plus, pendant la période de la thèse, j'ai effectué une mobilité de 6 mois consécutifs du 1er septembre 2019 au 28 février 2020 au *WBK Institut für Produktionstechnik* à *Karlsruhe Institut für Technologie* (KIT). Cette mobilité s'est effectuée dans le cadre du Collège doctoral franco-allemand entre Arts et Métiers et KIT, et elle visait à combiner les connaissances de WBK, et celles du LCFC, afin de travailler sur un projet commun qui consiste à proposer une nouvelle approche pour l'amélioration de la qualité d'un procédé de fabrication additive.

Introduction

Dans le contexte de l'industrie 4.0, en raison de l'évolution rapide de la demande, les systèmes de production et de fabrication sont confrontés à divers enjeux tels que l'estimation des coûts, la gestion de la chaîne d'approvisionnement, la prévision de la demande, la planification des ressources de l'entreprise, l'optimisation des processus, l'ordonnancement, le séquençage, l'organisation des cellules et le contrôle de la qualité, etc. Ces systèmes doivent être flexibles et réactifs aux demandes du marché afin de rester compétitifs, tout en répondant à différents objectifs en termes de coût, de temps et de qualité.

La gestion de la qualité est l'un des sujets les plus importants qui doivent être pris en compte pour répondre aux besoins uniques, améliorer la satisfaction des clients, identifier les goulets d'étranglement de la production, assurer une gestion efficace et une administration efficiente, du fait que la gestion de la qualité influence tous les domaines de chaque industrie manufacturière. De nos jours, la qualité doit être améliorée en continu et de nouveaux processus d'évaluation de la conformité doivent être acceptés dans la production, car les attentes des clients augmentent. Diverses stratégies ont été proposées pour gérer et améliorer la qualité des systèmes de production, telles que l'amélioration continue, la gestion de la qualité totale, le système de production Toyota, la fabrication de classe mondiale, le juste à temps et la méthode Six-Sigma, qui encouragent les responsables à collecter des données pour traiter les problèmes de qualité. Toutefois, ces concepts n'ont pas été conçus pour faire face à un contexte aussi dynamique, évolutif et complexe. Cela a incité les industries manufacturières à rechercher de nouvelles solutions basées sur l'innovation et la mise en œuvre de nouvelles technologies.

En conséquence, les industries manufacturières ont opté pour des technologies innovantes de collecte et d'analyse de données afin d'améliorer la qualité de leurs systèmes de production. Un grand volume de données est collecté chaque jour dans chaque système de production. Cependant, le stockage et l'analyse d'une telle quantité de données présente de nombreux nouveaux obstacles et défis. Cela a conduit à la naissance d'un domaine de recherche appelé "découverte de connaissances à partir de bases de données" *(KDD)*, communément appelé *"Data Mining" (DM)*, qui vise à extraire des données des informations inconnues et potentiellement précieuses afin de résoudre les problèmes de qualité et de contrôle. Le processus de KDD est basé sur des techniques de prétraitement, des algorithmes d'apprentissage automatique et des méthodes de visualisation pour la découverte et la présentation de connaissances utiles et de patterns cachés. Parmi les diverses tâches de qualité, la prédiction et la classification des données constituent la partie la plus importante des différentes étapes et sous-activités du processus KDD. Une analyse est donc obligatoire pour l'identification et la sélection de méthodes appropriées pour l'analyse des données de qualité.

Le domaine du *DM* est très diversifié, et le fait de disposer de nombreuses méthodes et algorithmes pourrait être considéré comme une arme à double tranchant. D'une certaine

manière, ces méthodes peuvent offrir un large éventail de solutions aux problèmes de qualité, mais il pourrait être difficile de trouver une méthode adaptée parmi celles qui existent. Ces techniques visent à identifier la relation entre les paramètres de qualité (entrées et sorties), à prédire la qualité de la sortie, à classifier la qualité des produits et à optimiser les entrées pour obtenir une sortie optimale. Ces différentes tâches peuvent être classées en quatre tâches principales, qui sont : la description de la qualité, la prédiction de la qualité, **la classification de la qualité** et l'optimisation des paramètres.

Dans ce travail, nous nous concentrons sur la tâche de classification de la qualité, puisque la plupart des industries manufacturières peuvent fournir des données étiquetées. Plusieurs méthodes et outils de classification ont été utilisés ces dernières années afin de contrôler la qualité du processus et d'augmenter la productivité des systèmes de fabrication, tel que les arbres de décisions *(DT)*, les réseaux de neurones *(ANN)*, et les machine à vecteur de support *(SVM)*. Ce travail de recherche se focalise particulièrement sur l'application de la SVM pour la classification de la qualité. La SVM a montré de bonnes performances lorsqu'il s'agit de traiter divers problèmes de qualité dans différentes industries manufacturières. De plus, la SVM est connue pour être un outil facile à utiliser, capable de traiter de grands ensembles de données avec de nombreux paramètres.

Cependant, les données manufacturières sont généralement sujettes à des incertitudes qui affectent les valeurs des paramètres ou les valeurs des classes. L'incertitude peut être causée par une perception ou une compréhension limitée de la réalité, par exemple, les limitations des ressources pour collecter, stocker, traiter, analyser ou comprendre les données. Ainsi, la précision de la prédiction de la méthode SVM sera influencée. Un défi central de cette thèse est d'étudier et d'analyser l'impact des incertitudes de mesures sur la qualité de la classification.

En résumant les observations de la section précédente, l'objective global peut être défini comme suit :

Comment évaluer l'impact des incertitudes de mesure sur les performances prédictives de la SVM, afin d'améliorer la qualité des systèmes de production ?

Pour atteindre cet objectif général, trois questions de recherche doivent être abordées :

1- *Comment les algorithmes d'apprentissage machine et les approches de classification peuvent-ils être utiles pour l'amélioration/évaluation de la qualité au sein des systèmes de fabrication ?*
2- *Comment quantifier l'impact des incertitudes de mesure sur les performances de la méthode de classification du SVM ?*
3- *Comment améliorer la robustesse de la classification SVM en ce qui concerne les incertitudes de mesure ?*

Chapitre 1 : la gestion de la qualité des systèmes de production par les outils d'apprentissage automatique : Machine à Vecteurs de Support (SVM)

1.1 - Machine learning au sein des industries manufacturières

L'apprentissage automatique (*machine learning ou simplement ML en anglais*) est un domaine d'étude de l'intelligence artificielle qui utilise des approches mathématiques et statistiques pour donner aux ordinateurs la capacité d'apprendre à partir de données, c'est-à-dire d'améliorer leurs performances dans la résolution de tâches sans être explicitement programmés pour chacune d'entre elles. L'histoire de l'apprentissage machine remonte aux années 1950, lorsque la machine d'apprentissage Turning a été introduite, en même temps que les premiers ordinateurs à réseau neuronal. Depuis, de nombreux progrès ont été réalisés dans le domaine du ML. De nombreuses industries ont reconnu que l'apprentissage automatique peut augmenter la capacité de calcul, et elles y consacrent donc davantage de recherches pour rester en tête de la concurrence.

Le domaine de la ML a atteint différents domaines au cours des deux dernières décennies. En médecine et en soins de santé, les techniques de ML ont été utilisées pour prévenir ou guérir certaines maladies. Dans la finance, les algorithmes de ML sont utilisés dans le commerce, la gestion des risques et l'automatisation des processus. Les techniques de ML sont également utilisées pour relever les défis de la durabilité en agriculture. En outre, dans le secteur manufacturier, des méthodes d'apprentissage automatique sont utilisées pour satisfaire efficacement la demande complexe et fluctuante de produits de haute qualité. Autres domaines d'application des techniques de ML sont l'éducation, la bio-informatique et la neuro-imagerie, la cyber-sécurité, la vente en détail, les réseaux sociaux, les énergies renouvelables, les prévisions météorologiques, le contrôle des épidémies, etc.

Aujourd'hui, l'industrie manufacturière est l'un des principaux secteurs qui bénéficient de l'apprentissage machine. Avec l'émergence du paradigme de l'industrie 4.0, les industries manufacturières collectent davantage de données de leurs systèmes de production. Par conséquent, il devient essentiel de trouver des outils pour organiser, synthétiser et analyser ces données afin d'obtenir des informations précieuses et efficaces. Ce traitement des données est par conséquent basé sur des outils de ML.

Plusieurs classifications des algorithmes d'apprentissage automatique ont été proposées dans le contexte de la fabrication. L'une de ces classifications consiste à diviser les algorithmes ML en trois catégories principales, à savoir :

- ***L'apprentissage non supervisé*** : se réfère à la situation d'apprentissage machine dans laquelle les données ne sont pas étiquetées. Les algorithmes tentent de découvrir des modèles utiles qui se cachent derrière ces données, ou d'identifier des groupes à partir de celles-ci.

- ***L'apprentissage par renforcement*** : consiste à former un agent autonome à prendre une série de décisions dans un environnement complexe et incertain. En fonction de la décision prise, l'environnement fournit à l'agent soit des récompenses, soit des pénalités. Par conséquent, l'agent vise à trouver une solution optimale en maximisant les récompenses totales.
- ***L'apprentissage supervisé*** : consiste à apprendre un modèle de prédiction à partir d'exemples étiquetés, par opposition à l'apprentissage non supervisé. On distingue les problèmes de régression et les problèmes de classification, où une variable quantitative est prédite dans la régression, et une variable qualitative est prédite dans les problèmes de classification.

L'apprentissage supervisé est surtout utilisé dans le secteur manufacturier, car les évaluations des experts sont disponibles, les données étiquetées sont souvent déjà établies, et les performances des algorithmes peuvent être facilement évaluées en utilisant certaines mesures comme la précision. Les avantages des aspects prédictifs des algorithmes d'apprentissage supervisé dans le secteur manufacturier ont été résumés en trois points principaux :

- ***Réduction des coûts :*** les informations sur l'état des systèmes de production peuvent être utilisées pour déterminer quand la maintenance est la plus nécessaire. De cette façon, les utilisateurs peuvent maximiser l'utilisation des composants et des consommables des machines.
- ***Efficacité opérationnelle :*** en connaissant les modes de dégradation, on peut déduire les événements de défaillance. Cette connaissance permettra aux experts de la production et de la maintenance de planifier leurs activités de manière collaborative, maximisant ainsi la disponibilité des équipements.
- ***Amélioration de la qualité des produits :*** la qualité des produits peut être maintenue à des niveaux acceptables en connaissant la façon dont les performances des systèmes de production dérivent dans le temps, et en intégrant ces connaissances dans les contrôles de processus, évitant ainsi les produits défectueux et les rebuts inutiles.

En raison de ces avantages, les algorithmes d'apprentissage supervisé sont abordés dans ce travail doctoral, et en particulier, les méthodes de classification sont utilisées afin d'améliorer la qualité des systèmes de production.

1.2 - Evaluation de la qualité des systèmes de production par machine learning

Avec l'émergence de l'industrie 4.0, les approches basées sur l'apprentissage automatique ont été de plus en plus utilisées afin d'améliorer la qualité des systèmes de production. Différents travaux de recherche ont présenté les problèmes de qualité dans les industries manufacturières et la façon dont ils sont traités à l'aide des approches d'apprentissage automatique. Ils ont examiné quatre tâches principales en matière de qualité, qui sont les suivantes :

- ***Description de la qualité*** : en identifiant les facteurs ou les variables d'entrée qui affectent de manière significative la qualité du produit.
- ***Prévision de la qualité*** : en établissant des modèles qui prédisent la valeur du paramètre de sortie sur la base des valeurs des paramètres d'entrée.
- ***Classification de la qualité*** : en prédisant la classe de la qualité dans le cas où elle est indiquée comme variable binaire ou nominale (par exemple : classification des défauts de qualité).
- ***Optimisation des paramètres*** : en définissant les niveaux optimaux des paramètres clés qui affectent la qualité du processus/produit.

Sur la base d'une recherche bibliographique, il a été déduit que les techniques de ML les plus utilisées pour la classification de la qualité au sein des industries manufacturières sont les arbres de décision (DT), les réseaux de neurones artificiels (ANN) et les machines à vecteurs de support (SVM). L'application de ces trois méthodes pour la gestion de la qualité a été étudiée et analysée, ce qui a permis de tirer les conclusions suivantes :

- ***Arbres de décision*** : leurs principales utilisations peuvent être résumées comme étant l'identification des causes des défaillances des systèmes de production, ainsi que l'identification des paramètres de production optimaux. L'adoption des arbres de décision est due à leur facilité d'interprétation et de mise en œuvre. Toutefois, on peut constater que les arbres de décision sont rarement utilisés seuls, où ils sont généralement complétés par d'autres méthodes de prétraitement et d'élagage afin d'éviter la génération de modèles surajustés.
- ***Machines à vecteurs de support*** : en analysant différents travaux, la SVM a montré son potentiel pour traiter de divers problèmes de qualité au sein de différents systèmes de production. L'efficacité de la SVM est due à sa capacité à obtenir de bons résultats sur des données non linéairement séparables, ce qui est généralement le cas des données de qualité, ainsi qu'à sa capacité à éviter la création de modèles surajustés. En outre, la SVM est considérée comme une méthode facile à utiliser qui permet de traiter de grands ensembles de données dimensionnelles ayant des caractéristiques différentes, ce qui représente un avantage par rapport aux autres méthodes de classification.
- ***Réseaux de neurones artificiels (MLP)*** : les différentes études de recherche ci-dessus démontrent l'efficacité des MLPs dans la réalisation de différentes tâches de qualité liées aux systèmes de production. Nombre de ces travaux se sont concentrés sur la capacité des MLPs à détecter les conditions de processus défectueuses, à identifier les produits défectueux et à classer les types de défauts. Les MLPs ont également montré de grandes performances dans la modélisation de problèmes non linéaires complexes, ainsi qu'une bonne capacité de généralisation.

Malgré ces avantages, les classificateurs mentionnés précédemment supposent que tous les éléments d'un jeu de données sont précis. Cette hypothèse peut être violée dans de nombreuses applications du monde réel en raison des imperfections des outils de mesure, de l'imprécision des informations des experts, etc. Il est donc nécessaire d'analyser et de

contrôler l'impact de l'incertitude/du bruit sur la performance prédictive des méthodes de classification afin de garantir une performance optimale dans le pire des cas.

1.3 – SVM sous incertitudes.

Dans le monde réel, les connaissances sont fondamentalement incertaines ; il en va de même dans le secteur manufacturier : les données mesurables pour l'apprentissage automatique et la fouille de données sont généralement inexactes, incomplètes ou bruyantes. Les systèmes d'apprentissage automatique doivent traiter des données imparfaites et sont donc censés raisonner dans des conditions d'ignorance.

De nombreux formalismes pour expliquer l'incertitude ont été proposés au fil des ans. Ils comprennent essentiellement des méthodes numériques basées sur la théorie des probabilités, la logique floue et la théorie des possibilités, ainsi que des méthodes largement symboliques telles que la logique par défaut et l'argumentation.

Deux formes de sources de bruit sont définies dans le cadre des méthodes de classification, à savoir le bruit d'attribut et le bruit de classe. Le premier peut être soit : des valeurs d'attributs erronées, des valeurs d'attributs manquantes, une distribution de données incohérente et des données redondantes. D'autre part, il existe deux sources de bruit de classe. La première source est l'étiquetage des mêmes exemples qui se produisent plusieurs fois avec des étiquettes différentes, tandis que la seconde source consiste en l'étiquetage d'un exemple avec une étiquette erronée. La seconde source de bruit de classe est couramment rencontrée lorsque deux classes présentent des symptômes similaires. Ces différentes sources de bruit doivent être correctement prises en compte, et leur impact sur la classification doit être étudié et analysé.

Selon les résultats de la littérature, le bruit de classe a un effet plus important sur l'efficacité des méthodes de classification. Malgré cela, on sait que le bruit d'attribut est plus difficile à gérer. Cette difficulté est due à la difficulté d'identifier les cas ayant des valeurs inexactes et à la complexité de leur traitement. Par conséquent, dans ce travail, nous nous concentrons sur l'étude du bruit d'attribut (incertitudes de mesure) et son impact sur les performances prédictives de la SVM.

Les raisons d'étudier la méthode SVM sont nombreuses. Dans un premier temps, la SVM présente un avantage par rapport aux arbres de décision. Les arbres de décision sont considérés comme des séparateurs de marges durs qui essaient de classifier parfaitement les données du jeu d'apprentissage. D'autre part, la marge souple de la SVM permet de tolérer les erreurs de classification lors de l'ajustement du modèle et donc d'éviter le surapprentissage. Cela rend le modèle SVM moins sensible au bruit et plus efficace en généralisation. En outre, les MLPs sont comparativement moins efficaces que les SVMs. L'efficacité des SVMs par rapport aux MLPs est due à plusieurs raisons. Premièrement, les MLPs sont basés sur la minimisation empirique du risque, tandis que les SVMs utilisent le principe de minimisation du risque structurel (SRM) qui traite le problème du surapprentissage en équilibrant la complexité du modèle et son succès à ajuster les données

d'apprentissage. Un autre grand avantage de la SVM est sa formulation qui conduit à un problème d'optimisation quadratique. Il réduit considérablement le nombre d'opérations dans la phase d'apprentissage, ce qui rend la SVM généralement beaucoup plus rapide sur dmes grands jeux de données. Enfin, le réglage d'un modèle MLP est considéré comme étant plus difficile car de nombreux hyperparamètres (nombre de couches, nombre de neurones par couche, optimiseur, fonction d'activation, etc.) doivent être calibrés simultanément.

Plusieurs approches robustes ont été proposées pour traiter le problème de la classification sous incertitudes avec la méthode SVM. Ces approches visent à évaluer l'impact des incertitudes sur les performances de la SVM, puis à développer/sélectionner des modèles SVM plus robustes. Ces travaux sont généralement basés sur une optimisation robuste qui modifie la formulation du SVM en introduisant des mesures probabilistes, ou sur une combinaison de différents outils pour assurer une performance optimale sur des jeux de données bruitées, comme l'introduction de la logique floue, ou l'utilisation d'approches de prétraitement adaptées à la méthode SVM.

En conséquence, l'objectif principal de cette thèse est d'améliorer la robustesse de la SVM face aux incertitudes de mesure. Ce défi est pertinent pour l'amélioration de la stratégie de mesure, les performances de production intelligente, le déploiement industriel de *IoT*, la *Digital Twin accuracy*, etc. Pour ce faire, un schéma de deux étapes principales est adopté :

1. L'impact des incertitudes de mesure (liées aux données de production) sur la précision des prédictions de SVM est tout d'abord quantifié. Cela a déjà été fait dans plusieurs autres travaux. Néanmoins, le principal défi est l'identification et la quantification des paramètres de production ayant un impact significatif sur la précision des prédictions du modèle SVM.
2. La deuxième étape consiste à introduire de nouvelles approches et de nouveaux modèles pour l'amélioration de la robustesse de la SVM face aux incertitudes de mesure, et ce en guidant l'algorithme SVM à se concentrer davantage sur la réduction de l'impact des incertitudes de mesure des paramètres qui affectent significativement la robustesse de la SVM.

Chapitre 2 : Identification des paramètres de production clés qui affectent la précision des prédictions de la SVM, en vue de l'évaluation de la qualité.

2.1 - Quantification de l'impact des incertitudes de mesure sur la robustesse de SVM.

Il a été démontré que les incertitudes de mesure ont un impact négatif sur la performance prédictive des méthodes de classification, en particulier des modèles SVM. Cet impact peut se manifester par une diminution de la capacité du classificateur à généraliser ou une augmentation de la complexité du modèle créé. La question scientifique abordée dans le deuxième chapitre consiste à quantifier l'effet de ces incertitudes de mesure sur la performance prédictive de la SVM et donc sur la généralisation de la SVM.

Afin de quantifier l'impact global des incertitudes de mesure sur la précision de la SVM, il est nécessaire de comprendre ce qui se passe lors de la prédiction d'un point de données qui est soumis à des incertitudes de mesure. Les incertitudes de mesure sont équivalentes à des translations qui font bouger un point dans toutes les directions. De tels mouvements peuvent faire en sorte qu'un point franchisse la frontière de décision et peut donc être mal classé. Ces mouvements deviennent beaucoup plus compliqués lorsqu'il s'agit de frontières de décision non linéaires, et en particulier dans les espaces de grandes dimensions. Pour cette raison, l'impact des incertitudes de mesure sur chaque point de données doit être pris en compte lors de la quantification.

En conséquence, dans cette étude, la probabilité qu'un point de données soumis à des incertitudes de mesure puisse franchir la limite de décision est calculée. Une approche basée sur la simulation de Monte-Carlo est donc proposée afin de quantifier l'impact des incertitudes de mesure sur les performances de prédiction de SVM.

Les simulations de Monte-Carlo sont utilisées dans ce travail pour calculer la chute en précision de la SVM due à la perturbation des données avec des incertitudes. Pour ce faire, chaque ensemble de données doit être divisé en un jeu d'apprentissage (2/3 du jeu de données) et un jeu de test (1/3 du jeu de données). Cela est nécessaire pour ajuster les hyperparamètres de la SVM à l'aide d'un algorithme génétique qui permet d'identifier un modèle SVM avec une précision de prédiction maximale sur le jeu de test. Ensuite, des incertitudes artificielles aléatoires sont générées pour perturber le jeu de test initial et ainsi définir les jeux de test bruités. Dans cette étude, l'impact des incertitudes de mesure gaussiennes est étudié, car elles sont considérées comme l'une des incertitudes les plus couramment rencontrées.

Cette approche été appliquée à quatre jeux de données, et ce en utilisant 1000 jeux de tests bruités. Les niveaux de bruit ont été fixés à 2,5% pour les trois premiers jeux de données,

et à 15% pour le dernier jeu de données. Ces niveaux de bruit ont été estimés en tenant compte de l'impact sur l'environnement, ainsi que de la précision et de l'exactitude des capteurs nécessaires dans les industries manufacturières.

Les résultats ont montré que le fait de bruiter le jeu de tests par des incertitudes de mesure a entraîné une chute de la précision de la prédiction de SVM, c'est-à-dire que la robustesse de la généralisation de SVM est affectée par les incertitudes. Pourtant, les résultats ne permettent pas d'identifier les incertitudes des paramètres qui ont les plus grands impacts sur la robustesse des prédictions de SVM. En conséquence, la question de recherche suivante a été formulée :

"Quelles sont les incertitudes des paramètres qui ont un impact significatif sur la précision des prédictions de la SVM ?

Pour répondre à cette question scientifique, trois approches sont proposées pour l'évaluation de l'impact des incertitudes de chaque paramètre sur la robustesse de la SVM, ce qui permet d'identifier les paramètres clés (appelés : *key measurement uncertainties*) avec des impacts significatifs.

2.2 - Identification des key measurement uncertainties

Dans cette partie, trois approches sont proposées pour l'identification des *key measurement uncertainties*.

Approche 1 : basée sur la simulation de Monte-Carlo

La première approche pour l'identification des *key measurement uncertainties* est similaire à la quantification effectuée dans la partie 2.1, où des simulations de Monte-Carlo sont utilisées pour évaluer la robustesse de SVM face aux incertitudes de mesure gaussiennes. Toutefois, dans cette approche, les paramètres sont bruités un par un au lieu de tous en même temps. Ainsi, il faut autant de simulations de Monte-Carlo que le nombre de paramètres. Cela permettrait de quantifier la chute en précision de la SVM due à la perturbation d'un paramètre avec des incertitudes de mesure gaussiennes. Ces simulations permettent de classer les paramètres de production en fonction de leurs impacts sur la SVM, ce qui permet en retour d'identifier les *key measurement uncertainties.*

Approche 2 : basée sur l'analyse de sensibilité de Sobol

Dans cette étude, la méthode de sensibilité de Sobol est utilisée pour évaluer la robustesse des prédictions de la SVM. Le choix de cette méthode est justifié par le fait que l'analyse de Sobol est connue pour être la seule approche capable de prendre en compte diverses distributions de paramètres d'entrée, de tenir compte des interactions multidimensionnelles des paramètres lorsque tous les paramètres varient simultanément, et d'englober à la fois les effets non linéaires et non additifs lorsque les interactions des paramètres sont prises en compte.

L'analyse de sensibilité de Sobol s'appuie sur des techniques de décomposition de la variance pour fournir un coefficient quantitatif de la contribution de chaque variance de paramètre d'entrée à la variance de la sortie. En conséquence, cette deuxième approche calcule ce que l'on appelle les indices totaux de sensibilité. Ces indices sont estimés par des simulations de Monte-Carlo, et ils permettent de mesurer l'effet des incertitudes de mesure de tout paramètre sur la précision du SVM. Le classement des indices de Sobol permet d'identifier les paramètres qui affectent de manière significative la précision du SVM et donc d'identifier les *key measurement uncertainties.*

Aussi, pour estimer les indices de Sobol, la simulation de Monte Carlo est utilisée en raison de sa convergence résultant de la forte loi des grands nombres. Cette estimation nécessite la définition des entrées, des sorties et du modèle qui les relie. Les entrées de l'analyse de Sobol sont généralement définies comme des scalaires, ce qui n'est pas le cas dans cette étude où les paramètres d'entrée sont des vecteurs. La méthode de Sobol a été donc rectifiée pour prendre en compte cette situation :

- ***Les entrées*** $(\boldsymbol{x_{n1}}, \dots, \boldsymbol{x_{np}})$: x_{nj} est le $j^{ème}$ paramètre du $n^{ème}$ jeu de test bruité.
- ***La sortie*** $\boldsymbol{Y_n}$: est la diminution de la précision de la SVM due aux incertitudes de mesure du $n^{ème}$ jeu de test bruité.
- ***Le modèle*** $\boldsymbol{f}$: est le modèle SVM optimisé; $[Y_n = \boldsymbol{f}(x_{n1}, \dots, x_{np})]$.

Approche 3 : basée sur une analyse de corrélations

Les deux premières approches pour l'identification des *key measurement uncertainties* sont basées sur des simulations de Monte-Carlo qui sont généralement coûteuses en termes de temps et de ressources. Pour surmonter ce problème, une troisième approche basée sur des outils statistiques est proposée. Cette approche estime les *key measurement uncertainties* en analysant statistiquement la corrélation d'un paramètre de production avec le paramètre ‘‘classe’’. L'objectif de cette approche est d'identifier les paramètres de production qui sont bien corrélés avec le paramètre ‘‘classe’’. Cela permet de classer les paramètres de production et donc d'estimer les *key measurement uncertainties*.

Dans cette troisième approche, le coefficient de corrélation *point biserial* est adopté dans cette approche. Ce coefficient est utilisé pour calculer la corrélation entre une variable continue et une variable binaire, dans une plage qui varie de -1 à +1. En conséquence, dans cette section, le coefficient de corrélation *point biserial* entre le paramètre de classe et chaque paramètre de production est calculé, en utilisant la formule exprimée dans l'équation ci-dessus :

$$r_{X_j \sim Y} = \frac{m_1 - m_2}{\sigma_n} \sqrt{\frac{n_1 n_2}{n^2}}$$

Où :

- X_j: le j^{eme} paramètre de production.
- Y: le paramètre ‘‘classe’’.
- m_1 : valeur moyenne de tous les points de données de la classe 1 du paramètre X_j.
- m_2 : valeur moyenne de tous les points de données de la classe 2 du paramètre X_j
- σ_n: écart-type de X_j.
- n_1: nombre de points de données de la classe 1.
- n_2: nombre de points de données de la classe 2.
- n: nombre de points du jeu de données.

Analyse de résultats

Les trois approches ont été testées sur quatre jeux de données industriels, et les résultats ont montré que :

1. L’identification des *key measurement uncertainties* des deux premières approches est similaire.
2. L’approche basée sur l’analyse de corrélation permet d'estimer un certain nombre des *key measurement uncertainties* identifiées par les deux premières approches. On peut donc conclure que les incertitudes des paramètres de production qui sont bien corrélées au paramètre ‘‘classe’’ sont plus susceptibles d'avoir un impact significatif sur la robustesse des prédictions de la SVM. En outre, cette estimation est imprécise et approximative, car les paramètres ayant de faibles coefficients de corrélation et des impacts significatifs sont négligés et ne sont pas estimés comme des *key measurement uncertainties*.
3. L’impact des *key measurement uncertainties* sur la robustesse des prédictions de la SVM est aussi important que l'impact des incertitudes de mesure de tous les paramètres.

2.3 - Conclusion

Le problème des incertitudes de mesure et leur impact sur la précision des prédictions de la SVM ont été abordés dans ce chapitre. Dans un premier temps, une simulation de Monte Carlo a été réalisée pour évaluer et quantifier l'impact des incertitudes de mesure. Ensuite, trois nouvelles approches ont été proposées pour l'identification des paramètres dont les incertitudes affectent de manière significative la robustesse de la SVM. La première approche, basée sur la simulation de Monte Carlo, évalue la manière dont la précision de la SVM est diminuée par le incertitudes d'un seul paramètre. Dans la seconde approche, l'analyse de sensibilité de Sobol a été modifiée afin d'évaluer la sensibilité des prédictions de la SVM par rapport aux incertitudes de mesure. Ces deux approches ont permis de fournir des mesures quantitatives qui représentent l'ampleur de l'impact de chaque paramètre. Néanmoins, les deux approches sont basées sur des simulations de Monte-Carlo qui sont coûteuses en termes de temps et de ressources. Pour surmonter ce problème, une nouvelle approche basée sur

une analyse de corrélation est proposée pour estimer les *key measurement uncertainties*. Les trois approches ont été appliquées à quatre jeux de données afin d'évaluer leurs performances. De plus, les résultats ont montré que les deux premières approches sont plus précises et exactes dans l'identification des *key measurement uncertainties*, tandis que la dernière approche a permis de donner une intuition sur les incertitudes de mesure qui sont suspectes d'avoir un impact significatif sur la précision de prédiction de la SVM.

Ce chapitre a permis de mieux appréhender la robustesse et la sensibilité des performances prédictives de la SVM face aux incertitudes de mesure. Ce travail a également permis d'identifier de manière robuste les paramètres clés conduisant à des problèmes de qualité. Par conséquent, la qualité d'un système de production peut être améliorée en contrôlant les *key measurement uncertainties.*

Chapitre 3 : Vers des modèles SVM robustes

Alors que l'objectif du deuxième chapitre est de quantifier les impacts des incertitudes sur la précision des prédictions d'un modèle SVM visant à évaluer la qualité des systèmes de production, ce chapitre vise à améliorer sa robustesse face aux incertitudes de mesure tout en maintenant des performances prédictives optimales. Cet objectif peut être formulé comme un problème d'optimisation multi-objectifs où la précision prédictive de la SVM doit être maximisée et l'impact des incertitudes de mesure sur les performances prédictives de la SVM doit être minimisé.

Deux idées principales sont développées pour atteindre cet objectif :

1. Comme le séparateur est défini par l'espace noyau de la SVM, la première idée est d'identifier les hyperparamètres optimaux de la SVM qui permettent une précision prédictive optimale et une robustesse optimale à l'impact des incertitudes de mesure.
2. Le séparateur SVM est construit à l'aide de produits scalaires. Par conséquent, ce séparateur est défini par les valeurs numériques des attributs. Un ajustement de ces valeurs en tenant compte de l'impact de chaque paramètre sur la robustesse de la SVM est donc prometteur.

Ces deux idées principales ont été développées à travers différentes approches. Deux approches basées sur un algorithme génétique visent à identifier les espaces noyaux de la SVM ayant une performance prédictive optimale et une robustesse optimale aux incertitudes de mesure. En outre, deux autres approches améliorent la robustesse du SVM en redimensionnant les paramètres d'entrée sur la base de l'analyse de sensibilité de Sobol. Enfin, une approche attribue un poids à chaque paramètre, où ils sont optimisés en même temps que les hyperparamètres de la SVM à l'aide d'un algorithme génétique.

3.1 - Sélection robuste

Approche 1 : optimisation bi-objectif pour la sélection de modèles SVM robustes

Afin de sélectionner un modèle SVM robuste, une approche d'optimisation est nécessaire, visant à maximiser la précision de la SVM tout en minimisant l'impact des incertitudes de mesure. Par conséquent, pour optimiser les deux objectifs conflictuels, il faut obtenir les solutions optimales de Pareto par rapport au problème rencontré.

Ainsi, la fonction objective que la première approche vise à optimiser peut être exprimée comme suit :

$$\textbf{max} \qquad f_1(\text{noyau}) - f_2(\text{noyau})$$

où : $\boldsymbol{f_1}$ est la précision de prédiction du SVM, et $\boldsymbol{f_2}$ est la diminution de la précision du SVM due aux incertitudes de mesure, évaluées dans le même espace du noyau.

Un algorithme génétique est utilisé pour l'optimisation de la fonction objective. L'algorithme génétique évalue chaque solution potentielle en évaluant sa précision de prédiction sur le jeu de test et en calculant sa robustesse par la simulation de Monte-Carlo mentionnée au chapitre 2.

Approche 2 : Optimisation à deux niveaux pour la sélection de modèles SVM robustes

La seconde approche pour la sélection de modèles SVM robustes comprend deux étapes principales : une première optimisation est effectuée pour l'identification de tous les espaces noyaux qui permettent d'avoir une précision de prédiction maximale. La deuxième étape vise à évaluer tous les espaces noyaux identifiés afin de sélectionner celui qui minimise l'impact des incertitudes de mesure.

Par conséquent, l'optimisation à deux niveaux proposée peut être formulée comme suit :

$$\begin{aligned} &\textbf{min} && f_2(\text{noyau}) \\ &\text{s. c} && max\ \ f_1(\text{noyau}) \end{aligned}$$

Un algorithme génétique est utilisé pour identifier tous les espaces noyaux optimaux dans la première étape. Ensuite, la robustesse de chaque espace noyau est évaluée en utilisant la même simulation de Monte-Carlo.

3.2 – Modèles SVM pondérées

Approche 3 : Approche basée sur les indices de Sobol et la recherche par grille

Comme les incertitudes de mesure sont additives et afin de redimensionner les différentes caractéristiques en fonction de leur impact sur la précision des prédictions de la SVM, différents poids basés sur les indices de Sobol' sont associés. Ce redimensionnement modifie l'orientation de la frontière de décision et affecte donc la performance prédictive de la SVM ainsi que sa robustesse aux incertitudes de mesure. Les poids sont définis comme exprimés dans l'équation suivante.

$$w_j = 1 + R.ST_j$$

Où : R est le nouveau paramètre à calibrer, et ST_j et l'indice de Sobol total du $j^{ème}$ paramètre.

Il est nécessaire d'utiliser de petites valeurs de R afin d'éviter de grands changements au niveau des valeurs des données et donc d'éviter la perte des informations contenues dans ces données.

Cette approche vise à améliorer la robustesse de la SVM face aux incertitudes de mesure en optimisant uniquement le paramètre R et en utilisant les mêmes hyperparamètres optimaux des modèles SVM identifiés au chapitre précèdent. Une recherche par grille est effectuée pour évaluer les différentes valeurs du paramètre R. La valeur R qui maintient la performance prédictive de la SVM tout en améliorant sa robustesse face aux incertitudes est sélectionnée.

Approche 4 : Approche basée sur les indices de Sobol et l'algorithme génétique

Dans l'algorithme précédent, la robustesse de la SVM est améliorée en optimisant uniquement le paramètre R. Afin d'améliorer davantage la robustesse du modèle SVM, le hyperparamètres de la SVM seront également inclus dans l'optimisation. Cette approche se compose de deux étapes. La première étape est basée sur l'algorithme génétique pour l'identification de toutes les solutions optimales (hyperparamètres de la SVM et paramètre R) qui permettent une précision maximale des prédictions de la SVM. Ensuite, dans la deuxième étape, la robustesse de chacune de ces solutions optimales est évaluée par simulation de Monte Carlo. Ainsi, la solution avec une précision de prédiction maximale, et une robustesse maximale à l'impact des incertitudes de mesure est sélectionnée.

Approche 5 : Amélioration de la robustesse du SVM - optimisation des hyperparamètres et des poids des paramètres

Cette approche est proposée pour déterminer un ensemble optimal de poids de caractéristiques et d'hyperparamètres de la SVM à l'aide d'un algorithme génétique. Contrairement aux approches précédentes basées sur Sobol, cette approche ne nécessite aucune information sur les poids des paramètres. Le réglage des poids et des hyperparamètres est guidé par les performances du SVM, c'est-à-dire que l'algorithme génétique reçoit le retour d'information du classificateur SVM pour déterminer les directions de recherche. Toutes les solutions optimales sont sélectionnées à la fin de l'exécution de l'algorithme génétique. La robustesse de ces solutions aux incertitudes de mesure est ensuite évaluée à l'aide de simulations de Monte Carlo. Ce problème d'optimisation est défini par :

$$\begin{aligned} &\textbf{min} && f_2(\text{noyau}, \text{poids}) \\ &\text{s.c} && \textbf{max}\ f_1(\text{noyau}, \text{poids}) \\ &\text{s.c} && 0 \leq \text{poids} \leq 1 \end{aligned}$$

où ''noyau'' désigne l'espace noyau défini par les hyperparamètres de la SVM, et ''poids'' désigne les poids des caractéristiques.

3.3 - Analyse et conclusion

Analyse de résultats

Les trois approches ont été testées sur quatre jeux de données industriels, et les résultats ont montré que :

1. Dans l'ensemble, pour tous les jeux de données industriels, les modèles SVM associés aux différentes approches ont fourni une classification avec une précision de prédiction de plus de 80%.
2. Les différentes approches ont permis d'assurer une prédiction optimale, tout en optimisant la robustesse de la SVM face aux incertitudes de mesure.
3. Les approches 1 et 2 basées sur la sélection robuste permettent de maintenir une prédiction optimale et d'améliorer la robustesse de la SVM par rapport aux incertitudes de mesure.
4. Les approches 3, 4 et 5 basées sur la pondération des paramètres améliorent la robustesse de la SVM tout en améliorant également ses performances prédictives, surtout la cinquième approche où des améliorations supérieures à 6 % ont été constatées.

Conclusion

Pour conclure, l'objectif principal de ce chapitre est d'améliorer la robustesse d'un modèle SVM lorsque les données sont considérées avec des incertitudes. Cinq approches ont été proposées. Les deux premières approches sont mathématiquement formulées comme des problèmes d'optimisation bi-objectifs qui visent à maximiser la précision de prédiction de la SVM ainsi que sa robustesse aux incertitudes de mesure. Ensuite, deux autres approches sont définies en tenant compte des indices de Sobol pour la définition de nouveaux poids de paramètres permettant d'améliorer la robustesse des modèles SVM. Enfin, la dernière approche identifie une solution optimale en optimisant simultanément les poids des paramètres et les hyperparamètres de la SVM. En se basant sur leurs applications à plusieurs jeux de données industriels, ces approches montrent leur capacité et leur efficacité à améliorer la robustesse des modèles SVM. En particulier, l'approche 5 a permis d'améliorer efficacement la précision et la robustesse des prédictions des modèles SVM.

Chapitre 4 : Conclusion générale et perspectives

Pour rappel et comme indiqué dans l'introduction, l'objectif global de ce travail de recherche est formulé comme suit :

> Comment évaluer l'impact des **incertitudes de mesure** sur la performance prédictive de la **méthode de classification SVM**, en vue d'améliorer la **qualité des systèmes de production** ?

Trois objectifs de recherche ont ensuite été dérivés pour répondre à cet objectif principal :

1. Évaluation et amélioration de la qualité des systèmes de production basés sur des algorithmes d'apprentissage automatique et des approches de classification.
2. Quantification de l'impact des incertitudes de mesure sur les performances de la méthode de classification SVM.
3. Amélioration de la robustesse de la classification SVM lors de la prise en compte des incertitudes de mesure.

4.1 - Évaluation et amélioration de la qualité des systèmes de production par les outils de machine learning.

Afin de répondre au premier objectif de recherche, une analyse documentaire a d'abord été réalisée. Cette revue a permis de mettre en évidence le potentiel des approches d'apprentissage automatique à améliorer et à évaluer la qualité au sein des systèmes de production. Trois classificateurs principaux ont été abordés : les arbres de décision, les machines à vecteur de support et le perceptron multicouche, où leurs applications dans les systèmes de production ont été discutées, analysées et valorisées.

En outre, pour évaluer la performance de ces classificateurs, ils ont été appliqués à des données industrielles, cela a démontré la capacité de la SVM et du MLP à prédire différents niveaux de qualité, et la capacité de coupler l'arbre de décision C4.5 avec des coordonnées parallèles pour identifier les causes des défauts et les paramètres de production optimaux d'une manière interprétable. Sur la base des résultats de ces applications et de la recherche documentaire, la SVM a été choisie comme classifieur à étudier.

Enfin, afin d'étudier la robustesse des modèles SVM à l'impact des incertitudes de mesure, une étude bibliographique des différentes approches développées a été réalisée. Cela a permis d'identifier les différentes méthodologies pour améliorer la robustesse des modèles SVM.

Cependant, ce premier objectif a certaines limites, car il ne concerne que les données structurées, en particulier les attributs numériques. Par ailleurs, en ce qui concerne les données bruitées, seules les incertitudes de mesure sont considérées dans cette étude.

4.2 - Quantification de l'impact des incertitudes de mesure sur les performances de la SVM.

Pour atteindre le deuxième objectif de recherche de ce travail, une première expérience a été réalisée pour quantifier l'impact des incertitudes de mesure gaussiennes sur la performance prédictive de la SVM. Cette expérience est basée sur une simulation de Monte-Carlo, et a montré que la perturbation d'un jeu de données avec des incertitudes gaussiennes entraîne une chute en précision d'un modèle SVM et donc une diminution de la généralisation de cette méthode. Ensuite, trois approches ont été proposées pour quantifier l'impact des incertitudes de chaque paramètre d'entrée sur la prédiction d'un modèle SVM, et donc l'identification des *key measurement uncertainties*. Les deux premières approches, basées sur la simulation de Monte-Carlo et l'analyse de sensibilité de Sobol, permettent de calculer pour chaque paramètre un coefficient représentant l'impact de ses incertitudes sur la robustesse de la SVM. D'autre part, la troisième approche, qui s'appuie sur des outils statistiques, permet d'estimer les incertitudes des paramètres qui peuvent influencer la robustesse du modèle SVM. Ces trois approches ont ensuite été appliquées à plusieurs jeux de données industrielles, et les résultats montrent leur efficacité.

Néanmoins, les deux premières approches proposées n'ont été appliquées qu'en tenant compte des incertitudes de mesure gaussiennes. Par ailleurs, l'approche statistique estime les *key measurement uncertainties* en analysant uniquement la corrélation entre un paramètre d'entrée et le paramètre de classe. Cette approche pourrait être étendue et améliorée en considérant la corrélation entre un paramètre et le reste des paramètres.

4.3 - Amélioration de la robustesse de la classification SVM lors de la prise en compte des incertitudes de mesure.

L'objectif de cette recherche est d'améliorer la robustesse des modèles SVM lors du traitement de données soumises à des incertitudes. Pour atteindre cet objectif, cinq approches ont été développées. Les deux premières approches se concentrent sur la sélection de modèles SVM présentant une précision prédictive optimale et une robustesse optimale aux incertitudes de mesure. Ces deux approches permettent la sélection d'un modèle SVM optimal qui permet une prédiction de qualité robuste sans avoir besoin d'un modèle SVM modifié. Les approches 3 et 4 sont basées sur les connaissances statistiques recueillies grâce à l'application de l'analyse de sensibilité de Sobol pour quantifier l'impact des *key measurement uncertainties*. Des poids de paramètres qui sont fonction des indices totaux de Sobol et d'un paramètre dénommé R ont donc été définis, et des modèles qui intègrent la notion de poids de paramètres sont établis pour permettre une meilleure robustesse aux impacts des incertitudes de mesure. Enfin, la dernière approche consiste à attribuer à chaque paramètre un poids (variant entre 0 et 1), où ils sont optimisés en même temps que les hyperparamètres de la SVM. Cette dernière approche a permis d'obtenir des modèles SVM

pondérés avec une meilleure performance prédictive et une meilleure robustesse à l'impact des incertitudes de mesure. Les résultats de l'application des approches proposées à divers jeux de données de production ont démontré l'efficacité des différentes approches.

Finalement, les questions des incertitudes rencontrées dans les données de qualité et leur impact sur al SVM ont été abordées dans ce travail de thèse, où plusieurs approches ont été proposées. Ce travail de recherche permet de mieux comprendre la robustesse et la sensibilité des modèles SVM lorsqu'ils traitent des données soumises à des incertitudes de mesure, car il permet d'identifier avec précision les paramètres clés (appelés *key measurement uncertainties*) à l'origine des problèmes de qualité. Par conséquent, la surveillance des *key measurement uncertainties* et la gestion de l'impact des incertitudes de mesure permettraient par la suite aux industries manufacturières d'améliorer la qualité de leurs systèmes et de prendre des décisions plus robustes.

4.4 - Perspectives et travaux futurs

En fonction des objectifs de recherche de ce travail, des perspectives et des développements futurs peuvent être proposés.

Premièrement, l'algorithme génétique développé pour l'optimisation des hyperparamètres de la SVM peut être amélioré en tenant compte de plusieurs critères, c'est-à-dire qu'au lieu de sélectionner les modèles SVM en ne considérant que la précision des prédictions du modèle, on peut inclure davantage de critères tels que la complexité du modèle, la séparabilité, la robustesse. En conséquence, des modèles plus légers avec une performance prédictive optimale peuvent être sélectionnés pour gérer la classification de la qualité.

Pour le deuxième chapitre, différents types d'incertitudes peuvent être ajoutés lorsqu'il s'agit de l'approche de Monte-Carlo ou de l'approche de Sobol. Cela permettra la généralisation des approches et l'extension de leurs domaines d'application. La troisième approche statistique pourrait bénéficier de la définition d'une nouvelle mesure de corrélation qui comprendrait à la fois : corrélation simple d'un paramètre avec le paramètre de classe, et la corrélation multiple d'un paramètre avec le reste des paramètres d'entrée.

Dans le troisième chapitre, les performances des approches basées sur Sobol peuvent être améliorées en définissant des pondérations par l'acquisition de plus de connaissances statistiques. Des travaux supplémentaires devraient être consacrés à la définition et à la sélection de ces pondérations.

D'autre part, même si les données considérées dans ce travail sont considérées au départ comme structurées, cela n'est généralement pas le cas pour les cas réels. Une analyse plus approfondie est nécessaire pour permettre l'exploration d'autres types de bruit liés aux données structurées et non structurées, afin d'analyser et de gérer leurs impacts sur les différentes méthodes d'apprentissage automatique et d'apprentissage profond.

Table of contents

Table of Figures

List of Tables

Glossary

1D-CNN	One-Dimensional Convolutional Neural Network
ANN	Artificial Neural Networks
BOW	Breadth-Oblivious-Wrapper
BSPO	Binary Particle Swarm Optimization
CART	Classification And Regression Tree
CCA	Curvilinear Component Analysis
CCCP	Concave-Convex Procedure
DA	Discriminant Analysis
DCA	Convex Algorithm Difference
DM	Data Mining
DMDB	Decision Making DataBase
DPMO	Defects Per Million of Opportunities
DT	Decision Tree
EDIR	Error Detection and Impact-sensitive instance Ranking
FFT	Fast Fourier Transform
GA	Genetic Algorithm
IoT	Internet of Things
KDD	Knowledge Discovery from Databases
KML-SVM	Kernel-based hybrid Manifold Learning and Support Vector Machine
L-PBF	Laser-Powder Bed Fusion
LS-SVM	Least Squares Support Vector Machines
MES	Manufacturing Execution System
ML	Machine Learning
MLP	Multi-Layer Perceptron
MOPs	Multi-Objective Optimization Problems
OEM	Original Equipment Manufacturer

OVO	One-Vs-One
PANDA	Pairwise Attribute Noise Detection
PCA	Principal Component Analysis
PSO	Particle Swarm Optimization
RBF	Basic Radial Function
RF	Random Forests
RFC	Regularized Fisher's Criterion
RLS-SVM	A Robust Least Square Support Vector Machine
RWS	Resistance Spot Welding
SA	Sensitivity Analysis
SRM	Structural Risk Minimization
SVM	Support Vector Machines
WEDM-MS	Wire Electrical Discharge Machining-Middle Speed
WLS-SVM	Weighted Least Squares Support Vector Machines

Preface

The purpose of this preface is to define the context of the research study discussed in this manuscript. A brief introduction of the doctoral school is therefore given, accompanied by a thorough review of the research laboratory and its fields of expertise. After, the emphasis is put on the industrial context of the PhD thesis as part of an industrial research chair that aims to address four major industrial challenges related to production systems and the Industry of the Future.

This PhD thesis' research work has been conducted at **Arts et Métiers Institute of Technology** campus Metz, at LCFC (Laboratory of Design, Manufacturing, and Control). The research activities of LCFC consist of developing future production systems in the service and manufacturing sectors. LCFC focuses particularly on the integration of human factors into the system, the robotization of manufacturing processes, and the design and development of innovative approaches for product simulation and optimization.

In the context of encouraging close collaborations between research and industry, the Arts et Métiers Institute of Technology, and LCFC laboratory have co-founded an industrial research chair entitled “Systèmes de production reconfigurables – sûrs – performants” in partnership with Thyssenkrupp, the European Regional Development Fund “Programme opérationnel FEDER-FSE Lorraine et Massif de Vosges 2014- 2020”, UIMM F2I, and UIMM Lorraine. The purpose of this chair is to provide industrial solutions that enhance flexibility and responsiveness of manufacturing systems to fluctuating demand, improve their efficiency, facilitate their maintenance, manage operational risks, and improve their performances by taking into account all factors, including human factors. One of the challenges of this research chair is:

Characterization of the performance of complex and modular production systems: how to enable a better understanding of the performance of production systems via machine learning methods?

This PhD thesis focuses on this challenge, and thus focuses on how machine learning can be used to improve the quality of production systems, taking into account the uncertainties encountered in those systems. It should be acknowledged that the thesis research work has been carried out in collaboration with industrial partners, using their industrial expertise and data for the development and assessment of new robust approaches. Our contributions in these collaborations consist of proposing machine learning approaches that aim to improve the quality of their manufacturing systems. These contributions are:

- **Fives Nordon**: Fives Nordon designs, manufactures, installs, and maintains its customers' piping equipment and networks in highly demanding fields (nuclear, chemical, marine, automotive/aerospace, …). This first experimentation consisted in using machine learning classifiers for the identification of the various welding defects as well as their causes.

- **TDV**: The experimentation conducted with **TDV** consists of reducing the number of manufacturing sequences in a system for the production of cold-rolled profiled wire made from stainless steels, carbon steels and non-ferrous metals.
- Also, we have proposed a training material that guides the industrial partners in the implementation of the C4.5 decision tree using the python programming language.

Additionally, during my PhD thesis, I did a mobility of 6 consecutive months from 01 September 2019 to 28 February 2020 at the WBK Institut für Produktionstechnik at the Karlsruhe Institut für Technologie (KIT). This mobility was carried out within the framework of the Franco-German Doctoral College between Arts et Métiers and KIT, and it aimed to combine both the WBK knowledge, and the LCFC knowledge, in order to work on a common project. The project consisted of proposing new optical quality monitoring approaches for the Laser powder bed fusion process, by using supervised learning techniques.

Finally, the works conducted during the PhD thesis led to the writing and publication of two conference papers and one journal article, namely:

- Zouhri, Wahb, Hamideh Rostami, Lazhar Homri, and Jean-Yves Dantan. **A Genetic-Based SVM Approach for Quality Data Classification.** In International Conference on Artificial Intelligence & Industrial Applications, pp. 15-31. Springer, Cham, 2020.
- Zouhri, W., Dantan, J.Y., Häfner, B., Eschner, N., Homri, L., Lanza, G., Theile, O., Schäfer, M., 2020. **Characterization of laser powder bed fusion (L-PBF) process quality: A novel approach based on statistical features extraction and support vector machine**. 14th CIRP Conference on Intelligent Computation in Manufacturing Engineering.
- Zouhri, W., Dantan, J.Y., Häfner, B., Eschner, N., Homri, L., Lanza, G., Theile, O., Schäfer, M., 2020. **Optical process monitoring for Laser-Powder Bed Fusion (L-PBF)**. CIRP Journal of Manufacturing Science and Technology S1755581720301061. https://doi.org/10.1016/j.cirpj.2020.09.001

Introduction

In the context of industry 4.0, due to the fast evolution of demand, production and manufacturing systems are facing various topics such as cost estimations, supply-chain management, human factors, demand forecasting, material requirements and enterprise resource planning, process optimization, scheduling, sequencing, cell organization, and quality control, etc. These systems must be flexible and responsive to market demands in order to remain competitive, while at the same time addressing different objectives in terms of cost, time and quality. Therefore, to cope with this challenge, manufacturing industries rely on innovation and implementation of new technologies. The use of the latest technologies is essential to the success of an industry, as they enable production systems to be more efficient and effective (Mittal et al., 2018).

In manufacturing industries, in order to remain profitable and maintain a competitive edge, achieving a high level of quality in products, processes and services is a vital issue today; therefore, manufacturing industries cannot survive without assessing quality and providing high quality products. Quality management is one of the most important topics that must be considered to meet unique needs, improve customer satisfaction, identify production bottlenecks, ensure effective management and efficient administration, due to the fact that quality assessment/management influences all areas of every manufacturing organization. Nowadays, quality must be continuously improved and new conformity assessment processes must be accepted in manufacturing, as customer expectations are increasing. Various strategies have been proposed for managing and improving the quality of production systems, such as Continuous Improvement, Total Quality Management, Toyota Production System, World-Class Manufacturing, Just-in-Time and Six-Sigma, which encourage quality managers to collect data to address quality issues, as new advances in automation and computer systems and data from manufacturing processes become more available. However, these concepts have not been designed to cope with such a dynamic, evolving, and complex context (Colledani et al., 2014). This has driven the manufacturing industries to search for new solutions.

Recently, and to deal with automated industries, the Internet of Things (IoT) has become the key to connect and exchange huge amounts of data using software and sensor-like devices. These huge amounts of data, i.e. Big Data, are constantly being gathered and stored. Thus, the IoT and Big Data can bring new understanding and valuable knowledge to manufacturing systems by analyzing and visualizing the available data.

Consequently, manufacturing industries have opted for innovative data collection and analysis technologies to improve the quality of their production systems. A large volume of quality data is collected every day in each production system. However, even storing, let alone analyzing, such a large amount of data presents many new obstacles and challenges when drowning in data and starving for knowledge. This has led to the birth of a field of research

called "Knowledge Discovery from Databases" (KDD), commonly referred to as "Data Mining" (DM), which aims at extracting previously unknown and potentially valuable information from data to solve quality and control problems. The KDD's process is based on preprocessing techniques, machine learning (ML) algorithms, and visualization methods for the discovery and the presentation of useful knowledge and hidden patterns (Rostami et al., 2015). Among the various quality tasks, data prediction and classification are the most important part of the various steps and sub-activities of the KDD process as shown in Figure 0.1. An analysis is therefore mandatory for the identification and the selection of suitable methods for the analysis of quality data.

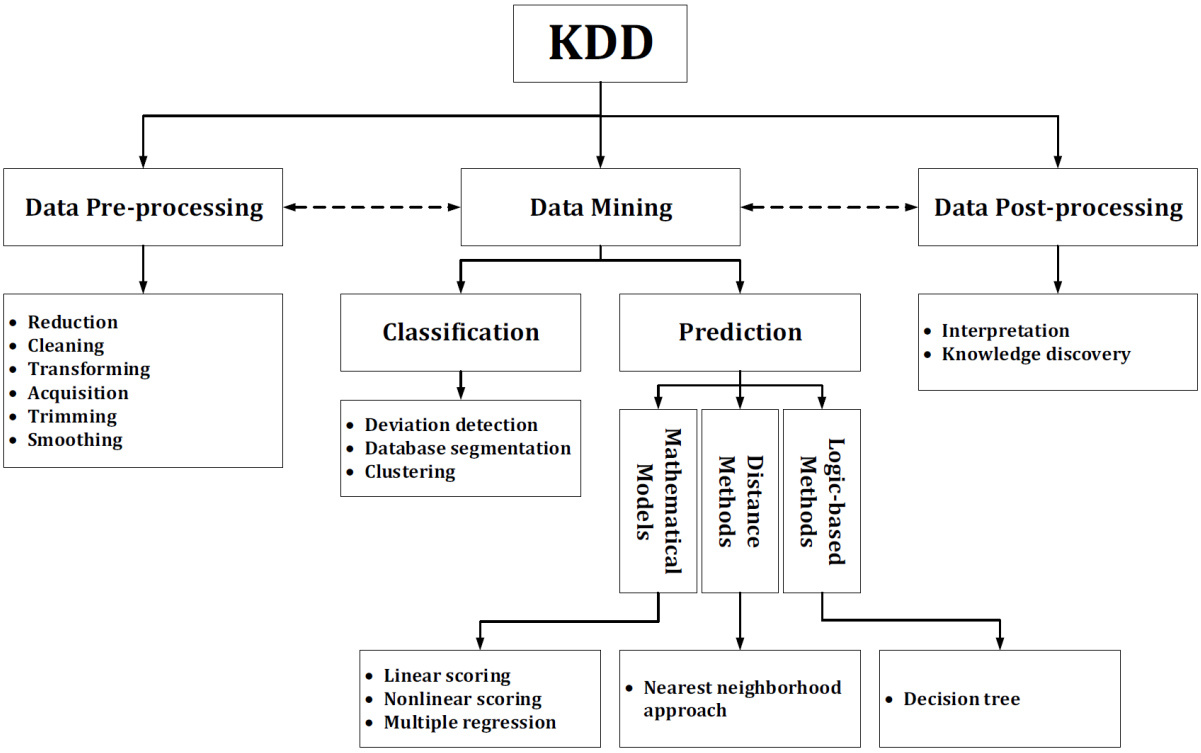


Figure 0.1: KDD main process and sub activities

In this work, machine learning algorithms are considered for the assessment and the improvement of quality within manufacturing systems. However, the field of ML is very diverse, and having many methods and algorithms at one's disposal could be considered a double-edged sword. In a way, these methods may offer a large set of solutions that deal with manufacturing issues, but consequently, identifying a suitable ML method among the existing ones might be challenging. These techniques aim at identifying the relation between the quality parameters (inputs and outputs), to predict the output quality, to classify the quality of the products, and to optimize the inputs to obtain an optimal output. (Köksal et al., 2011) summarized those quality tasks within manufacturing industries into four main ones, which are: quality description, quality prediction, **quality classification**, and parameter optimization (see Figure 0.2). In this work, we focus on the quality classification task.

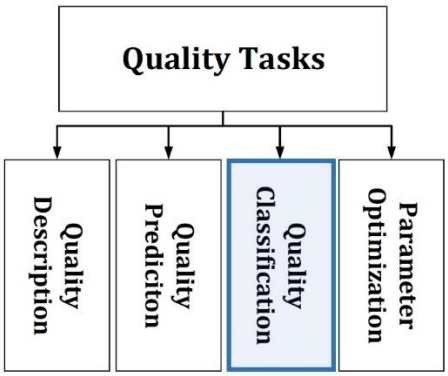


Figure 0.2: Quality tasks in the manufacturing industry

Since most manufacturing industries can supply labelled data, several classification methods and tools have been employed over the last few years in order to control the quality of the process and to increase the productivity of manufacturing systems (Wuest et al., 2016). For example, Decision trees (DT) classifiers have been used to identify the optimal manufacturing settings and the different causes of different defects (Siltepavet et al., 2012). Neural networks (ANN) and Support Vector Machines (SVM) have been applied for quality improvement allowing the prediction of product conformity as well as the prediction of quality levels (Wu et al., 2019; Chen et al., 2011). This research work focuses particularly on the application of the **SVM** for **quality classification**. SVMs showed good performance when addressing various quality issues in different manufacturing industries (Rostami et al., 2015; Wuest et al., 2016). In addition, the SVM has been known to be an easy-to-use tool capable of handling large dimensional datasets with different features.

Manufacturing data are generally subject to uncertainties: uncertainties affecting parameter values or affecting class values (Hickey 1996). Uncertainty can be caused by limited perception or understanding of reality, for example: limitations of observation equipment, limited resources to collect, store, process, analyze or understand the data, e.g. sensor error (Leung 2011). Thus, the accuracy of the prediction of the SVM method will be influenced. A central challenge of this thesis is to study and analyze the impact of parameter uncertainties on the quality of the classification.

Scientific issues and research methodology

One research question is associated to the thesis objective:

*"How to assess the impact of **measurement uncertainties** on the predictive performance of **SVM**, in order to improve the **manufacturing systems quality**?"*

The main objective is illustrated in Figure 0.3.

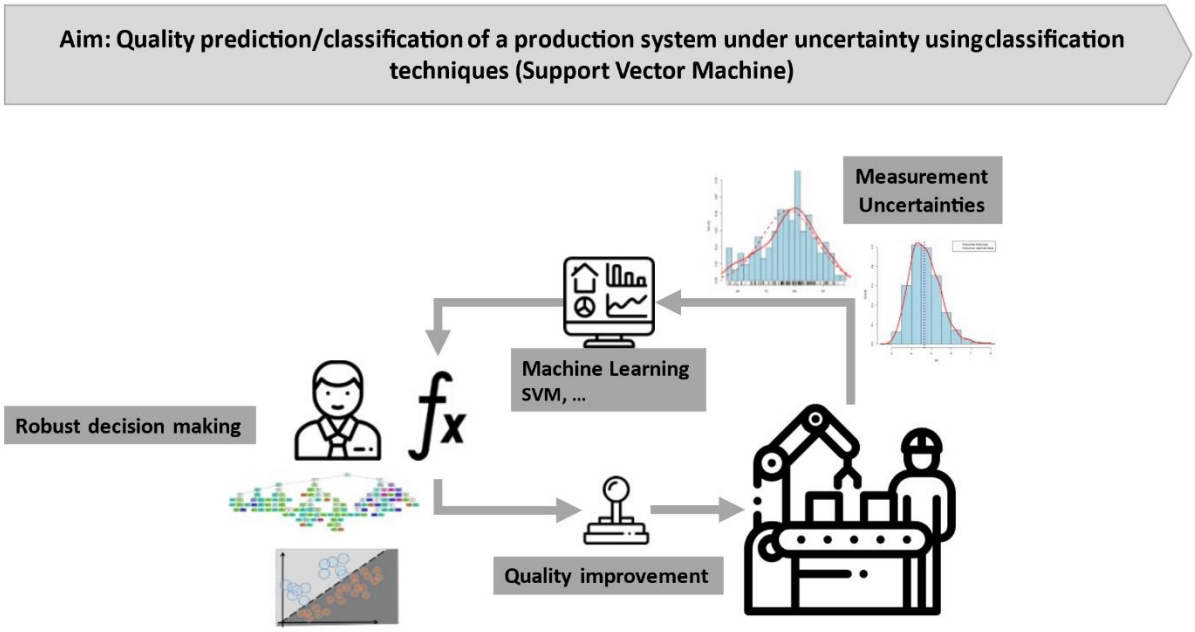


Figure 0.3: Main objective of the PhD thesis

To meet the overall objective, three main issues need to be addressed:

1- *How can machine learning algorithms and classification approaches be useful for quality improvement/assessment within manufacturing systems?*
2- *How to quantify the impact of measurement uncertainties on the performances of the SVM classification method?*
3- *How to improve the robustness of SVM classification regarding measurement uncertainties?*

These issues are being addressed through an iterative research methodology with the following steps:

- **Develop**: identification and development of methods and approaches to deal with each issue.
- **Experiment**: implementation and application of the methods and approaches on industrial and academic datasets.
- **Analyze**: effectiveness and efficiency assessment, approaches comparison and conclusions.

This proposed research methodology is illustrated in Figure 0.4.

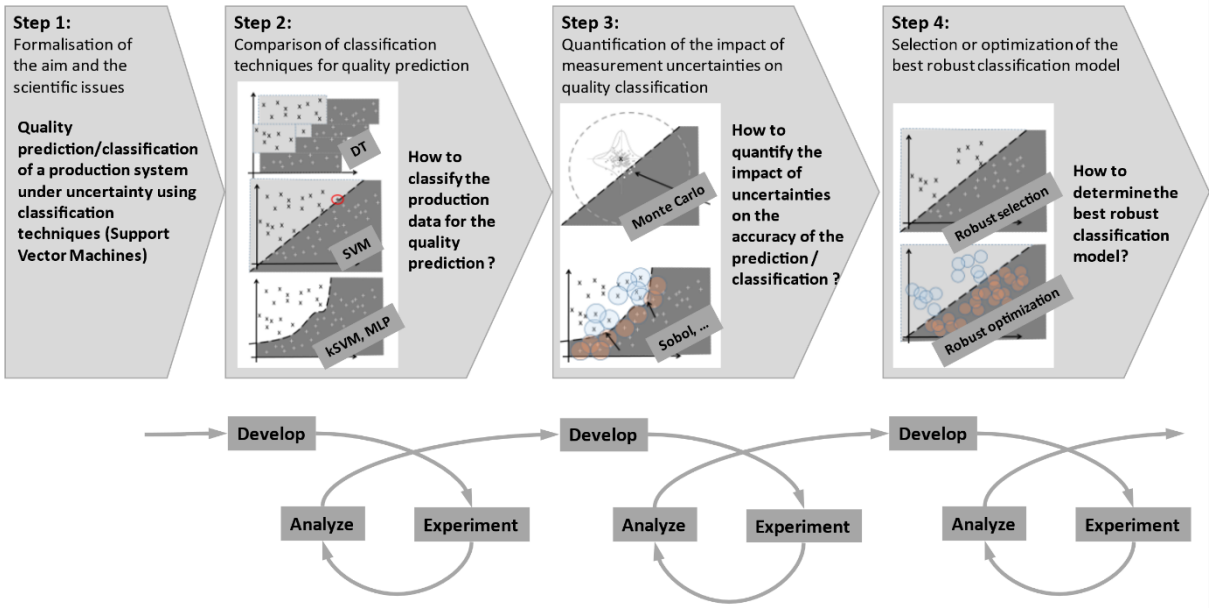


Figure 0.4: Research methodology

Organization of the thesis

According to Figure 0.4, the thesis manuscript is organized as follows:

Chapter 1 presents the different machine learning techniques and their applications in the manufacturing industry. In particular, it focuses on a comprehensive review of classification methods and their applications for quality management/assessment. In order to analyze their performance, some of the methods examined are then applied to several industrial datasets to solve quality assessment problems. Finally, the approaches developed in the literature that deal with the impacts of parameter uncertainties on the robustness of the SVM are particularly explained and studied.

In chapter 2, three approaches are developed to assess the impact of measurement uncertainties on the predictive performances of the SVM. Two main approaches which are based on Monte-Carlo simulation are proposed for the identification of the measurement uncertainties of the parameters with significant impacts on the performances of SVM. A third correlation research-based approach is proposed to estimate the measurement uncertainties that significantly affect the predictive performance of SVM.

In addition, we propose different approaches that aim to manage the impact of measurement uncertainties on the predictive performance of SVM in Chapter 3. Their main goal is to improve the robustness of SVM to measurement uncertainties as well as their predictive performances.

In the final chapter, the different results of the approaches are discussed. Finally, a final synthesis and perspectives are presented.

Case studies

In order to apply the different approaches developed in this thesis work, several industrial datasets are considered:

- ***Chem_4/8*** and ***Chem_5/7***: two datasets are adopted from a chemical manufacturing example. The manufacturing system consists of 11 manufacturing parameters, and its produces different products with different quality levels.
- ***Mines_1/2***: the dataset is from a floatation plant of a mining process. Floatation is one of the most used methods for extracting minerals from their ores. The quality of this extraction is affected by many process parameters (22 in this example), thus, a good process setup is mandatory.
- ***Roll_0/1***: a manufacturing system that consists of 95 manufacturing parameters *-Confidential manufacturing process-.*
- ***Fives_Nordon:*** The data set consists of 948 data points, 23 parameters and 14 labels representing 13 types of welding defect; *- Confidential data. This case study is only handled in chapter 1.*
- ***Siemens_L-PBF***: a data defined form a Laser Powder Bed Fusion (L-PBF) process; *- Confidential data. These data are only considered for applications in chapter 1.*

Chapter I

Machine learning applications for quality assessment in manufacturing industry: Support Vector Machine (SVM)

In the literature, various methods have been studied and proposed to process numerical data that are very common in different applications. Machine learning provides a set of tools to analyze data collected and stored in data warehouses. These tools are designed to extract interesting information and hidden patterns. However, the different formulations of machine learning methods do not take into account the uncertainty of the data, which negatively affects their performances. Therefore, to discuss this issue, the first chapter has been divided into three main parts. First, a review is provided in which the different applications of classification methods within manufacturing industries are examined. Next, some industrial case studies are presented in which we evaluate the performances of different classification methods. Finally, the issue of data uncertainty, how it affects the predictive performance of classification, and the different works that manage the impact of uncertainties on the SVM in particular are discussed.

I.1 - Machine learning in manufacturing

I.1.1 - Introduction

Machine learning (ML) is a field of study of artificial intelligence that uses mathematical and statistical approaches to give computers the ability to learn from data, i.e., to improve their performance in solving tasks without being explicitly programmed for each one. The history of machine learning dates back to the 1950s, when the Turing learning machine was introduced, along with the first neural network machines. Still, ML did not prosper until 1997 when a chess computer named IBM computer Deep Blue, beat the world chess champion (Deng et al., 2020). Since then, much more progress has been made in the ML area. Many industries have acknowledged that machine learning can increase the calculation ability, so they are devoting more research to it to stay ahead of the competition.

The ML field has reached different areas during the last two decades. In medicine and healthcare, ML techniques have been used for diagnosis, as they suggest possible ways to prevent or cure some diseases based on patients data (Deo, 2015.; Ghassemi et al., 2019). In finance, ML algorithms are used in trading, risk management, and process automation (Aziz et al., 2019). ML is also used to tackle the increasing complex problems in agricultural by offering a set of tools that allow addressing the agricultural sustainability challenges (Sharma et al., 2020). In computer vision, ML is used for object detection and classification, as well as for the extraction of relevant information from images, and videos (Khan and Al-Habsi, 2020). Additionally, in manufacturing, machine learning methods are used to satisfy the complex and fluctuant demand for high-quality products in an efficient manner (Wuest et al., 2016). Other application areas of ML techniques include education (Johnson, 2018), bioinformatics and neuroimaging (Serra et al., 2018), cybersecurity (Xin et al., 2018), retail (Huber and Stuckenschmidt, 2020), social media (Arasu et al., 2020), renewable energies (Salcedo-Sanz et al., 2018), weather forecasting (Scher and Messori, 2018), epidemic control (Libin et al., 2020), etc.

Today, the manufacturing industry is one of the main sectors that benefits from machine learning (Sharp et al., 2018). With the emergence of the Industry 4.0 paradigm, manufacturing industries are collecting more data from their production systems. Therefore, it becomes essential to find tools to organize, synthesize and analyze this data in order to obtain valuable and efficient information. This data handling is consequently based on ML approaches.

Many research have been focused on the various advantages of the application of the ML approaches within manufacturing systems. The adoption of these approaches allows a clearer understanding of the customer voices (Tao et al., 2018). This can be converted into product features and quality requirements that make it possible to cope with a changing and dynamic market. In addition, ML techniques are used during production to monitor both the manufacturing process and the manufacturing equipment in real time. As a result, manufacturers can keep up to date with changes, thus, decisions can be taken to better adjust

the manufacturing process and equipment. ML techniques can make the process more efficient by reducing costs without compromising quality (Lade et al., 2017).

Still, one of the main challenges in machine learning is the identification of a suitable method for a specific task. Several classifications of machine learning algorithms have been proposed in the manufacturing context (Wuest et al., 2016), where the different ML algorithms have been divided into three main categories, namely:

- Unsupervised learning: refers to the machine learning situation where data is not labelled. The algorithms try to discover some valuable patterns that lie behind these data, or to identify clusters from them. Classic examples of unsupervised learning are clustering, dimensionality reduction, and association rules.
- Reinforcement learning: consists of training an autonomous agent to make a sequence of decisions in a complex and uncertain environment. According to the decision made, the environment provides the agent either with rewards or penalties. Therefore, the agent aims to come up with an optimal solution by maximizing the total rewards.
- Supervised learning: consist of learning a prediction model from labelled examples, as opposed to unsupervised learning. A distinction is made between regression problems and classification problems, where a quantitative variable is predicted in regression, and a qualitative variable is predicted in classification problems.

To sum up, Figure 1.1 shows the known machine learning algorithms used in manufacturing industry area. A code color is added to indicate which type of learning an algorithm belongs to. It should also be noted that an algorithm can be used in different types of learning, e.g., genetic algorithms are used for supervised, unsupervised, and reinforcement learning, while SVM is only used in supervised learning.

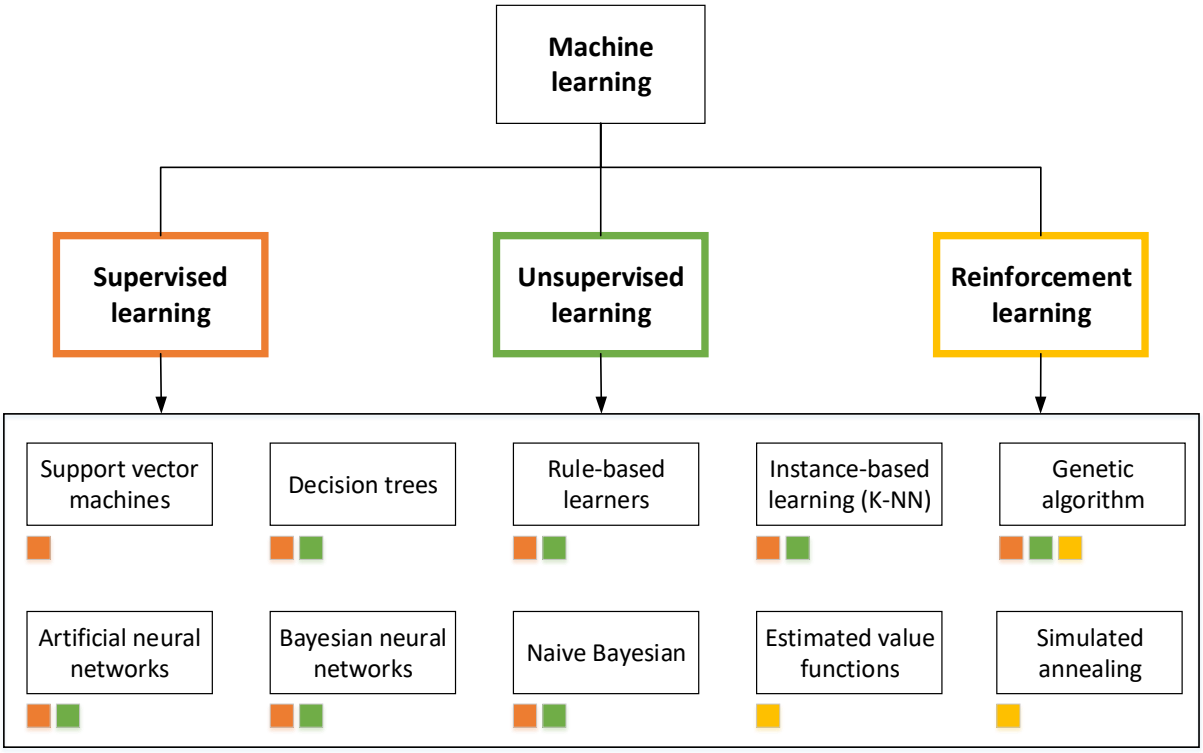


Figure 1.1: Structuring of ML techniques and algorithms (Wuest et al. 2016 – Edited)

When comparing these three categories of ML, unsupervised learning is useful when there is no expert feedback, which is not the case in manufacturing industries where expert knowledge is provided. That limits the use of unsupervised learning algorithms in manufacturing to some specific tasks as the identification of outliers, or the understanding of data before feeding it to a supervised algorithm (Hansson et al., 2016). Still, due to the fast increase of unlabeled data in manufacturing, the unsupervised machine learning might be so useful in the near future. Similarly, reinforcement learning is not widely used in manufacturing yet, and only find some applications in the heavy industry (e.g. Oil, mining, discrete manufacturing areas) (Nian et al., 2020). This is due to the many shortcomings that accompany the adoption of reinforcement learning algorithms since they are greedy and expensive in terms of data and can sometimes be replaced by easier mathematical programs. On the contrary, supervised learning is mostly used in manufacturing, as expert feedback is available, labeled data are already established for statistical process control purposes, and the algorithms performance can be easily evaluated by using some metrics as the accuracy or the F1-score (Cavalcante et al., 2019). The advantages of the predictive aspects of supervised learning algorithms in manufacturing have been summarized by Lee et al. (Lee et al., 2013) in three main points:

- Cost reduction: information on the status of production systems can be used to determine when maintenance is most needed. In this way, users can maximize the use of machine components and consumables.

- Operating efficiency: by knowing degradation modes, failure events can be deduced. This knowledge will enable production and maintenance supervisors to plan their activities collaboratively, maximizing thus the availability of equipment.
- Product quality Improvement: product quality can be maintained at acceptable levels by knowing how the performance of production systems drifts over time, and by integrating this knowledge into process controls, avoiding therefore defective products and unnecessary scrap.

Because of these advantages, supervised learning algorithms are discussed in this first chapter, and particularly, the classification methods are used in order to improve the quality of manufacturing systems. Three top performing classification algorithms are discussed in the following. A brief presentation of the structure of each algorithm is provided. The applications of each algorithm within manufacturing industries for quality management/improvement are presented after that.

I.1.2 - Classification methods for quality assessment in manufacturing

With the emergence of industry 4.0, machine learning based approaches have been increasingly used in order to improve the quality of manufacturing systems. Köksal et al., (2011) presented a comprehensive review paper discussing quality issues in manufacturing industries and how they are addressed using machine learning approaches. They considered four main quality tasks, which are:

- Quality description: by identifying which factors or input variables significantly affect the product quality.
- Quality prediction: by establishing models that predict the value of the output parameter based on the input parameters values.
- Quality classification: by predicting the class of quality in case of having it stated as binary or nominal variable (e.g.: classifying quality defects).
- Parameter optimization: by defining the optimal levels of the most important parameters that affect the process/product quality.

The authors (Köksal et al., 2011) concluded that when it comes to quality classification in manufacturing, the most used ML techniques are decision trees (DTs) and artificial neural networks (ANNs). Besides, similar work presented by Rostami et al., (2015) emphasized the great ability of support vector machines (SVMs) in handling the same quality tasks within production systems. Based on that, these three main classifiers are discussed in the following.

I.1.2.1 - Decision trees (DT)

DTs are popular approaches of machine learning used for the prediction of a qualitative variable from qualitative and/or quantitative variables. This flexibility in handling different types of data at the same time is considered to be an advantage over some other classification tools designed to treat a single type of variable. DTS provide a graphical, meaningful, and easy-

to-read representation. The graphical representation is in the form of a tree consisting of terminal leaves representing the classes of the instances, and nodes corresponding to a binary question using a variable of the dataset. A number of tasks could be addressed by the graphical interpretation of the tree, such as the estimation of the value of an attribute, the extraction of sets of classification rules, the interpretation of the relevance of the attributes, and the classification of new instances.

Several algorithms have been proposed in order to build decision trees, particularly, the CART technique (Grajski et al., 1986) , the Quinlan's ID3 algorithm (Quinlan, 1986), and the C4.5 decision tree (Quinlan, 1996). All these algorithms have the same structure but use different splitting criteria and pruning strategies in the tree construction. The C4.5 algorithm is considered in this work.

C4.5 algorithm firstly consists of identifying the root attribute of the tree, this root has as many branches as this attribute takes values if the attribute is qualitative, or two branches that represent a test if this attribute is quantitative. After, the tree is defined by a recursive way. The choice of the most discriminating attributes to be placed at the nodes of the tree is made using a probabilistic measure called entropy, which aims to evaluate the homogeneity of each node. The aim is to construct a tree in such a way that when the data reaches a certain node of the decision tree, it should be more homogeneous than the data that reaches an ancestor node. Thanks to this, the split is constructed at each node in such a way that it maximizes the gain of information provided by the question on the knowledge of the response variable.

There are therefore two fundamental concepts that are essential. The probabilistic entropy, which is a mathematical function (see Eq.1.1) described as a measure of disorder.

$$H(X) = -\sum_{i=1}^{n} p_i log_2(p_i) \tag{1.1}$$

where X is the data sample, n is the number of possible values of an attribute, and p_i is the proportion of examples whose attribute value is "i".

The entropy calculation is used to calculate a second measure called the gain ratio (see Eq.1.2). This measure allows to identify the most interesting attribute to place in the root, and this by calculating the difference between the entropy of the initial set of examples and the entropy associated with a specific attribute "a_i".

$$Gain\ ratio(X, a_i) = \frac{Gain(X,a_i)}{SplitInfo(X,a_i)} \tag{1.2}$$

where:

$$Gain(X, a_i) = H(X) - \sum_{v \in values(a_i)} \frac{card\left(X_{a_i=v}\right) H(X_{a_i=v})}{card(X)}$$

$$SplitInfo(X, a_i) = \sum_{v \in values(a_i)} \frac{card\left(X_{a_i=v}\right)}{card(X)} log_2\left(\frac{card\left(X_{a_i=v}\right)}{card(X)}\right)$$

In the case of a numerical attribute, it is necessary to define a threshold that divides the different values into two partitions, so that it is possible to determine whether the value of that attribute is higher or lower than that threshold. Similarly, the threshold is chosen in such a way as to maximize the information gain associated with the attribute.

Various studies based on decision trees have been proposed in the context of manufacturing. Siltepavet et al., (2012) have proposed an approach for quality improvement within a hard drive manufacturing industry. By using the C4.5 decision tree for the selection of the best manufacturing setting of controllable parameters, the approach allowed reducing the number of defects by 12%. Chen et al., (2010) have used the DT technique in a microchip manufacturing industry in order to identify the causes of defects, making immediate decisions and, eventually, reducing the cycle time required to solve problems related to quality. Besides, Bakır et al., (2008) demonstrated the ability of decision trees to identify the most influential variables that cause defects in a part produced by a casting company. The decision trees were compared to logistic regression, and the results showed that the former method provided significant results compared to the results of the later method.

In order to study the causes of defective products in manufacturing systems, Choi et al., (2017) proposed a key factor extraction approach based on the C4.5 method, considering data from different sources, i.e., critical to quality tables, handwritten data, in-process external air-conditioning sensors, meteorological data, static electricity, vibration data, etc. The key factors are then fed the C4.5 algorithm to investigate the defects causes and to increase the productivity in the manufacturing process. Tsironis et al., (2005) applied an approach based on decision trees and association rules to model the hidden knowledge of production systems. The approach enables quality to be improved by identifying the station responsible for the occurrence of the failure, while the association rules aim to find the relationships between the product characteristics and the different stations in the production line.

In another work Sun, (2010), DTs have been used to extract knowledge from manufacturing execution system (MES) to decision making database (DMDB). DTs allowed offering a structured and standardized way of transmitting data between programs involving different operating systems and different CPUs, making it possible to analyze data in a convenient manner. Chen et al., (2012) proposed an approach that combines K-means clustering, feature selection, and decision tree method for the evaluation of the quality of the performance of suppliers in the manufacturing industry.

DTs were also applied for the prediction of the resistance spot welding (RWS) quality using data collected from an automotive OEM (original equipment manufacturer) (Ahmed and Kim, 2017). The DT was used to extract decision rules for the weld nugget width prediction. That allowed determining the impact of design and process parameters on the response parameters as it allowed generating a model that analyzes and predicts the welding quality when new materials are considered for assembly. Çiflikli and Kahya-Özyirmidokuz, (2010) used the C4.5 decision tree to assess the carpet manufacturing quality. The authors preprocessed

the manufacturing data by cleaning incorrect instances, completing missing values, and more importantly, reducing the dimensionality of data using the binarization approach and the attribute relevance analysis.

To determine the root cause of HGA manufacturing failure, a C4.5 model was implanted to process and explore both machine parameters and product attributes that are susceptible to affect the yield the HGA production, see Figure 1.2 (Taetragool and Achalakul, 2009).

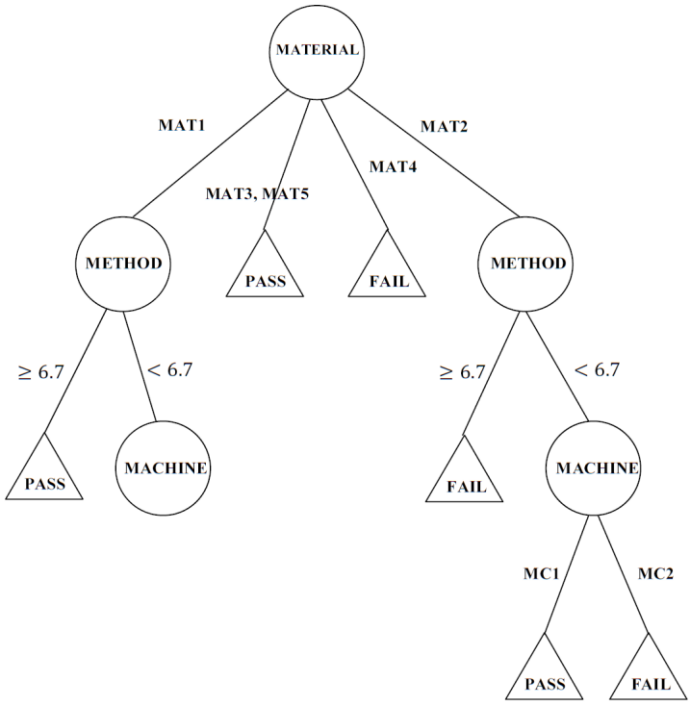


Figure 1.2: C4.5 decision tree for HGA failure identification (Taetragool et al., 2009)

To deal with the binary classification of manufacturing quality problems, Kim et al., (2018) proposed an approach based on cost-sensitive decision tree ensembles. The approach consists at first of preprocessing the manufacturing quality data, namely, time series data. Then, two methods for the labeling of the production lots into defective or normal were suggested. Afterwards, the authors used the C4.5 for the classification of the data as well as three cost-sensitive ensembles to address the problem of the imbalanced data encountered when dealing with manufacturing quality data. The CART decision tree has also been used within the pharmaceutical domain for the optimization of the manufacturing process of pellets (Ronowicz et al., 2015). The CART decision tree was applied therefore to explore the impact of the formulation composition and the process parameters on the pellet aspect ratio. The application of the CART method has resulted in the generation of a tree that can be easily interpreted, making it easier to find the optimal settings, thus, to obtain the desired spherical pellets.

Rokach and Maimon, (2006) have implemented a new iterative method called Breadth-Oblivious-Wrapper (BOW) which consists in selecting the most relevant characteristics to

create the decision tree. This approach makes it possible to create a less voluminous tree while keeping the quality of the extracted knowledge.

The above work focused on the use of decision trees in manufacturing industries. Their objectives can be summarized as the identification of the roots of failures in manufacturing systems, as well as the identification of optimal manufacturing parameters. The adoption of decision trees is due to their ease of interpretation and implementation. However, it can be noted that DTs are rarely utilized alone, where they are usually complemented by other pre-processing and pruning methods to avoid the generation of overfitting models. One of the most popular classifiers for dealing with overfitting problems is support vector machine (SVM). In what follows, the SVM method is presented along with its applications in the context of quality management.

I.1.2.2 – Support vector machine (SVM)

Support vector machine is a supervised learning machine that learns from a training data set and attempt to generalize and make correct predictions on new data by defining a model that assigns new examples to the different classes. the technique was introduced by Boser (Cortes and Vapnik, 1995) and was considered as an advanced classification method with high accuracy, ability to deal with high-dimensional data, and flexibility in modeling diverse sources of data. The technique was applied in different domains such as bioinformatics, environmental modeling, epidemiology, finance, marketing, medical diagnosis, and various scientific fields (Rostami et al., 2015).

SVM was first introduced for the classification of linearly separable binary problems. Given a set of N observations $D = \{(x_1, y_1), \dots, (x_N, y_N)\}$, where x_i represents the i^{th} training data and y_i is the class label of x_i, the SVM tries to find a hyperplane that separates the set space into two areas according to the two different classes. The separating hyperplane is given as $w^T x + b = 0$, where b is the bias of the hyperplane, x are the data located within the hyperplane, and w are the weights that determine the hyperplane's orientation. With an infinite number of hyperplanes ensuring proper data separation, SVM seeks to maximize the margin to make the model less sensitive to data noise. To ensure a maximum margin the SVM formulation (Eq. 1.3) has been established in order to define the parameters of the optimal hyperplane.

$$\begin{aligned} min \quad & \tfrac{1}{2}\|w\|^2 \\ s.t \quad & y_i(w^T.x_i + b) \geq 1 \qquad \forall\, i = 1, \dots, N \\ & y_i \in \{-1,1\} \\ & w \in \mathbb{R}^d; b \in \mathbb{R} \end{aligned} \tag{1.3}$$

Additionally, the linear SVM formulation might be useful to deal with the classification of some non-linearly separable problems. This can be done by allowing the constraint (of formulation 1) to be violated by some datapoints. To this end, the formulation of the SVM is modified so that it can accept that the margin is crossed, and that certain datapoints are on

the wrong side of the hyperplane. A new variable "ξ_i" called slack variable is therefore added in order to inform how far this point violates the constraint. This variable can be either:

- $\xi = 0$: if the data point complies with the constraint.
- $\xi > 1$: if the data point is misclassified.
- $0 < \xi < 1$: if the data point has crossed the margin.

Consequently, the new formulation (Eq. 1.4) introduces a new regularization parameter "C" that manages the compromise between the margin width and the relaxation of the constraint.

$$\begin{aligned} min \quad & \tfrac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \\ s.t \quad & y_i(w^T.x_i + b) \geq 1 \qquad \forall\, i = 1, \dots, N \\ & \xi_i \geq 0 \\ & y_i \in \{-1,1\} \\ & w \in \mathbb{R}^d; b, \xi_i \in \mathbb{R} \end{aligned} \tag{1.4}$$

However, to address most of non-linearly separable problems, the kernel trick method is adopted to create non-linear SVM models. The kernel trick consists of mapping the datapoints to a larger dimensional space called the feature space $\phi: R^d \rightarrow R^p$ such as $z_i = \phi(x_i)$, the aim is to allow the separation of data in a high enough dimensionality space. The associated SVM formulation is given by Eq. 1.5.

$$\begin{aligned} min \quad & \tfrac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \\ s.t \quad & y_i(w^T.\phi(x_i) + b) \geq 1 \qquad \forall\, i = 1, \dots, N \\ & \xi_i \geq 0 \\ & y_i \in \{-1,1\} \end{aligned} \tag{1.5}$$

To apply the SVM model, the dual formulation is considered using Lagrange multipliers α_i:

$$L(w,b) = \tfrac{1}{2}w^T w + \sum_i \alpha_i(1 - y_i(w^T x_i + b)) \tag{1.6}$$

At the minimum, we can take the derivatives with respect to the primal variables "w" and "b"and set these to zero:

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{n}\alpha_i y_i = 0 \tag{1.7}$$

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{n}\alpha_i y_i x_i = 0 \tag{1.8}$$

By replacing the two previous formulas in the formulation (1.9):

$$\begin{aligned} min \quad & \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j x_i^T x_j - \sum_{i=1}^{n}\alpha_i \\ s.t \quad & \sum_{i=1}^{n} y_i \alpha_i = 0 \\ & \alpha_i \geq 0 \end{aligned} \tag{1.9}$$

This formulation has a major advantage when using the kernel trick because the functional form of the mapping $\phi(\mathrm{x}_i)$ does not need to be known since it is implicitly defined in the feature space by the choice of the kernel function $K(x_i . x_j) = \phi(x_i) . \phi(x_j)$.

After introducing the SVM and its different formulations for the classification of linearly and nonlinearly separable data, the application of SVM in manufacturing are presented in the following.

Over the last two decades, SVM has been widely used for quality management and improvement within different manufacturing companies. An SVM-based approach for diagnosing bearing defects has been proposed by Zhi-qiang et al., (2005). Based on the vibrations emitted by the bearings, the model has been used to classify the different defects, which proves the good capability of SVM for multi-classification in mechanical systems. The approach has also been used to classify the feature of the cutting force signals for the prediction of tool breakage in face milling (Hsueh and Yang, 2008).

Gryllias and Antoniadis, (2012) proposed a two-stage hybrid approach for the automated diagnosis of defective rolling element bearings. The proposed approach consists at first of training an SVM model using simulation data, eliminating therefore the need of training the SVM model with experimental data. Afterwards, vibration measurements resulting from the machine under condition monitoring are imported and processed directly by the trained SVM, which allows identifying the normal condition signals from the faulty ones. Besides, Zhang et al., (2015) presented a new method based on digital image processing for on-line monitoring discharge pulse in wire electrical discharge machining-middle speed (WEDM-MS) process. At first, different techniques were applied for the extraction waveform image features and for the reduction of image dimension. These features were then used in a two-stage classification technique that employs support vector machine (SVM) and random forests (RF) for pulse classification and identification.

Baccarini et al., (2011) used vibration as a basis for diagnosing induction motor faults. Four SVM-based models have been developed to classify three recurring defects using the ''one against all'' approach. Similarly, Fernández-Francos et al., (2013) presented an automatic method for bearing fault detection and diagnosis. The method uses the information contained in the vibration signals and an one-class v-SVM to discriminate between normal and faulty conditions. Moreover, by using vibration signals and classification methods, Jegadeeshwaran and Sugumaran, (2015) proposed an approach for diagnosing hydraulic brake system failure. The C4.5 decision tree was used to select the relevant statistical features extracted from the acquired signals, then classification was performed using two types of SVMs with different kernels, whose basic radial function (RBF) ensured better accuracy.

A product quality improvement method in manufacturing process called KML-SVM (**K**ernel-based hybrid **M**anifold **L**earning and **S**upport **V**ector **M**achine) was proposed by Wei et al., (2017). The method consists of applying KML at first in order to solve the problems of manufacturing process quality data dimension curse, while an optimized SVM model is

adopted afterwards to classify and predict low-dimensional embedded data. The method was applied on data collected from AVIC Liming Aero-Engine Group Co., and the results show its efficiency in providing more accurate results of quality classification no matter which kernel function is adopted. Diao et al., (2015) proposed a quality control approach by improving the most influential factors. The approach relies on a principal component analysis (PCA) to identify the factors that lead to quality problems, known as dominant factors (DFs). Then, a quality prediction model to improve DFs is proposed based on the SVM. An additional weight is introduced into the SVM model to improve and increase the prediction accuracy, and to ensure high quality products.

Li et al., (2018) proposed a SVM based approach for the inspection of flip chips based on vibration. 34 features were extracted from the signals, including 18 time domain features and 16 frequency domain features. These features where then employed in order to create a SVM model for the classification of the quality of chips. The performance of the SVM model were optimized using a genetic algorithm, which made it possible to identify good flip chips from detective flip chips with an accuracy of 92.67%. Besides, to reduce the breakdowns of components suffering from failures such as bearings and gears, a vibration-based approach for fault detection has been proposed by Ziani et al., (2017). The RFC (regularized Fisher's criterion) is used for the selection of sensitive features. This criterion is used as fitness function for the optimization of SVM with BSPO (binary particle swarm optimization algorithm).

In order to avoid setting up manually the machines in a milling manufacturing system, Ay et al., (2019) applied the SVM with an objective to identify and to model the dynamic behavior of the machine tool. The SVM based model allowed controlling the process reliably, which resulted in improving the productivity of the process while respecting the product quality restrictions. When compared to the previous control strategy, the proposed monitoring approach showed better performance with 15% shorter manufacturing time. Another process quality monitoring approach based on SVM has been introduced by Escobar and Morales-Menendez, (2019). The approach is addressed to the organizations that generate only a few Defects Per Million of Opportunities (DPMO). The data generated by these organizations are usually complex are unbalanced, which results in training overfitting models. To cope with this problem, the proposed approach considers three criteria - prediction, separability, complexity- for the optimization and the selection of the SVM model.

By analyzing different works, the SVM has shown its potential in dealing with various quality issues within different manufacturing systems. The SVM efficiency is due to its ability to perform well on non-linearly separable data, which is generally the case of quality data, and due to its ability in avoiding the training of overfitting models. In addition, SVM is considered an easy-to-use method that is used to handle large dimensional dataset with different features, which represents an advantage compared to other classification methods. However, other approaches have also shown great performance in dealing with nonlinear classification problems, such as the Multilayer Perceptron (MLP), which allows

multiple classification tasks (not just binary classification) to be performed using a single structure.

I.1.2.3 – Multilayer perceptron (MLP)

A Multilayer Perceptron (MLP) is an artificial neural network model organized in several layers in which information flows from the input layer to the output layer. Each layer is made up of a variable number of neurons, the neurons of the output layer being the outputs of the overall system. In MLP, all neurons have connections between them, and these connections are identified by associated weights. A MLP consists of at least three layers of nodes: an input layer, a hidden layer, and an output layer (see Figure 1.3). A j^{th} neuron can be defined mathematically by Eq. 1.10.

$$O_j = f(u_j + b_j) \tag{1.10}$$

$$u_j = \sum_{i=1}^{n} w_{ij} x_i$$

where x_i denotes the i^{th} input feature, w_{ij} is the j^{th} connection weight of the neuron, u_j is the linear output of the linear combination among weighted inputs, b_j is the bias term, f is the activation function, and finally O_j is the output signal of the neuron.

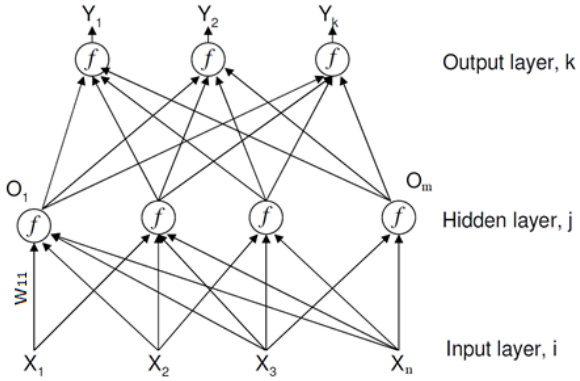


Figure 1.3: A multilayer perceptron with one hidden layer

Training an MLP is still considered one of the main topics in neural networks research. Finding the right network for a problem means defining a good architecture (number of nodes and number of hidden layers) as well as the best learning parameters. On the one hand, an oversized network needs more time for training and has higher chances to overfit the training data, while an undersized network may not converge at all (Ramchoun et al., 2016).

The learning phase of MLPs consists of identifying a set of weights to model the relationship between the inputs and the outputs in an accurate manner. The main goal is to find the combination of weights which result in the smallest error. One way to tune the weights of a MLP is the use of the backpropagation algorithm, that tries to find the minimum of the error surface using gradient descent. This technique consists in correcting errors

according to the importance of the elements that have precisely participated in the realization of these errors. In the case of neural networks, the weights that contribute to generating a large error will be modified more significantly than the weights that generated a marginal error. Additionally, the performance of the backpropagation algorithm depends on two parameters, a momentum that assist the training phase in avoiding local minima, and a learning rate that represent the step taken by the iterative gradient descent. Finally, it should be noted that MLP networks requires several training epochs (iterations) before achieving a sufficient error level defined by the addressed problem. In the following, different applications of MLP within manufacturing systems are presented.

In different works, MLPs have been used for modeling and optimizing the quality of manufacturing systems. Shen et al., (2007) proposed an approach that combines the MLP neural network and the genetic algorithm to assess the optimal process conditions for the injection of high-quality molded plastic parts. This allowed modeling the complex non-linear relationship between the process conditions and the quality indexes of the injection molded parts. Establishing this relationship has made it possible to improve the product quality by mainly improving the quality index of the volumetric shrinkage variation in the part. Besides, a novel approach based on principal component analysis (PCA) and artificial neural network (ANN) was presented by Mirapeix et al., (2007) for the automatic detection and classification of arc-welding defects.

A new approach for fault detection of three-phase induction motors based on MLP has been proposed by Ghate and Dudul, (2011). The proposed MLP model takes as inputs some statistical features extracted from the stator current. A principal component analysis (PCA) was performed on the input data in order to reduce the number of the features and remove the redundant or irrelevant information. It has been shown that the MLP model is able to detect the different faults in induction motors with an accuracy exceeding 98%. Delgado et al., (2012) presented a new methodology for the detection and the classification of defective bearings of electrical machines. The methodology is used for the classification of six different bearing conditions using information in time-domain from the vibration data. To do so, the vibrations signals from accelerometers are acquired at first, then a total of 15 features from time-domain were extracted from each acquired signal. Afterwards, a discriminant analysis (DA) was applied to evaluate the significance of the features, and a curvilinear component analysis (CCA) was used to project the data into a lower dimensionality space while preserving the structure of inter-point distances. Finally, data were fed to a MLP neural network, allowing therefore the distinction between the six bearings conditions. The accuracy of the classification by MLP reached 94%, where all points corresponding to healthy machine were correctly classified.

In order to ensure and guaranty the required quality of products within a lacquerer company, Noyel et al., (2013) proposed an approach based on an MLP network to determine the optimal setting for production machines. Even if the studied manufacturing system is free of human factors, the production quality is unpredictable and fluctuates. Consequently, the

authors used the MLP model to define the lower and upper limits for each controllable factor, taking into account both controllable and non-controllable factors. The proposed approach has made it possible to better monitor the quality of the process, thereby allowing a more efficient product flow and a process that is adapted to change. In the case of a diagnosis of rolling-element bearing faults, Unal et al., (2014) has introduced an approach combining the Fast Fourier Transform (FFT) and the back-propagation ANN algorithm. At first, the signals emitted by bearings (including faulty ones) were recorded, then features were extracted from the signals using envelope analysis accompanied by Hilbert Transform and Fast Fourier Transform (FFT); then, a MLP was trained for the classification of faulty bearings into three kinds of defects with an accuracy reaching 98%.

MLP neural network was also used for the early detection of faults in bevel gearboxes Jedliński and Jonak, (2015). The wavelet transform method was applied in order to preprocess the vibrations collected from triaxial vibration acceleration sensor to bring out the informative part of the signals. The preprocessed data were trained using a MLP network, allowing thus the identification of good gears from the damaged ones with an accuracy higher than 92%. Casalino et al., (2016) implemented a MLP neural network for the investigation of the main effects of the process parameters on the laser welding process quality. The ANN has made it possible to overcome the difficulty of predicting the characteristics of the bead in the laser welding process of the aluminum alloy by establishing a model that links the process parameters to the bead characteristics.

Al-kharaz et al., (2019) have proposed an MLP-based approach for product quality inspection in semiconductor manufacturing processes. The approach consists of collecting data from the process alarm system and then building an MLP to model the relationship between the extracted features and the product quality. By comparing the model's performance to other models, the proposed MLP network allows better prediction of defective products.

The different research studies above demonstrate the effectiveness of MLPs in addressing different quality tasks in manufacturing systems. Many of these works focused on the ability of MLPs in the detection of faulty process conditions, the identification of defective products, and the classification of types of defects. The MLPs showed also great performance in modeling complex non-linear problems, as well as a good generalization ability.

In this first part, three machine learning classifiers (DT, SVM, and ANN) were introduced. Each classifier has been successfully implemented and used to solve different quality tasks within different production systems. Accordingly, the three algorithms have been applied on some case studies as presented in the following section.

I.2 - Case studies

I.2.1 - Product quality assessment by SVM – application on a manufacturing process (Roll_0/1 data)

In order to assess the quality of the manufacturing system (related to *Roll_0/1* data), a dataset is considered with 874 instances and 95 parameters in which 438 instances are defined of class 0 (poor quality product) and 436 instances of class 1 (good quality product).

Since the quality issue encountered in this case study is a binary classification problem and the number of the process parameters is relatively high, the support vector machines was chosen as the machine learning method to address this problem, as SVM showed great performance in handling with large dimensional datasets.

In general, two data sets are required to perform a classification task: a training set to build a model by pairing the inputs with the corresponding labels (outputs), and a test set to estimate the accuracy of the model on the basis of new data. In our case, the dataset to be analyzed is divided into two subsets. A training set representing 2/3 of the initial data, while the test set presents 1/3 of the data.

Subsequently, and before starting the training phase, the SVM hyper-parameters need to be identified and optimized. Hyperparameters are defined as the parameters to be initialized before model training, as well as the characteristics that regulate the entire training process. Such parameters could impact the efficiency of the model, such as its complexity or its learning rate (Mantovani et al., 2019). Three main SVM hyper-parameters are defined:

- The kernel used to map the data into the new space to perform non-linearly separations.
- The kernel's parameters.
- The "C" parameter that is essentially a regularization parameter that controls the tradeoff between achieving a low error on the training data and maximizing the width of the margin.

Once identified, the SVM hyper-parameters are tuned using a genetic algorithm (GA). Accordingly, to optimize the SVM's predictive performances, different evolutionary algorithms have been utilized such as genetic algorithm (Zouhri et al., 2020), particle swarm optimization (Lin et al., 2008), artificial bee colony algorithm (Yang et al., 2015), and firefly algorithm (Olatomiwa et al., 2015). Genetic algorithms (GA) have been widely used to solve optimization problems and are therefore considered in this work. The GA optimization can be summed up by Algorithm 1.

Algorithm 1 Optimization of SVM hyper-parameters by genetic algorithm.

Inputs: training set ***Ml***, test set ***Mt***, population size ***pop_size***, number of generations ***max_gen***

Output: optimal SVM kernel space $\boldsymbol{Ker}$

01: Encode the solution using an integer representation
02: Generate an initial random population of size ***pop_size: POP***
03: for $i \in \{1, \dots, \boldsymbol{pop\_size}\}$ do
04: Train an SVM model using the hyperparameters encoded in ***POP***[***i***]***: Model_SVM***
05: Evaluate the fitness of ***POP***[***i***] by predicting ***Mv,*** using ***Model_SVM***
06: end for
07: Pick the solution with the highest fitness score: $\boldsymbol{Ker}$
08: for $i \in \{1, \dots, \boldsymbol{max\_gen}\}$ do
09: Apply 3-way tournament selection to create the off-spring: ***POP***
10: Apply uniform crossover to ***POP: POP***
11: Apply random resetting mutation to ***POP: POP***
12: for $j \in \{1, \dots, \boldsymbol{pop\_size}\}$ do
13: Train an SVM model using the hyperparameters encoded in ***POP***[***j***]***: Model_SVM***
14: Evaluate the fitness of ***POP***[***j***] by predicting ***Mt,*** using ***Model_SVM***
15: end for
16: Pick the fittest solution of the i^{th} generation: $\boldsymbol{Ker_i}$
17: if $\boldsymbol{Ker_i}$ is fitter than $\boldsymbol{Ker}$
18: $\boldsymbol{Ker} \leftarrow \boldsymbol{Ker_i}$
19: end if
20: end for
21: return $\boldsymbol{Ker}$

Algorithm 1 was run by considering a population size equal to 100, and a number of generations equal to 50. The kernel functions were either polynomial, sigmoid, or radial basis function (RBF), and the "C" parameter was varied between 0.1 and 1000. Also, for the RBF kernel, the "σ" parameter was varied between 0.001 and 10. Table 1.1 details the results of the application of the genetic algorithm with the considered specifications on the *Roll_0/1* data. Details of the GA are provided in Appendix A.

Table 1.1: Optimal SVM hyperparameters – *Roll_0/1* data

Dataset	Kernel	σ	C	Accuracy of test set
Roll_0/1	RBF	0.317	883.4	91.43%

Based on these results, the product quality level can be predicted with high confidence, allowing for improved quality and efficiency of the process.

I.2.2 - Quality assessment using SVM

By following the same approach developed in the previous case study, two chemical and one mining process datasets are utilized to establish SVM models for the prediction of the quality levels. A summary of these datasets is provided in Table 1.2.

Table 1.2: Summary of the considered *chemical* datasets

Name of dataset	Number of parameters	Size of class 1	Size of class 2	Manufacturing system	Use of SVM
Chem_4/8	11	163	175	Chemical industry	Quality level prediction
Chem_5/7	11	1457	880	Chemical industry	Quality level prediction
Mines_0/1	22	5940	5940	Mining industry	Quality level prediction

By splitting the datasets into training sets (2/3) and test sets (1/3), two SVM models were tuned using the same genetic algorithm detailed in Algorithm 1, and by considering the SVM hyperparameters. The different results are given in Table 1.3.

Table 1.3: SVM prediction accuracy

Dataset	Kernel	σ	C	Accuracy of test set
Chem_4/8	RBF	0.292	15.0	89.47%
Chem_5/7	RBF	0.340	231.8	85.77%
Mines_0/1	RBF	0.364	23.4	82.34%

The SVM models have made it possible to establish a relationship between the process parameters and the final quality level. This can therefore be used to improve the efficiency of the process by predicting whether or not the setting used would result in a good quality production.

I.2.3 – Identification of defects causes by C4.5 – application on a piping manufacturing process

We considered in this application a dataset provided by an industrial partner: FIVES Nordon company. FIVES Nordon's activity is the production of industrial pipe elements, the installation and maintenance of fluid networks and specific equipment.

The aim of this application is to identify the various welding defects and their causes. The data set consists of 948 instances, 23 parameters and 14 labels representing 13 types of welding defect.

The decision tree C4.5 was therefore used. The approach aims at identifying the parameters leading to weld defects and to extract the rules leading to a non-defective weld.

The tree generated from the dataset contains 47 rules as shown in Figure 1.4.

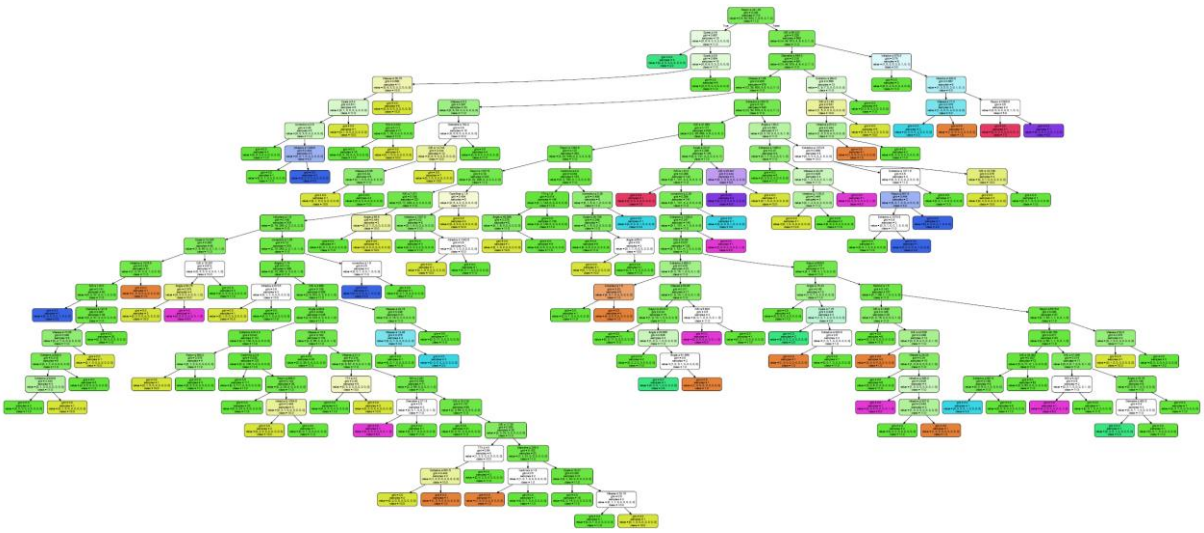


Figure 1.4: Application of C4.5 on the welding data

As it can be seen, the built tree is difficult to read and interpret. This is mainly due to the significant number of process parameters that define the dataset. To address this issue, the parallel coordinates method (Roberts et al., 2019) was used to better visualize the results.

Parallel coordinates have been proposed by d'Ocagne as a powerful technique that can be easily manipulated and interpreted. This technique enables the visualization of a large amount of information (e.g. multidimensional data) in a clear manner and without needing additional calculations. The method represents data as a set of lines passing through parallel axes representing the variables. Each data point is represented by a set of lines called polygon chains, see Figure 1.5.

Data			
	Variable A	Variable B	Variable C
Item 1	50	100	2.0
Item 2	30	115	0.5

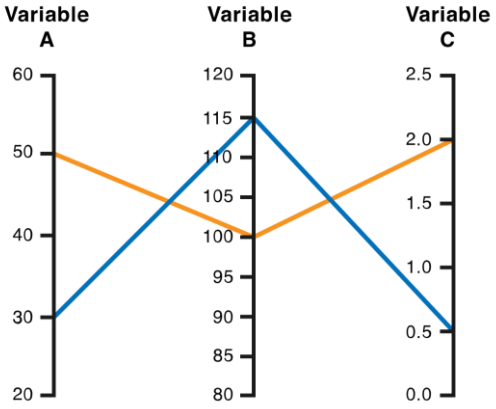


Figure 1.5: Illustration of a data visualization by parallel coordinates

Therefore, in order to simplify the visualization of the extracted rules, we propose to visualize each branch of the decision tree by two polygonal chains representing the minimum and maximum values of each parameter, including the class parameter as shown in Figure 1.6 (results anonymous for confidentiality reasons). In this example, an extracted rule related to the defect type "3" is shown. Experts can easily analyze the discriminant parameters that cause this defect through parallel coordinate representation. In this case, it is the set of operating parameters whose range of values is restricted (identified by ellipses in Figure 1.6). The experts were able to compare this analysis based on C4.5 with their practices. Only 5 (yellow ellipses) of the 7 parameters identified by C4.5 were already known to the experts. This combination of C4.5 techniques and parallel coordinates makes it possible to predict the different defects and to identify the discriminating parameters.

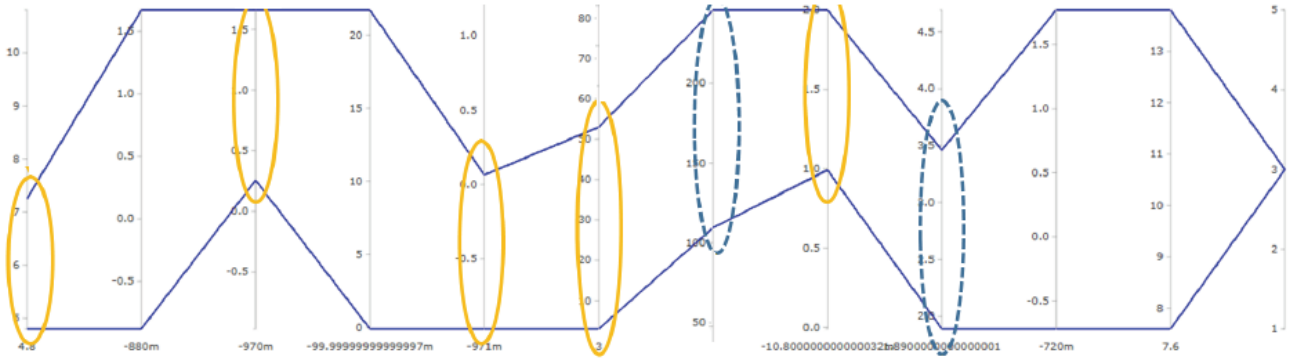


Figure 1.6: Representation of a branch by parallel coordinates

I.2.4 – Optical process monitoring for Laser Powder Bed Fusion (L-PBF)

We aim in this case study to propose approaches for the evaluation of the quality of the Laser-Powder Bed Fusion (L-PBF) process. The L-PBF still lacks process quality and reproducibility. For this reason, robust process monitoring needs to be developed to reduce the process variation and ensure quality. Accordingly, two approaches have been developed to predict the quality of the L-PBF products from optical signals. The first approach consists of extracting statistical features from the signals, then feeding them to machine learning classifiers to identify the quality of the different manufactured parts. The approach is compared afterwards to a deep learning approach that tries to perform the quality classification from raw optical signals. This comparison indicates if the human knowledge (statistical features extraction) is important or not. The comparison is illustrated in Figure 1.7.

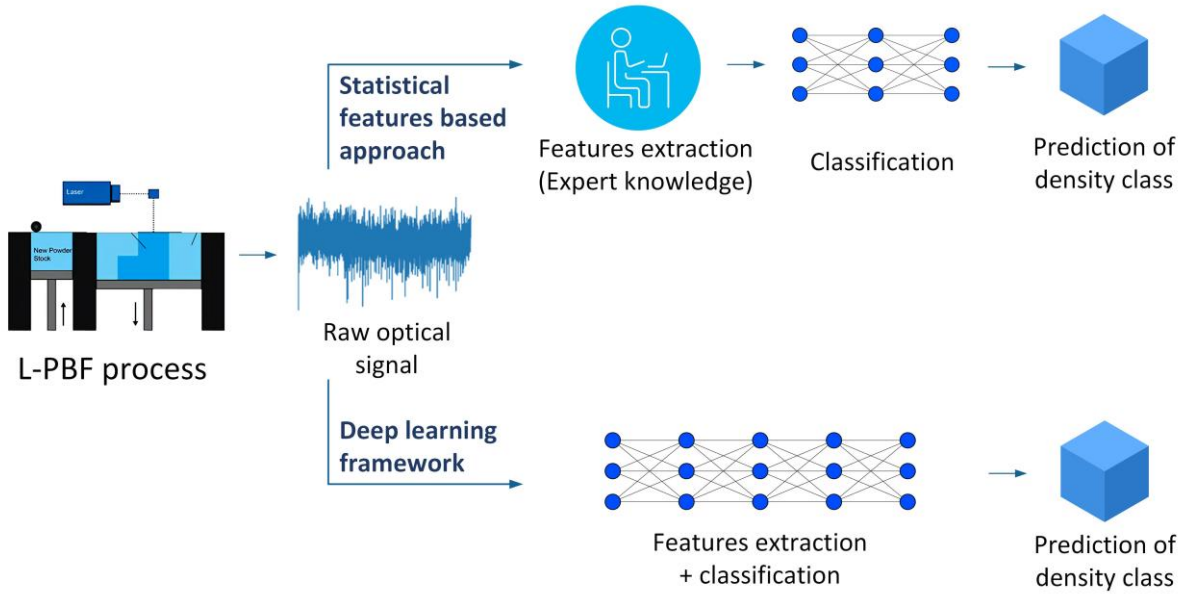


Figure 1.7: Optical quality monitoring approaches for L-PBF process

The following approaches are applied to a real-world data set provided by Siemens Corporate Technology. The different steps for the production of different parts and the collection of data are presented at first. The two approaches are then described and the results of their application to Siemens data are presented and analyzed.

Instrumentation and Experiment

In this part, the quality measure is defined at first, then, the followed Design of Experiments and the monitoring setup are depicted, finally, the quality measurement technique used is presented.

The basic concept of quality is to meet customer standards by maintaining the productivity of resources throughout the life-cycle of the product, process and production system. It is therefore necessary to identify an appropriate measure to assess the efficiency of the manufacturing process. Accordingly, the part density is chosen as a quality measure of the L-PBF process in this paper. This choice is due to the fact that porosity affects the mechanical strength, fatigue strength, and the elongation to rupture of a layer-wise part, thus, the component should be sufficiently dense to prevent failure during service (Slotwinski et al., 2014).

In our work, the parameters of the L-PBF process were intentionally adjusted in order to invoke different pore concentrations within the manufactured parts. To do so, 18 different process parameters (18 configurations) were defined at first with the help provided by domain experts, then for each configuration, two 316L stainless steel cubes of 123 layers were produced. A total of 4402 optical signals have been collected by recording the corresponding signal of each layer using the "Kleiber KGA 740-LO" pyrometer. These signals represent the emissions occurring at the infrared wavelength range from the process zone.

Subsequently, the density of the 36 cubes was measured using the Archimedes method because of its advantages when compared to other measurement techniques (Spierings et al., 2011).

To summarize, Figure 1.8 illustrates the monitoring approach followed, while Table 1.4 presents the different measured densities of the different cubes.

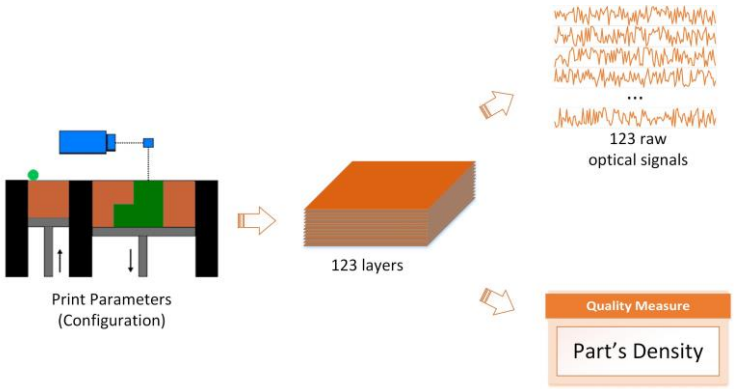


Figure 1.8: Illustration of the L-PBF optical monitoring approach

Table 1.4: Densities of the manufactured cubical specimens

Configuration	Density cube 1	Density cube 2	Configuration	Density cube 1	Density cube 2
1	97.83 %	97.78 %	10	97.65 %	97.76 %
2	95.89 %	95.46 %	11	95.28 %	95.38 %
3	94.58 %	95.73 %	12	95.80 %	95.90 %
4	98.76 %	98.88 %	13	98.55 %	98.56 %
5	98.33 %	98.75 %	14	98.01 %	98.20 %
6	96.68 %	96.35 %	15	95.62 %	95.56 %
7	95.09 %	93.61 %	16	98.16 %	97.14 %
8	98.55 %	99.03 %	17	98.56 %	98.61 %
9	98.13 %	98.00 %	18	97.77 %	97.88 %

At this stage, all the data needed are collected. The two approaches are then used to establish a link between the signal of the layer and the density of the final part (cube).

Statistical Feature-based Approach for Quality Classification of L-PBF Parts

The approach consists of defining 10 statistical features as given in Table 1.5. These features are then extracted from each optical signal, which allows having a dataset of 4402 observations and 10 features. Later, data is labeled in a way that generates a balanced dataset. Three classes are defined (high, medium and low density), each of which consists of 12 cubes. The cubical specimens with a density higher than 98.15% were considered to be high-quality products (Class 1), while those with a density lower than 96.13% were considered to be low-

quality products (Class 3), finally, those with a density between 96.13% and 98.15% are medium-quality parts (Class 2). Once the data were labeled, the dataset was normalized and randomly divided into three balanced sets as described in the following:

- Training set: 18 cubes (50%) ⇒ 2205 signals.
- Validation set: 9 cubes (25%) ⇒ 1099 signals.
- Test set: 9 cubes (25%) ⇒ 1098 signals.

Table 1.5: List of extracted statistical features

List of extracted statistical features	
sum of the absolute values of consecutive changes	$\sum \lvert x_{i+1} - x_i \rvert$
time series complexity	$\sqrt{\sum_{i=0}^{N-lag} (x_{i+1} - x_i)^2}$
number of values higher than the mean	-
first location of the maximum value	-
mean	-
median	-
mean over the differences between subsequent time series of values	$\frac{1}{N}\sum (x_{i+1} - x_i)$
number of peaks	-
entropy	-
variance	$\frac{\sum (x_i - \overline{x})^2}{N-1}$

The generated dataset is then fed to two machine learning classifiers, i.e., SVM and MLP. This would make it possible to evaluate the effectiveness of the statistical features-based approach, as it would allow a comparison of the predictive performance of SVM with the predictive performance of MLP.

Classification by SVM

With three classes, three SVM models were trained and optimized to perform a one-vs-one (OVO) multi-classification task. Thus, the genetic algorithm (Algorithm 1) was run three times. Multi-classification by SVM was applied to the entire dataset in order to predict the different density classes. Table 1.6 presents the optimal SVM hyperparameters and Table 1.7 presents the different prediction accuracies, where the overall accuracy is the accuracy of the entire dataset.

Table 1.6: Optimal hyperparameters of the SVM

SVM model	Kernel	σ	C	Accuracy of validation set
Class1 vs Class2	RBF	0.07	79	99.18%
Class1 vs Class3	RBF	1.851	983	100.00%
Class2 vs Class3	RBF	5.651	16	99.45%

Table 1.7: Prediction accuracy of SVM

Accuracy training set	Accuracy validation set	Accuracy test set	Overall accuracy
93.78%	99.09%	90.07%	94.18%

Classification by MLP

Unlike SVMs that perform binary classifications, MLP structures allow multi-classification tasks to be carried out, i.e., only one structure is trained and optimized. Therefore, MLP was applied in this case study for quality classification. At first, the different MLP hyperparameters are identified as well as their variation ranges, see Table 1.8.

Table 1.8: MLP *hyperparameters* to tune

Hyperparameter	Variation range
Hidden layers	1-3
Nodes	3-11
Batch size	{32,64,128,256,512}
Batch normalization	{On,Off}
Epochs	200 -with early stopping-
Weight initiliazer	He-normal
Activation function	Relu
Optimizer	Adam

Afterwards, a grid search is performed to test all the different combinations of these hyperparameters (8190 combinations). This grid search allowed identifying an optimal MLP artificial neural network with the following characteristics:

- Hidden layers = 2
- Nodes of hidden layer 1 = 8
- Nodes of hidden layer 2 = 5
- Batch normalization = On
- Batch size = 128

This MLP structure allowed getting the following results, see Table 1.9.

Table 1.9: Prediction accuracy of MLP

Accuracy on training set	Accuracy on validation set	Accuracy on test set	Overall accuracy
93.46%	98.36%	91.98%	94.32%

Both SVM and MLP are very good at learning from statistical features, with a similar prediction accuracy over 93%. The generalization of these classifiers is also great, as the accuracy of the validation set, and the test set is approximately 98% and 90%, respectively. These results prove how density is well correlated to optical signals. These results also validate the selected statistical features and their ability to predict the different density classes. In the following, the application of deep learning frameworks for density prediction is presented.

Deep Learning Approach for Quality Classification of L-PBF Parts Based on Raw Signals

The main objective of the approach is the prediction of the density of the manufactured parts from the raw optical signals. An one-dimensional convolutional neural network (1D-CNN) was therefore tested. Having signals with different lengths, a linear interpolation was performed in order to up-sample all the optical signals. This interpolation allowed having a dataset of 4402 observations and 500000 columns, where each one of these columns represents a timestep.

Two 1D-CNN architectures derived from AlexNet and VGG models were tested and evaluated. The deep learning architectures were defined as shown in Figure 1.9. The architectures were then trained by varying the number of filters between 8 and 32, the strides between 1 and 4, and the filter size between 2x1 and 5x1. Additionally, the RELU activation function, the ADAM optimizer, and the He-normal weights initializer were used to train the networks. The best out of all these networks is depicted in Figure 1.10.

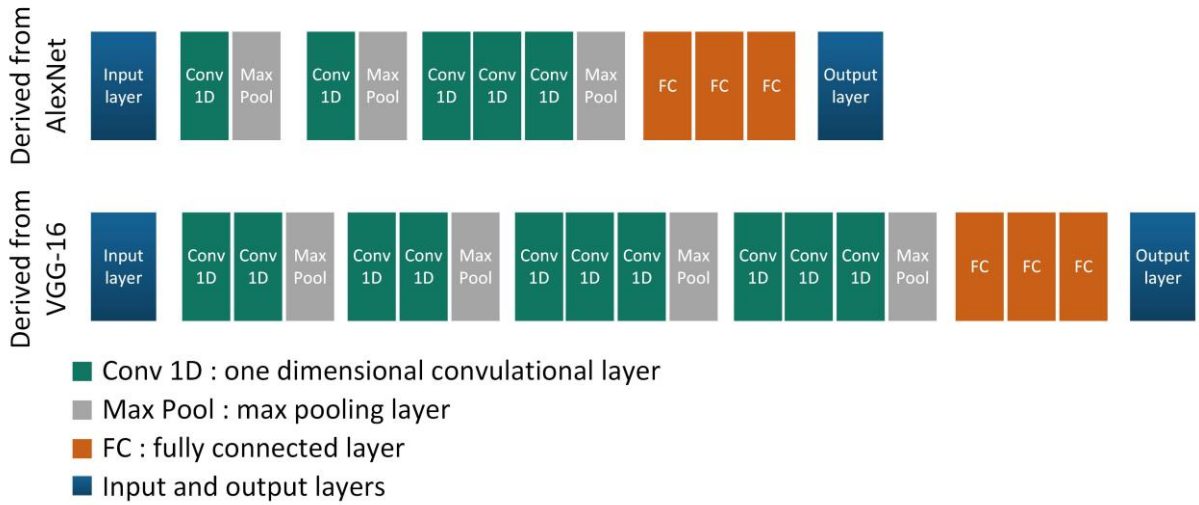


Figure 1.9: 1D-CNN architectures tested

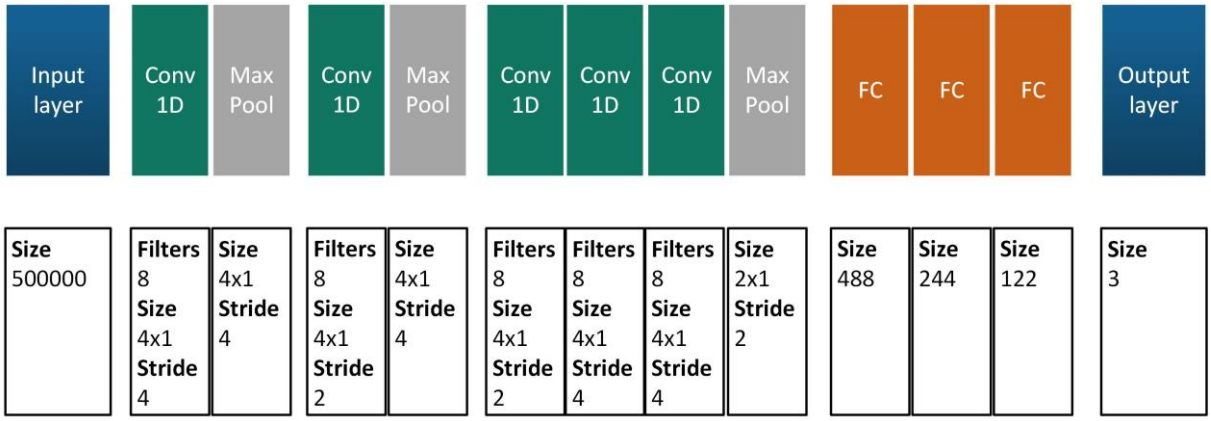


Figure 1.10: Optimal 1D-CNN model

The model illustrated in Figure 1.10 was able to identify the different density classes with an accuracy of 94.88% on the training set, and with accuracies of 88.54% and 81.24% on the validation set and the test set, respectively. Still, these results are inferior to the ones found by the first approach, which emphasize the importance of the expert knowledge in data preprocessing and assisting the machine learning methods for a better model performance.

To summarize, in these case studies, the great ability of classification methods in addressing different quality issues within manufacturing systems has been demonstrated. In the first two case studies (rolling and chemical processes) the SVM showed great ability in predicting the quality levels of different products with accuracies exceeding 85%. While in the third case study, a coupling of DTs and parallel coordinates allowed the identification of the main parameters leading to different welding defects in an interpretable and easy way. Finally, in the last case study, MLP and SVM made it possible to establish models that link the optical signals of layers to the density of the final parts, which can be used for improving the efficiency and for monitoring the quality of the L-PBF process by preventing end-of-line quality assurance.

Nonetheless, the classifiers utilized earlier assume that all the elements of a dataset are precise. This assumption may be violated in many real-world applications due to imperfections of measurement tools, imprecision of expert information, etc. (Utkin and Zhuk, 2017). Therefore, the impact of uncertainty/noise on the predictive performance of classification methods must be analyzed and controlled in order to guarantee an optimal performance under the worst-case scenario.

For this reason, the definition of noise, its sources and its effect on classification methods are discussed in the following section. Subsequently, different groups of methods used to deal with uncertain data are identified. Finally, the choice of SVM as a classifier to study is justified and the different works and models that use SVM under uncertainties are presented.

I.3 – Support vector machines under uncertainties

I.3.1 Data noise: definition, sources, and impacts

In the real world, knowledge is fundamentally uncertain; the same is true in manufacturing: measurable data for machine learning and data mining is usually inaccurate, incomplete or noisy. Machine learning systems are required to process imperfect data and are thus supposed to reason under conditions of ignorance (Wickramasinghe, 2017). In order to better manage this ignorance, researchers often try to construct general taxonomies about the various forms of ignorance, such as uncertainty, incompleteness, ambiguity and confusion (see Figure 1.11).

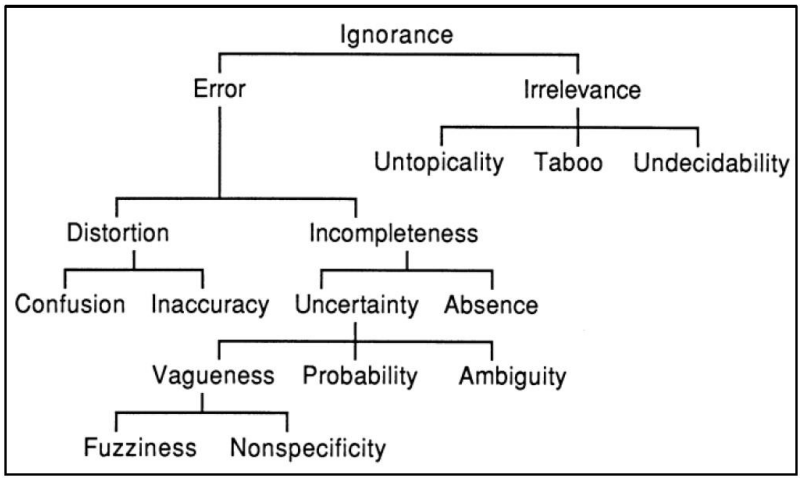


Figure 1.11: Taxonomy of ignorance (Smithson 1989)

Numerous formalisms for explaining uncertainty have been proposed over the years. They include essentially numerical methods based on probability theory, fuzzy logic and possibility theory, as well as largely symbolic methods such as default logics and argumentation (Smithson, 1989). Hunter and Parsons, (1998) have classified uncertainty management formalisms into two broad families:

- The first family contains all numerical approaches such as probability theory, evidence theory, and possibility theory based on fuzzy logic.
- The second family includes approaches based on non-monotonic reasoning where one finds the default reasoning methods and the methods of auto-epistemic reasoning.

Two forms of noise sources are defined in the context of the classification methods, i.e., attribute noise and class noise (Wickramasinghe, 2017). The former may be either: erroneous values of attributes, missing values of attributes, inconsistent data distribution, and redundant data. On the other hand, there are two sources of class noise. The first source is the labeling of the same examples that occur more than once with different labels, while the second source consists of the labeling of an example with a wrong label. The second source of class

noise is commonly encountered when two classes have similar symptoms. These different sources of noise should be properly considered, and their impact on the classification must be studied and analyzed.

Accordingly, Zhu et al., (2006) assessed the impact of both attribute noise and class noise on the performance of classification methods. Their experiments concluded that the accuracy of the classifiers is better if noisy instances are removed or filtered out. Besides, Yin and Dong, (2011) identified three groups of methods to control noise in classification: The first group is based on robust models that take into account data noise. The other two groups aim at building classifiers from "cleaner" datasets, by removing instances that are supposed to be noisy or by replacing their values with more appropriate ones. Two examples of these methods are EDIR and PANDA. The EDIR (Error Detection and Impact-sensitive instance Ranking) system corrects the values of the attributes of suspicious instances that have an influence on the learning of the classifier, while the PANDA (Pairwise Attribute Noise Detection Algorithm) detects instances that contain noise in one or more attributes by identifying those that have values with significant deviations from the normal.

According to the literature findings, class noise has a greater effect on the efficiency of classification methods. Despite this, it is known that the attribute noise is more difficult to manage (Frenay and Verleysen, 2014). This difficulty is due to the challenge of identifying the instances with inaccurate values and the complexity of handling them (Wickramasinghe, 2017). Therefore, in this work, we concentrate on studying attribute noise (measurement uncertainties) and its impact on the performance of SVM.

There are many motives for studying the SVM method. At first, SVM has an advantage when compared to DTs. As shown in Figure 1.12, DTs are considered as hard margin separators that try to perfectly classify the data of the learning set. On the other hand, the soft margin of SVM allows to tolerate classification errors when fitting the model and thus avoid overfitting. This makes the SVM model less noise-sensitive and more efficient in generalization. Besides, MLPs are comparatively less effective than SVMs (Zanaty, 2012). The effectiveness of SVM over MLP is due to several reasons. Firstly, MLPs are based on empirical risk minimization while SVMs utilize the Structural Risk Minimization (SRM) principle that addresses the problem of overfitting by balancing the model's complexity and its success at fitting the training data. Another great advantage of SVM is the formulation of its learning problem which leads to a quadratic optimization problem. It greatly reduces the number of operations in the training phase, which makes it usually much quicker for large data sets (Shawe-Taylor et al., 1998). Finally, tuning an MLP model is considered more challenging because many hyperparameters (number of layers, number of neurons per layer, optimizer, activation function, etc.) need to be calibrated simultaneously.

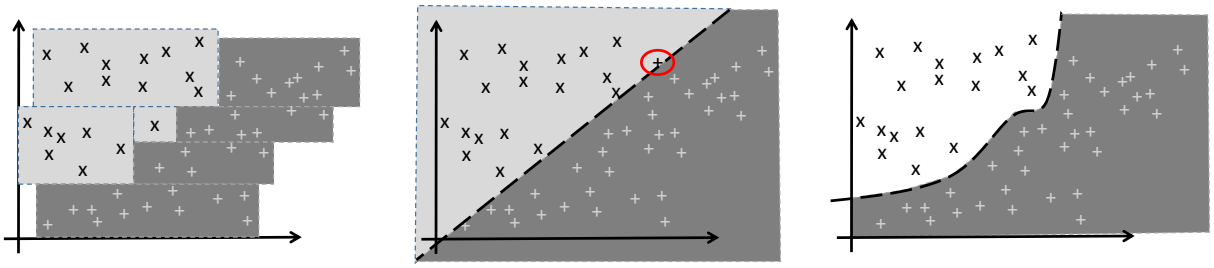


Figure 1.12: Illustration of the classification performed by the studied classifiers

In the following the different robust models and approaches of SVM under uncertainties are presented.

I.3.2 - SVM under uncertainties

Several works have been proposed to deal with the classification with SVM under uncertainties. These works generally rely on robust optimization or a combination of different tools to ensure an optimal performance on noisy datasets. The works relying on robust optimization are presented at first, where several improved SVM models are proposed to tackle the impact of different types of uncertainties. After that, various approaches that consider uncertainties when using SVM for classification are described.

Accordingly, Bi and Zhang, (2005) studied a new probabilistic model that takes into account data subject to measurement uncertainties, see Eq. 1.11. The authors have assumed that the data points are subject to additive noise. This resulted in the expression of the data point by $x_i = x'_i + \Delta x$ and the addition of a new constraint to the SVM formulation, which consists of bounding the noise.

$$\begin{aligned} min \quad & \frac{1}{2}\|w\|^2 + C\sum \xi_i && (1.11) \\ s.t \quad & y_i(w^T.x'_i + b) \geq 1 - \xi_i && \forall\, i = 1, \dots, n \\ & \xi_i \geq 0 && \forall\, i = 1, \dots, n \\ & \|\Delta x_i\|_2 \leq \delta_i && \forall\, i = 1, \dots, n \end{aligned}$$

In addition, when using the kernel trick, the authors proposed to use a first-degree Taylor's polynomial approximation $(K(x_i + \Delta x_i, .) = K(x_i, .) + \Delta x_i.K'(x_i, .))$ in order to avoid mapping uncertainties to the new kernel space. This new model would therefore allow an uncertain data to be free in the circle of center "x_i" and radius "δ_i".

Similarly, Niaf et al., (2011) defined a new formulation of SVM called the P-SVM, to simultaneously deal with certain and uncertain classes, see Eq.1.12. They defined a new probabilistic variable "p_i" that expresses the uncertainty of a datapoint, and allows forcing the right classification by setting limits $\{z_i^-;\ z_i^+\}$ that depend on "p_i".

$$min \quad \frac{1}{2}\|w\|^2 \tag{1.12}$$

$$s.t \quad y_i(w^T.x'_i + b) \geq 1 - \xi_i \qquad \forall\, i = 1, \dots, n$$

$$z_i^-(p_i) \leq w^T.x_i + b \leq z_i^+(p_i)$$

Pant et al., (2010) have developed a robust SVM model for the classification of data with bounded measurement uncertainties. Their work consists of reformulating the slack variable ξ_i in terms of x_i and restating the SVM classification problem as an unconstrained optimization problem as expressed by Eq. 1.13. This would make it easier to classify noisy data, as it would make it possible to get a computationally portable problem.

$$min \quad \frac{1}{2C}\|w\|^2 + \sum_{i=1}^{N}[1 - y_i(w^T.x_i + b)]_+ \tag{1.13}$$

Also, for unbalanced datasets, the authors proposed to separate the SVM formulation into two parts representing the two classes. This separation helps to control the perturbation on samples, which can be critical in the classification of important samples the minority class.

$$\min \lambda\|w\|^2 + \sum_{y_i=1}[1 - y_i(w^T.x_i + b) + \|\delta_i\|_p\|w\|_q]_+ + \sum_{y_i=-1}[1 - y_i(w^T.x_i + b) + \|\delta_i\|_p\|w\|_q]_+ \tag{1.14}$$

Robust optimization was also used by Jeyakumar et al., (2014). Their study consisted of formulating a new robust version of Farkas' lemma, as well as reformulating the SVM problem as a standard quadratic optimization problem in order to control uncertain data during the SVM classification.

Ghaoui, (2003) proposed a robust model when the uncertainty is expressed as intervals. To that end, a hyper-rectangle is defined for each data point x_i where each side of that hyper-rectangle is defined by an upper bound and a lower bound,$(l_{ij} \leq x_{ij} \leq u_{ij})$. Therefore, in order to ensure a robust optimization, the hyper-rectangle should satisfy $y_i(w^T.x_i + b) \geq 1 - \xi_i$. Besides, Fan et al., (2014) addressed a more general case by proposing an SVM model that deals with data prone to polyhedral uncertainty. Generally, a datapoint $(x_i \in \mathbb{R}^m)$ is considered to be polyhedral uncertain when it satisfies $D_i x_i \leq d_i$, with $D_i \in \mathbb{R}^{q\times m}$ and $d_i \in \mathbb{R}^q$. Thus, each datapoint is bounded by "q" inequalities where "q" represents the largest dimension of the uncertainties of all the points. This refers us to the new SVM bi-level optimization formulation, as given by Eq.1.15.

$$min \quad \frac{1}{2}\|w\|^2 + C\sum \xi_i \tag{1.15}$$

$$s.t \quad min_{\{x_i:\, D_i x_i \leq d_i\}} \; y_i(w^T.x_i + b) \geq 1 - \xi_i \qquad \forall i = 1, \dots, n$$

$$\xi_i \geq 0 \qquad \forall i = 1, \dots, n$$

On the other hand, several studies focus on proposing new approaches that handle uncertainties without resorting to robust optimization. Choi et al., (2016) proposed a weighted SVM model for the classification of noisy data. This SVM model assigns to each training instance a weight based on how uncertain this instance is. This allows the creation of hyperplanes that are robust against the effect of noise and the effects of small datasets. The

model is called RC-margin SVM, and it combines the maximum-margin criterion with a penalty based on reduced convex hulls. The model was tested on a MNIST dataset, and the results show that the RC-margin SVM outperformed the basic SVM model by 8%.

A Robust Least Square Support Vector Machine (RLS-SVM) for the classification under uncertainties was presented by Yang et al., (2014). The proposed approach consists of analyzing the reason why the robustness of Weighted Least Squares Support Vector Machines (WLS-SVM) is higher than the robustness of Least Squares Support Vector Machines (LS-SVM). The analysis proved that WLS-SVM are significantly influenced by the different weights and therefore a good choice of weights is necessary. Accordingly, to avoid choosing the weights values of the training samples, the RLS-SVM optimizes the SVM model by considering these weights and their impact on the model robustness. This new SVM model combines the Concave-Convex Procedure (CCCP) and the Newton algorithm, and the results show that the RLS-SVM is the most robust of the three classifiers compared. Le Thi et al., (2014) proposed an approach for feature selection for SVM in the presence of measurement uncertainties based on the zero-norm and an appropriate Convex Algorithm Difference (DCA) in order to select relevant features from noisy data.

Fuzzy logic was also used to cope with the impact of uncertainties on SVM. Wu and Law, (2011) introduced new SVM models based on fuzzy logic, defining all inputs and outputs as triangular fuzzy members. The performance of the fuzzy SVM models have been evaluated, and the results show that they are effective in handling uncertain data and finite samples. Also, in comparison with other models such as v-SVM and Fv-SVM, the fuzzy-SVM models showed better generalization performance as well as greater robustness to some types of noise and singular data points. On the other hand, Heo and Gader, (2009) introduced a robust membership calculation method for the classification of noisy data with fuzzy-SVM. The method measures the similarity between the overall data structure and a data point using the reconstruction error concept. This makes it possible to represent the degree of outlier-ness of a datapoint, and thus, to achieve noise robustness. Experimental results have shown that the proposed method has a lower error rate than the original SVM and its variant, which demonstrates its usefulness.

At this stage, the main objective of the first chapter has been achieved. The different classification methods for quality improvement within production systems have been examined, which served as a basis for the selection of the SVM as the method to study in the rest of the thesis. Subsequently, case studies were conducted where different classification tools were evaluated. These case studies represent the main contributions of this first chapter, as summarized in Figure 1.13.

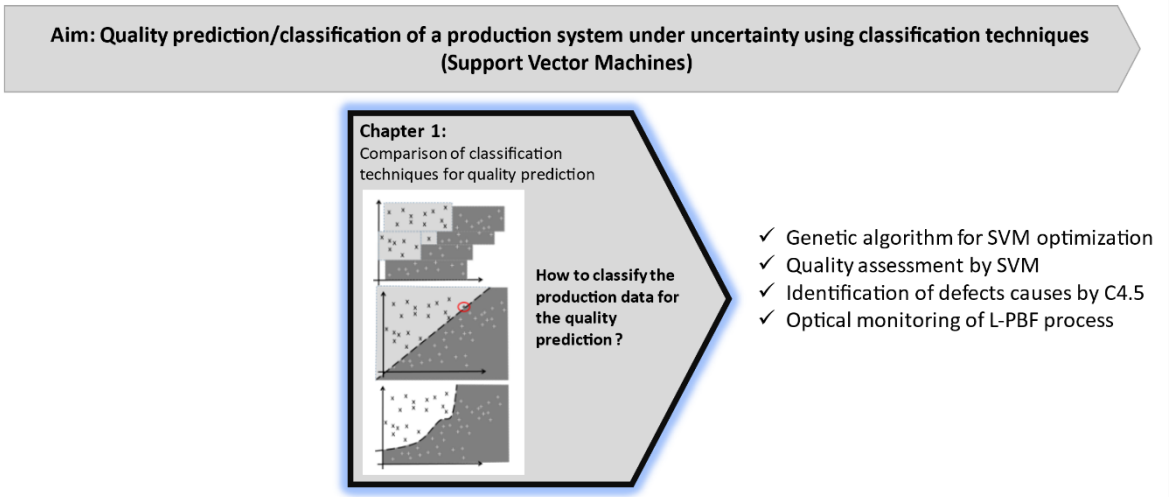


Figure 1.13: Main contributions of chapter I

Finally, several approaches addressing the optimization of the robustness of the SVM prediction accuracy under uncertainties were reviewed. In general, these approaches aim at assessing the impact of uncertainties on the performance of SVM and then to develop new SVM models that are more robust. This allows offering new SVM classifiers with better resilience to uncertainties, therefore allows offering robust solutions. Accordingly, **the main objective of this thesis is to improve the robustness of SVM regarding measurement uncertainties. This challenge is relevant for improvement of the measurement strategy, the smart manufacturing performances, industrial IoT deployment, Digital twin accuracy, etc. To do so, a scheme of two main steps is followed:**

- The impacts of measurement uncertainties (related to manufacturing data) on the prediction accuracy of SVM is firstly quantified. This has been performed previously in several other works. Nonetheless, the main challenge is the identification and the quantification of the manufacturing parameters with significant impacts on the prediction accuracy of the SVM model.
- The second step consists on introducing new approaches and models for the improvement of the robustness of SVM to measurement uncertainties, and that by guiding the SVM algorithm to focus more on reducing the impact of the measurement uncertainties of the parameters that affect greatly the robustness of SVM.

Chapter II

Identification of the key manufacturing parameters impacting the prediction accuracy of SVM model for quality assessment

In this second chapter, the issue of attribute noise (or measurement uncertainties), as well as its impact on the robustness of SVM prediction accuracy, is discussed. Two main tasks are carried out accordingly. At first, the robustness of SVM accuracy regarding measurement uncertainties is assessed. The robustness assessment consists of quantifying the decrease in the accuracy of the SVM prediction due to perturbing the data with artificial measurement uncertainties. Afterwards, three approaches are proposed to identify which parameters' uncertainties contribute the most to the SVM accuracy decrease. These approaches allow assigning to each parameter a quantitative coefficient that represents the magnitude of the impact of the uncertainties of each parameter on the robustness of SVM. On the one hand, the first two approaches rely on Monte Carlo simulations for the quantification of these coefficients. On the other hand, simple statistical tools are used in the third approach to estimate the impact of the parameters' uncertainties on the SVM robustness. The proposed approaches would eventually make it possible to identify the uncertainties of the parameters that mostly affect the SVM. Such parameters are referred to as key measurement uncertainties. Identifying the ***key measurement uncertainties*** *would provide a better understanding of how the SVM is affected by measurement uncertainties, as it would provide a strong basis for improving the robustness of SVM. In the following, the proposed approaches are described, then the results of applying them to four datasets are discussed and compared.*

II.1 - Introduction

It has been shown that measurement uncertainties have a negative impact on the predictive performance of classification methods, particularly SVM models. This impact can be manifested as a decrease in the ability of the classifier to generalize or an increase in the complexity of the model created (Wickramasinghe, 2017). As shown in Figure 2.1, the issue addressed in the second chapter is to quantify the effect of these measurement uncertainties on the predictive performance of SVM and thus on the generalization of SVM.

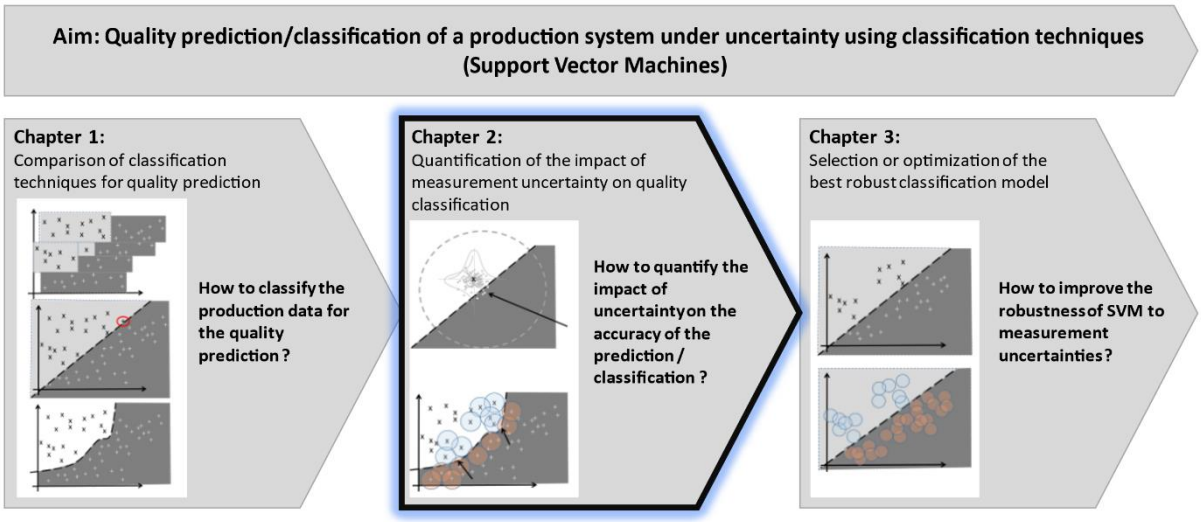


Figure 2.1: Thesis positioning

In order to quantify the overall impact of measurement uncertainties on the accuracy of the SVM, it is necessary to understand what is happening when predicting a data point that is subject to measurement uncertainties. As shown in Figure 2.2, measurement uncertainties are equivalent to translations that make one point move in all directions. Such movements may result in a point crossing the decision boundary and may therefore be misclassified. These movements become much more complicated when dealing with non-linear decision boundaries, as shown in Figure 2.3, and especially in high-dimensional spaces. For this reason, the impact of measurement uncertainties on **each datapoint** should be considered during the quantification task.

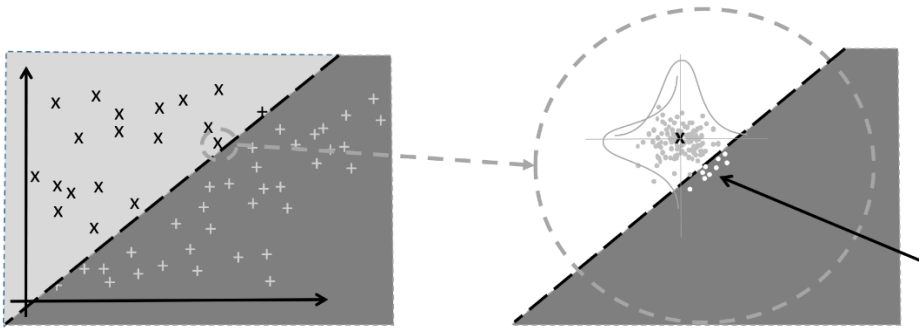


Figure 2.2: Illustration of a datapoint subject to measurement uncertainties

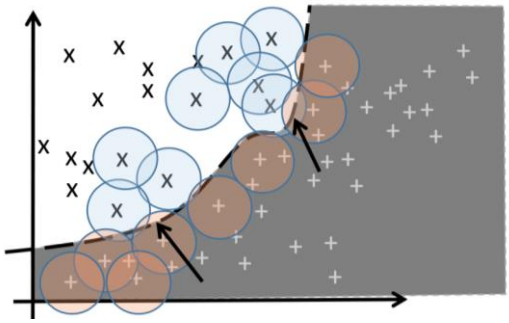


Figure 2.3: Non-linear SVM decision boundary

Accordingly, in this study, the probability that a datapoint that is subject to measurement uncertainties may cross the decision boundary is computed. An approach based on Monte-Carlo simulation is therefore proposed in order to quantify the impact of measurement uncertainties on the prediction performances of SVM.

II.2 - Assessment of the SVM robustness by Monte-Carlo simulation

In this section, we proposed an approach based on a Monte-Carlo simulation in order to assess and quantify the impact of measurement uncertainties on SVM 's robustness.

The term Monte-Carlo refers to a family of algorithmic methods which use random processes to calculate an approximate numeric value. The main objective of Monte-Carlo simulation is to conduct a large number of experiments on random samples and then to draw conclusions about the model outputs (Metropolis and Ulam, 1949).

Monte Carlo simulations are considered as effective mathematical analysis methods used to solve complex engineering problems due to their ability to manage a large number of random variables, different types of distribution, and highly nonlinear engineering models (Mohammadi et al., 2015). Some of the particular applications of Monte-Carlo methods include computing integrals in dimensions greater than 1 (e.g. for computing areas and volumes). They are also widely used in particle physics, where probabilistic simulations are used to estimate a signal's form, or detector sensitivity.

In this work, Monte-Carlo simulations are used to calculate the decrease in the accuracy of the SVM due to the perturbation of data with uncertainties. To do so, each dataset should be split into a training set (2/3 of the dataset) and a test set (1/3 of the dataset). That is necessary to tune the SVM hyperparameters by using the GA presented in Algorithm 1. Afterwards, random artificial uncertainties are generated to perturb the initial test set and thus define the noisy test sets. In this study, the impact of ***gaussian*** measurement uncertainties is studied, as they are considered to be one of the most commonly encountered uncertainties.

To generate artificial gaussian measurement uncertainties, a gaussian distribution $N_i(0, p_i\sigma_i)$ is associated to each input parameter, where σ_i is the standard

deviation of the i^{th} manufacturing parameter, and p_i denotes the noise level. Once the gaussian distributions are established, the inverse transform sampling method is used to generate the different measurement uncertainties (Özdemir and Çavuş 2016). Afterwards, an initial experiment is carried out to assess the robustness of SVM to gaussian measurement uncertainties. This evaluation of SVM robustness is depicted in Algorithm 2.

Algorithm 2 assessment of SVM robustness by Monte Carlo simulation – noising all manufacturing parameters with gaussian uncertainties.

Inputs: Dataset ***M***, N-sample of artificial gaussian measurement uncertainties sets ***MU***
Output: SVM accuracy decrease due to the uncertainties of all manufacturing parameters: $\Delta\boldsymbol{ACC}$

01: Split ***M*** into training set ***Ml***, and test set ***Mt***
02: Use Algorithm 1 to optimize the SVM model: ***Model_SVM***
03: Predict ***Mt*** using ***Model_SVM***: ***Acc***
04: Initialize the SVM accuracy decrease: $\Delta\boldsymbol{ACC} \leftarrow 0$
05: for $n \in \{1, \dots, N\}$ do
06: Pick the n^{th} set of ***MU***: $\boldsymbol{MU^{(n)}}$
07: Perturb ***M*** with $\boldsymbol{MU^{(n)}}$: $\boldsymbol{M^{(noised)}} \leftarrow \boldsymbol{M} + \boldsymbol{MU^{(n)}}$
08: Split $\boldsymbol{M^{(noised)}}$ into training set $\boldsymbol{Ml^{(noised)}}$, and test set $\boldsymbol{Mt^{(noised)}}$
09: Predict $\boldsymbol{Mt^{(noised)}}$ using ***Model_SVM***: $\boldsymbol{Acc^{(noised)}}$
10: Update $\Delta\boldsymbol{ACC}$: $\Delta\boldsymbol{ACC} \leftarrow \frac{\mathbf{1}}{\boldsymbol{n}} * ((\boldsymbol{Acc} - \boldsymbol{Acc^{(noised)}}) + (\boldsymbol{n} - \mathbf{1}) * \Delta\boldsymbol{ACC})$
11: end for
12: return $\Delta\boldsymbol{ACC}$

Algorithm 2 was applied on the four datasets, previously described in chapter1 and that using 1000 noisy test sets.

The noise levels were set at 2.5% for the three first datasets, and at 15% for the last dataset. These noise levels were estimated by taking into account the environment impact, and the precision and the accuracy of the sensors needed within the studied manufacturing/process industries. Table 2.1 gives the different results of the application of Algorithm 2.

Table 2.1: SVM accuracies considering Gaussian measurement uncertainties

Datasets	**Accuracy of test set**	**Average accuracy - 1000 noisy test sets -**	**SVM accuracy decrease**
Chem_4/8	89.47%	87.51%	-1.96%
Chem_5/7	85.77%	85.12%	-0.65%
Mine_1/2	82.34%	81.94%	-0.40%
Roll_0/1	91.43%	90.89%	-0.55%

As shown in Table 2.1, noising the test set with measurement uncertainties resulted in a decrease in the accuracy of the SVM prediction, i.e., the robustness of the generalization of SVM is affected by uncertainties. The results do not allow identifying which parameter uncertainties have the greatest impact on the SVM accuracy as explained in Figure 2.4 and Figure 2.5. In Figure 2.4, it is shown that the uncertainties of the first parameter have no effect on the classification of the datapoint; on the other hand, the uncertainties associated with the second parameter may result in a misclassification of this datapoint. Accordingly, the following research question was formulated:

"Which parameter uncertainties have a significant impact on the accuracy of the SVM prediction?"

To address this scientific question, three approaches for the evaluation of the impact of the uncertainties of each manufacturing parameter, and the identification of the key measurement uncertainties are proposed.

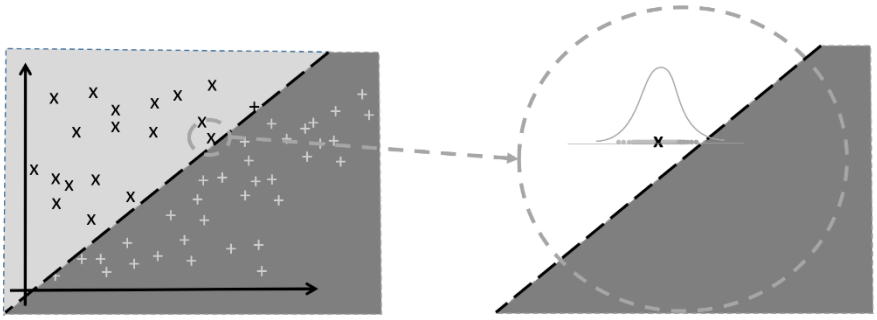


Figure 2.4: Impact of the uncertainties of the first parameter

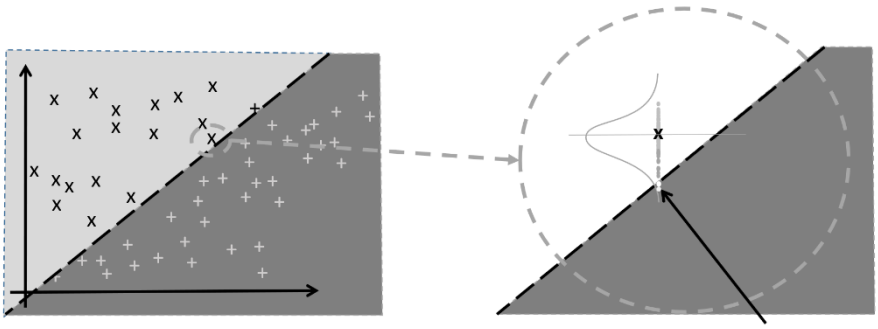


Figure 2.5: Impact of the uncertainties of the second parameter

II.3 - Identification of the key measurement uncertainties by Monte Carlo simulation

The first approach for the identification of the key measurement uncertainties is similar to the quantification performed in part **II.2**, where Monte-Carlo simulations are used to assess the robustness of SVM to gaussian measurement uncertainties. However, in this approach, the parameters are noised one by one instead of all at once. Thus, as many Monte-Carlo simulations as the number of parameters are required. This would allow quantifying the

decrease in SVM accuracy due to the perturbation of **one** parameter with gaussian measurement uncertainties. These simulations allow ranking the manufacturing parameters according to their impacts on SVM, which allows identifying the key measurement uncertainties in return. The global description of the approach could be illustrated by Figure 2.6.

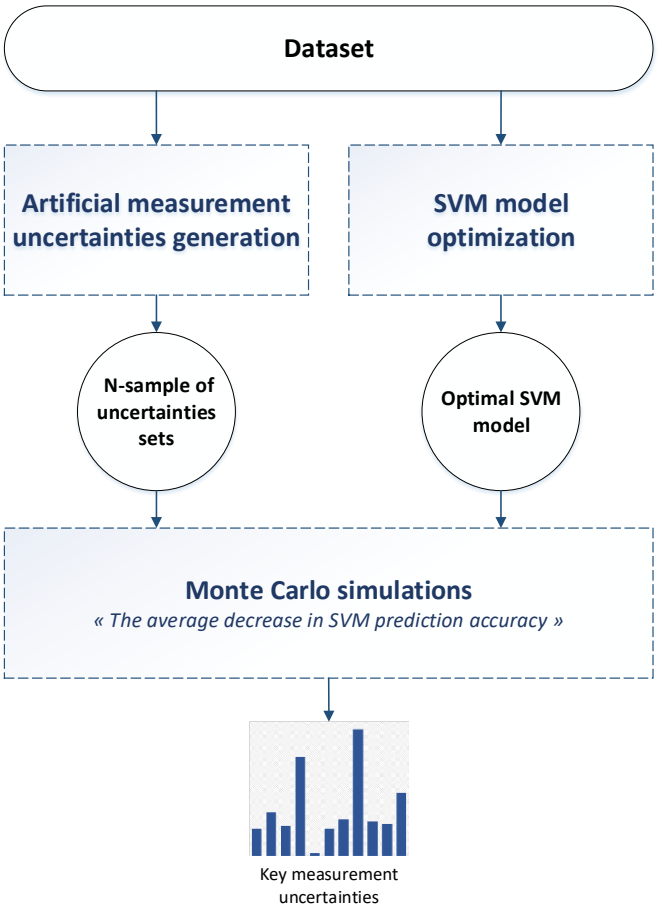


Figure 2.6: Identification of key measurement uncertainties by Monte Carlo simulations

The key measurement uncertainties identification is performed afterwards by using the same 1000 sets of measurement uncertainties used earlier. The different steps of this identification are presented in Algorithm 3.

Algorithm 3 assessment of the impact of a parameter's measurement uncertainties on SVM accuracy based on Monte Carlo simulation

Inputs dataset ***M***, N-sample of gaussian measurement uncertainties sets ***MU***

Output decrease of SVM accuracy due to the uncertainties of the j^{th}parameter $\{\Delta \boldsymbol{acc_j}\}_{\boldsymbol{j}\in[\mathbf{1},\ldots,\boldsymbol{P}]}$

01: Split ***M*** into training set ***Ml***, and test set ***Mt***
02: Optimize the SVM model based on Algorithm 1: ***Model_SVM***
03: Predict ***Mt*** using ***Model_SVM***: ***Acc***
04: Get the number of parameters in ***M: P***
05: for $j \in \{1, \ldots, \boldsymbol{P}\}$ do
06: Initialize the SVM accuracy decrease: $\Delta \boldsymbol{acc_j} \leftarrow 0$
07: for $n \in \{1, \ldots, N\}$ do
08: Pick the n^{th} set of MU: $\boldsymbol{MU^{(n)}}$
09: Pick the j^{th} column of $\boldsymbol{MU^{(n)}}$: $\boldsymbol{MU_j^{(n)}}$
10: Perturb the j^{th} parameter of ***M*** with $\boldsymbol{MU_j^{(n)}}$: $\boldsymbol{M^{(noised)}}$
11: Split $\boldsymbol{M^{(noised)}}$ into training set $\boldsymbol{Ml^{(noised)}}$, and test set $\boldsymbol{Mt^{(noised)}}$
12: Predict $\boldsymbol{Mt^{(noised)}}$ using ***Model_SVM***: $\boldsymbol{Acc^{(noised)}}$
13: Update $\Delta \boldsymbol{acc_j}$: $\Delta \boldsymbol{acc_j} \leftarrow \frac{\mathbf{1}}{\boldsymbol{i}} * ((\boldsymbol{Acc} - \boldsymbol{Acc^{(noised)}}) + (\boldsymbol{i} - \mathbf{1}) * \Delta \boldsymbol{acc_j})$
14: end for
15: end for
16: return $\{\Delta \boldsymbol{acc_j}\}_{\boldsymbol{j}\in[\mathbf{1},\ldots,\boldsymbol{P}]}$

Algorithm 3 was applied on the four datasets using the same 1000 sets of artificial Gaussian measurement uncertainties as used in Algorithm 2. The different results are presented in Tables 2.2, 2.3, 2.4, and synthetized by Figure 2.7.

Table 2.2: Decreases in SVM accuracy (-%) due to the uncertainties of parameter X_j - Chemical data

Datasets	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$	$\mathbf{X_5}$	$\mathbf{X_6}$	$\mathbf{X_7}$	$\mathbf{X_8}$	$\mathbf{X_9}$	$\mathbf{X_{10}}$	$\mathbf{X_{11}}$
Chem_4/8	0	0.73	0.21	0.47	0	0.51	0	0.58	0.33	0.20	0.41
Chem_5/7	0.11	0.18	0.14	0.35	0	0.08	0.07	0.44	0.03	0.16	0.17

Table 2.3: Decreases in SVM accuracy (-%) due to the uncertainties of parameter X_j - *Mine_1/2*

Dataset	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$	$\mathbf{X_5}$	$\mathbf{X_6}$	$\mathbf{X_7}$	$\mathbf{X_8}$	$\mathbf{X_9}$	$\mathbf{X_{10}}$	$\mathbf{X_{11}}$
	0.23	0.08	0.08	0.02	0.02	0.07	0.11	0.06	0.08	0	0.05
Mine_1/2	$\mathbf{X_{12}}$	$\mathbf{X_{13}}$	$\mathbf{X_{14}}$	$\mathbf{X_{15}}$	$\mathbf{X_{16}}$	$\mathbf{X_{17}}$	$\mathbf{X_{18}}$	$\mathbf{X_{19}}$	$\mathbf{X_{20}}$	$\mathbf{X_{21}}$	$\mathbf{X_{22}}$
	0.03	0.02	0.04	0.04	0.06	0.08	0.06	0.01	0.02	0.03	0.20

Table 2.4: Non-zero decreases in SVM accuracy due to the uncertainties of parameter X_j - *Roll_0/1*

Dataset	$\mathbf{X_{45}}$	$\mathbf{X_{49}}$	$\mathbf{X_{66}}$	$\mathbf{X_{67}}$	$\mathbf{X_{78}}$	$\mathbf{X_{85}}$	$\mathbf{X_{90}}$	$\mathbf{X_{92}}$	$\mathbf{X_{95}}$
Roll_0/1	0.03	0.10	0.07	0.01	0.04	0.11	0.05	0.08	0.11

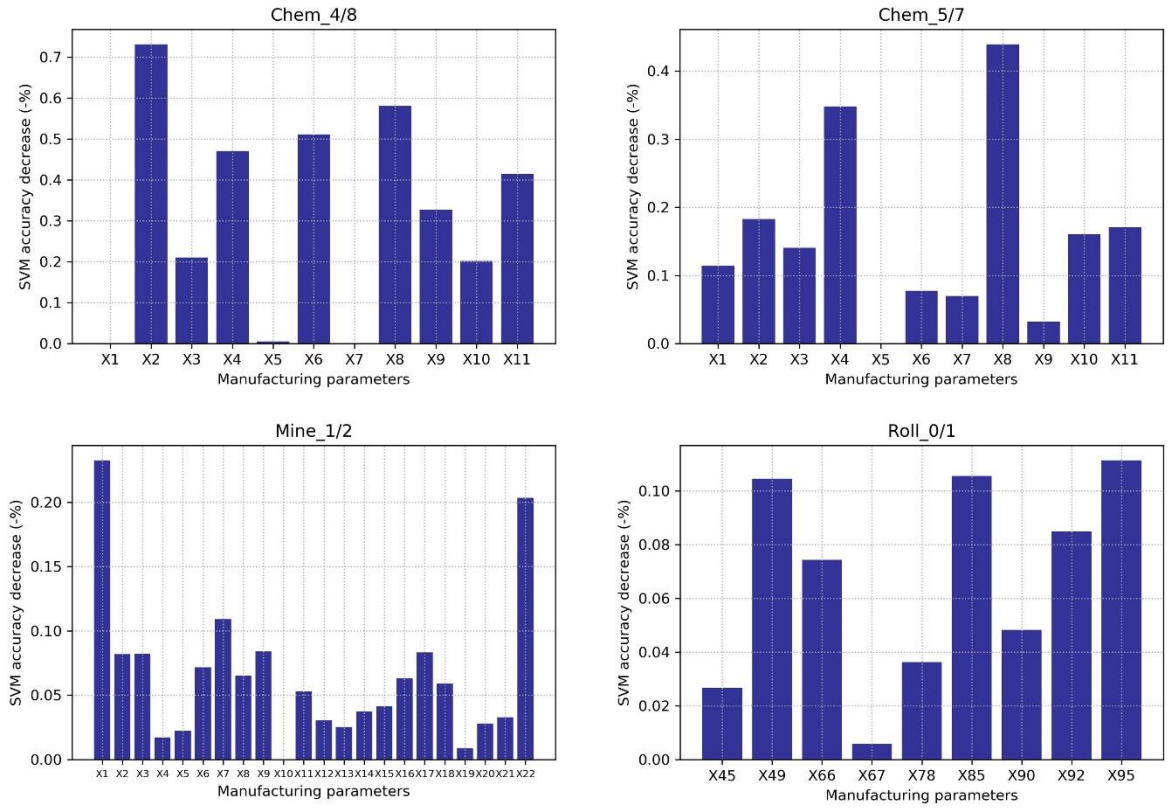


Figure 2.7: Average of decreases of SVMs accuracies due to the uncertainties of the manufacturing parameters

The results show that the measurement uncertainties of the parameters have different impacts on the SVM accuracy. Such impacts could be ranked, and the parameters with significant impacts could be considered as the key measurement uncertainties. We can then consider:

1. For the 1st dataset (*Chem_4/8*), the Gaussian uncertainties applied on the parameters ***X2, X4, X6, X8,*** and ***X11*** have greater impacts. These parameters may be thus regarded as key measurement uncertainties and hence their uncertainties should be monitored when dealing with the robustness of the SVM prediction accuracy.
2. In the 2nd dataset (*Chem_5/7*), the SVM accuracy is mostly impacted by the uncertainties of the manufacturing parameters ***X4*** and ***X8***. These two parameters are considered as keys and controlling their uncertainties would allow a robust prediction of the product quality levels.

3. In the case of mining process dataset (*Mine_1/2*), only the measurement uncertainties of the parameters ***X1*** (% Iron Feed), ***X7*** (Ore Pulp Density), and ***X22*** (% Iron Concentrate) have impacts greater than -0.1%. we will consider them key parameters.
4. Finally, for the *Roll_0/1* dataset, the SVM accuracy is only affected by 9 manufacturing parameters among a total of 95 parameters. Consequently, a robust quality prediction could be achieved by monitoring the uncertainties of the identified parameters.

By assigning to each parameter a quantitative coefficient, the proposed approach allowed identifying the parameters that affect significantly the SVM prediction accuracy. The different results show the efficiency of the proposed approach in identifying the key measurement uncertainties in an easy and understandable manner. More analyses of the results are presented at the end of the chapter.

In the following, a second approach for the identification of the key measurement uncertainties is proposed. This approach is based on the Sobol sensitivity analysis, which decomposes the output variance into fractions that can be attributed to the variances of the input variables. For a better understanding, the approach based on Sobol sensitivity analysis is described, and the different results of its application on the four considered datasets are presented.

II.4 - Assessment of the SVM sensitivity by Sobol analysis

Sensitivity analysis (SA) is the study of how perturbations on the input variables of the model generate perturbations on the response variable. A variety of SA methods have been developed in various disciplines to explore the content of different models (Kristensen and Petersen, 2016). These methods can be divided into three classes: screening methods, consisting of a qualitative study of the sensitivity of the output variable to an input variable, local study methods that quantitatively measure the effect of a small variance on a given input value, and finally global sensitivity analysis methods that analyze the variability of the input variables. While the local sensitivity analysis focuses more on the value of the response variable, the global sensitivity analysis focuses on the variability of the response variable.

Sensitivity analysis can be useful for many applications, such as the understanding of the relationships between the inputs and the output, the identification of the variables or groups of variables that interact with others, the obtention of lighter models by fixing inputs that have negligible effects on the output, **the assessment of the robustness of a model in the presence of uncertainty, the reduction of uncertainty through the identification of the inputs that contribute significantly to uncertainty in the output,** etc. (Jacques, 2011).

In this study, the Sobol Sensitivity Method is used to assess the robustness of the SVM prediction accuracy. The use of the Sobol method can be justified by referring to Figure 2.8, where Sobol analysis is known to be the only approach capable of taking into account various input parameter distributions, taking into account multi-dimensional parameter influences

where all parameters vary simultaneously, and embracing both non-linear and non-additive effects when parameter interactions are taken into account (Kristensen and Petersen, 2016).

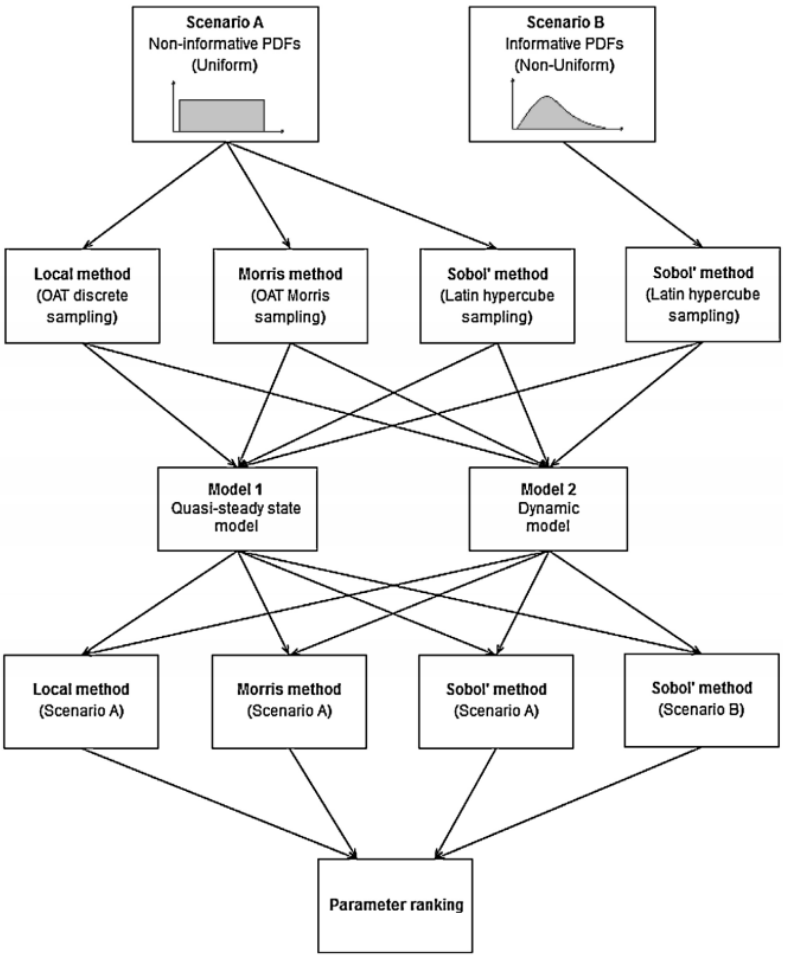


Figure 2.8: Methodical approach for the selection of a SA method (Kristensen et al. 2016)

Sobol sensitivity analysis relies on variance decomposition techniques to provide a quantitative coefficients of the contribution of each input parameter variance to the output variance (Glen and Isaacs, 2012). Accordingly, the second approach of our study calculates what are called Sobol total-effect indices. These indices are estimated by Monte-Carlo simulations, and they allow measuring the effect of the measurement uncertainties of any parameter on the SVM accuracy. The ranking of the Sobol indices makes it possible to identify which parameters significantly affect the accuracy of the SVM and thus to identify the key measurement uncertainties. In the following, a complete description of the proposed approach is presented. Figure 2.9 gives a global illustration of the second approach.

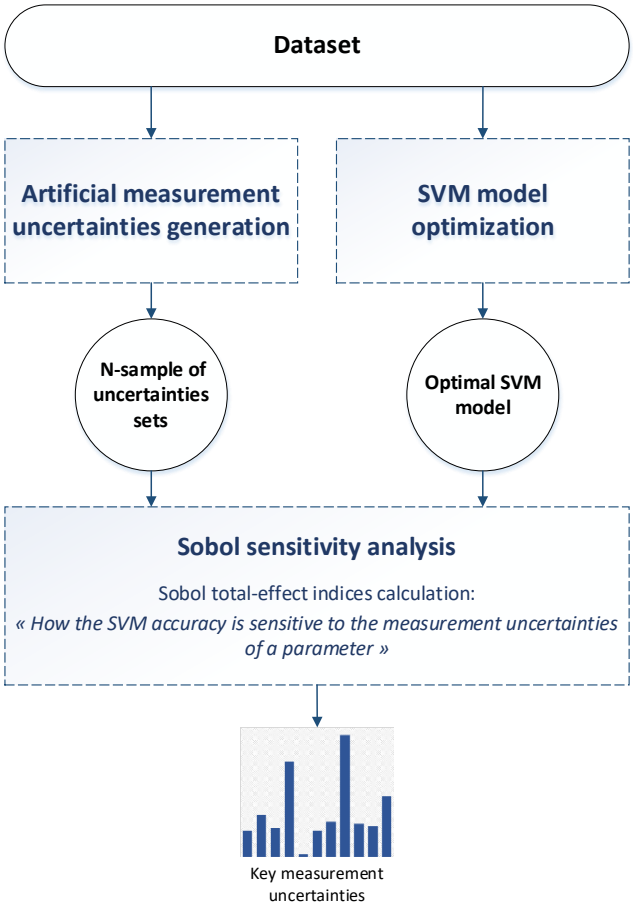


Figure 2.9: Identification of key measurement uncertainties by Sobol sensitivity analysis

Generally, Sobol sensitivity analysis allows providing two quantitative indices, namely:

- First order sensitivity index "S_i": The contribution of the main effect of the input parameter Xi to the output variance.
- Total-effect index "S_{T_i}": The sum of the contributions caused by an input parameter Xi and by its interactions -of any order- with any other input parameter.

To estimate the Sobol indices, a Monte Carlo simulation is used due to its convergence resulting from the strong law of large numbers (Jacques et al. 2006). This estimation requires the definition of the input(s), the output(s), and the model that links them together, see. The inputs of the Sobol analysis are usually defined as scalars, which is not the case in this study where the input parameters are vectors. The Sobol method was reshaped to deal with this issue as shown in Figure 2.10:

- The inputs $(x_{n1}, \dots, x_{np})$: x_{nj} is the j^{th} parameter of the n^{th} noisy test set.
- The output Y_n : is the decrease in the SVM accuracy due to measurement uncertainties in the n^{th} noisy test set.
- The model $\boldsymbol{f}$: is the optimized SVM model; $[Y_n = \boldsymbol{f}(x_{n1}, \dots, x_{np})]$

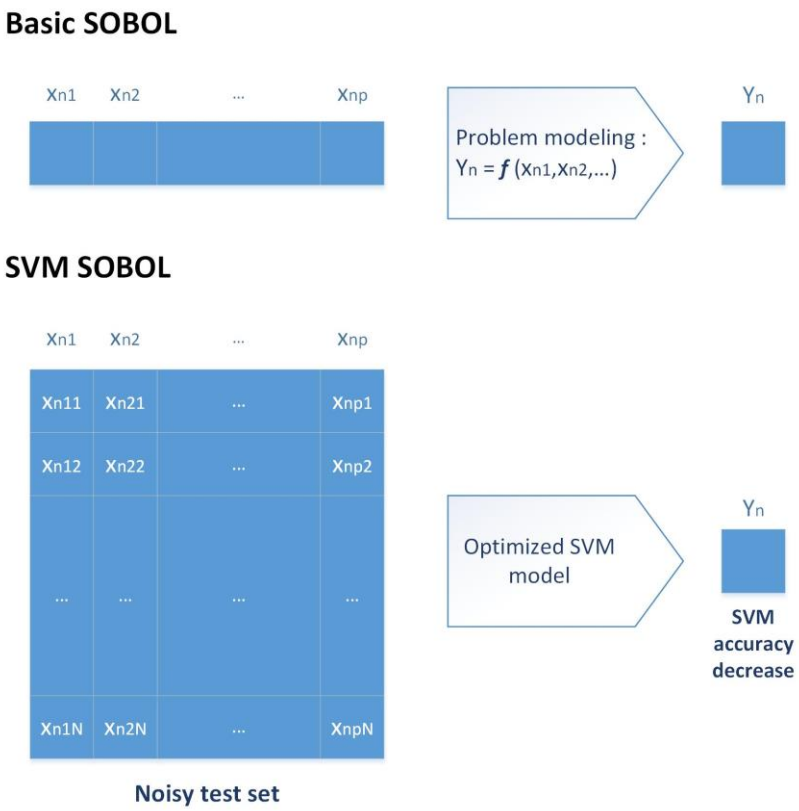


Figure 2.10: Sobol method for analyzing the sensitivity of SVM accuracy

This approach can be explained in Algorithm 4.

Algorithm 4 analyzing the SVM sensitivity using SOBOL indices

---

Inputs: dataset M***,*** two N-samples of gaussian measurement uncertainties sets ***MU1***; ***MU2***.

Output: SOBOL total-effect indices of all the parameters $\{\boldsymbol{STj}\}_{\boldsymbol{j}\epsilon\{\mathbf{1},\ldots,\boldsymbol{P}\}}$

01: Split **M** into training set ***Ml***, and test set ***Mt***.

02: Apply the GA algorithm to optimize the SVM model: ***Model_SVM.***

03: Predict ***Mt*** using ***Model_SVM***: ***Acc***

04: Initialize the expected value $\boldsymbol{E} \leftarrow 0$

05: for $n \in \{1, \ldots, N\}$ do

06: Pick the n^{th} set of ***MU1***: $\boldsymbol{MU1}^{(\boldsymbol{n})}$

07: Perturb ***M*** with $\boldsymbol{MU1}^{(\boldsymbol{n})}$: $\boldsymbol{M}^{(\boldsymbol{noised})}$

08: Split $\boldsymbol{M}^{(\boldsymbol{noised})}$into training set $\boldsymbol{Ml}^{(\boldsymbol{noised})}$, and test set $\boldsymbol{Mt}^{(\boldsymbol{noised})}$

09: Predict the accuracy of $\boldsymbol{Mt}^{(\boldsymbol{noised})}$ using ***Model_SVM***: $\boldsymbol{Acc}^{(\boldsymbol{noised})}$

10: $\Delta\boldsymbol{acc} \leftarrow \boldsymbol{Acc} - \boldsymbol{Acc}^{(\boldsymbol{noised})}$

11: $\boldsymbol{E} \leftarrow \frac{1}{n} * (\Delta\boldsymbol{acc} + (\boldsymbol{n} - 1) * \boldsymbol{E})$

12: end for

13: Initialize the variance $\boldsymbol{V} \leftarrow 0$

14: for $n \in \{1, \ldots, N\}$ do

15: Pick the n^{th} set of ***MU1***: $\boldsymbol{MU1}^{(\boldsymbol{n})}$

16: Perturb $\boldsymbol{M}$ with $\boldsymbol{MU1^{(n)}}$: $\boldsymbol{M^{(noised)}}$

17: Split $\boldsymbol{M^{(noised)}}$into training set $\boldsymbol{Ml^{(noised)}}$, and test set $\boldsymbol{Mt^{(noised)}}$

18: Predict the accuracy of $\boldsymbol{Mt^{(noised)}}$ using Model_SVM: $\boldsymbol{Acc^{(noised)}}$

19: $\Delta \boldsymbol{V} \leftarrow ((\boldsymbol{Acc} - \boldsymbol{Acc_{noised}}) - \boldsymbol{E})^2$

20: $\boldsymbol{V} \leftarrow \frac{1}{n} * (\Delta \boldsymbol{V} + (\mathrm{n} - 1) * \boldsymbol{V})$

21: end for

22: Get the number of parameters in $\boldsymbol{M}$: $\boldsymbol{P}$

23: for $j \in \{1, \dots, \boldsymbol{P}\}$ do

24: Initialize the quantity $\boldsymbol{U_j} \leftarrow 0$

25: for $n \in \{1, \dots, N\}$ do

26: Pick the n^{th} set of $\boldsymbol{MU1}$: $\boldsymbol{MU1^{(n)}}$

27: Pick the n^{th} set of $\boldsymbol{MU2}$: $\boldsymbol{MU2^{(n)}}$

28: Perturb $\boldsymbol{M}$ with $\boldsymbol{MU1^{(n)}}$: $\boldsymbol{M1^{(noised)}}$

29: Perturb $\boldsymbol{M}$ with $\boldsymbol{MU2^{(n)}}$: $\boldsymbol{M2^{(noised)}}$

30: Split $\boldsymbol{M1^{(noised)}}$into training set $\boldsymbol{M1l^{(noised)}}$, and test set $\boldsymbol{M1t^{(noised)}}$

31: Split $\boldsymbol{M2^{(noised)}}$into training set $\boldsymbol{M2l^{(noised)}}$, and test set $\boldsymbol{M2t^{(noised)}}$

32: Create a duplicate of $\boldsymbol{M1t^{(noised)}}$: $\boldsymbol{m1t^{(noised)}}$

33: Replace the j^{th} parameter of $\boldsymbol{m1t^{(noised)}}$ with the j^{th} parameter of $\boldsymbol{M2t^{(noised)}}$: $\boldsymbol{m1t^{(noised)}}$

34: Predict the accuracy of $\boldsymbol{M1t^{(noised)}}$ using Model_SVM: $\boldsymbol{Acc_1^{(noised)}}$

35: Predict the accuracy of $\boldsymbol{m1t^{(noised)}}$ using Model_SVM: $\boldsymbol{Acc_2^{(noised)}}$

36: $\Delta \boldsymbol{acc}_{\boldsymbol{1}} \leftarrow \boldsymbol{Acc} - \boldsymbol{Acc_1^{(noised)}}$

37: $\Delta \boldsymbol{acc}_{\boldsymbol{2}} \leftarrow \boldsymbol{Acc} - \boldsymbol{Acc_2^{(noised)}}$

38: $\hat{\mathrm{u}} \leftarrow \Delta \boldsymbol{acc}_{\boldsymbol{1}} * \Delta \boldsymbol{acc}_{\boldsymbol{2}}$

39: $\boldsymbol{U_j} \leftarrow \frac{(\boldsymbol{n}-\boldsymbol{1}) * \boldsymbol{U_j} + \hat{\mathrm{u}}}{\boldsymbol{n}}$

40: end for

41: $\boldsymbol{ST_j} \leftarrow (\boldsymbol{1} - \frac{\boldsymbol{U_j} - \boldsymbol{E^2}}{\boldsymbol{V}})$

42: end for

43: return $\{\boldsymbol{STj}\}_{\boldsymbol{j} \epsilon \{\boldsymbol{1}, \dots, \boldsymbol{P}\}}$

Algorithm 4 is then applied on the four datasets. Two 1000-samples of gaussian measurement uncertainties sets were used for each manufacturing dataset. As given previously, the noise level is set at 2.5% for the three first datasets, and at 15% for the fourth dataset. The approach allowed measuring the total-effect sensitivity indices as given in Tables 2.5, 2.6, 2.7 and an illustrative summary is presented in Figure 2.11.

Table 2.5: Sobol total effect indices – Chemical data

Datasets	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$	$\mathbf{X_5}$	$\mathbf{X_6}$	$\mathbf{X_7}$	$\mathbf{X_8}$	$\mathbf{X_9}$	$\mathbf{X_{10}}$	$\mathbf{X_{11}}$
Chem_4/8	0.13	0.54	0.09	0.42	0.05	0.41	0.01	0.49	0.23	0.11	0.34
Chem_5/7	0.14	0.20	0.11	0.44	0.02	0.14	0.16	0.56	0.09	0.12	0.23

Table 2.6: Sobol total effect indices – *Mine_1/2* data

Dataset	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$	$\mathbf{X_5}$	$\mathbf{X_6}$	$\mathbf{X_7}$	$\mathbf{X_8}$	$\mathbf{X_9}$	$\mathbf{X_{10}}$	$\mathbf{X_{11}}$
	0.54	0.27	0.07	0.13	0	0.24	0.38	0.11	0.13	0.04	0.10
Mine_1/2	$\mathbf{X_{12}}$	$\mathbf{X_{13}}$	$\mathbf{X_{14}}$	$\mathbf{X_{15}}$	$\mathbf{X_{16}}$	$\mathbf{X_{17}}$	$\mathbf{X_{18}}$	$\mathbf{X_{19}}$	$\mathbf{X_{20}}$	$\mathbf{X_{21}}$	$\mathbf{X_{22}}$
	0.01	0.07	0.06	0.06	0.09	0.15	0.08	0.02	0.06	0.07	0.36

Table 2.7: Sobol total effect indices with a variation ratio higher than 0.2 – *Roll_0/1* data

Dataset	$\mathbf{X_{45}}$	$\mathbf{X_{49}}$	$\mathbf{X_{66}}$	$\mathbf{X_{78}}$	$\mathbf{X_{85}}$	$\mathbf{X_{90}}$	$\mathbf{X_{92}}$	$\mathbf{X_{95}}$
Roll_0/1	0.21	0.32	0.30	0.31	0.42	0.29	0.32	0.44

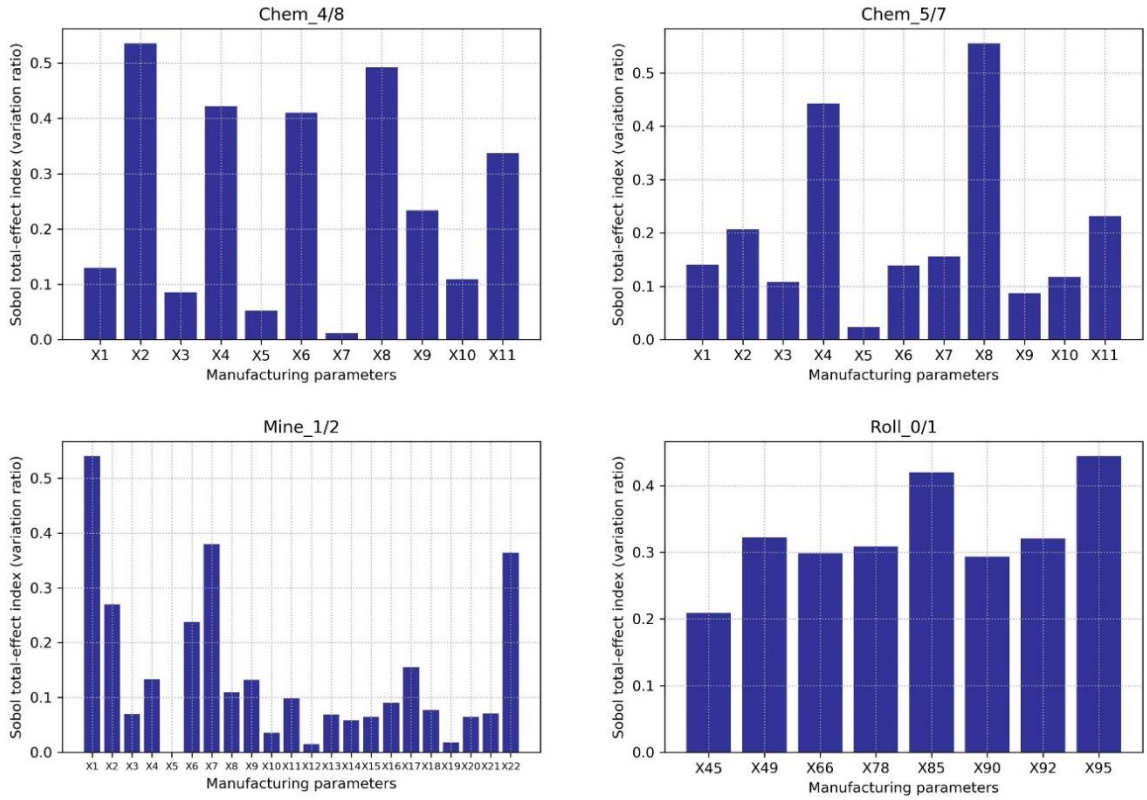


Figure 2.11: Sobol total effect indices of the manufacturing datasets

These results enable the identification of parameters with significant Sobol total effect indices and thus allow the identification of key measurements uncertainties. We give a synthesis on this identification:

1. In the case of the first chemical dataset (*Chem_4/8*), the manufacturing parameters ***X2, X4, X6, X8***, and ***X11*** have higher Sobol indices, and therefore they could be regarded as key measurement uncertainties.
2. By application on the second chemical dataset (*Chem_5/7*), the Sobol indices of the parameters ***X4*** and ***X8*** are clearly prominent, and thus considered as the key measurement uncertainties.
3. Considering the mining process dataset (*Mine_1/2*), three leading parameters are identified (***X1*** :% Iron Feed, ***X7***:Ore Pulp Density and ***X22*** :% Iron Concentrate). They are considered as key measurement uncertainties when dealing with the floatation process.
4. For the last case study (*Roll_0/1*), 8 parameters with a Sobol index higher than 0.2 are considered as key measurement uncertainties.

This second approach allows the identification of key measurement uncertainties in various manufacturing systems. Eventually, this identification would allow for a robust prediction of quality levels and a robust assurance of quality. A more in-depth comparison of the results of the two approaches is made at the end of this second chapter. But before that, a third approach based on simple statistical tools for the **estimation** of the key measurement uncertainties is presented.

II.5 - Estimation of key measurement uncertainties by correlation research

The two first approaches for the identification of the key measurement uncertainties are based on Monte-Carlo simulations that are generally time and resource consuming. To overcome this issue, a third approach based on statistical tools is proposed. This approach estimates the key measurement uncertainties by statistically analyzing the correlation of a manufacturing parameter with the class parameter. The objective of the approach is to identify the manufacturing parameters that are well correlated with the class parameter. This enables the ranking of the manufacturing parameters and thus the estimation of the key measurement uncertainties. Figure 2.12 gives a graphical description of the approach.

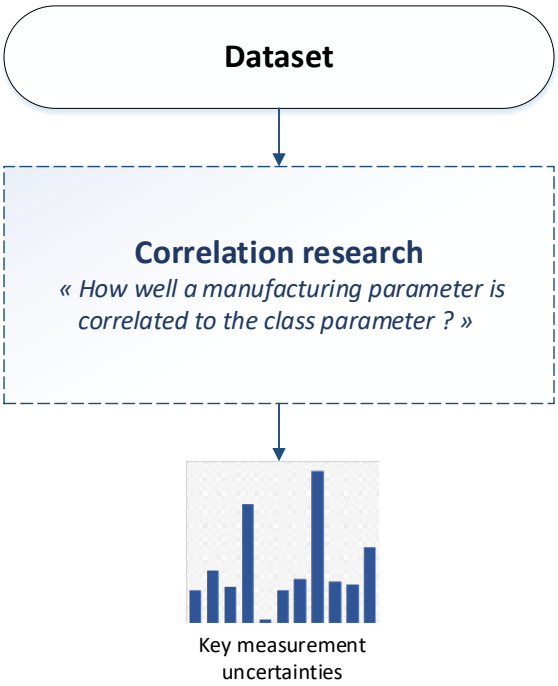


Figure 2.12: Estimation of the key measurement uncertainties by correlation research

The point-biserial correlation coefficient is adopted in this approach. This correlation coefficient is used calculate the strength of association between a continuous variable and a binary variable, in a range varying from -1 to +1. Accordingly, in this section, the point-biserial correlation coefficient between the class parameter and each manufacturing parameter is calculated, using the formula expressed in Eq.2.1.

$$r_{X_j \sim Y} = \frac{m_1 - m_2}{\sigma_n} \sqrt{\frac{n_1 n_2}{n^2}} \tag{2.1}$$

Where:

- X_j: the j^{th} manufacturing parameter of the dataset.
- Y: the class parameter.
- m_1: mean value on X_j for all data points of class 1.
- m_2: mean value on X_j for all data points of class 2.
- σ_n: standard deviation of X_j.
- n_1: number of data points of class 1.
- n_2: number of data points of class 2.
- n: number of data points of the dataset.

The calculation of the correlation coefficients for all the datasets allowed getting the results presented in Tables 2.8, 2.9, 2.10 and illustrated in Figure 2.13.

Table 2.8: Point-biserial correlation coefficients – Chemical data

Datasets	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$	$\mathbf{X_5}$	$\mathbf{X_6}$	$\mathbf{X_7}$	$\mathbf{X_8}$	$\mathbf{X_9}$	$\mathbf{X_{10}}$	$\mathbf{X_{11}}$
Chem_4/8	-0.27	-0.34	0.06	0.03	-0.32	0.34	-0.05	-0.43	0.21	0.04	0.58
Chem_5/7	-0.11	-0.20	-0.03	-0.23	-0.26	-0.06	-0.28	-0.47	0.17	0.12	0.61

Table 2.9: Point-biserial correlation coefficients – *Mine_1/2* data

Dataset	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$	$\mathbf{X_5}$	$\mathbf{X_6}$	$\mathbf{X_7}$	$\mathbf{X_8}$	$\mathbf{X_9}$	$\mathbf{X_{10}}$	$\mathbf{X_{11}}$
	0.58	-0.56	-0.19	-0.34	0.04	-0.29	-0.39	-0.06	-0.41	-0.0	0.04
Mine_1/2	$\mathbf{X_{12}}$	$\mathbf{X_{13}}$	$\mathbf{X_{14}}$	$\mathbf{X_{15}}$	$\mathbf{X_{16}}$	$\mathbf{X_{17}}$	$\mathbf{X_{18}}$	$\mathbf{X_{19}}$	$\mathbf{X_{20}}$	$\mathbf{X_{21}}$	$\mathbf{X_{22}}$
	-0.04	-0.30	-0.08	0.42	0.25	0.40	0.18	0.36	0.41	0.24	-0.57

Table 2.10: Point-biserial correlation coefficients greater than ± 0.5 – *Roll_0/1* data

Dataset	$\mathbf{X_2}$	$\mathbf{X_{11}}$	$\mathbf{X_{18}}$	$\mathbf{X_{24}}$	$\mathbf{X_{41}}$	$\mathbf{X_{46}}$	$\mathbf{X_{48}}$	$\mathbf{X_{63}}$
	-0.61	0.82	0.62	-0.57	-0.59	0.84	0.59	-0.71
Roll_0/1	$\mathbf{X_{64}}$	$\mathbf{X_{66}}$	$\mathbf{X_{67}}$	$\mathbf{X_{75}}$	$\mathbf{X_{78}}$	$\mathbf{X_{90}}$	$\mathbf{X_{92}}$	$\mathbf{X_{95}}$
	-0.50	0.83	-0.80	-0.50	0.65	-0.79	0.83	0.63

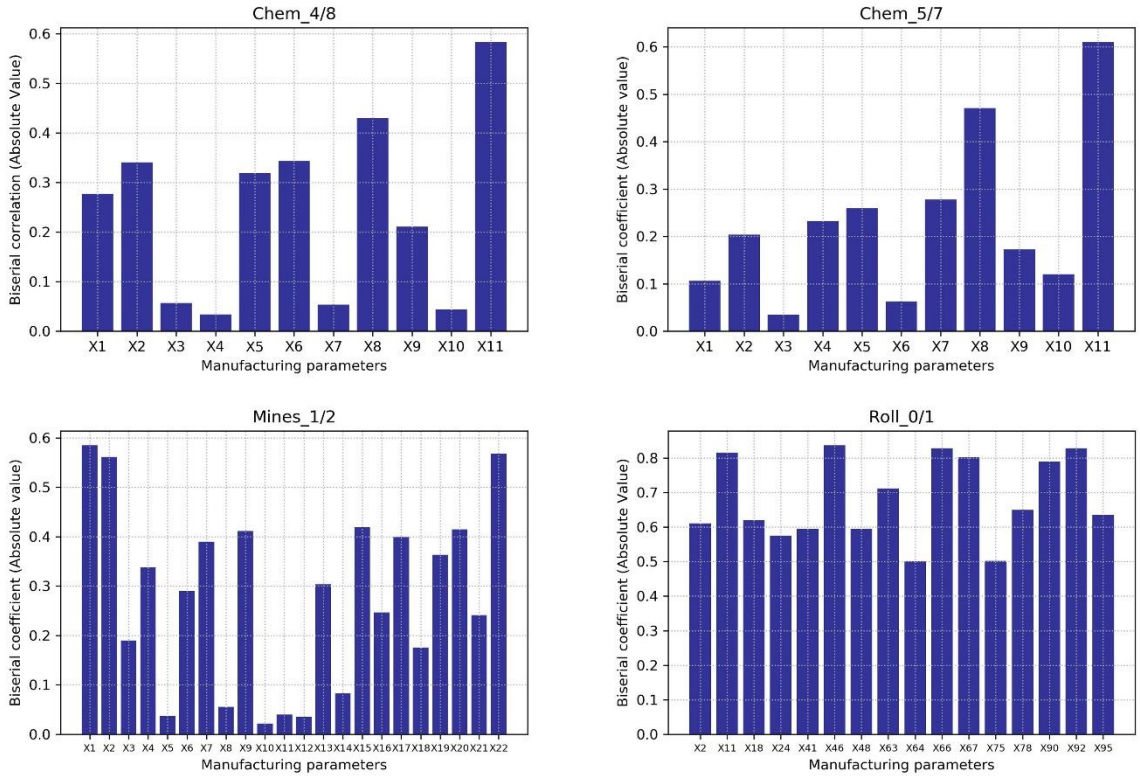


Figure 2.13: Point-biserial correlation coefficients of the manufacturing datasets

According to these results:

1. Six parameters (***X1, X2, X5, X6, X8,*** and ***X11)*** of the chemical dataset *Chem_4/8* are well correlated to the class parameter. Therefore, they are estimated to be the key measurement uncertainties.
2. the two parameters ***X8*** and ***X11 in the Chem_5/7 data*** are more correlated to the class than the other manufacturing parameters. These two parameters could be estimated as the main key measurement uncertainties of the chemical dataset.
3. In the mining process, the leading parameters that could be estimated as the key measurement uncertainties are ***X1***, ***X2*** and ***X22***.
4. For le last application, the parameters with a correlation coefficient greater than ± 0.5 could be estimated as key measurement uncertainties. 15 key measurement uncertainties are identified for the *Roll_0/1* dataset.

The third approach estimated the different key measurement uncertainties by performing correlation research and without the need for artificial gaussian measurement uncertainties. To assess how accurate this estimation is, a comparison of the results of the three proposed approaches is provided in the following. Further analysis of the results is also presented, as well as some concluding remarks.

II.6 - Discussion and conclusion

In this section, the results of the last three sections are analyzed and discussed carefully. First of all, the results of the approaches are compared. This comparison makes it possible to assess the similarities between the different results, as well as to evaluate the estimation ability of the third approach. After that, the combined impact of the key measurement uncertainties on SVM accuracy is quantified. Finally, the time complexities of the approaches are presented, and a conclusion is given.

When comparing the results of the **first approach based on Monte Carlo simulation**s to the results of **the second approach based on Sobol sensitivity analysis**, it can be noticed that:

1. *First study case (Chem_4/8)*: the ranking defined in the first approach is quite similar to that found in the second approach, where the uncertainties that significantly affect the accuracy of the SVM are the uncertainties of the manufacturing parameters X2 and X8 at first degree, then X4, X6 and X11 at second degree, and finally the remaining manufacturing parameters at last degree. The main difference is on the level of the parameter X6, where in the first approach it is ranked before X4, and in the second approach after X4. Nevertheless, in both approaches, the impacts of these two parameters are very similar and can thus be monitored with equal importance.
2. *Second case study (Chem_5/7):* the results found by both approaches are the same for the second dataset. The measurement uncertainties of the parameter X8 have the biggest

impact on the accuracy of the SVM followed by the impact of the measurement uncertainties of the parameter X4. These two impacts are prominent and therefore, to provide a robust control over the quality of the manufacturing system, the measurement uncertainties of these two parameters should be controlled carefully.

3. *Third case study (Mine_1/2):* the most affecting uncertainties found by the first approach are the uncertainties of the parameters X1 (% Iron Feed), X22 (% Iron Concentrate), and X7 (Ore Pulp Density), respectively. It can be noticed that the quality level (Silica level) is sensitive to iron, thus a good monitoring of the iron parameters would allow a better control of the quality of the floating process. For the second approach, the uncertainties of X1, X7 and X22 have the greatest effects on the accuracy of the SVM. For the rest of the parameters, it could be found that the rank of some of these parameters varies from the first approach to the second approach. This is due to the very similar impacts of the uncertainties of certain manufacturing parameters. Though, when considering the magnitude of the different results, the identification of the key measurement uncertainties by the two approaches is very much the same.
4. *Fourth case study (Roll_0/1):* the first approach allowed the identification of nine parameters that decrease the accuracy of the SVM. In the second approach, and by selecting the manufacturing parameters with Sobol indices greater than 0.2, eight key measurement uncertainties were defined. All manufacturing parameters identified by the first approach have also been identified by the second approach, with the exception of the parameter X67. This parameter has a Sobol index lower than 0.2 and affects the accuracy of the SVM by a decrease of -0.006%, and therefore it can be ignored.

After comparing the results of the two first approaches, it can be concluded that the identification made using these approaches is mostly similar. Another comparison consists of comparing the results of the approach based on correlation research to the results of the first two approaches. This makes it possible to evaluate the estimation ability of the correlation-based approach.

1. *First study case (Chem_4/8)*: Based on the results of the third approach, six key measurement uncertainties could be considered: ***X1, X2, X5, X6, X8***, and ***X11***. These key measurement uncertainties are different from the ones found by the two first approaches, where ***X4*** was included instead of ***X1*** and ***X5***.
2. *Second case study (Chem_5/7)*: two parameters ***X8*** and ***X11*** are well correlated to the class parameter. Thus, these two parameters can be considered as the main key measurement uncertainties of the second dataset. However, compared to the previous results, the manufacturing parameter ***X4*** was included instead of the parameter ***X11***. This shows that the manufacturing parameter with the highest correlation coefficient does not necessarily have the most impact on SVM accuracy.

3. *Third case study (Mine_1/2)*: the three parameters with correlation coefficients greater than ± 0.5 are ***X1***, ***X2*** and ***X22***. Two of these parameters (***X1*** and ***X22***) have also been identified by the other approaches.
4. *Fourth case study (Roll_0/1)*: here, only the parameters with a correlation greater than $\pm$ 0.5 are selected. This resulted in estimating 15 key measurement uncertainties. Six of these fifteen parameters have been identified previously by the two first approaches.

This comparison showed that the correlation-based approach allows estimating a number of the key measurement uncertainties identified by the two first approaches. It can therefore be concluded that the uncertainties of the manufacturing parameters that are well correlated to the class parameter are more likely to have a significant impact on the robustness of the SVM accuracy. Moreover, this estimate is imprecise and rough, where parameters with weak correlation coefficients and significant impacts are overlooked and not estimated as key measurement uncertainties.

In the following, Monte Carlo simulations are used to quantify the decrease in SVM accuracy due to the uncertainties of the parameters identified as key measurement uncertainties. This allows assessing the combined impact of the key measurement uncertainties on the SVM accuracy. The key measurement uncertainties considered in this analysis are:

- 1st case study *(Chem_4/8):* ***X2, X4, X6, X8, X11.***
- 2nd case study *(Chem_5/7):* ***X4, X8.***
- 3rd case study *(Mine_1/2):* ***X1, X7, X22.***
- 4th case study *(Roll_0/1):* ***X45, X49, X66, X78, X85, X90, X92, X95.***

The different results are presented in Figure 2.14.

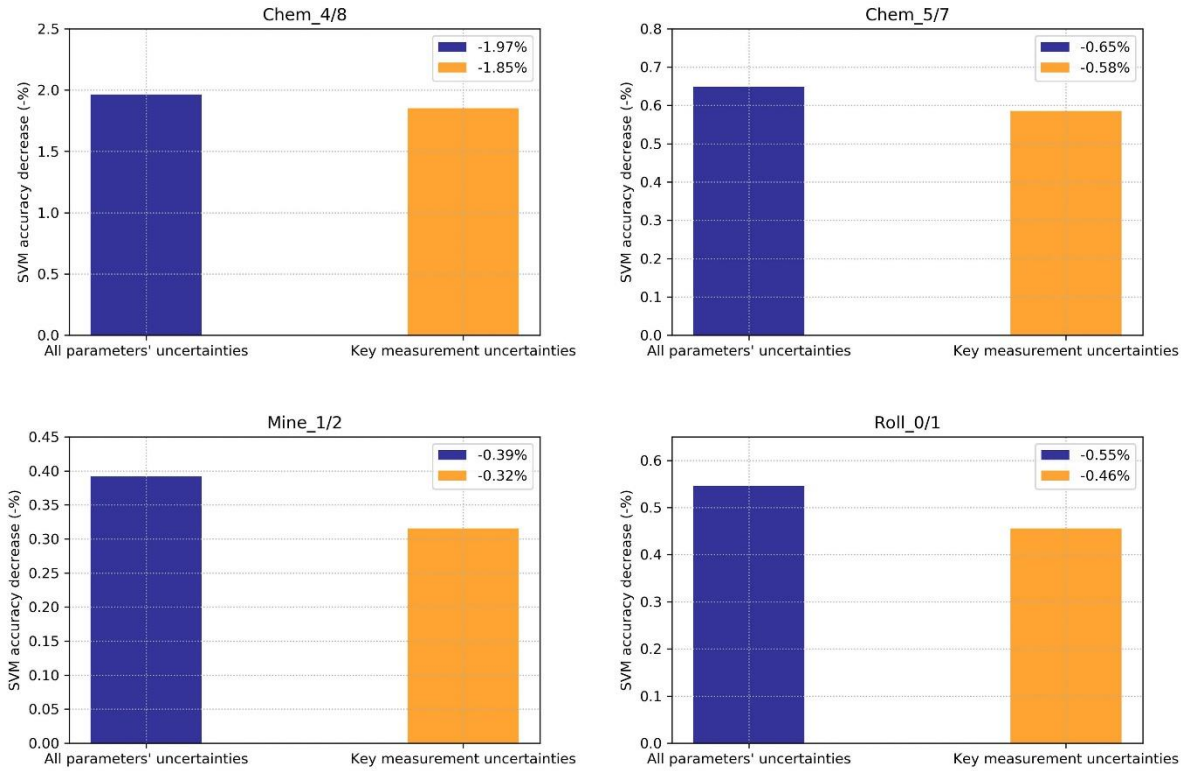


Figure 2.14: Decrease of SVM accuracies due to the key measurement uncertainties

These results show that the impact of the key measurement uncertainties on the SVM accuracy is as significant as the impact of the measurement uncertainties of all the parameters. A robust monitoring of these key measurement uncertainties will therefore allow for robust quality prediction and better control over various manufacturing processes.

Furthermore, the complexity and the computation times of the three approaches are discussed. The analysis performed by Abdiansah and Wardoyo (2015) has proved that the complexity of SVM is $O(n^3)$, either for training the model or predicting new data. The SVM model is trained once in the first approach and used 11001 times for prediction. For the second approach, the SVM model is also trained once and used 12001 times to predict noisy test sets. These approaches were run on Python 3.7 using a computer with the following specifications: Intel(R) Core (TM) i7-8700 CPU @ 3.20GHz, RAM – 16,0 Go. Additionally, the Scikit-learn library was used to train the SVM models, the Pandas library to handle the data, and the NumPy library to perform the various mathematical operations. The computation times needed to get the different results are presented in Table 2.11.

Table 2.11: Computation times of the three proposed approaches

Dataset	Approach	Computation time
Chem_4/8	1st: Monte Carlo simulations	85 s
	2nd: Sobol analysis	128 s
	3rd: Correlation research	3 s
Chem_5/7	1st: Monte Carlo simulations	916 s
	2nd: Sobol analysis	1216 s
	3rd: Correlation research	4 s
Mine_1/2	1st: Monte Carlo simulations	4644 s
	2nd: Sobol analysis	6368 s
	3rd: Correlation research	9 s
Roll_0/1	1st: Monte Carlo simulations	1747 s
	2nd: Sobol analysis	2385 s
	3rd: Correlation research	4 s

To conclude, the problem of measurement uncertainties and their impact on the prediction accuracy of SVM were addressed in this chapter. At first, a Monte Carlo simulation was conducted to assess and to quantify the impact of measurement uncertainties. Afterwards three **novel** approaches were proposed for the identification of the parameters with uncertainties affecting significantly the SVM robustness, i.e., the identification of the key measurement uncertainties. The first approach based on Monte Carlo simulation evaluates how the accuracy of SVM is affected by noising a single parameter. In the second approach, the Sobol sensitivity analysis was reshaped in order to assesses the sensitivity of the SVM accuracy regarding the measurement uncertainties. These two approaches allowed providing quantitative measures that represent the magnitude of the impact of each parameter. Still, the two approaches are based on Monte-Carlo simulations which are time and resource consuming. To overcome this issue, a new approach based on a correlation research is proposed to estimate the key measurement uncertainties. The three approaches were applied to four datasets in order to evaluate their performances. This evaluation can be illustrated by the results of the mining process, where the uncertainties of three parameters (% Iron Feed, Ore Pulp Density and %Iron Concentrate) were identified as the one impacting the robustness of the SVM prediction, and thus impacting the robustness of the Silica level (quality level) prediction. Also, the results did show that the first two approaches are more accurate and precise in identifying the key measurement uncertainties, while the last approach did allow giving an intuition of the measurement uncertainties that are suspicious to have a significant impact on the SVM prediction accuracy.

To summarize, Figure 2.15 presents the main contributions presented and discussed in the second chapter of this PhD thesis.

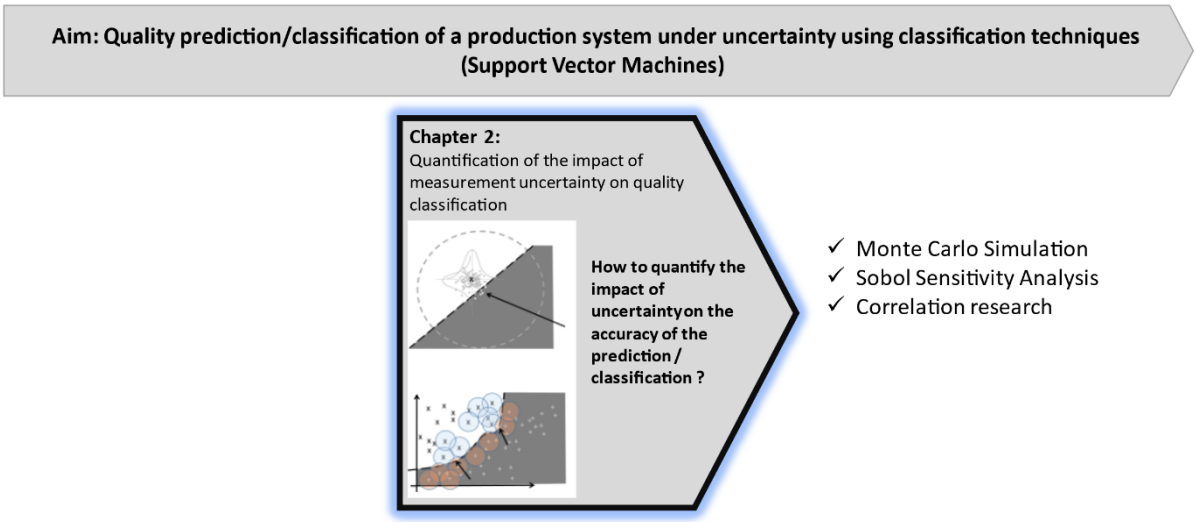


Figure 2.15: The scientific contributions presented in chapter II

This chapter allowed getting a better understanding of the robustness and the sensitivity of the predictive performances of SVM regarding measurement uncertainties. This work has also provided a robust identification of the key parameters leading to quality issues. Therefore, the quality of a manufacturing system can be improved by controlling the key measurement uncertainties.

Chapter III

Towards Robust SVM Model

The main objective of this work is to develop robust SVM models when considering measurements/parameters uncertainties. In this chapter, two groups of different approaches are proposed. One group introduces two new GA-based approaches that allow the selection of SVM models which are more robust to measurement uncertainties. In addition, three approaches are proposed in the second group. The three approaches are based on the definition of feature-weights in order to increase the robustness of the SVM.

III.1 – Introduction

While the second chapter objective is to quantify the impacts of the uncertainties on the prediction accuracy of a SVM model assessing the quality of manufacturing systems, this chapter aims to improve its robustness regarding measurement uncertainties while maintaining optimal predictive performances. This aim can be formulated as a multi-objective optimization problem where the predictive accuracy of SVM needs to be maximized and the impact of measurement uncertainties on the predictive performance of SVM has to be minimized.

Multi-Objective Optimization Problems (MOPs) are an essential part of optimization activities with practical importance in almost all real-world optimization problems (Liu et al., 2020). They concern mathematical optimization problems involving several objective functions that need to be optimized simultaneously, which can be formulated as follows:

$$\min \quad F(x) = \left(f_1(x), f_2(x), \dots, f_m(x)\right) \qquad (3.1)$$
$$s.t \quad x \in D$$

Where: x is a n-dimensional candidate solution and $F(x)$ is a m-dimensional objective space.

Generally, it is hard to optimize multiple objectives simultaneously since they are conflicting. The concept of Pareto dominance is therefore adopted to obtain a set of optimal solutions instead of a single optimal solution. Additionally, a solution y is dominating a solution x only if $\forall i \in \{1,2,\dots,m\},\ f_i(x) \leq f_i(y)$ and $\exists j \in \{1,2,\dots,m\},\ f_j(x) < f_j(y)$. Therefore, a solution is considered as a Pareto optimal if it is not dominated by any other solution.

Several multi-objectives optimizations approaches have been developed, which can be divided into two main categories (Liu et al., 2020), i.e., **classical approaches** like the weighted sum method, the goal-programming method, two phase approaches, etc., and **metaheuristics** such as genetic algorithms and particle swarm optimization (PSO). Multi-objective metaheuristics are widely used as they allow providing optimum solutions within reasonable computation times.

Based on that, and in order to meet the main objective of this study, it is necessary to understand how the robustness of the SVM is defined.

Given a specific kernel space, the SVM searches for an optimal hyperplane that separates the two classes while maximizing the distance between the data points of the two classes. This maximization makes the SVM model less sensitive to uncertainties which results in a better SVM robustness. It was also noted that the robustness of the SVM varies from one kernel space to another. This is illustrated in Figure 3.1, which shows how the probability (blue and red circles) of an uncertain data point to be well predicted is determined by the choice of

separator, and thus by the choice of the kernel space. Accordingly, the main objective can be met by identifying separators that are robust to the impact of measurement uncertainties.

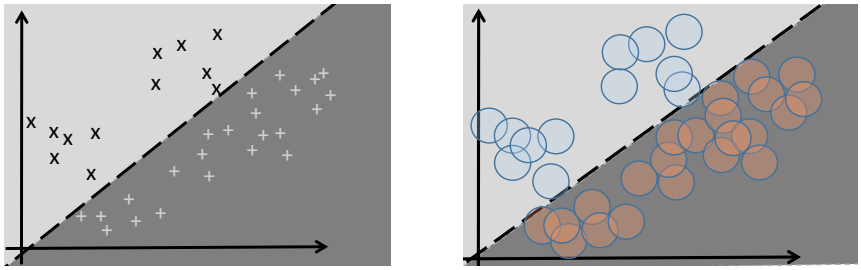


Figure 3.1: Probability of uncertain datapoints to be well classified

Two main ideas are developed to meet this objective:

1- As the separator is defined by the SVM kernel space, the first idea is to identify optimal SVM hyperparameters that allow an optimal predictive accuracy and an optimal robustness to the impact of measurement uncertainties.

2- The SVM separator is built using scalar products. Therefore, this separator is defined by the numerical values of the attributes. A rescaling of these values taking into account the impact of each parameter on the robustness of the SVM is therefore promising.

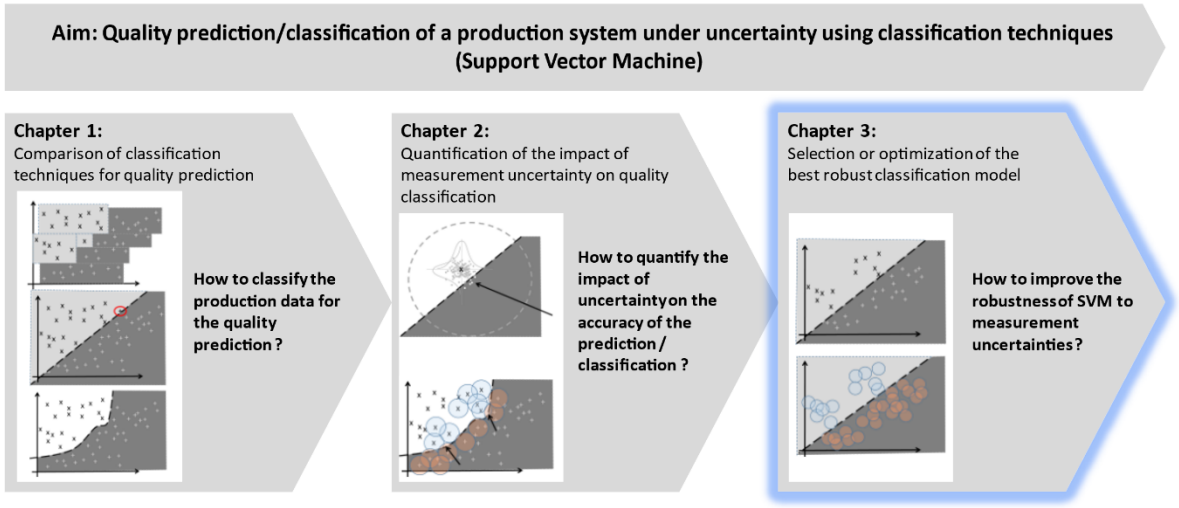


Figure 3.2: Thesis positioning

These two main ideas have been developed through different approaches. Two GA-based approaches aim at identifying SVM kernel spaces with optimal predictive performance and optimal sensitivity to measurement uncertainties. In addition, two other approaches improve the robustness of the SVM by re-scaling the input parameters based on the Sobol sensitivity analysis. Finally, one approach assigns a weight to each parameter, where they are tuned along with the SVM hyperparameters using a genetic algorithm. These different approaches are detailed and discussed in the following.

III.2 – Selection of robust SVM models

In this section, two approaches are proposed for the selection of robust SVM models. These approaches do not modify the SVM formulation but introduce new systematic procedures to identify kernel spaces with optimal predictive performance and optimal robustness to measurement uncertainties. Two objective functions are thus associated with the optimization problem: maximizing the robustness of the SVM and the accuracy of the SVM. A bi-objective optimization is defined. For the resolution, two different approaches are proposed.

III.2.1 – Bi-objective optimization for the selection of robust SVM models

In order to select a robust SVM model, an optimization approach is required aiming at maximizing the SVM accuracy **and at the same time** minimizing the impact of measurement uncertainties. Therefore, to optimize both conflict objectives, it requires to obtain the Pareto optimal solutions with respect to the problem encountered. In this matter, Figure 3.3 illustrates the Pareto optimal solutions found.

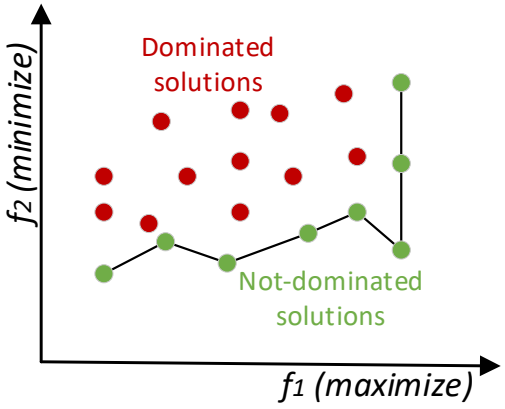


Figure 3.3: Illustration of pareto optimal solutions

Likewise, the objective function that the first approach aims to optimize can be expressed as:

$$\max: \quad f_1(kernel) - f_2(kernel) \tag{3.2}$$

where: $\boldsymbol{f_1}$ is the prediction accuracy of SVM, and $\boldsymbol{f_2}$ is the decrease in the SVM accuracy due to measurement uncertainties, evaluated in the same kernel space.

A genetic algorithm is used for the resolution of the Eq. 3.2. The genetic algorithm evaluates every potential solution by assessing its predictive accuracy on a test set and by calculating its robustness using the Monte-Carlo simulation described in Chapter II (Algorithm 2). This approach is described by Figure 3.4 and Algorithm 5.

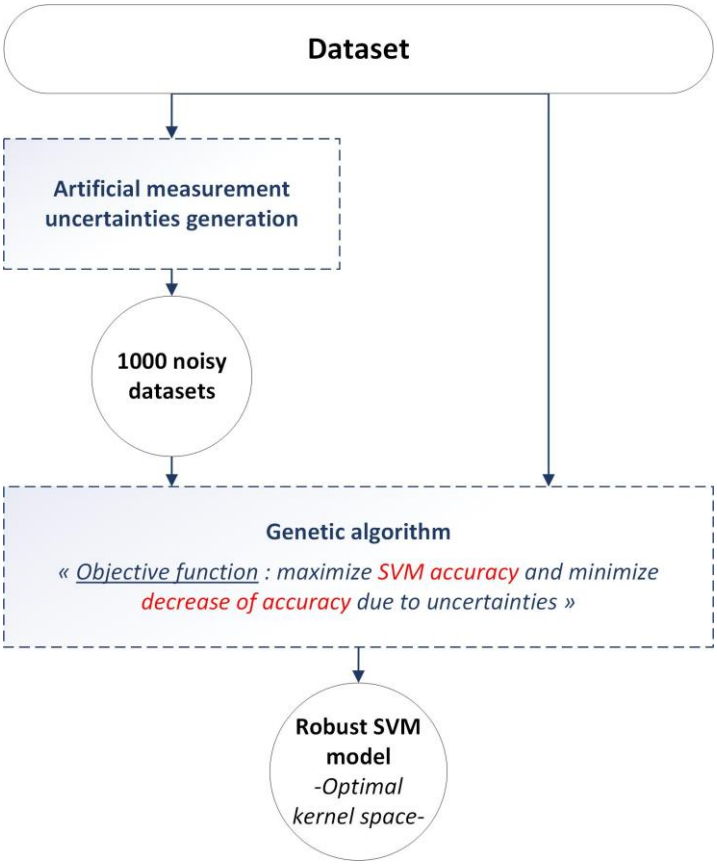


Figure 3.4: Bi-objective optimization approach for the selection of robust SVM models

Algorithm 5 was applied to the datasets using 1000 sets of measurement uncertainties. The different results as well as the robustness improvements due the identification of the new kernel spaces are given in Table 3.1. **The robustness improvement is defined as the difference between the accuracy on noisy data found by the proposed approach and the accuracy on noisy data as computed in chapter II**. In the case of *Chem_4/8*, for example, the proposed approach improved the accuracy on noisy sets from 87.51% to 89.01%.

Algorithm 5 Bi-objective optimization of SVM hyperparameters by genetic algorithm.

Inputs: Dataset ***M***, population size ***pop_size***, number of generations ***max_gen***, N-sample of measurement uncertainties sets ***MU***

Output: optimal SVM kernel space $\boldsymbol{Ker}$

01: Split ***M*** into training set ***Ml***, and test set ***Mt***
02: Encode the solution using an integer representation
03: Generate an initial random population of size ***pop_size: POP***
04: for $i \in \{1, \dots, \boldsymbol{pop\_size}\}$ do
05: Train an SVM model using the hyperparameters encoded in ***POP***[*i*]***: Model_SVM***
06: Predict ***Mt*** using ***Model_SVM: Acc***
07: Evaluate the robustness of ***Model_SVM*** using **Algorithm 2***:* $\Delta \boldsymbol{Acc}$
08: Define the fitness score of ***POP[i]*** as: ***Acc-***$\Delta \boldsymbol{Acc}$
09: end for
10: Pick the solution with the highest fitness score: $\boldsymbol{Ker}$
11: for $j \in \{1, \dots, \boldsymbol{max\_gen}\}$ do
12: Apply 3-way tournament selection to create the off-spring: ***POP***
13: Apply uniform crossover to ***POP: POP***
14: Apply random resetting mutation to ***POP: POP***
15: Evaluate the fitness of ***POP***
16: Pick the fittest solution of the j^{th} generation: $\boldsymbol{Ker_j}$
17: if $\boldsymbol{Ker_j}$ is fitter than $\boldsymbol{Ker}$
18: $\boldsymbol{Ker} \leftarrow \boldsymbol{Ker_j}$
19: end if
20: end for
21: return $\boldsymbol{Ker}$

Table 3.1: Results of the bi-objective optimization (Algorithm 5)

Dataset	Kernel	σ	C	Accuracy on test set	Accuracy on noisy sets	Robustness improvement
Chem_4/8	RBF	0.043	924.0	88.60%	89.01% (+0.40%)	+1.50%
Chem_5/7	RBF	0.35	215.8	85.77%	85.12% (-0.65%)	+0.00%
Mines_1/2	RBF	0.494	15.9	82.32%	81.93% (-0.39%)	-0.01%
Roll_0/1	RBF	0.37	0.3	91.44%	91.22% (-0.22%)	+0.33%

In this table, it is shown that Algorithm 5 is able to improve the robustness to the impact of measurement uncertainties, particularly for the *Chem_4/8* and *Roll_0/1* datasets. However, in the case of *Chem_4/8* data, the accuracy on the initial test is 88.60%, which is less than the accuracy previously determined (89.47%). Therefore, to ensure a maximal accuracy on the initial test set, a bi-level optimization is proposed.

III.2.2 – Bi-level optimization for the selection of robust SVM models

The second approach for the selection of robust SVM models consists of two main steps: one first optimization is performed for the identification of all the kernel spaces that allow having a maximal SVM prediction accuracy. The second step aims to assess all the identified kernel spaces in order to select the one that minimizes the impact of measurement uncertainties. As shown in Figure 3.5, this would ensure identifying optimal Pareto solutions, with a maximal SVM accuracy, but not necessarily an optimal SVM robustness.

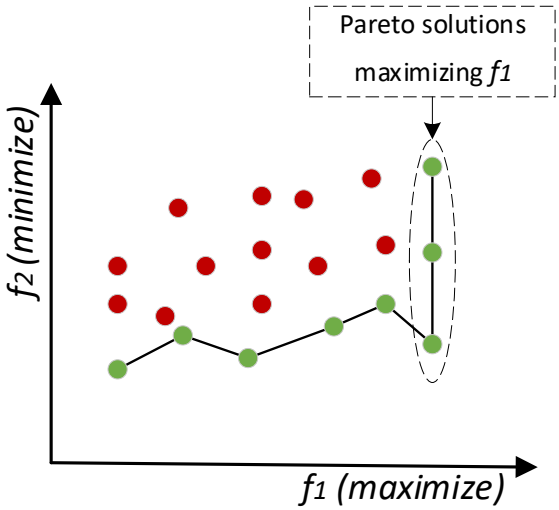


Figure 3.5: Illustration of bi-level optimization solutions

Consequently, the proposed bi-level optimization can be formulated as follows:

$$\begin{aligned} &\min && f_2(kernel) \\ &\text{s.t} && max\ \ f_1(kernel) \end{aligned} \tag{3.3}$$

A genetic algorithm is used to identify all the optimal kernel spaces in the first step. Then the robustness of every solution is assessed through a Mont Carlo simulation based on Algorithm 2. The second approach is described by Figure 3.6 and Algorithm 6.

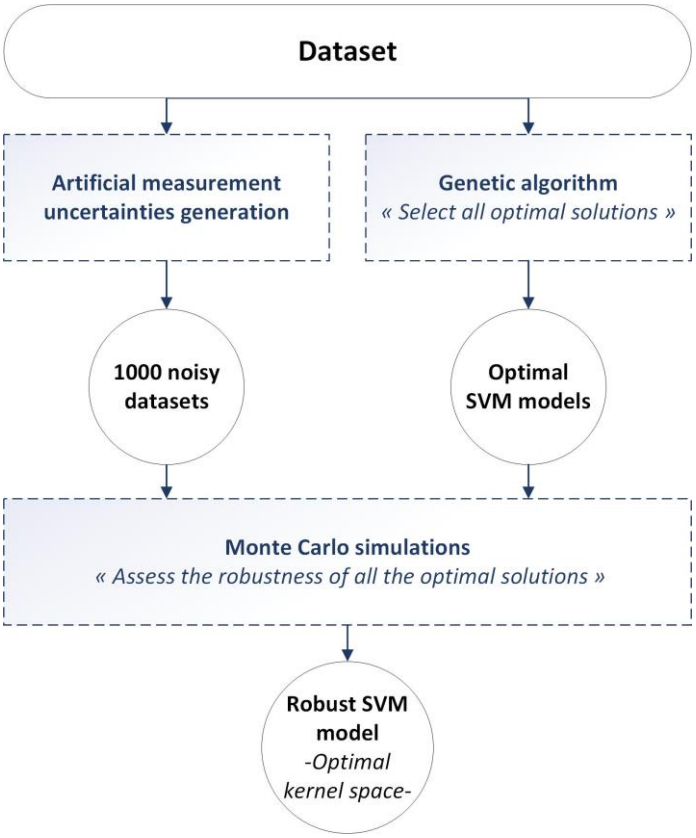


Figure 3.6: Bi-level optimization approach for the selection of robust SVM models

Algorithm 6 Bi-level optimization of SVM hyperparameters.

Inputs: Dataset ***M,*** N-sample of noisy datasets ***MU.***

Output: optimal SVM hyperparameters: $\boldsymbol{Ker}$

01: Split ***M*** into training set ***Ml*** and test set ***Mt.***

02: Optimize SVM model based on **Algorithm 1.**

03: Pick all solutions that maximize the prediction accuracy on test set: $\{\boldsymbol{Ker_i}\}_{i\epsilon\{1,\dots,P\}}$

04: for $i \in \{1, \dots, P\}$ do

05: Evaluate the robustness of $\boldsymbol{Ker_i}$ using **Algorithm 2***:* $\Delta\boldsymbol{Acc_i}$

06: end for

07: Choose $\boldsymbol{Ker_k}$ that minimizes $\Delta\boldsymbol{Acc}$

08: return $\boldsymbol{Ker_k}$

The results of applying Algorithm 6 to the datasets are given in Table 3.2.

Table 3.2: Results of the bi-level optimization

Dataset	Kernel	σ	C	Accuracy on test set	Accuracy on noisy sets	Robustness improvement
Chem_4/8	RBF	0.108	207.1	89.47%	88.63% (-0.85%)	+1.11%
Chem_5/7	RBF	0.333	243.8	85.77%	85.12% (-0.65%)	+0.00%
Mines_1/2	RBF	0.494	15.9	82.34%	81.95% (-0.39%)	+0.01%
Roll_0/1	RBF	0.356	0.3	91.44%	91.22% (-0.22%)	+0.33%

In the case of *Chem_4/8* and the *Roll_0/1* datasets, Algorithm 6 allowed identifying SVM models with better robustness to measurement uncertainties. The main difference between the results of Algorithm 5 and Algorithm 6 is noticed at the level of the *Chem_4/8* dataset, where the former algorithm allowed a prediction accuracy of 88.60% on the initial test set, and the latter maintained an optimal accuracy of 89.47%. On the other hand, we can notice that the SVM models selected for the *Chem_5/7* and *Mines_1/2* datasets did not allow a better robustness to measurement uncertainties. A modification of the basic SVM formulation might therefore be necessary to furtherly improve the robustness of the technique.

III.3 – Feature weighting for the improvement of SVM robustness

In this section, three novel approaches based on the introduction of feature-weights are proposed in order to improve the robustness of SVM models. The first two approaches assign different weights to the input parameters considering Sobol total-effect indices and a new parameter denoted by R. On the other hand, the third approach allows simultaneous optimization of the SVM hyperparameters and feature-weights in order to increase the performances of SVM.

III.3.1 – Approach based on Sobol sensitivity indices

This sub-section introduces the approaches based on Sobol sensitivity analysis. First, the definition of the new weights is provided and the impact of these weights on the SVM decision boundary is analyzed. Then, the two algorithms to tune the parameter R are presented and their application to the datasets are discussed.

Weights definition

Since the measurement uncertainties are additive and in order to rescale the different features according to their impact on the SVM prediction accuracy, different weights based on Sobol' indices are associated. This rescaling modifies the orientation of the decision boundary and therefore affects the predictive performance of SVM as well as its robustness to the measurement uncertainties. The weights are defined as expressed in Eq. 3.4:

$$w_j = 1 + R.ST_j \tag{3.4}$$

where R is a parameter to tune and ST_j is the calculated Sobol index of the j^{th} feature.

It is necessary to use small values of R in order to avoid large changes at the level of data values and therefore avoid the loss of information contained within these data. The impact of the defined weights on the SVM decision boundary could be illustrated in Figure 3.7, where the basic SVM model (R=0) is compared to another SVM model generated with a R=-0.2.

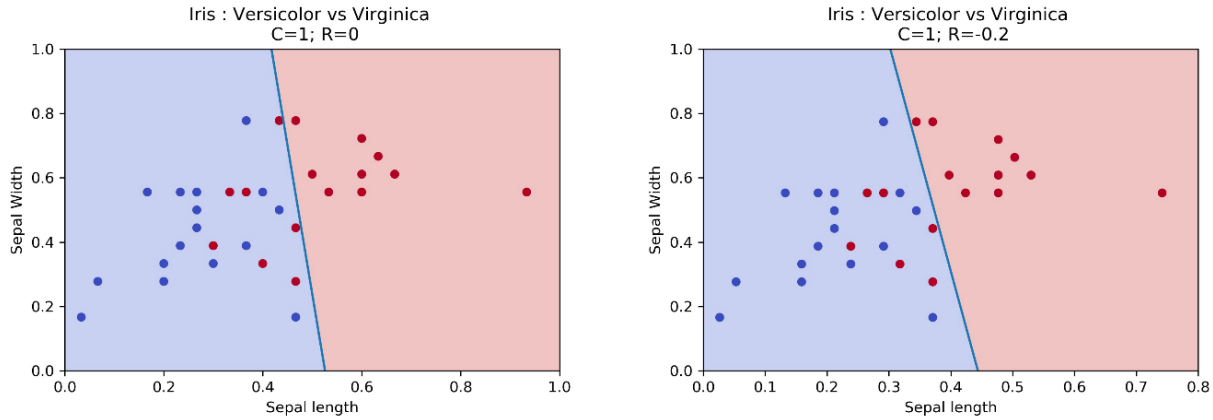


Figure 3.7: Impact of the parameter R on the linear SVM decision boundary

Figure 3.7 represents two linear SVM models for the classification of the Iris data. The estimated Sobol indices are: 1.01 for the “sepal length” feature and 0.02 for the “sepal width” feature. The first feature is mostly affecting the robustness of the SVM accuracy, and therefore it needs to be rescaled in order to reduce its impact. In Figure 3.7 and with R=-0.2, the orientation of the decision boundary has been changed, improving the prediction accuracy from 76.47% to 79.41%, and reducing the sensitivity of SVM accuracy from +0.3% to -0.01%. These results emphasize the potential of the proposed approach. Consequently, both approaches are proposed to identify an optimal value for the parameter R in order to avoid setting it randomly.

Parameter R tuning by a grid search

The first Sobol based approach aims to improve the robustness of SVM regarding measurement uncertainties by only tuning the parameter R, and by using the same optimal SVM hyperparameters identified in **chapter I**. A grid search is performed to evaluate different values for the parameter R. The R-value that maintains the predictive performance of the SVM while improving its robustness to uncertainties is selected.

Figure 3.8 provides a graphical description of the first approach, while Algorithm 7 presents the different steps to implement it.

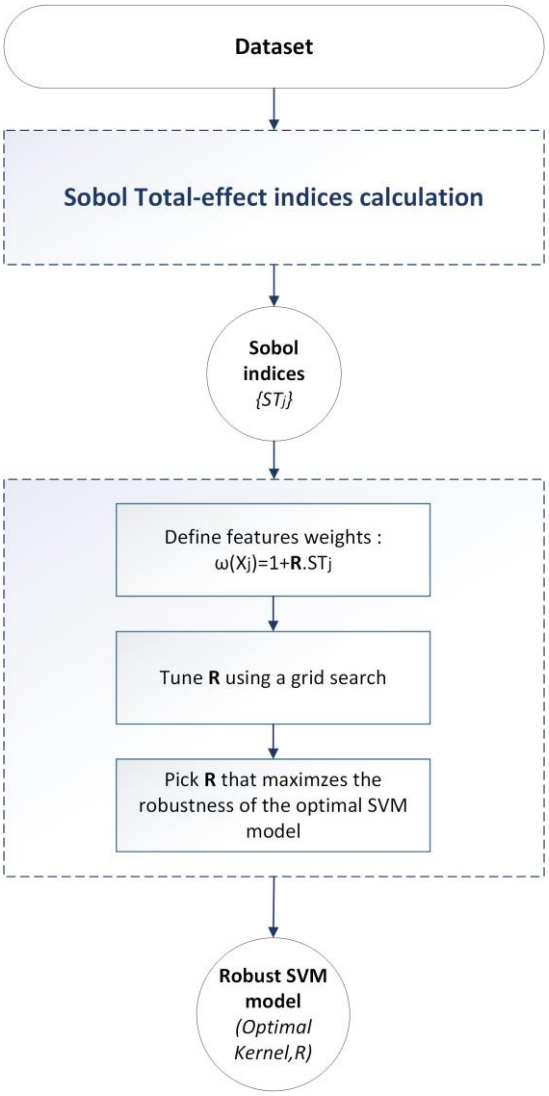


Figure 3.8: A grid search algorithm for tuning the parameter R

Algorithm 7 was applied to the datasets by considering 81 values of R between -0.2 and 0.2and a sample of 1000 noisy datasets. Table 3.3 presents the different results as well as the different optimal R values that increase the robustness and maintain the predictive performance of the SVM models.

Algorithm 7 Optimizing SVM robustness by tuning the parameter R – Grid Search

Inputs: Dataset ***M*,** SOBOL total-effect indices **{ST_j},** N-sample of noisy datasets ***MU*,** optimal SVM hyperparameters ***Ker.***

Output: optimal coefficient ***R.***

01: Get the number of features in ***M***: ***Cols***
02: Define a variation range of ***R: [Rmin, Rmax]***
03: Define a variation step: ***r***
04: ***R*** ← ***Rmin***
05: while ***R*** ≤ ***Rmax*** do
06: for $j \in \{1, \dots, \boldsymbol{Cols}\}$ do
07: $\boldsymbol{\omega}(\boldsymbol{X_j}) \leftarrow \mathbf{1} + \boldsymbol{R}.\boldsymbol{ST_j}$
08: $\boldsymbol{W\_X_j} \leftarrow \boldsymbol{\omega}(\boldsymbol{X_j}).\boldsymbol{X_j}$
09: end for
10: $\boldsymbol{W\_M} \leftarrow \{\boldsymbol{W\_X_j}\}_{\boldsymbol{j} \in \{\mathbf{1},\dots,\boldsymbol{Cols}\}}$
11: Split ***W_M*** into training set ***W_Ml***, and test set ***W_Mt.***
12: Use ***W_ Ml*** and the hyperparameters ***ker*** to train the SVM model: ***Model_SVM_R***
13: Predict ***W_Mt*** using ***Model_SVM_R: Acc***
14: $\Delta\boldsymbol{acc} \leftarrow \mathbf{0}$
15: for $i \in \{1, \dots, N\}$ do
16: Rescale the i^{th} dataset of ***MU*** using the same weights $\{\boldsymbol{\omega}(\boldsymbol{X_j})\}$**: *W_MU***
17: Split ***W_MU*** into training set ***W_MUl***, and test set ***W_MUt.***
18: Predict ***W_ MUt*** using ***Model_SVM_R:*** $\boldsymbol{Acc_{noised}}$
19: $\Delta\boldsymbol{acc} \leftarrow \frac{1}{i} * ((\boldsymbol{Acc} - \boldsymbol{Acc_{noised}}) + (\boldsymbol{i} - 1) * \Delta\boldsymbol{acc})$
20: end for
21: $\boldsymbol{R} \leftarrow \boldsymbol{R} + \boldsymbol{r}$
22: end while
23: Pick ***R*** that minimizes $\Delta\boldsymbol{acc}$ and maintain the prediction accuracy of SVM***: R***
24: return ***R***

Table 3.3: Results of Algorithm 7

Dataset	Kernel	σ	C	R	Accuracy on test set	Accuracy on noisy sets	Robustness improvement
Chem_4/8	RBF	0.292	15.0	0.005	89.47%	87.70% (-1.77%)	+0.19%
Chem_5/7	RBF	0.340	231.8	-0.03	85.77%	85.16% (-0.61%)	+0.04%
Mines_1/2	RBF	0.364	23.4	0.015	82.34%	81.99% (-0.35%)	+0.05%
Roll_0/1	RBF	0.317	883.4	-0.04	91.44%	90.91% (-0.53%)	+0.02%

These results show that Algorithm 7 allowed improving **slightly** the robustness of SVM to the impact of measurement uncertainties while maintaining the same prediction accuracies. It can be noticed that in the four datasets, the parameter R was never equal to zero, which emphasize the potential of this parameter in offering SVM models that are more robust.

Parameter R and SVM hyperparameters tuning by a genetic algorithm

In the previous algorithm, the robustness of the SVM is improved only by tuning the parameter R. To better improve the SVM model robustness, SVM hyperparameters will be included in the optimization. Algorithm 8 is therefore developed. This algorithm consists of two steps. The first step is based on genetic algorithm for the identification of all the optimal solutions $\{(Kernel\ space)_i, R_i\}$ that allow a maximal SVM prediction accuracy. Then, in the second step, the robustness of each of these optimal solutions is evaluated through Monte Carlo simulation. For a better understanding, Figure 3.9, and Algorithm 8 depict the different steps of the approach.

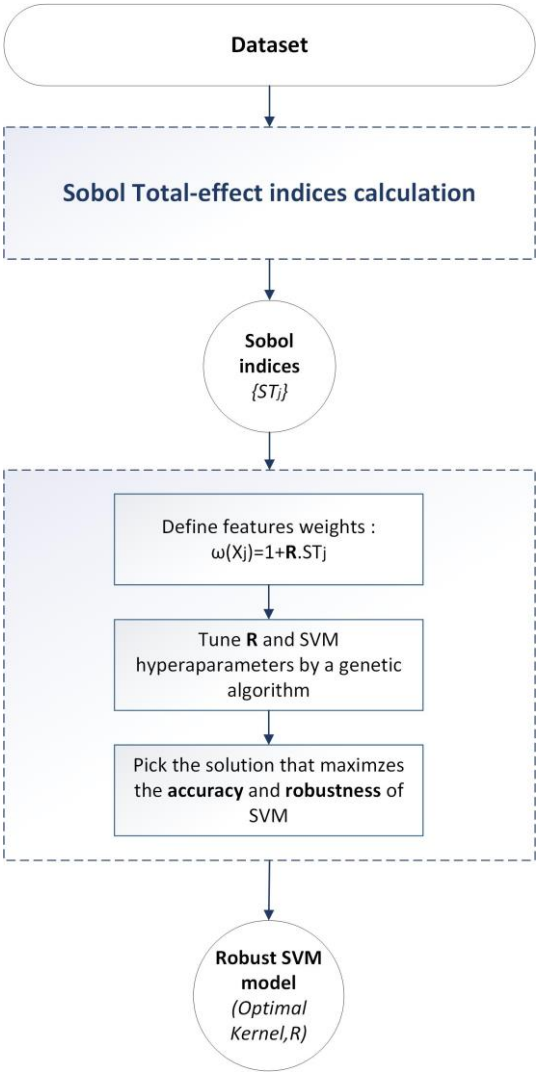


Figure 3.9: Optimization of the parameter R and the SVM hyperparameters

Algorithm 8 Optimizing SVM robustness by tuning the parameter R and the SVM hyperparameters – Genetic Algorithm

Inputs: Dataset ***M*,** SOBOL total-effect indices **$\{ST_j\}$,** N-sample of noisy datasets ***MU*.**

Output: optimal SVM hyperparameters and ***R*** coefficient: $\{\boldsymbol{Ker}, \boldsymbol{R}\}$

01: Get the number of features in ***M***: ***Cols***

02: for $j \in \{1, \dots, \boldsymbol{Cols}\}$ do

03: Assign a weight ***ω(Xj)*** to the j^{th} feature ***Xj*** of ***M***

04: $\boldsymbol{\omega}(\boldsymbol{X_j}) \leftarrow \mathbf{1} + \boldsymbol{R}.\boldsymbol{ST_j}$

05: end for

06: Optimize ***R*** and ***SVM hyperparameters*** based on ***Algorithm 1***

07: Pick solutions that maximize the accuracy on the **weighted** test set:

$\{\boldsymbol{Ker_i}, \boldsymbol{R_i}\}_{\ i\in\{1,\dots,P\}}$

08: for $i \in \{1, \dots, P\}$ do

09: for $j \in \{1, \dots, \boldsymbol{Cols}\}$ do

10: $\boldsymbol{\omega}(\boldsymbol{X_j}) \leftarrow \mathbf{1} + \boldsymbol{R_i}.\boldsymbol{ST_j}$

11: $\boldsymbol{W\_X_j} \leftarrow \boldsymbol{\omega}(\boldsymbol{X_j}).\boldsymbol{X_j}$

12: end for

13: $\boldsymbol{W\_M} \leftarrow \{\boldsymbol{W\_X_j}\}_{\boldsymbol{j}\in\{\mathbf{1},\dots,\boldsymbol{Cols}\}}$

14: Split ***W_M*** into training set ***W_Ml*** and test set ***W_Mt.***

15: Train ***W_ Ml*** using $\{\boldsymbol{Ker_i}, \boldsymbol{R_i}\}$***: Model_SVM_*** $\boldsymbol{R_i}$

16: Predict ***W_Mt*** using ***Model_SVM_*** $\boldsymbol{R_i}$***: Acc***

17: $\Delta\boldsymbol{acc} \leftarrow \mathbf{0}$

18: for $n \in \{1, \dots, N\}$ do

19: Rescale the n^{th} dataset of ***MU*** using the same weights $\{\boldsymbol{\omega}(\boldsymbol{X_j})\}$**: W_MU**

20: Split ***W_MU*** into training set ***W_MUl*** and test set ***W_MUt.***

21: Predict ***W_ MUt*** using ***Model_SVM_*** $\boldsymbol{R_i}$***:*** $\boldsymbol{Acc_{noised}}$

22: $\Delta\boldsymbol{acc} \leftarrow \frac{1}{n} * ((\boldsymbol{Acc} - \boldsymbol{Acc_{noised}}) + (n-1) * \Delta\boldsymbol{acc})$

23: end for

24: end for

25: Choose $\{\boldsymbol{Ker}, \boldsymbol{R}\}$ that minimizes $\Delta\boldsymbol{acc}$

26: return $\{\boldsymbol{Ker}, \boldsymbol{R}\}$

Algorithm 8 was applied to the datasets by varying R between -0.2 and 0.2. This allowed getting the different results given in Table 3.4.

Table 3.4: Results of Algorithm 8

Dataset	Kernel	σ	C	R	Accuracy on test set	Accuracy on noisy sets	Robustness improvement
Chem_4/8	RBF	0.046	866.5	-0.168	89.47%	87.70% (-1.55%)	+1.41%
Chem_5/7	RBF	0.286	306.9	0.125	85.77%	85.32% (-0.45%)	+0.20%
Mines_1/2	RBF	0.489	0.6	-0.052	82.55%	81.99% (-0.25%)	+0.15%
Roll_0/1	RBF	0.420	0.1	0.007	91.44%	91.22% (-0.22%)	+0.33%

Algorithm 8 provides better solutions than the previous one, where the optimal SVM models are more robust to measurement uncertainties. Improvement at the level of the prediction accuracy of SVM is also observed for the *Mines_1/2* data. This demonstrates that considering the R parameter (thus feature weighting) while optimizing the SVM hyperparameters allow providing models that are less sensitive to measurement uncertainties. The following subsection proposes an approach that further investigates the ability of feature-weights in improving the performance of SVM.

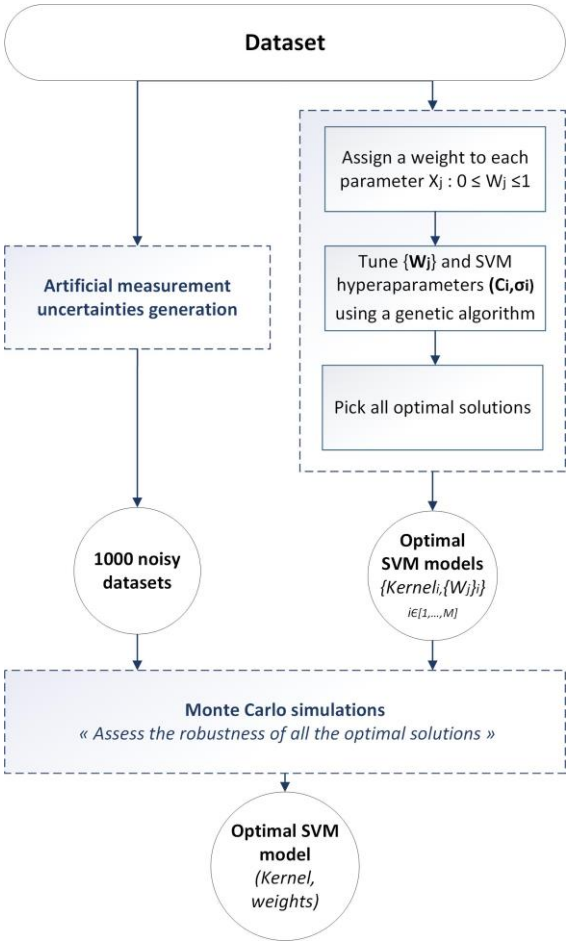
III.3.2 – Improvement of SVM robustness: feature weighting and SVM hyperparameters optimization

This approach is proposed to determine an optimal set of feature weights and SVM hyperparameters using a genetic algorithm. In contrast with the previous Sobol based approaches, this approach needs no information about the weights of features. The tuning of weights and of the hyperparameters is guided by the performances of SVM, i.e., the genetic algorithm receives the feedback of the SVM classifier to determine the searching directions. All the optimal solutions are selected at the end of the running of the genetic algorithm. The robustness of these solutions to measurement uncertainties is evaluated afterwards using Monte Carlo simulations. This optimization problem is defined by Eq. (3.5).

$$\begin{aligned} &\min && f_2(\text{kernel}, \text{weights}) \\ &\text{s.t} && \max\ f_1(\text{kernel}, \text{weights}) \\ &&& \text{s.t} \quad 0 \leq \text{weights} \leq 1 \end{aligned} \tag{3.5}$$

where $kernel$ refers to kernel space defined by SVM hyperparameters, and $weights$ to feature-weights.

The approach followed to solve the optimization problem is described in Figure 3.10 and Algorithm 9.



3.10: Robust optimization of feature weights and SVM hyperparameters

Algorithm 9 Robust optimization of feature weights and SVM hyperparameters

Inputs: Dataset ***M,*** N-sample of measurement uncertainties set ***MU.***

Output: optimal SVM hyperparameters and feature weights: $(\boldsymbol{Ker}, \{\boldsymbol{\omega_j}\})$.

01: Get the number of features in ***M***: ***Cols***

02: for $j \in \{1, \dots, \boldsymbol{Cols}\}$ do

03: Associate a weight $\boldsymbol{\omega_j}$ to each input parameter $\boldsymbol{X_j}$ of ***M***

04: $0 \leq \boldsymbol{\omega_j} \leq 1$

05: $\boldsymbol{W\_X_j} \leftarrow \boldsymbol{\omega_j}.\boldsymbol{X_j}$

06: end for

07: $\boldsymbol{W\_M} \leftarrow \{\boldsymbol{W\_X_j}\}_{\boldsymbol{j}\in\{\mathbf{1},\dots,\boldsymbol{Cols}\}}$

08: Split ***W_M*** into training set ***W_Ml*** and test set ***W_Mt.***

09: Optimize $\{\boldsymbol{\omega_j}\}$ and the SVM hyperparameters based on **Algorithm 1**

10: Pick the solutions that maximize the accuracy of the associated weighted test sets: $(\boldsymbol{Ker}, \{\boldsymbol{\omega_j}\})_{i\epsilon\{1,\dots,P\}}$

11: for $i \in \{1, \dots, P\}$ do

12: Evaluate the robustness of $(\boldsymbol{Ker}, \{\boldsymbol{\omega_j}\})_{i\epsilon\{1,\dots,P\}}$ using **Algorithm 2***:* $\Delta\boldsymbol{Acc_i}$

13: end for

14: Choose $\{\boldsymbol{Ker}, \{\boldsymbol{\omega_j}\}\}$ that minimizes $\Delta\boldsymbol{acc}$

15: return $\{\boldsymbol{Ker}, \{\boldsymbol{\omega_j}\}\}$

Algorithm 9 is applied to the several sets, which allows getting the results given in Tables 3.5 ,3.6 and 3.7.

Table 3.5: Optimal sets of weights– *Chemical data*

Datasets	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9	ω_{10}	ω_{11}
Chem_4/8	0.32	0.21	0.65	0.23	0.08	0.11	0.74	0.06	0.23	0.94	0.40
Chem_5/7	0.16	0.83	0.70	0.86	0.52	0.83	0.4	0.03	0.34	0.08	0.01

Table 3.6: Optimal set of weights– *Mine_1/2 data*

Dataset	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9	ω_{10}	ω_{11}
Mine_1/2	0.91	0.71	0.43	0.97	0.83	0.48	0.13	0.36	0.55	0.42	0.65
	ω_{12}	ω_{13}	ω_{14}	ω_{15}	ω_{16}	ω_{17}	ω_{18}	ω_{19}	ω_{20}	ω_{21}	ω_{22}
	0.29	0.46	0.22	0.65	0.31	0.05	0.84	1.00	0.66	0.12	1.00

The weights related to the *Roll_0/1* are given in Appendix C, since this dataset consists of 95 manufacturing parameters.

Table 3.7: Algorithm 9 application

Dataset	**Kernel**	**σ**	**C**	**weights**	**Accuracy on test set**	**Accuracy on noisy sets**	**Robustness improvement**
Chem_4/8	RBF	1.23	133.2	Table 3.5	94.74%	94.03% (-0.7%)	+6.52%
Chem_5/7	RBF	1.636	293	Table 3.5	88.08%	87.73% (-0.34%)	+2.61%
Mines_1/2	RBF	8.986	584.7	Table 3.6	95.53%	95.35% (-0.18%)	+13.41%
Roll_0/1	RBF	1.341	898.9	Given in Appendix C	92.47%	92.43% (-0.04%)	+1.45%

The results provided in Table 3.7 indicate that this approach provided good results regarding its ability to improve the predictive performance of the SVM as well as the robustness to measurement uncertainties. For example, the SVM model generated for the *Mine_1/2* data increases the prediction accuracy by +13.19% and limits the impact of measurement uncertainties to only -0.18%. In the following, the different approaches developed in the chapter are discussed, and some concluding remarks are given.

III.4 - Discussion and conclusions

The results of the application of the approaches proposed in this chapter III are carefully analyzed.

At first, the approaches are compared to the SVM models studied in **chapter I** and **chapter II**. This comparison makes it possible to evaluate the performance of the proposed approaches and to identify algorithms that best optimize the robustness of SVM to measurement uncertainties. In addition, the computation times of the algorithms are discussed, which would allow the evaluation of their efficiency.

In the rest of this manuscript the following nomenclature is proposed to refer to the different approaches:

1. **Basic_SVM**: SVM classifiers identified by approaches described in chapter I and chapter II.
2. **Bi_Objective**: **bi-objective** optimization associated to 1st approach in this chapter.
3. **Bi_Level: bi-level** optimization associated to the 2nd approach in this chapter.
4. **Sobol_GS: Sobol base-approach** using **grid search**.
5. **Sobol_GA: Sobol based approach** using **genetic algorithm**.
6. **Weighted_SVM:** to refer to the last approach that tunes SVM hyperparameters and **feature-weights** simultaneously.

Overall, for all the case studies, the SVM models provided a classification with a prediction accuracy of more than 80%. In addition, it can be observed that:

1. *Case of Chem_4/8 dataset*: all approaches allowed the identification of SVM model that improve the robustness of the **Basic_SVM** models. The robust SVM model was provided by the **Weighted_SVM** approach, where the prediction accuracy was improved from 89.47% to 94.74% and the robustness to measurement uncertainties from -1.96% to -0.7%. Another interesting remark can be made about the **Bi_Objective** approach where the prediction accuracy of SVM on noisy data is greater than the accuracy on the initial test set. The accuracy can therefore be improved by the perturbation by measurement uncertainties. However, the **Bi_Objective** SVM model allows the lowest prediction accuracy on the initial test set.
2. *Case of Chem_5/7 dataset*: the **Bi_Objective** and **Bi_Level** approaches did not allow improving the robustness of the SVM. Besides, small improvements were observed when using the **Sobol_GS** and the **Sobol_GA** approaches. Unlike, the **Weighted_SVM** approach that led to the generation of an SVM model with a greater prediction accuracy and improved robustness.
3. *Mines_1/2 dataset*: **Bi_Objective, Bi_Level, and Sobol_GS** approaches allowed some robustness improvements of the SVM model despite the measurement uncertainties. The **Sobol_GA** approach allowed the generation of a model that increases the SVM

accuracy by +0.20% and improve the SVM robustness (from -0.40% to -0.25%). Effective results in terms of prediction accuracy and improved robustness have been achieved using the **Weighted_SVM** approach, which increased prediction accuracy by +13.19% and reduced the impact of measurement uncertainties to -0.18%.

4. *Roll_0/1 dataset:* the **Sobol_GS** approach has only increased the robustness of SVM by +0.02, all other approaches allowed significant improvements. Figures 3.11 and 3.12 summarize the different results discussed above.

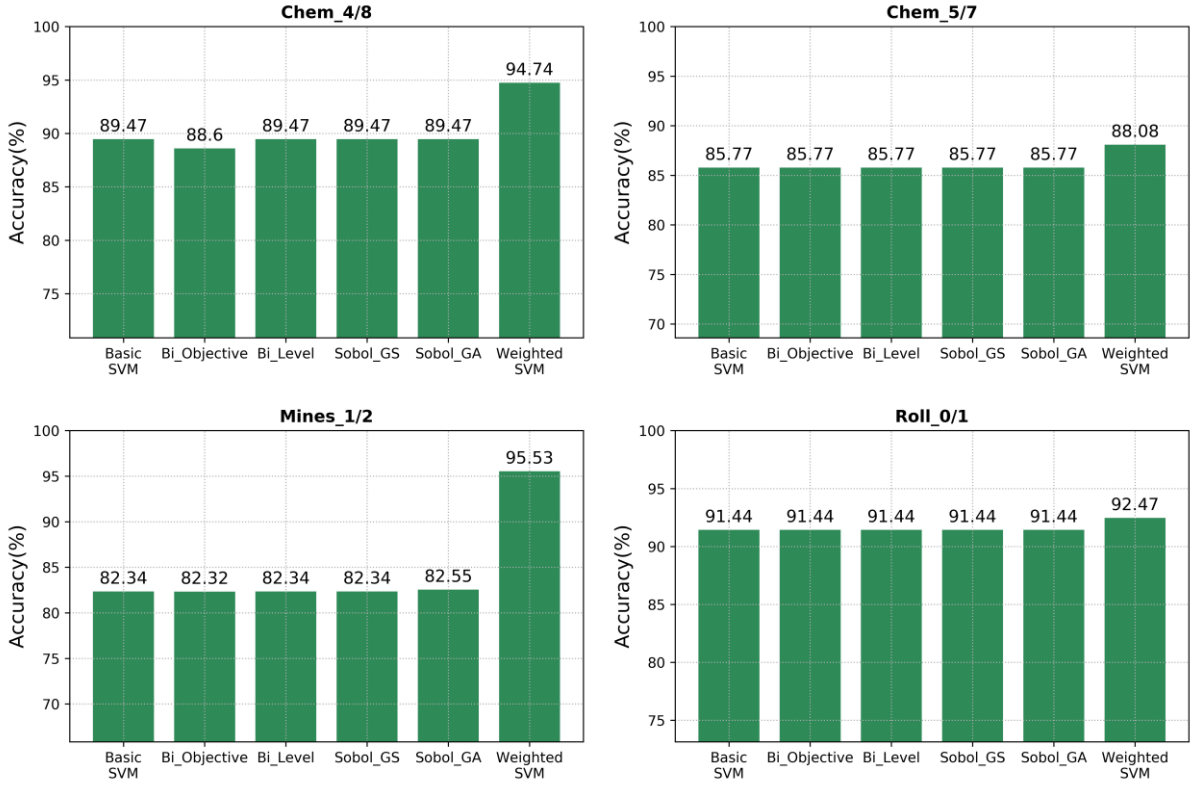


Figure 3.11: Prediction accuracies of the proposed approaches for the four datasets

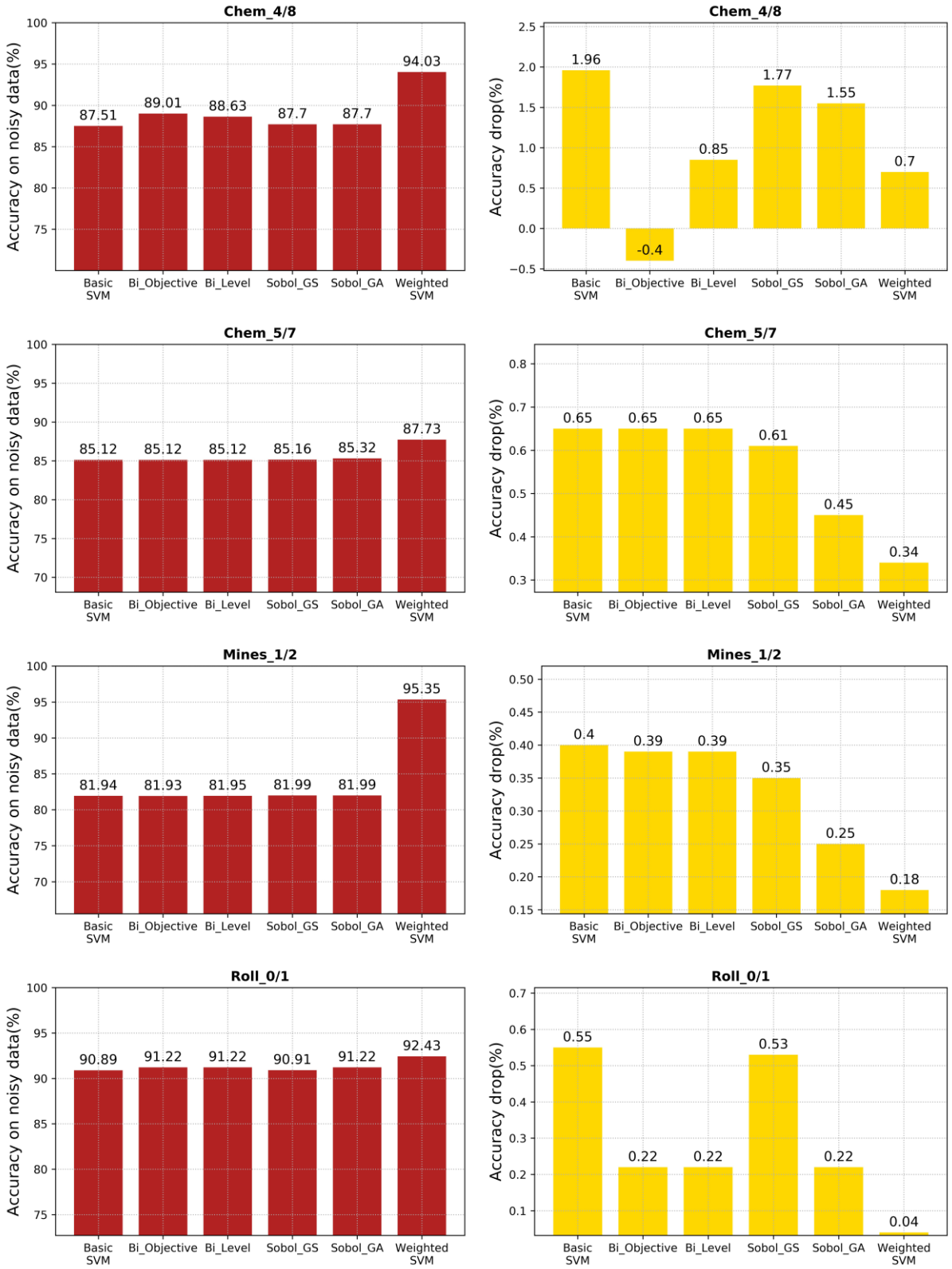


Figure 3.12: Prediction robustness of the proposed approaches

Finally, the approaches were compared in terms of computing time. Accordingly, the computation times of the proposed approaches are given in Table 3.8 and illustrated in Figure 3.13.

Table 3.8: computation time for the identification of optimal robust solutions

Dataset	Bi_Objective	Bi_Level	Sobol_GS	Sobol_GA	Weighted_SVM
Chem_4/8	64 h 48 min	18 min	1 h 10 min	2h 26min	03 min
Chem_5/7	295 h 12 min	44 min	8 h 50 min	19h 59min	24 min
Mines_1/2	805 h 26 min	1 h 50 min	63 h 03 min	11h 31min	3 h 36 min
Roll_0/1	527 h 29 min	34 h 20 min	17h 15 min	9 h 01 min	6 h 17 min

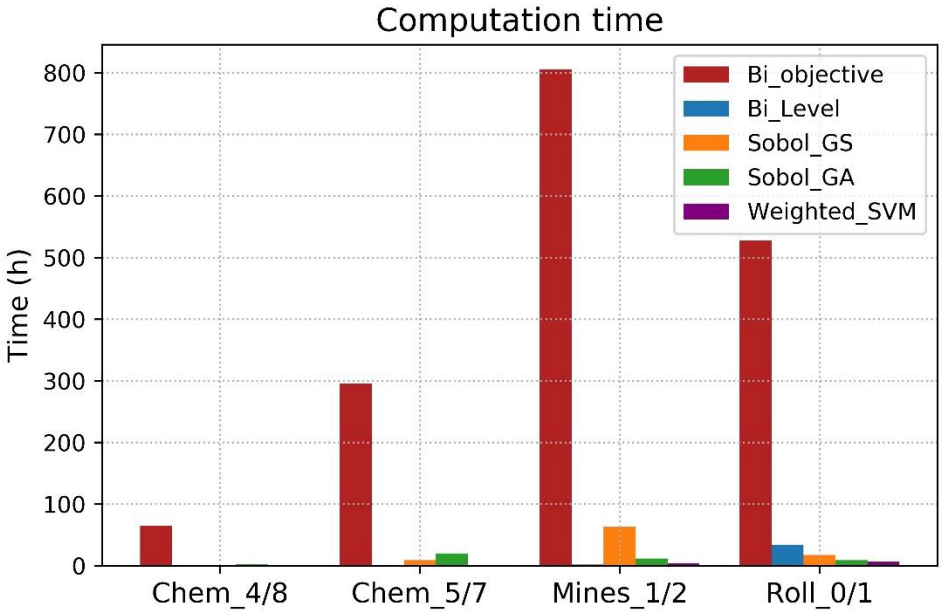


Figure 3.13 (a): Computation times of the proposed approaches

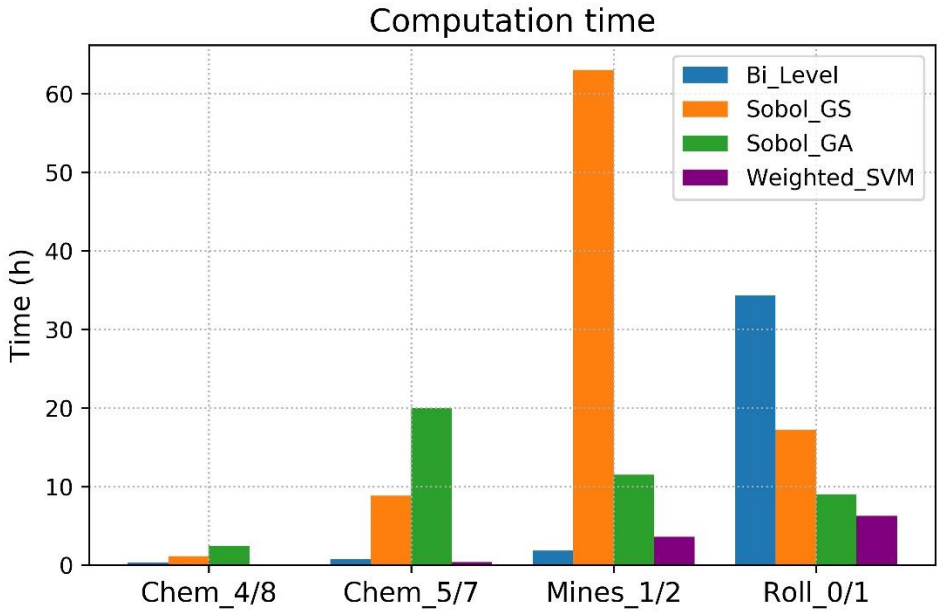


Figure 3.13 (b): Computation times (Bi-Objective results excluded)

The five approaches can be divided into two main groups. The first group of approaches consists of applying the same number of Monte-Carlo simulations for the identification of a robust solution. This is the case of the **Bi_Objective** and **Sobol_GS** approaches where the first one evaluates the robustness of all potential solutions of the genetic algorithm, and the second one evaluates 81 values of the parameter R, and thus performs 81 Monte-Carlo simulations. Fixing the number of Monte-Carlo simulations could be computationally exhaustive when dealing with large datasets, for example, the computation times related to the *Mines_1/2* dataset were 805 hours for the **Bi_Objective** approach and 63 hours for the **Sobol_GS** approach. On the other hand, the computation times for the second group of approaches (**Bi_Level, Sobol_GA** and **Weighted_SVM**) depend on the number of solutions identified in the first step. It can be noticed that these approaches generally require less computing time than the approaches of the first group. In particular, the **Weighted_SVM** method requires less time, because taking into account the feature-weights in the optimization results in the identification of very few SVM models with optimal prediction accuracy.

To conclude, the main objective of this chapter is to improve the robustness of a SVM model when data are considered with uncertainties. Five approaches have been proposed. The two first approaches are mathematically formulated as bi-objective optimization problems that aim to maximize the prediction accuracy of SVM as well as its robustness to the measurement uncertainties. Then, two other approaches are defined taking into account the Sobol indices for the definition of new feature-weights allowing the improvement of the robustness of SVM models. Finally, the last approach identifies an optimal solution by optimizing simultaneously the feature-weights and the SVM hyperparameters. Based on their applications to several datasets, these approaches show their ability and efficiency to improve the robustness of SVM models. In particular, the **Weighted_SVM** approach has effectively improved the prediction accuracy and robustness of SVM models.

To summarize, the main contributions of the third chapter are presented in Figure 3.13.

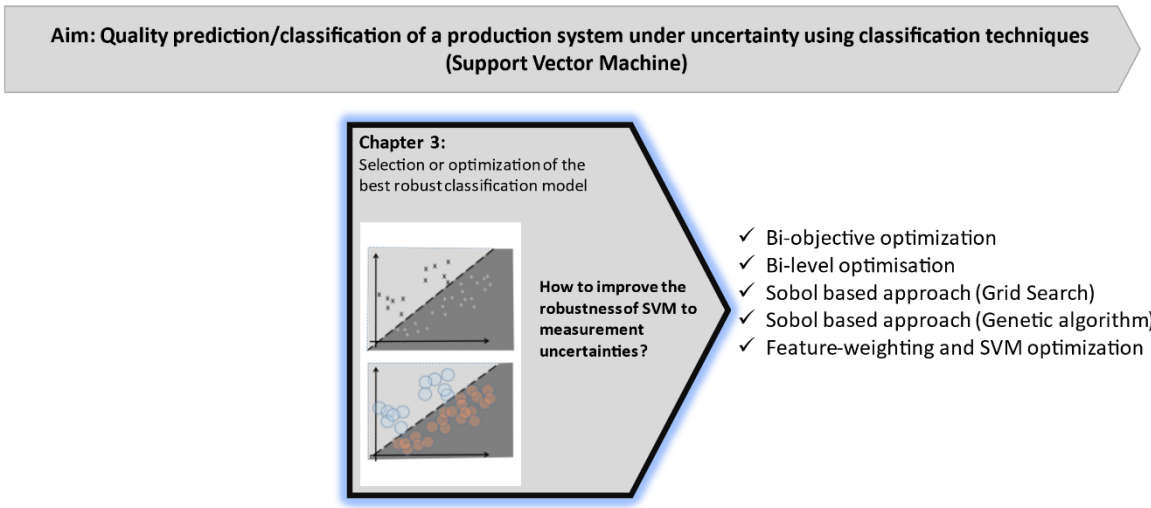


Figure 3.13: Main contributions of chapter III

Chapter IV

General conclusion and perspectives

This chapter is dedicated to the general conclusion of the presented research work and its perspectives. The conclusion highlights the different developments and proposed approaches that have been covered in the three previous chapters, as well as the limits of the research works. The perspectives of these works are proposed in order to propose new research horizons.

The aim of this research is to propose new robust approaches for the prediction and the classification of manufacturing systems quality under uncertainties. These approaches have been applied at the same time to academic and real-world industrial data in order to assess their efficiency.

IV.1 - Conclusion

As a reminder and as indicated in the introduction of the manuscript, the overall objective of this research work is formulated as follows

How to assess the impact of **measurement uncertainties** on the predictive performance of the SVM **classification method**, in order to improve the **manufacturing systems quality**?

Three research objectives have been then derived to meet the main objective. The structure of the conclusion follows theses aims:

- Assessment and improvement of the quality of manufacturing systems based on machine learning algorithms and classification approaches
- Quantification of the impact of measurement uncertainties on the performances of the SVM classification method
- Improvement of the robustness of SVM classification when considering measurement uncertainties

IV.1.1 - Assessment and improvement of the quality of manufacturing systems based on machine learning algorithms and classification approaches

In order to meet the first research objective, a literature review was first conducted. This review made it possible to emphasize the potential of machine learning approaches to improve the quality in many areas and, in particular, the ability of classification tools to improve and to assess the quality within manufacturing systems. Three main classifiers have been addressed: decision trees, support vector machines, and multilayer perceptron, where their applications in manufacturing systems have been discussed and can therefore be summarized as: the identification of the roots of failures, the identification of optimal manufacturing settings, the detection of faulty process conditions, the identification of defective products and the classification of types of defects.

In addition, to evaluate the performance of these classifiers, they have been applied to industrial data. These industrial case studies demonstrated the ability of SVM and MLP to predict different levels of quality, and the ability of coupling the C4.5 decision tree with parallel coordinates to identify the causes of defects and the optimal manufacturing settings in an interpretable manner.

Based on the findings of the literature review and the results of the case studies, SVM was selected as the classifier to be studied. The selection of SVM over MLP and DT is due to:

- good performances of the SVM to deal with non-linearly separable data,
- handling of large dimensional data sets with different features,
- a better robustness of the soft margin of the SVM to classification errors during the training phase compared to the hard margin of the DT,
- easy adjustment of the hyperparameters in comparison to the MLP,
- reduced number of operations in the training phase compared to MLP,
- better predictive performances compared to MLP (Zanaty, 2012).

Finally, in order to study the robustness of the SVM models to the impact of measurement uncertainties, a review on the proposed approaches has been conducted. This allowed identifying the different methodologies for improving the robustness of the SVM models.

Still, this first aim has some limits, as it deals only with structured data, especially with numerical attributes. Besides, when considering noisy data, only measurement uncertainties that affect the values of the various input parameters are considered in this study.

IV.1.2 - Quantification of the impact of measurement uncertainties on the performances of the SVM classification method

To reach the second research objective of this work, a first experiment has been carried out to quantify the impact of Gaussian measurement uncertainties on the predictive performance of SVM. This experiment is based on a Monte-Carlo simulation, and has shown that the perturbation of a data set with Gaussian uncertainties leads to a decrease in the accuracy of the SVM model and thus a decrease in the generalization of this method. Then, three approaches were proposed for quantifying the impact of the uncertainties of each input parameter on the prediction performance of an SVM model and thus the identification of the key measurement uncertainties. The first two approaches, based on Monte-Carlo simulation and Sobol sensitivity analysis, allow calculating for each parameter a coefficient representing the impact of its uncertainties on the robustness of the SVM. On the other hand, the third approach, which is based on statistical tools, allows the estimation of parameters' uncertainties that may influence the robustness of the SVM model. All three approaches were then applied to several industrial datasets. Some remarks can be made:

- The first two approaches allow a precise identification of the main measurement uncertainties, as well as the quantification of the impacts each parameter uncertainties. Thus, parameters with significant impacts should be carefully monitored, in order to obtain a robust quality prediction within manufacturing systems.
- The statistical approach estimates the parameters whose uncertainties have a significant impact on the robustness of the SVM. It can therefore be concluded that

parameters with a high correlation with the class parameter are more likely to have significant impacts on the robustness of the SVM.

- Monte Carlo and Sobol-based approaches allow a precise identification and quantification of the impacts of the key measurement uncertainties. However, these two approaches are time and resource consuming, which is not the case of the third approach.

The first two proposed approaches were only applied considering Gaussian measurement uncertainties. In addition, the statistical approach estimates the key measurement uncertainties by analyzing only the correlation between an input parameter and the class parameter. This could be extended and improved by considering the correlation between a parameter and the rest of the parameters.

IV.1.3 - Improvement of the robustness of SVM classification

The aim of this research is to improve the robustness of the SVM models when handling data subject to uncertainties. To meet this objective, three approaches have been developed. The first two approaches focus on the selection of SVM models with optimal predictive accuracy and optimal robustness to measurement uncertainties. These two approaches allow the selection of an optimal SVM model that allows robust quality prediction without the need for a modified SVM model. Approaches 3 and 4 are based on statistical knowledge gathered through the application of Sobol's sensitivity analysis to quantify the impact of the key measurement uncertainties. Weights of characteristics that are a function of Sobol's total effect indices and a parameter called R have therefore been defined, and models that incorporate the notion of characteristic weights are ultimately generated to be more robust to the impacts of measurement uncertainties. Finally, the last approach consists of assigning to each parameter a weight (varying between 0 and 1), where they are tuned at the same time with the SVM hyperparameters. This last approach has resulted in weighted SVM models with better predictive performance and better robustness to the impact of measurement uncertainties. The results of the application of the proposed approaches to various manufacturing datasets allowed concluding that the robustness of the SVM can be improved by:

- selection of SVM models according to two criteria, which are: maximum prediction accuracy of the SVM, and maximum robustness of the SVM to the impact of measurement uncertainties.
- definition of the feature weights to improve the SVM robustness and its prediction accuracy.

Globally, the issues of uncertainties encountered in quality data and their impact on SVM have been addressed in this work, where several approaches have been proposed. This research work provides a better understanding on the robustness and the sensitivity of SVM models when encountering data subject to measurement uncertainties, as it allows a precise

identification of the main parameters (called key measurement uncertainties) leading to quality problems. Therefore, monitoring the key measurement uncertainties, and managing the impact of measurement uncertainties would ultimately enable manufacturing industries to improve the quality of their systems and make more robust decisions.

IV.2 - Perspectives and future works

The presented research work aims to develop data driven approaches for manufacturing quality management. Various machine learning techniques have been reviewed then applied to different datasets. The SVM classification method has been considered and its prediction accuracy and robustness to the measurements uncertainties impacts have been analyzed and improved, respectively. According to the research objectives of this work, perspectives for further developments can be proposed.

Firstly, the genetic algorithm developed for the tuning of SVM hyperparameters can be improved by taking into account several criteria, i.e., instead of selecting SVM models by considering only the prediction accuracy of the model, more criteria such as model complexity, separability, robustness can be included. As a result, lighter models with optimal predictive performance can be selected to manage quality classification. Further analysis can also help to understand how these criteria influence each other, for example, the robustness of the SVM can be synonymous with model complexity.

In addition, for the case study entitled "Optical monitoring of the L-PBF process", it has been shown that the identified statistical features allow the prediction of the different density/quality classes. Another idea that can improve this approach is to add frequency domain features. It is believed that by adding these new features, more knowledge is gathered, and thus a better classification of quality can be obtained.

The second aspect of the research works concerns the quantification of the impacts of uncertainties on the prediction accuracy of a classification method. Different types of uncertainties could be added when dealing with the Monte-Carlo based approach or the Sobol based approach. That will allow the generalization of the approaches and the extension of their applicability domains. The third approach could benefit from the definition of a new correlation measure that includes both:

- Simple correlation of a parameter with the class parameter
- Multiple correlation of a parameter with the rest of the input parameters.

This could be promising, as many manufacturing parameters are not independent. Therefore, the definition of a new metric including both correlation coefficients can lead to a better estimation of the key measurement uncertainties.

The third aim of the research work enables to improve the performance of Sobol-based approaches by defining weights through the acquisition of more statistical knowledge. Further works should be dedicated on the definition and the selection of these weights. In fact,

models denoted weighted-SVM models in this work have demonstrated their good performance and their robustness when data are subject to uncertainties.

Even if the data considered in this work were initially considered as structured, which is generally not the case for real cases, a more in-depth analysis is necessary to allow the exploration of other types of noise related to structured or unstructured data, through the analysis and management of their impacts on the different machine learning and deep learning methods.

References

Abdiansah, A., Wardoyo, R., 2015. Time Complexity Analysis of Support Vector Machines (SVM) in LibSVM. IJCA 128, 28–34. https://doi.org/10.5120/ijca2015906480

Ahmed, F., Kim, K.-Y., 2017. Data-driven Weld Nugget Width Prediction with Decision Tree Algorithm. Procedia Manufacturing 10, 1009–1019. https://doi.org/10.1016/j.promfg.2017.07.092

Al-kharaz, M., Ananou, B., Ouladsine, M., Combal, M., Pinaton, J., 2019. Quality Prediction in Semiconductor Manufacturing processes Using Multilayer Perceptron Feedforward Artificial Neural Network *, in: 2019 8th International Conference on Systems and Control (ICSC). Presented at the 2019 8th International Conference on Systems and Control (ICSC), IEEE, Marrakesh, Morocco, pp. 423–428. https://doi.org/10.1109/ICSC47195.2019.8950664

Angelova, M., Pencheva, T., 2011. Tuning Genetic Algorithm Parameters to Improve Convergence Time. International Journal of Chemical Engineering 2011, 1–7. https://doi.org/10.1155/2011/646917

Arasu, B.S., Seelan, B.J.B., Thamaraiselvan, N., 2020. A machine learning-based approach to enhancing social media marketing. Computers & Electrical Engineering 86, 106723. https://doi.org/10.1016/j.compeleceng.2020.106723

Ay, M., Stemmler, S., Schwenzer, M., Abel, D., Bergs, T., 2019. Model Predictive Control in Milling based on Support Vector Machines. IFAC-PapersOnLine 52, 1797–1802. https://doi.org/10.1016/j.ifacol.2019.11.462

Aziz, S., Dowling, M.M., Hammami, H., Piepenbrink, A., 2019. Machine Learning in Finance: A Topic Modeling Approach. SSRN Journal. https://doi.org/10.2139/ssrn.3327277

Baccarini, L.M.R., Rocha e Silva, V.V., de Menezes, B.R., Caminhas, W.M., 2011. SVM practical industrial application for mechanical faults diagnostic. Expert Systems with Applications 38, 6980–6984. https://doi.org/10.1016/j.eswa.2010.12.017

Baker, B.M., Ayechew, M.A., 2003. A genetic algorithm for the vehicle routing problem. Computers & Operations Research 30, 787–800. https://doi.org/10.1016/S0305-0548(02)00051-5

Bakır, B., Batmaz, İ., Güntürkün, F.A., İpekçi, İ.A., Köksal, G. and Özdemirel, N.E., 2006. Defect cause modeling with decision tree and regression analysis. World Acad Sci Eng Technoly, 24, pp.1-4.

Bi, J. and Zhang, T., 2005. Support vector classification with input data uncertainty. In Advances in neural information processing systems (pp. 161-168).

Casalino, G., Facchini, F., Mortello, M., Mummolo, G., 2016. ANN modelling to optimize manufacturing processes: the case of laser welding. IFAC-PapersOnLine 49, 378–383. https://doi.org/10.1016/j.ifacol.2016.07.634

Cavalcante, I.M., Frazzon, E.M., Forcellini, F.A., Ivanov, D., 2019. A supervised machine learning approach to data-driven simulation of resilient supplier selection in digital manufacturing. International Journal of Information Management 49, 86–97. https://doi.org/10.1016/j.ijinfomgt.2019.03.004

Chen, R.S. and Wu, R.C., 2006, July. Using data mining technology to design an intelligent quality analysis control system for semiconductor packaging industry. In Proceedings of the 10th WSEAS international conference on Computers (pp. 185-191). World Scientific and Engineering Academy and Society (WSEAS).

Chen, Y.-S., Cheng, C.-H., Lai, C.-J., 2012. Extracting performance rules of suppliers in the manufacturing industry: an empirical study. J Intell Manuf 23, 2037–2045. https://doi.org/10.1007/s10845-011-0530-8

Choi, M.-K., Lee, H.-G., Lee, S.-C., 2016. Weighted SVM with classification uncertainty for small training samples, in: 2016 IEEE International Conference on Image Processing (ICIP). Presented at the 2016 IEEE International Conference on Image Processing (ICIP), IEEE, Phoenix, AZ, USA, pp. 4438–4442. https://doi.org/10.1109/ICIP.2016.7533199

Choi, Sungsu, Battulga, Lkhagvadorj, Nasridinov, Aziz, Yoo, Kwan-Hee, 2017. A Decision Tree Approach for Identifying Defective Products in the Manufacturing Process. International Journal of Contents 13, 57–65. https://doi.org/10.5392/IJOC.2017.13.2.057

Çiflikli, C., Kahya-Özyirmidokuz, E., 2010. Implementing a data mining solution for enhancing carpet manufacturing productivity. Knowledge-Based Systems 23, 783–788. https://doi.org/10.1016/j.knosys.2010.05.001

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach Learn 20, 273–297. https://doi.org/10.1007/BF00994018

Delgado, M., Cirrincione, G., García, A., Ortega, J.A. and Henao, H., 2012, October. Accurate bearing faults classification based on statistical-time features, curvilinear component analysis and neural networks. In IECON 2012-38th Annual Conference on IEEE Industrial Electronics Society (pp. 3854-3861). IEEE.

Deng, C., Ji, X., Rainey, C., Zhang, J., Lu, W., 2020. Integrating Machine Learning with Human Knowledge. iScience 101656. https://doi.org/10.1016/j.isci.2020.101656

Deo, R.C., 2015. Machine learning in medicine. Circulation, 132(20), pp.1920-1930.

Diao, G., Zhao, L., Yao, Y., 2015. A dynamic quality control approach by improving dominant factors based on improved principal component analysis. International Journal of Production Research 53, 4287–4303. https://doi.org/10.1080/00207543.2014.997400

Escobar, C.A., Morales-Menendez, R., 2019. Process-Monitoring-for-Quality — A Model Selection Criterion for Support Vector Machine. Procedia Manufacturing 34, 1010–1017. https://doi.org/10.1016/j.promfg.2019.06.094

Fan, N., Sadeghi, E., Pardalos, P.M., 2014. Robust Support Vector Machines with Polyhedral Uncertainty of the Input Data, in: Pardalos, P.M., Resende, M.G.C., Vogiatzis, C., Walteros, J.L. (Eds.), Learning and Intelligent Optimization, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 291–305. https://doi.org/10.1007/978-3-319-09584-4_26

Fang, Y., Li, J., 2010. A Review of Tournament Selection in Genetic Programming, in: Cai, Z., Hu, C., Kang, Z., Liu, Y. (Eds.), Advances in Computation and Intelligence. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 181–192. https://doi.org/10.1007/978-3-642-16493-4_19

Fernández-Francos, D., Martínez-Rego, D., Fontenla-Romero, O., Alonso-Betanzos, A., 2013. Automatic bearing fault diagnosis based on one-class v-SVM. Computers & Industrial Engineering 64, 357–365. https://doi.org/10.1016/j.cie.2012.10.013

Frenay, B., Verleysen, M., 2014. Classification in the Presence of Label Noise: A Survey. IEEE Trans. Neural Netw. Learning Syst. 25, 845–869. https://doi.org/10.1109/TNNLS.2013.2292894

El Ghaoui, L., Lanckriet, G.R.G. and Natsoulis, G., 2003. Robust classification with interval data.

Ghassemi, M., Naumann, T., Schulam, P., Beam, A.L., Chen, I.Y. and Ranganath, R., 2020. A Review of Challenges and Opportunities in Machine Learning for Health. AMIA Summits on Translational Science Proceedings, 2020, p.191.

Ghate, V.N., Dudul, S.V., 2011. Cascade Neural-Network-Based Fault Classifier for Three-Phase Induction Motor. IEEE Trans. Ind. Electron. 58, 1555–1563. https://doi.org/10.1109/TIE.2010.2053337

Glen, G., Isaacs, K., 2012. Estimating Sobol sensitivity indices using correlations. Environmental Modelling & Software 37, 157–166. https://doi.org/10.1016/j.envsoft.2012.03.014

Grajski, K.A., Breiman, L., Di Prisco, G.V., Freeman, W.J., 1986. Classification of EEG Spatial Patterns with a Tree-Structured Methodology: CART. IEEE Trans. Biomed. Eng. BME-33, 1076–1086. https://doi.org/10.1109/TBME.1986.325684

Gryllias, K.C., Antoniadis, I.A., 2012. A Support Vector Machine approach based on physical model training for rolling element bearing fault detection in industrial environments. Engineering Applications of Artificial Intelligence 25, 326–344. https://doi.org/10.1016/j.engappai.2011.09.010

Hansson, K., Yella, S., Dougherty, M. and Fleyeh, H., 2016. Machine learning algorithms in heavy process manufacturing. American Journal of Intelligent Systems, 6(1), pp.1-13. https://doi.org/10.5923/j.ajis.20160601.01

Heo, G., Gader, P., 2009. Fuzzy SVM for noisy data: A robust membership calculation method, in: 2009 IEEE International Conference on Fuzzy Systems. Presented at the 2009 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, Jeju Island, South Korea, pp. 431–436. https://doi.org/10.1109/FUZZY.2009.5277191

Hsueh, Y.-W., Yang, C.-Y., 2008. Prediction of tool breakage in face milling using support vector machine. Int J Adv Manuf Technol 37, 872–880. https://doi.org/10.1007/s00170-007-1034-8

Huber, J., Stuckenschmidt, H., 2020. Daily retail demand forecasting using machine learning with emphasis on calendric special days. International Journal of Forecasting S0169207020300224. https://doi.org/10.1016/j.ijforecast.2020.02.005

Hunter, A. (Ed.), 1998. Applications of uncertainty formalisms, Lecture notes in computer science Lecture notes in artificial intelligence. Springer, Berlin.

Jacques, J., 2011. Pratique de l'analyse de sensibilité: comment évaluer l'impact des entrées aléatoires sur la sortie d'un modèle mathématique. Lille: sn.

Jacques, J., Lavergne, C., Devictor, N., 2006. Sensitivity analysis in presence of model uncertainty and correlated inputs. Reliability Engineering & System Safety 91, 1126–1134. https://doi.org/10.1016/j.ress.2005.11.047

Jedliński, Ł., Jonak, J., 2015. Early fault detection in gearboxes based on support vector machines and multilayer perceptron with a continuous wavelet transform. Applied Soft Computing 30, 636–641. https://doi.org/10.1016/j.asoc.2015.02.015

Jegadeeshwaran, R., Sugumaran, V., 2015. Fault diagnosis of automobile hydraulic brake system using statistical features and support vector machines. Mechanical Systems and Signal Processing 52–53, 436–446. https://doi.org/10.1016/j.ymssp.2014.08.007

Jeyakumar, V., Li, G., Suthaharan, S., 2014. Support vector machine classifiers with uncertain knowledge sets via robust optimization. Optimization 63, 1099–1116. https://doi.org/10.1080/02331934.2012.703667

Johnson, W.G., 2018. Data Mining and Machine Learning in Education with Focus in Undergraduate CS Student Success, in: Proceedings of the 2018 ACM Conference on International Computing Education Research. Presented at the ICER '18: International Computing Education Research Conference, ACM, Espoo Finland, pp. 270–271. https://doi.org/10.1145/3230977.3231012

Khan, A.I., Al-Habsi, S., 2020. Machine Learning in Computer Vision. Procedia Computer Science 167, 1444–1451. https://doi.org/10.1016/j.procs.2020.03.355

Kim, A., Oh, K., Jung, J.-Y., Kim, B., 2018. Imbalanced classification of manufacturing quality conditions using cost-sensitive decision tree ensembles. International Journal of Computer Integrated Manufacturing 31, 701–717. https://doi.org/10.1080/0951192X.2017.1407447

Köksal, G., Batmaz, İ., Testik, M.C., 2011. A review of data mining applications for quality improvement in manufacturing industry. Expert Systems with Applications 38, 13448–13467. https://doi.org/10.1016/j.eswa.2011.04.063

Kristensen, M.H., Petersen, S., 2016. Choosing the appropriate sensitivity analysis method for building energy model-based investigations. Energy and Buildings 130, 166–176. https://doi.org/10.1016/j.enbuild.2016.08.038

Lade, P., Ghosh, R., Srinivasan, S., 2017. Manufacturing Analytics and Industrial Internet of Things. IEEE Intell. Syst. 32, 74–79. https://doi.org/10.1109/MIS.2017.49

Le Thi, H.A., Vo, X.T., Pham Dinh, T., 2014. Feature selection for linear SVMs under uncertain data: Robust optimization based on difference of convex functions algorithms. Neural Networks 59, 36–50. https://doi.org/10.1016/j.neunet.2014.06.011

Lee, J., Lapira, E., Yang, S., Kao, A., 2013. Predictive Manufacturing System - Trends of Next-Generation Production Systems. IFAC Proceedings Volumes 46, 150–156. https://doi.org/10.3182/20130522-3-BR-4036.00107

Li, K., Wang, L., Wu, J., Zhang, Q., Liao, G., Su, L., 2018. Using GA-SVM for defect inspection of flip chips based on vibration signals. Microelectronics Reliability 81, 159–166. https://doi.org/10.1016/j.microrel.2017.12.032

Libin, P., Moonens, A., Verstraeten, T., Perez-Sanjines, F., Hens, N., Lemey, P. and Nowé, A., 2020. Deep reinforcement learning for large-scale epidemic control. arXiv preprint arXiv:2003.13676.

Lin, S.-W., Ying, K.-C., Chen, S.-C., Lee, Z.-J., 2008. Particle swarm optimization for parameter determination and feature selection of support vector machines. Expert Systems with Applications 35, 1817–1824. https://doi.org/10.1016/j.eswa.2007.08.088

Liu, Q., Li, X., Liu, H., Guo, Z., 2020. Multi-objective metaheuristics for discrete optimization problems: A review of the state-of-the-art. Applied Soft Computing 93, 106382. https://doi.org/10.1016/j.asoc.2020.106382

Mantovani, R.G., Rossi, A.L.D., Alcobaça, E., Vanschoren, J., de Carvalho, A.C.P.L.F., 2019. A meta-learning recommender system for hyperparameter tuning: Predicting when tuning improves SVM classifiers. Information Sciences 501, 193–221. https://doi.org/10.1016/j.ins.2019.06.005

Metropolis, N., Ulam, S., 1949. The Monte Carlo Method. Journal of the American Statistical Association 44, 335–341. https://doi.org/10.1080/01621459.1949.10483310

Mirapeix, J., García-Allende, P.B., Cobo, A., Conde, O.M., López-Higuera, J.M., 2007. Real-time arc-welding defect detection and classification with principal component analysis and artificial neural networks. NDT & E International 40, 315–323. https://doi.org/10.1016/j.ndteint.2006.12.001

Mittal, S., Khan, M.A., Romero, D., Wuest, T., 2018. A critical review of smart manufacturing & Industry 4.0 maturity models: Implications for small and medium-sized enterprises (SMEs). Journal of Manufacturing Systems 49, 194–214. https://doi.org/10.1016/j.jmsy.2018.10.005

Mohammadi, M., Siadat, A., Dantan, J.-Y., Tavakkoli-Moghaddam, R., 2015. Mathematical modelling of a robust inspection process plan: Taguchi and Monte Carlo methods. International Journal of Production Research 53, 2202–2224. https://doi.org/10.1080/00207543.2014.980460

Niaf, E., Flamary, R., Lartizien, C., Canu, S., 2011. Handling uncertainties in SVM classification, in: 2011 IEEE Statistical Signal Processing Workshop (SSP). Presented at the 2011 IEEE Statistical Signal Processing Workshop (SSP), IEEE, Nice, France, pp. 757–760. https://doi.org/10.1109/SSP.2011.5967814

Nian, R., Liu, J., Huang, B., 2020. A review On reinforcement learning: Introduction and applications in industrial process control. Computers & Chemical Engineering 139, 106886. https://doi.org/10.1016/j.compchemeng.2020.106886

Noyel, M., Thomas, P., Charpentier, P., Thomas, A., Beauprêtre, B., 2013. Improving production process performance thanks to neuronal analysis. IFAC Proceedings Volumes 46, 432–437. https://doi.org/10.3182/20130522-3-BR-4036.00055

Olatomiwa, L., Mekhilef, S., Shamshirband, S., Mohammadi, K., Petković, D., Sudheer, C., 2015. A support vector machine–firefly algorithm-based model for global solar radiation prediction. Solar Energy 115, 632–644. https://doi.org/10.1016/j.solener.2015.03.015

Özdemir, Ö.; Çavuş, M. 2016. "Performance of the Inverse Transformation Method for Extreme Value Distributions." Xth International Statistics Days Conference (ISDC'2016), Giresun, Turkey. 8.

Pant, R., Trafalis, T.B. and Barker, K., 2011, July. Support vector machine classification of uncertain and imbalanced data using robust optimization. In Proceedings of the 15th WSEAS international conference on computers (pp. 369-374). World Scientific and Engineering Academy and Society (WSEAS) Stevens Point, Wisconsin, USA.

Quinlan, J.R., 1996. Improved Use of Continuous Attributes in C4.5. jair 4, 77–90. https://doi.org/10.1613/jair.279

Quinlan, J.R., 1986. Induction of decision trees. Mach Learn 1, 81–106. https://doi.org/10.1007/BF00116251

Ramchoun, H., Amine, M., Idrissi, J., Ghanou, Y., Ettaouil, M., 2016. Multilayer Perceptron: Architecture Optimization and Training. IJIMAI 4, 26. https://doi.org/10.9781/ijimai.2016.415

Roberts, R.C., Laramee, R.S., Smith, G.A., Brookes, P., D'Cruze, T., 2019. Smart Brushing for Parallel Coordinates. IEEE Trans. Visual. Comput. Graphics 25, 1575–1590. https://doi.org/10.1109/TVCG.2018.2808969

Rokach, L., Maimon, O., 2006. Data Mining for Improving the Quality of Manufacturing: A Feature Set Decomposition Approach. J Intell Manuf 17, 285–299. https://doi.org/10.1007/s10845-005-0005-x

Ronowicz, J., Thommes, M., Kleinebudde, P., Krysiński, J., 2015. A data mining approach to optimize pellets manufacturing process based on a decision tree algorithm. European Journal of Pharmaceutical Sciences 73, 44–48. https://doi.org/10.1016/j.ejps.2015.03.013

Rostami, H., Dantan, J.-Y., Homri, L., 2015. Review of data mining applications for quality assessment in manufacturing industry: support vector machines. Int. J. Metrol. Qual. Eng. 6, 401. https://doi.org/10.1051/ijmqe/2015023

Salcedo-Sanz, S., Cornejo-Bueno, L., Prieto, L., Paredes, D., García-Herrera, R., 2018. Feature selection in machine learning prediction systems for renewable energy applications. Renewable and Sustainable Energy Reviews 90, 728–741. https://doi.org/10.1016/j.rser.2018.04.008

Scher, S., Messori, G., 2018. Predicting weather forecast uncertainty with machine learning. Q.J.R. Meteorol. Soc. 144, 2830–2841. https://doi.org/10.1002/qj.3410

Serra, A., Galdi, P., Tagliaferri, R., 2018. Machine learning for bioinformatics and neuroimaging. WIREs Data Mining Knowl Discov 8. https://doi.org/10.1002/widm.1248

Sharma, R., Kamble, S.S., Gunasekaran, A., Kumar, V., Kumar, A., 2020. A systematic literature review on machine learning applications for sustainable agriculture supply chain performance. Computers & Operations Research 119, 104926. https://doi.org/10.1016/j.cor.2020.104926

Sharp, M., Ak, R., Hedberg, T., 2018. A survey of the advancing use and development of machine learning in smart manufacturing. Journal of Manufacturing Systems 48, 170–179. https://doi.org/10.1016/j.jmsy.2018.02.004

Shawe-Taylor, J., Bartlett, P.L., Williamson, R.C., Anthony, M., 1998. Structural risk minimization over data-dependent hierarchies. IEEE Trans. Inform. Theory 44, 1926–1940. https://doi.org/10.1109/18.705570

Shen, C., Wang, L., Li, Q., 2007. Optimization of injection molding process parameters using combination of artificial neural network and genetic algorithm method. Journal of Materials Processing Technology 183, 412–418. https://doi.org/10.1016/j.jmatprotec.2006.10.036

Siltepavet, A., Sinthupinyo, S. and Chongstitvatana, P., 2012. Improving quality of products in hard drive manufacturing by decision tree technique. International Journal of Computer Science Issues (IJCSI), 9(3), p.29.

Slotwinski, J.A., Garboczi, E.J., Hebenstreit, K.M., 2014. Porosity Measurements and Analysis for Metal Additive Manufacturing Process Control. J. RES. NATL. INST. STAN. 119, 494. https://doi.org/10.6028/jres.119.019

Smithson, M., 1989. Ignorance and uncertainty: Emerging paradigms.

Spierings, A.B., Schneider, M., Eggenberger, R., 2011. Comparison of density measurement techniques for additive manufactured metallic parts. Rapid Prototyping Journal 17, 380–386. https://doi.org/10.1108/13552541111156504

Sun, J., 2010. Application of Data Mining for Decision Tree Model of Multi-variety Discrete Production and Manufacture, in: 2010 Third International Symposium on Intelligent Information Technology and Security Informatics. Presented at the 2010 Third International Symposium on Intelligent Information Technology and Security Informatics (IITSI), IEEE, Jian, China, pp. 724–728. https://doi.org/10.1109/IITSI.2010.13

Taetragool, U., Achalakul, T., 2009. Applying Decision Tree in Fault Pattern Analysis for HGA Manufacturing, in: 2009 International Conference on Complex, Intelligent and Software Intensive Systems. Presented at the 2009 International Conference on Complex, Intelligent and Software Intensive Systems (CISIS), IEEE, Fukuoka, Japan, pp. 83–89. https://doi.org/10.1109/CISIS.2009.139

Tao, F., Qi, Q., Liu, A., Kusiak, A., 2018. Data-driven smart manufacturing. Journal of Manufacturing Systems 48, 157–169. https://doi.org/10.1016/j.jmsy.2018.01.006

Tsironis, L., Bilalis, N., Moustakis, V., 2005. Using machine learning to support quality management: Framework and experimental investigation. The TQM Magazine 17, 237–248. https://doi.org/10.1108/09544780510594207

Unal, M., Onat, M., Demetgul, M., Kucuk, H., 2014. Fault diagnosis of rolling bearings using a genetic algorithm optimized neural network. Measurement 58, 187–196. https://doi.org/10.1016/j.measurement.2014.08.041

Utkin, L.V., Zhuk, Y.A., 2017. Interval SVM-Based Classification Algorithm Using the Uncertainty Trick. Int. J. Artif. Intell. Tools 26, 1750014. https://doi.org/10.1142/S0218213017500142

Walchand College of Engineering, A.J., U., P.D., S., Government College of Engineering, Karad, 2015. CROSSOVER OPERATORS IN GENETIC ALGORITHMS: A REVIEW. IJSC 06, 1083–1092. https://doi.org/10.21917/ijsc.2015.0150

Wei, Z., Feng, Y., Hong, Z., Qu, R., Tan, J., 2017. Product quality improvement method in manufacturing process based on kernel optimisation algorithm. International Journal of Production Research 55, 5597–5608. https://doi.org/10.1080/00207543.2017.1324223

Wickramasinghe, R.I.P., 2017. Attribute Noise, Classification Technique, and Classification Accuracy, in: Palomares Carrascosa, I., Kalutarage, H.K., Huang, Y. (Eds.), Data Analytics and Decision Support for Cybersecurity, Data Analytics. Springer International Publishing, Cham, pp. 201–220. https://doi.org/10.1007/978-3-319-59439-2_7

Wu, Q., Law, R., 2011. The complex fuzzy system forecasting model based on fuzzy SVM with triangular fuzzy number input and output. Expert Systems with Applications 38, 12085–12093. https://doi.org/10.1016/j.eswa.2011.02.094

Wuest, T., Weimer, D., Irgens, C., Thoben, K.-D., 2016. Machine learning in manufacturing: advantages, challenges, and applications. Production & Manufacturing Research 4, 23–45. https://doi.org/10.1080/21693277.2016.1192517

Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H., Gao, M., Hou, H., Wang, C., 2018. Machine Learning and Deep Learning Methods for Cybersecurity. IEEE Access 6, 35365–35381. https://doi.org/10.1109/ACCESS.2018.2836950

Yang, D., Liu, Y., Li, S., Li, X., Ma, L., 2015. Gear fault diagnosis based on support vector machine optimized by artificial bee colony algorithm. Mechanism and Machine Theory 90, 219–229. https://doi.org/10.1016/j.mechmachtheory.2015.03.013

Yang, X., Tan, L., He, L., 2014. A robust least squares support vector machine for regression and classification with noise. Neurocomputing 140, 41–52. https://doi.org/10.1016/j.neucom.2014.03.037

Yin, H., Dong, H., 2011. The problem of noise in classification: Past, current and future work, in: 2011 IEEE 3rd International Conference on Communication Software and Networks. Presented at the 2011 IEEE 3rd International Conference on Communication Software and Networks (ICCSN), IEEE, Xi'an, China, pp. 412–416. https://doi.org/10.1109/ICCSN.2011.6014597

Zanaty, E.A., 2012. Support Vector Machines (SVMs) versus Multilayer Perception (MLP) in data classification. Egyptian Informatics Journal 13, 177–183. https://doi.org/10.1016/j.eij.2012.08.002

Zhang, Z., Ming, W., Zhang, G., Huang, Y., Wen, X., Huang, H., 2015. A new method for on-line monitoring discharge pulse in WEDM-MS process. Int J Adv Manuf Technol 81, 1403–1418. https://doi.org/10.1007/s00170-015-7261-5

Zhi-qiang, J., Hang-guang, F., Ling-jun, L., 2005. Support Vector Machine for mechanical faults classification. J. Zheijang Univ.-Sci. A 6, 433–439. https://doi.org/10.1631/jzus.2005.A0433

Zhu, X., Wu, X., Chen, Q., 2006. Bridging Local and Global Data Cleansing: Identifying Class Noise in Large, Distributed Data Datasets. Data Min Knowl Disc 12, 275–308. https://doi.org/10.1007/s10618-005-0012-8

Ziani, R., Felkaoui, A., Zegadi, R., 2017. Bearing fault diagnosis using multiclass support vector machines with binary particle swarm optimization and regularized Fisher's criterion. J Intell Manuf 28, 405–417. https://doi.org/10.1007/s10845-014-0987-3

Zou, X., Zhao, X., Li, G., Li, Z., Sun, T., 2017. Sensitivity analysis using a variance-based method for a three-axis diamond turning machine. Int J Adv Manuf Technol 92, 4429–4443. https://doi.org/10.1007/s00170-017-0394-y

Zouhri, W., Rostami, H., Homri, L., Dantan, J.-Y., 2020. A Genetic-Based SVM Approach for Quality Data Classification, in: Masrour, T., Cherrafi, A., El Hassani, I. (Eds.), Artificial Intelligence and Industrial Applications, Advances in Intelligent Systems and Computing. Springer International Publishing, Cham, pp. 15–31. https://doi.org/10.1007/978-3-030-51186-9_2

Appendices

The appendix section contains:

- Appendix A: includes more details about the genetic algorithm (Algorithm 1) developed in the first chapter.
- Appendix B: includes an explanation of the Sobol sensitivity analysis method.
- Appendix C: includes the weights related to the application of the **Weighted_SVM** approach to the *Roll_0/1* dataset.

Appendix A – Genetic algorithm for SVM optimization

Appendix A describes the design of the proposed algorithm to optimize the SVM model for quality data classification.

In the following, the different parameters and steps of the proposed genetic-based SVM approach are presented. The genetic representation and the initialization of the first population are discussed at first. After that, the fitness function is described, then, the selection, the crossover, and the mutation techniques used in the approach are presented.

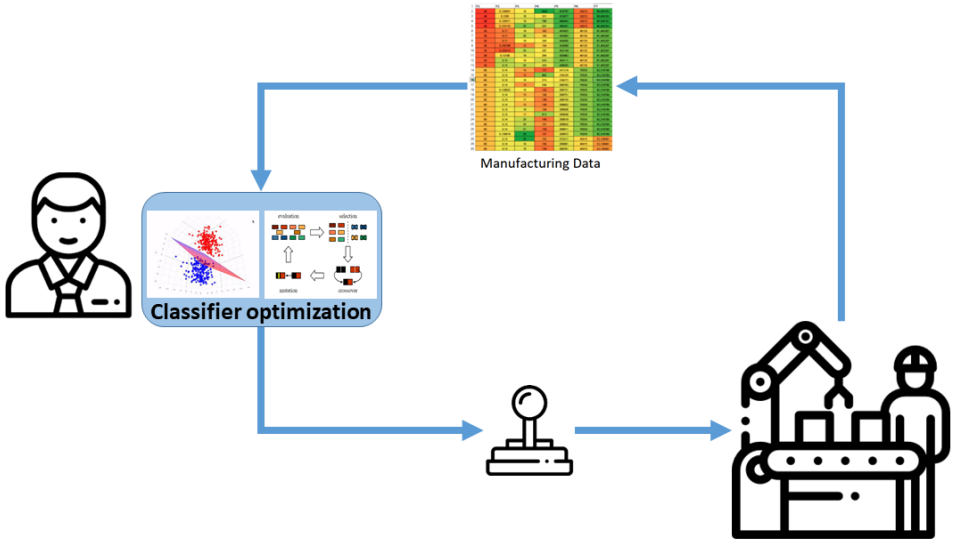


Figure A.1: GA-SVM approach for Quality data classification- a graphical description

A.1 - Genetic representation and initialization

One of the most critical decisions that significantly affects the performance of a genetic algorithm while implementing it, is deciding the representation to use to represent the solutions. It has been observed that unsuitable genetic representation can lead to poor performance of a GA, hence, choosing a proper genetic representation is necessary for the success of a GA. The representation of the solution must be complete and must contain the information needed to represent a solution to the problem, otherwise, the search will be either larger or poorer than necessary.

Consequently, in order to define a proper genetic representation, the different parameters to optimize need to be defined. Table A.1 represents the different SVM kernel functions and their different parameters that need to be optimized.

Table A.1: Kernel functions and their parameters

Kernel	Expression	Parameters to tune	Range of variation
Polynomial	$(\alpha x^T y + r)^d$	Penalty C Constant r Slope α Degree d	[1, 1000] [-49, 50] [1, 50] [1, 10]
Sigmoid	$tanh(\alpha x^T y + r)$	Penalty C Constant r Slope α	[1, 1000] [-49, 50] [1, 50]
RBF	$e^{\left(-\frac{\|\|x-y\|\|^2}{2\sigma^2}\right)}$	Penalty C Parameter σ	[1, 1000] [0.001, 10]

Since different kernel functions have different parameters, the representation of the solution varies for each kernel. In this thesis, an integer representation is used, where each parameter is represented by a set of genes depending on its range of variation. Consequently, the representation of each kernel functions, and the links that map the genotypes to the phenotypes, are shown in the next figure:

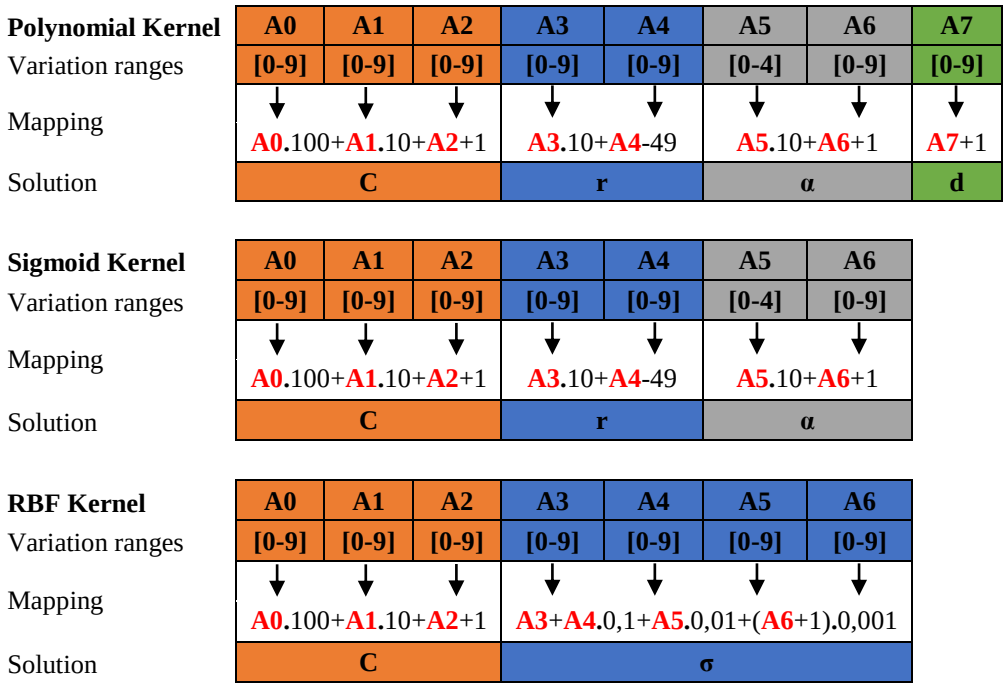


Figure A.2: Genetic representations of Kernels functions

Once the three representation are defined, a first population needs to be initialized. The first population should be diverse to avoid premature convergence, at the same time, its size

should not be too large as it can slow down the algorithm. Many methods can be found in the literature to initialize the population in a GA, including random generation, structured generation, and combination of random and structured generation. The random generation was applied in this work, as it has been noticed that a randomly generated population leads to optimality due to its diversity (Baker and Ayechew, 2003).

A.2 - GA-SVM parameters

- **Fitness evaluation**

Through any GA, it is necessary to be able to evaluate how good a potential solution is, compared to other solutions. To do so, a fitness function is defined. The fitness values are then used for the selection of individuals (parents) on which crossover and mutation operations will be applied.

For the proposed approach, to get the fitness value of a potential solution, the following steps should be followed:

1. Split the dataset into training (train), and test (test) sets.
2. Train the SVM model using one potential solution (chromosome).
3. Get the SVM accuracy on the test set, which represents the fitness value of the potential solution.

- **Selection operation**

Selection operation in a genetic algorithm consists of choosing individuals (parents) for reproduction. Parent selection is very important to the convergence rate of the genetic algorithm (GA), as good parents generate and yield better and fitter solutions.

In this GA, a 3-way tournament selection is used. This selection method consists of selecting randomly 3 individuals from the population, and then the one that guarantees the best accuracy on the validation set is selected to become a parent. The choice of this selection method is due to its simplicity, efficiency in parallel and non-parallel architecture, and also to the non-necessity to sort the population (Fang and Li, 2010).

- **Crossover operator**

By this operator, individuals (parents) share information by crossing their genotypes to create better individuals (children). Three different types of crossover are mainly used in the literature: one-point crossover, two-point crossover, and uniform crossover (Walchand College of Engineering et al., 2015).

The uniform crossover was considered in this work. In a uniform crossover, each gene is treated separately. In this, a coin is flipped for each gene to decide whether or not it'll be included in the off-spring. This operator does not require the definition of the crossover rate

parameter, as it avoids the genes interchanging of two different parameters (e.g. C genes with σ genes). Figure A.3 depicts the concept of the uniform crossover.

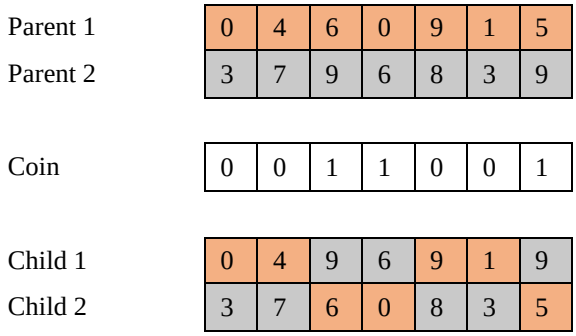


Figure A.3: Uniform crossover example

- **Mutation operator**

Mutation may be defined as a small random modification in the chromosome to get a new solution. It is used to introduce diversity in the genetic population and to release from local minima (Angelova and Pencheva, 2011). To that end, a random resetting mutation is used, where a random value from the set of permissible values is assigned to randomly chosen genes.

Appendix B - Sobol sensitivity analysis

Sobol's analysis is one the different methods to analyse the global sensitivity of a model. It is based on variance decomposition techniques to provide a quantitative measure of the contributions of the input to the output variance. The decomposition of the output variance in a Sobol sensitivity analysis employs the same principal as the classical analysis of variance in a factorial design, which allows representing the output's variance as follow:

$$V = \sum_{i=1}^{p} V_i + \sum_{1 \le i \le j \le p}^{p} V_{ij} + \cdots + V_{1\ldots p} \qquad (B.1)$$

Where:

V : Total variance of the model output.

V_i : The first order contribution of the i^{th}model parameter.

V_{ij}: The contribution of the interaction of the i^{th} and j^{th}parameters.

Based on this decomposition, Sobol defines first order sensitivity indices, Eq. B.2, as well as higher-order sensitivity indices Eq. B.3, such as:

$$S_i = \frac{V_i}{V} \qquad (B.2)$$

$$S_{ij} = \frac{V_{ij}}{V} \; ; \quad S_{ijk} = \frac{V_{ijk}}{V} \qquad (B.3)$$

Also, to measure the total sensitivity of the variance Y due to a variable Xi, Homma and Saltelli (Zou et al. 2017) introduced Total-effect indices Eq. B.4 defined as the sum of the contribution caused by Xi and by its interactions -of any order- with any other input variables.

$$S_{T_i} = \sum_{k \# i} S_k \qquad (B.4)$$

Where #i represents all sets containing the index i.

One of the ways to estimate Sobol indices is by using the Monte-Carlo simulation. The Monte-Carlo estimation consists on estimating: the output expected value (E[Y]), the output variance (V[Y]), and both quantities ($\hat{U}_i$) and ($\hat{U}_{\sim i}$). The estimation of these four quantities requires two N-samples $\tilde{X}_{(N)}^{(1)} = (x_{k1}^{(1)}, \ldots, x_{kp}^{(1)})_{k=1..N}$ and $\tilde{X}_{(N)}^{(2)} = (x_{k1}^{(2)}, \ldots, x_{kp}^{(2)})_{k=1..N}$, using the following formulas:

$$\hat{E} = E[Y] = \frac{1}{N} \sum_{k=1}^{N} f(x_{k1}, \ldots, x_{kp}) \qquad (B.5)$$

$$\hat{V} = V[Y] = \frac{1}{N}\sum_{k=1}^{N} f^2(x_{k1}, \dots, x_{kp}) \; - \; \hat{\mathrm{E}}^2 \tag{B.6}$$

$$\hat{\mathrm{U}}_i = \frac{1}{N}\sum_{k=1}^{N} f(x_{k1}^{(1)}, \dots, x_{k(i-1)}^{(1)}, x_{ki}^{(1)}, x_{k(i+1)}^{(1)}, \dots, x_{kp}^{(1)}).f(x_{k1}^{(2)}, \dots, x_{k(i-1)}^{(2)}, x_{ki}^{(1)}, x_{k(i+1)}^{(2)}, \dots, x_{kp}^{(2)}) \tag{B.7}$$

$$\hat{\mathrm{U}}_{\sim i} = \frac{1}{N}\sum_{k=1}^{N} f(x_{k1}^{(1)}, \dots, x_{k(i-1)}^{(1)}, x_{ki}^{(1)}, x_{k(i+1)}^{(1)}, \dots, x_{kp}^{(1)}).f(x_{k1}^{(1)}, \dots, x_{k(i-1)}^{(1)}, x_{ki}^{(2)}, x_{k(i+1)}^{(1)}, \dots, x_{kp}^{(1)}) \tag{B.8}$$

Where “f” is the model that links the inputs to the output.

Based on the previous quantities, the Sobol indices can be estimated in the following manner:

First order sensitivity indices: $\hat{S}_i = \frac{\hat{\mathrm{U}}_i - \hat{\mathrm{E}}^2}{\hat{V}}$

Total-effect indices: $\hat{S}_{Ti} = 1 - \frac{\hat{\mathrm{U}}_{\sim i} - \hat{\mathrm{E}}^2}{\hat{V}}$

Appendix C - Optimal sets of weights: *Roll_0/1* data

Table C.1: Optimal sets of weights - *Roll_0/1* data

ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9	ω_{10}	ω_{11}	ω_{12}
0,9	0,1	0,21	1	0,29	0,33	0,73	0,74	0,65	0,81	0,2	0,68
ω_{13}	ω_{14}	ω_{15}	ω_{16}	ω_{17}	ω_{18}	ω_{19}	ω_{20}	ω_{21}	ω_{22}	ω_{23}	ω_{24}
0,16	0,1	0,66	0,16	0,37	0,33	0,15	0,5	0,62	0,3	0,98	0,2
ω_{25}	ω_{26}	ω_{27}	ω_{28}	ω_{29}	ω_{30}	ω_{31}	ω_{32}	ω_{33}	ω_{34}	ω_{35}	ω_{36}
0,54	0,76	0,41	0,8	0,64	0,37	0,76	0,34	0,17	0,44	0,86	0,24
ω_{37}	ω_{38}	ω_{39}	ω_{40}	ω_{41}	ω_{42}	ω_{43}	ω_{44}	ω_{45}	ω_{46}	ω_{47}	ω_{48}
0,27	0,4	0,13	0,27	0,3	0,1	0,29	0,26	0,92	0,56	0,3	0,59
ω_{49}	ω_{50}	ω_{51}	ω_{52}	ω_{53}	ω_{54}	ω_{55}	ω_{56}	ω_{57}	ω_{58}	ω_{59}	ω_{60}
0,82	0,52	0,46	0,66	0,92	0,01	1	0,17	0,37	0,3	0,94	0,57
ω_{61}	ω_{62}	ω_{63}	ω_{64}	ω_{65}	ω_{66}	ω_{67}	ω_{68}	ω_{69}	ω_{70}	ω_{71}	ω_{72}
0,13	0,37	0,24	0,59	0,76	0,08	0,87	0,07	0,18	0,64	0,34	0,03
ω_{73}	ω_{74}	ω_{75}	ω_{76}	ω_{77}	ω_{78}	ω_{79}	ω_{80}	ω_{81}	ω_{82}	ω_{83}	ω_{84}
0,1	0,82	0,29	0,59	0,56	0,22	0,52	0,85	0,09	0,88	0,76	0,14
ω_{85}	ω_{86}	ω_{87}	ω_{88}	ω_{89}	ω_{90}	ω_{91}	ω_{92}	ω_{93}	ω_{94}	ω_{95}	
0,1	0,42	0,75	0,11	0,03	0,69	0,7	1	0,3	0,84	0,99	

Wahb ZOUHRI

Quality prediction/ classification of a production system under uncertainty based on Support Vector Machine

Résumé

Avec l'émergence des techniques d'IoT, les industries manufacturières adoptent de nouvelles technologies d'analyse de données afin d'améliorer la qualité de leurs systèmes de production. Les méthodes de classification offrent diverses solutions aux problèmes de management de la qualité, comme la détection des défauts et la prédiction de la conformité. Cependant, les données de production sont entachées d'incertitudes qui affectent les performances de ces méthodes. Ces travaux visent à étudier l'impact des incertitudes de mesure sur les performances des machines à vecteurs supports (SVM). Deux groupes d'approches sont proposés, le premier visant à quantifier l'impact des incertitudes de mesure sur la précision de prédiction des SVM via des techniques de propagation d'incertitudes et d'analyse de données, et le second visant à améliorer la robustesse de la SVM via des approches d'optimisation robuste intrusives et non intrusives. Les différentes approches permettent de mieux appréhender la robustesse de la SVM et la manière de l'améliorer. Ces approches proposées ont été évaluées à l'aide d'études de cas avec des partenaires industriels.

Mots clés : gestion de la qualité, machines à vecteurs supports, incertitudes de mesure, optimisation robuste.

Résumé en anglais

With the emergence of the IoT paradigm, manufacturing industries are opting for new technologies for data collection and analysis to evaluate the quality of their manufacturing systems. Machine learning and classification methods provide various solutions to quality management such as defect detection and conformity prediction. However, manufacturing data are affected by uncertainties, which affect the performances of classification techniques. Accordingly, the thesis aims to study and manage the impact of measurement uncertainties on the predictive performances of support vector machine (SVM). Two groups of approaches are thus proposed: the former aiming to quantify the impact of measurement uncertainties on the prediction accuracy of SVM using several propagation techniques and data mining techniques, and the latter aiming to improve the robustness of SVM to uncertainties using robust optimization techniques. The various approaches provide a better understanding of the SVM robustness and how to improve it. The proposed approaches are evaluated through case studies with industrial partners.

Keywords: quality management, support vector machines, measurement uncertainties, robust optimization.