THÈSE PRÉSENTÉE À
L'UNIVERSITÉ DU QUÉBEC EN OUTAOUAIS

COMME EXIGENCE PARTIELLE
À L'OBTENTION DU TITRE DE
DOCTEURE EN SCIENCES ET TECHNOLOGIES DE L'INFORMATION

PAR
SYLVIA ANDRIAMAHAROSOA

STRUCTURE CAUSALE DES RISQUES DANS LES SYSTÈMES INDUSTRIELS PAR LA MÉTHODE DES RÉSEAUX BAYÉSIENS DYNAMIQUES

GATINEAU, 27 AVRIL 2017

PRÉSENTATION DU JURY

CETTE THÈSE A ÉTÉ ÉVALUÉE

PAR UN JURY COMPOSÉ DE :

M. Ahmed Lakhssassi, Ph.D., Président du jury
Professeur titulaire
Département d'informatique et d'ingénierie, Université du Québec en Outaouais

M. Salim Lahmiri, Ph.D., Membre externe du jury
Professeur agrégé
ESCA, École de Management, Casablanca, Maroc

M. Raul Valverde, Ph.D., Membre externe du jury
Senior Lecturer, Director of BTM Program
John Molson School of Business, Concordia University, Montreal

M. Emmanuel Kengne, Ph.D., Membre interne du jury
Professeur associé
Département d'informatique et d'ingénierie, Université du Québec en Outaouais

M. Stéphane Gagnon, Ph.D., Directeur de la thèse
Professeur agrégé
Département des sciences administratives, Université du Québec en Outaouais

REMERCIEMENTS

Arrivant à l'échéance de ma formation de docteure en sciences et technologies de l'information, je tiens à exprimer ma reconnaissance à mon directeur de thèse Dr Stéphane Gagnon pour m'avoir proposé ce sujet, pour son encadrement et ses judicieux conseils.

Mes sincères remerciements s'adressent à mes enfants, les deux adorables jumelles, Andriana et Nirina Dubé pour leur temps, leur écoute et compréhension afin de mener à terme cette thèse.

J'exprime aussi ma profonde gratitude à tous les amis et amies de près et de loin qui m'ont apporté leur précieuse aide et collaboration pour atteindre mon objectif.

Je dédie ce travail à mes défunts parents bien-aimés Eloi Razafimaharo et Gabrielle Rasoarinoro et à toute ma famille.

STRUCTURE CAUSALE DES RISQUES DANS LES SYSTÈMES INDUSTRIELS PAR LA MÉTHODE DES RÉSEAUX BAYÉSIENS DYNAMIQUES

Sylvia Andriamaharosoa

RÉSUMÉ

La thèse porte sur la détection de la structure causale des risques dans les systèmes industriels. On se concentre en particulier sur la priorisation des risques sous formes de séquences d'évènements corrélés. Pour améliorer les méthodes de priorisation établies, l'application d'une nouvelle méthodologie utilisant les réseaux bayésiens dynamiques (dBN) est proposée. Pour démontrer cette méthode d'analyse de la structure causale des risques, le développement d'une nouvelle interface utilisateur pour les systèmes industriels de contrôle et d'acquisition de données mieux connus par leur appellation en anglais, Supervisory Control and Data Acquisition (SCADA) est étudié. Un test basé sur un jeu de données d'un SCADA est effectué, obtenu par des auteurs du Royaume-Uni dans une usine de fabrication de semi-conducteurs. Nous utilisons la plateforme R pour notre environnement de développement, et les algorithmes de classifications tels qu'implémentés dans l'outil Tanagra. Nos résultats démontrent que : (1) le réseau de variables avant et après la défaillance est représenté par un nombre limité et distinct de facteurs; (2) le réseau de variables avant et après la défaillance peut être représenté graphiquement de manière dynamique dans une interface utilisateur pour aider la prévention et le diagnostic des pannes; (3) les variables liées à la séquence d'évènements au moment de la défaillance peuvent être utilisées comme modèle pour prévoir son occurrence (dont la qualité de la prévision est évaluée par la mesure F1), et trouver la principale cause de celle-ci, permettant ainsi de prioriser les exigences du système de production sur les bonnes variables à surveiller et gérer en cas de panne. La fiabilité de nos prévisions de défaillances est évaluée grâce aux méthodes de "Train-Test", "Cross-Validation", et "Bootstrap". Ces résultats ont une valeur significative pour les ingénieurs industriels, travaillant en équipe via un SCADA durant l'exécution du système de production. Ainsi, en utilisant cette nouvelle interface plus intuitive, ils pourront plus aisément détecter la cause probable d'une panne du système, et pourront intervenir sur les bons facteurs avec un plus haut taux de confiance.

CAUSAL STRUCTURE OF INDUSTRIAL SYSTEMS RISK USING DYNAMIC BAYESIAN NETWORKS METHOD

Sylvia Andriamaharosoa

ABSTRACT

The thesis deals with detecting the causal structure of risk in industrial systems. We focus on the prioritization of risks in the form of correlated events sequences. To improve the established prioritization methods, the application of a new methodology using Dynamic Bayesian networks (DBN) is proposed. We are studying the development of a new user interface for industrial control systems and data acquisition, known as Supervisory Control and Data Acquisition (SCADA), to demonstrate the analysis method of risk causal structure. We perform a test based on a dataset of an actual SCADA system, obtained by UK authors from a semiconductor manufacturing plant. Our analysis use the R statistical software as a development platform, with classification algorithms implemented in the tool Tanagra. Our results show that: (1) the network of variables before and after the failure is represented by a limited and distinct number of factors; (2) the network of variables before and after the failure can be graphically represented dynamically in a user interface to assist in fault prevention and diagnosis; (3) variables related to the sequence of events at the time of failure can be used as a model to predict its occurrence (whose forecast quality is evaluated by the F1 measure), and find the main cause of it, thus making it possible to prioritize the requirements of the production system on the right variables to be monitored and manage in the event of a breakdown. The reliability of our fault forecasts is evaluated using the Train-Test, Cross-Validation and Bootstrap methods. These results have a significant value for industrial engineers, working as a team through a SCADA during the execution of the production system. Using this new, more intuitive interface, they will be able to more easily detect the probable cause of a system failure, and can intervene on the right factors with a higher confidence level.

TABLE DES MATIÈRES

Page

LISTE DES TABLEAUX

Page

LISTE DES FIGURES

Page

INTRODUCTION GÉNÉRALE

Dans les entreprises manufacturières, l'évolution technologique incessante dans les systèmes industriels, l'augmentation continuelle des coûts de maintenance des équipements et de main-d'œuvre, posent énormément des problèmes que le contexte de la compétitivité mondiale impose de résoudre. Le problème actuel est caractérisé par la complexité des systèmes industriels.

Les phénomènes de dégradation des systèmes conjugués aux défauts de fabrication ont pour impact de réduire les capacités des équipements ou de la machinerie à continuer dans les délais projetés à fournir les services pour lesquels ces derniers ont été conçus. La diversité des interactions de ces phénomènes rend le problème plus difficile et confère au bris ou panne une dimension aléatoire.

De nombreuses approches et solutions ont été proposées dans la littérature pour faire face à de tels phénomènes et obtenir ainsi les meilleurs compromis entre la disponibilité des équipements, les coûts de production, la qualité et la compétitivité du produit. Ces approches et solutions se trouvent dans les domaines de la maintenance, de l'analyse des risques, du diagnostic ou du pronostic des systèmes industriels.

Parmi les moyens aujourd’hui couramment utilisés dans les entreprises manufacturières, nous pouvons mentionner les approches statistiques de contrôle des procédés qui consistent à contrôler les chaînes de production à l'aide d'indicateurs de performance, et celles de détection et classification de défauts qui visent à suivre en temps réel l'évolution des paramètres des équipements ou de la machinerie. Ainsi, les différentes stratégies de maintenance qu’elles soient correctives, préventives ou prédictives font notamment partie des solutions appliquées.

Cependant, on remarque que certaines de ces approches, même combinées, ne permettent pas de maîtriser la forte variabilité liée à un équipement de production introduit dans un contexte industriel complexe et incertain.

Dans les systèmes industriels complexes, la tendance actuelle est d'utiliser les réseaux comme un outil de modélisation plus exhaustif que les approches analytiques. Ces outils ont plus de potentiel pour analyser le comportement des équipements et détecter les causes de leurs défaillances ou de pannes ainsi que pour mesurer leurs performances.

Dans le cadre de cette thèse, nous nous proposons d'appliquer une nouvelle approche utilisant la modélisation par raisonnement mathématique (statistique et probabiliste) focalisée sur le formalisme de la théorie de Bayes, particulièrement les réseaux bayésiens dynamiques. Notre but est de prioriser les risques et leur structure causale liés aux pannes des équipements, et ainsi aider à solutionner les problèmes de fiabilité des systèmes industriels complexes.

Comme objectif spécifique, nous étudions le design initial d'une nouvelle interface utilisateur pour les systèmes industriels de contrôle et d'acquisition de données. L'ensemble de jeu de données à la base de notre étude est issu d'une usine de fabrication des semi-conducteurs. À partir de ce jeu de données, notre démarche en laboratoire utilise le logiciel R comme plateforme de développement des composantes de base d’une éventuelle interface.

Les résultats obtenus amènent à des éléments de décision afin d'aider les opérateurs ou les ingénieurs de systèmes de repérer l'émergence spontanée des facteurs de risques et planifier les stratégies de maintenance des systèmes complexes, en particulier au niveau de la prédiction des pannes inattendues des équipements de production.

Cette thèse comporte cinq chapitres comme suit.

Dans le premier chapitre, nous présentons de manière générale la gestion et l'évolution des systèmes industriels ainsi que la priorisation de l'analyse des risques. Nous présentons le contexte d'un système de production complexe. Dans notre cas, nous prenons le système de fabrication des semi-conducteurs parce que leur processus fait face à l'incertitude de l'innovation technologique continuelle. De plus, les problématiques liées aux défaillances et aux pannes des technologies et équipements de production sont difficiles à résoudre. Nous introduisons également les méthodes d'analyses des risques des données dans ce contexte de production où la maîtrise des procédés et l'expertise des ressources impliquées sont fortement mises en avant. Aussi, nous présentons nos travaux de recherche au sein des systèmes d'information automatisés de production, plus précisément au niveau d'un système SCADA. Nous décrivons notre objectif de recherche ainsi que son positionnement en conformité avec les travaux de la communauté scientifique.

Le chapitre deux énonce le concept et démontre les notions et l'utilité des réseaux bayésiens dynamiques pour répondre à des problèmes liés à des représentations ou modélisations des systèmes complexes. Nous proposons une revue de littérature variée pour étudier les algorithmes d'apprentissage des réseaux. Ainsi, nous présentons plusieurs domaines d'applications pour évaluer la pertinence de différents types de réseaux.

La définition de chaînes de Markov et le modèle de Markov caché est abordé avant d'introduire les réseaux bayésiens dynamiques. Les différents algorithmes d'apprentissage des réseaux sont expliqués. La révision de la littérature dans un vaste domaine d'applications met en lumière comment ces types de réseaux sont fiables et performants comme outils de modélisation des systèmes complexes, et comment ils peuvent jouer un rôle important en apprentissage artificiel. Enfin, nous montrons un tableau de synthèse sur les grands domaines d'applications des réseaux bayésiens dynamiques qui permet ainsi d'évaluer les critères pertinents de recherches.

Le chapitre trois présente notre proposition d'étude. Nous commençons par la contribution scientifique attendue de la thèse. Par la suite, ce chapitre décrit la méthodologie détaillée utilisant les réseaux bayésiens dynamiques portant sur le développement d'interfaces pour un système SCADA dans le domaine des semi-conducteurs. Elle est structurée autour de trois étapes : 1) la préparation des données incluant l'analyse d'identification et de réduction des variables ; 2) la détermination des scénarios pertinentes ; 3) les tests de nos hypothèses de recherche avec les structures graphiques et les lois de probabilités associées.

Le chapitre quatre expose la méthodologie d'analyses des réseaux suivie par les évaluations des résultats. Les réseaux bayésiens statiques et dynamiques sont construits pour les différents états du système. Ainsi, nous vérifions la structure des relations causales entre les variables du système et en deux ou plusieurs périodes dans le temps. Nous soulignons l'importance des interfaces dynamiques pour le système SCADA. Les résultats obtenus sur les prévisions de séquences d'événements corrélés et leurs mesures sont vérifiés et validés.

En dernière partie, le chapitre cinq est consacré aux discussions et conclusions. Nous présentons la performance des réseaux bayésiens dynamiques. Nous discutons aussi des limites de notre recherche. Nous terminons avec les conclusions et les perspectives à venir pour ce travail de thèse.

CHAPITRE 1 - Analyse des risques des systèmes industriels

Introduction

Ce chapitre comporte trois sections et présente de façon globale la gestion et l'évolution des besoins dans les systèmes industriels. Dans la première section, nous présentons le contexte actuel d'un système de production, dit complexe, dédié à l'industrie des semi-conducteurs. Les défis technologiques dans ce système de production sont nombreux et demeurent les facteurs de risques importants de l'industrie. Nous introduisons les facteurs de risques sur la maîtrise et le contrôle des procédés de fabrication des semi-conducteurs pour mieux positionner les travaux de recherche.

La deuxième section porte sur l'analyse et la priorisation des risques dans les systèmes industriels. Nous présentons l'état de l'art sur l'analyse des risques industriels, surtout comment identifier les incertitudes et les risques d'échec des systèmes complexes. Nous présentons les différentes méthodes de priorisation des risques des systèmes industriels complexes, à partir d'une revue de la littérature.

La troisième section propose le réseau bayésien dynamique comme méthode d'analyse de la structure causale des risques dans les systèmes industriels. Nous introduisons les notions des systèmes industriels de contrôle et d'acquisition des données. Le réseau bayésien dynamique est étudié pour servir d'interfaces au sein d'un système SCADA dans le domaine des semi-conducteurs.

1.1 Gestion et évolution des besoins dans les systèmes industriels

Dans la plupart des secteurs d'activités industrielles, la gestion et l'évolution des besoins dans les systèmes sont de nature complexe, coûteuse et risquée. La complexité dans la gestion des besoins dans les systèmes industriels réside dans la diversité des facteurs à caractère aléatoire pour faire face à la compétitivité et les exigences de marchés des produits. En effet, les exigences d'affaires dans les systèmes industriels se caractérisent particulièrement dans la réduction des coûts et délais de production, ainsi que la disponibilité et la fiabilité des équipements pour augmenter davantage les performances des usines de production.

1.1.1 Contexte des systèmes industriels complexes

Le contexte du système de production, dans l'industrie des semi-conducteurs, est très complexe et concurrentiel. La complexité dans le système de production se caractérise par la combinaison des sous-systèmes de technologies différentes. Ils sont liés également à la difficulté de maîtriser les nouveaux procédés technologiques des produits miniaturisés, ainsi que d'assurer une disponibilité sans faille des équipements.

Le processus de fabrication des semi-conducteurs est constamment évolutif ou dynamique. C'est l'une des industries du secteur de la haute technologie où le marché est en évolution et à forte intensité [1]. L'obligation d'innover et de créer de nouveaux produits implique également la nécessité d'implémenter de nombreuses innovations de procédés dans le processus de fabrication.

Deux facteurs prédominent parmi les défis rencontrés par les usines de semi-conducteurs :

1. une complexité de production importante : Les composants ou les produits sont conçus et développés par un grand nombre d'opérations et souvent d'une durée assez longue. On constate que les interactions dynamiques entre les différentes technologies et outils de production sont très marquées. De plus, la complexité dans le processus de production est causée par la rapidité du rythme de renouvellement des technologies.

2. un milieu ou environnement très incertain : déterminé par la conjoncture du flux des produits, la poussière ou d'autres saletés qui peuvent entrainer des pertes de produits voire une propagation de particules qui corrompent les machines, le déréglage des machines, la pollution liée à la présence d'autres facteurs, et l'intervention plus ou moins appropriée des employés opérateurs d'équipements.

1.1.2 Défis technologiques et particularité du système

Nous offrons ici une vue d'ensemble des défis technologiques et la particularité dans les systèmes industriels. De nos jours, les principaux défis technologiques dans les systèmes industriels complexes résident dans la diversité des innovations technologiques [2] qui touchent l'industrie, dans le rythme de renouvellement sans cesse des changements technologiques [3] apportés dans les processus de fabrication, ainsi que dans la maîtrise et l'expérience [4] avec les procédés, technologies et équipements.

En ce qui concerne l'industrie des semi-conducteurs, face à une concurrence mondiale et une forte clientèle de plus en plus exigeante, les entreprises sont contraintes d'améliorer continuellement la performance, et donc le rendement à travers des idées innovantes et une qualité des plus élevées. La recherche constante d'une production à capacité maximale est l'un des principaux objectifs de ce secteur, afin d'assurer un retour sur investissement sur des machines très coûteuses caractérisées par un coût d'exploitation très important. Les principaux défis résident dans le processus de fabrication, en particulier dans le traitement des dégradations inattendues et difficilement contrôlables des équipements de production ce qui est l'une des causes de la grande quantité de rejets des produits.

Le processus de fabrication des semi-conducteurs se caractérise par un contexte soumis à de nombreux risques de fonctionnement ou opérationnels. Les différentes politiques de maintenance, le système de contrôle en ligne, les méthodes de détection et de classification, les tests paramétriques, les tests et mesures électriques, sont parmi les solutions qui permettent de réduire une partie de la variabilité, de mieux maîtriser les pannes des équipements et d'améliorer la qualité des produits. Mais de nombreux problèmes rencontrés dans les usines de production sont présents pour nous rappeler que tout n'est pas observé et observable, ni repéré et repérable, que le risque zéro n'existe donc pas.

Nous pouvons résumer la réalité existante dans les usines de semi-conducteurs comme suit :

- Le rejet de produits est fréquent et sa variabilité n'est pas maitrisée,
- La fréquence des interventions correctives imprévues est élevée,
- La planification des maintenances préventives ou prédictives n'est pas optimale,
- Le coût lié au processus de contrôle des produits est considérable.

Nous pouvons conclure que le domaine des semi-conducteurs est largement considéré comme un système industriel complexe. Les faits issus d'interactions et d'interférences entre plusieurs éléments de ce système sont souvent imprévisibles. Donc, il est nécessaire de développer et de mettre en place des méthodes et des outils permettant d'agir en fonction de cette complexité.

1.1.3 Maîtrise des procédés

Dans un système industriel complexe, la maîtrise des procédés technologiques et/ou techniques demeure toujours contraignante et insuffisante dans plusieurs secteurs industriels. La maîtrise d'un processus industriel est caractérisée par divers facteurs, dont les principaux sont le niveau de compétences des ressources humaines (opérateurs, ingénieurs, spécialistes) nécessaires aux différents niveaux de production, ainsi que les méthodes ou procédés de fabrication utilisés afin de maîtriser les activités/opérations. Les autres facteurs prépondérants concernent le degré de maturité des équipements, la présence des outils de production adéquats, et les technologies de fabrication innovantes. Ainsi, maîtriser un procédé industriel signifie maîtriser l'environnement de travail propice au bon déroulement des activités ou des opérations. Ces facteurs ont un impact direct sur la qualité et la conformité des produits.

Dans un contexte industriel concurrentiel et dynamique, ces facteurs représentent des risques pouvant affecter la maîtrise et le contrôle des procédés, et donc requièrent un processus de gestion et d'analyse du risque adapté à l'ampleur et la complexité de ces facteurs. En résumé, les processus opérationnels, la maîtrise et le contrôle des procédés font partie intégrale du processus d'analyse des risques dans un contexte de production complexe.

1.2 Sources de risques dans les systèmes industriels

La notion de risque est définie comme une mesure d'un danger associé à une mesure de l'occurrence d'un événement indésirable et une mesure de ses effets ou conséquences [5]. Dans le contexte industriel, le risque amène aussi à une probabilité d'occurrence d'une défaillance, vu l'interdépendance forte entre toutes les composantes du système de production complexe. Il représente les défaillances éventuelles qui peuvent avoir un effet sur le produit, les équipements et les technologies ou les processus de fabrication. Les dommages engendrés peuvent être des pertes de rendement, de temps de cycle, de coût, ou toute autre propriété prioritaire selon la stratégie priorisée par l'usine.

Le vrai défi à surmonter dans la priorisation de l'analyse des risques dans les systèmes industriels complexes reposent sur les différentes approches ou méthodes utilisées pour diagnostiquer et repérer les anomalies, les manques ou l'insuffisance concernant un procédé de production, ou encore les vulnérabilités des équipements en opération.

La priorisation des risques dans les systèmes industriels complexes est donc un défi important pour la gestion des usines de haute technologie.

Selon Thamhain [6] un traitement des risques efficace dans un environnement complexe et difficile nécessite une gestion d'interventions qui va au-delà des approches analytiques simples. Les résultats des travaux effectués sur plusieurs entreprises à grande échelle de la haute technologie démontrent que l'analyse et la détection des risques, tôt dans un cycle de vie d'un projet, est un facteur de succès pour les entreprises. De plus, les données obtenues sur le terrain indiquent que la gestion et la priorisation efficace des risques impliquent un ensemble des variables liées au processus opérationnel, à l'environnement organisationnel et aux facteurs humains.

1.2.1 Ressources humaines et gestion des risques industriels

Dans un système industriel moderne, l'intervention du personnel à tous les niveaux est requise pour assurer une bonne gestion des risques et des incertitudes. Les personnes doivent, d'abord, pouvoir maîtriser toute déviation des exigences dans les produits ou les technologies [7] ; doivent également s'adapter aux changements ou aux modifications technologiques [3] apportés ; doivent, enfin, contrôler les actions, les opérations et les expériences avec les technologies innovantes [8] .

Dans un environnement industriel incertain et dynamique, afin de bien évaluer et prévenir les facteurs de risques, il est nécessaire de mettre en œuvre et de développer des méthodes et ou des procédés efficaces et flexibles permettant d'agir face à des changements constants. Les ressources humaines ou l'équipe de production sont donc des éléments clé de succès pour l'analyse et la priorisation des risques dans les processus industriels complexes. La communication, la cohérence et l'évolution de l'information au sein d'une équipe de production sont problématiques dans un contexte industriel compétitif.

Un défi se pose, celui de prioriser les risques potentiels lorsque les ressources humaines n'ont pas toutes les mêmes compétences et expériences ou expertises dans un milieu d'usine de haute technologie. La réponse réside dans l'adoption d'une méthode de gestion des risques adaptative, satisfaisant les exigences d'un milieu complexe, pouvant ainsi le mieux représenter leur interdépendance et les interactions dynamiques entre les facteurs de risque, de façon à permettre à des personnes aux compétences variées d'agir dans la gestion des risques.

Les compétences de l’équipe de production dépendent directement de leur expérience avec les technologies et le système de production et ses procédés. D'autres éléments plus complexes à gérer aussi sont situés au niveau de la structure de l'organisation et des ressources humaines impliquées dans le processus. Une bonne gestion des risques est donc reliée à la qualité de l'équipe et aux personnes impliquées dans la gestion du système de production, et aussi dans l’équipe qui développe les produits et procédés. Il faudra pallier aux lacunes dans les connaissances des équipes de travail par des méthodes adaptatives pour l’analyse et la priorisation des risques, lorsque les individus en charge ne sont pas spécialisés dans le domaine ou le champ d'application concerné, et ne possèdent pas une expérience pertinente ainsi que le savoir nécessaire dans le processus de production.

1.2.3 Méthodes d’identification et de priorisation des risques

Dans cette section, nous présentons les approches développées pour l'identification et la priorisation des risques potentiels dans le système industriel complexe.

Selon Hu et al. [9] le risque potentiel d'un système industriel peut être évalué à travers la localisation de l'origine des dangers potentiels et en déduisant les possibles conséquences correspondantes. Dans cette étude, les chercheurs proposent des modèles de prévisions intégrés pour identifier les risques éventuels du système industriel. Les modèles concernent l’origine (localisation) et la propagation des défauts dans le système, l'intégration de la connaissance à priori des interactions et dépendances entre les sous-systèmes, des composants et de l'environnement du système, ainsi que les relations entre les causes et les effets de la défaillance.

Les risques potentiels sont nombreux dans un contexte de production fortement variable, liés aussi à la disponibilité aléatoire des équipements de pointe. Il faut alors cibler sur le comportement de ces équipements critiques. Medjaher et al. [10] ont proposé de préparer des actions d''améliorations du rendement de ces équipements en se concentrant sur la détection ou l’anticipation de l’occurrence d’une défaillance future ou sur la localisation des causes à l’origine d’une défaillance.

Les fortes exigences des clients en termes de qualité des produits et le développement continu des technologies ainsi que la complexité des procédés de fabrication sont à l'origine des évolutions des méthodes d'analyses des risques et de sûreté de fonctionnement des systèmes industriels. Les méthodes d'analyse des risques industriels varient selon les secteurs d'activités, la nature du système et le type de problème industriel à résoudre ainsi que les objectifs visés. Ces méthodes sont définies ainsi :

Méthodes inductives :

Les méthodes inductives ou ascendantes [11] sont très efficaces pour l'identification des risques potentiels dans un système caractérisé par un fonctionnement séquentiel.

Dans l'industrie des procédés de fabrication classiques ou peu complexes, les méthodes les plus utilisées sont :

- Analyse Préliminaire des Risques (APR), une technique très générale orientée vers le domaine de la sécurité des systèmes ; elle est utilisée durant les phases préliminaires de conception pour identifier et évaluer les risques des systèmes.
- Étude du danger et opérabilité *(Hazard and Operability Studies)* (*HAZOP*), cette méthode est utilisée pour identifier et évaluer les problèmes qui peuvent représenter un risque humain et matériel. L'application de cette méthode est souvent utilisée dans les industries chimiques et pétrolières.
- Analyse des Modes de Défaillance, de leurs Effets et de leur criticité (*AMDEC*); c'est la méthode plus détaillée pour analyser les systèmes présentant un aspect dynamique [12] et incertain.

Méthodes déductives :

Les méthodes de type déductives ou descendantes [13] sont appliquées pour les systèmes statiques ou dépendant de peu ou pas du temps, comme par exemple, les arbres de défaillance. Cette méthode est principalement utilisée dans le domaine de l'ingénierie de la sécurité et elle procède par une démarche causes-effets.

Méthodes stochastiques :

Les méthodes stochastiques [14; 15] sont appliquées pour les systèmes dynamiques. Les chaînes de Markov sont utilisées pour l'identification des différents états du système pendant l'exploitation, et lors de l'analyse des transitions de passage aléatoire d'un état à un autre. Les réseaux de Pétri sont des graphes bipartites orientés, utilisés pour les modèles dynamiques comportant un plus grand nombre de variables d'état. Les méthodes de Monte Carlo sont appliquées [16] avec un grand nombre de variables d'états complexes, si le processus n'est pas markovien (probabilité pas exponentielle). En dernier, les réseaux bayésiens généralisent les méthodes des arbres de défaillance et de décision.

On illustre les principales méthodes d'analyses de risque de sureté de fonctionnement industriel au Tableau 1. Par la suite, nous expliquons leurs principes de fonctionnement dans les systèmes industriels.

Tableau 1 : Méthode d'analyse de risque et de sureté de fonctionnement industriel

Méthodes inductives qualitatives	**Méthodes déductives quantitatives**	**Méthodes Stochastiques**
Analyse Préliminaire des Risques	Diagramme de fiabilité	Retour de l'expérience
Étude du danger et opérabilité (HAZOP)	Arbre de défaillances	Réseaux Bayésiens
Analyse des Modes de Défaillance, de leurs Effets et de leur Criticité AMDEC - AMDE)	Arbre de défaillance dynamique	Simulation Monte Carlo.
	Processus de Markov	
	Réseaux de Pétri.	

1.2.3.1 Analyse Préliminaire des Risques (APR)

L'Analyse Préliminaire des Risques (APR) est une méthode universelle couramment utilisée pour identifier les scénarios d'accidents en présence de danger/risque d'un système industriel complexe. Cette méthode permet aussi de repérer à priori les risques du système étudié.

Le principal avantage de l'APR est de permettre une étude de manière plus rapide et efficace des situations dangereuses ou risquées dans des systèmes industriels complexes. Cette méthode est avantageuse en termes de temps passé, et ne demande pas un niveau assez détaillé de description du système étudié. Elle permet d'éviter une analyse systématique du contexte. Un autre avantage de cette dernière permet aussi de cibler les accidents potentiels susceptibles d'affecter le système industriel et d'évaluer les causes envisageables des accidents potentiels.

De plus, elle permet d'estimer la probabilité d'occurrence des accidents potentiels et la gravité des dommages qu'ils pourraient causer. Aussi, elle peut déterminer les mesures qui permettent de diminuer la probabilité des accidents potentiels.

Cependant, l'APR ne permet pas de caractériser l'enchaînement des évènements susceptibles de conduire à un accident majeur pour des systèmes complexes. Cette méthode vise seulement à des évènements simples qui pourraient causer des accidents du système. Ainsi, elle permet d'identifier des points critiques devant faire l'objet d'études plus détaillées. Cette méthode s'applique particulièrement au domaine de la sécurité. La méthode conduit à l'identification et l'estimation des risques du système industriel.

Quant aux autres désavantages de la méthode APR, elle est souvent problématique pour chercher les définitions claires permettant de distinguer sans ambiguïté les événements causant les accidents potentiels particuliers. Par conséquent, la similitude entre les notions des accidents potentiels, des éléments dangereux, de situation risquée et sa conséquence sont souvent difficiles à cerner et à définir ainsi qu'utiliser pour une équipe de fabrication dans un système industriel complexe. Le coût est élevé pour mener des analyses avec cette méthode.

1.2.3.2 Étude du danger et opérabilité (HAZOP)

La méthode sur l'étude du danger et opérabilité permet d'identifier les dangers/risques suite à une déviation des paramètres d'un procédé ou un système de production. L'avantage de cette méthode permet de détecter précocement certains problèmes/ anomalies du système de production et les erreurs de conception du produit.
Cependant, la méthode consiste seulement à déterminer les déviations, par rapport aux valeurs des paramètres physiques (température, pression, etc.) régissant le système / procédé qui peuvent créer des dangers.

Un autre désavantage de cette méthode est qu'elle n'est pas conçue pour identifier les risques d'évènements à très faible probabilité d'occurrence. Cette méthode nécessite une équipe pluridisciplinaire pour chaque domaine d'application dans l'industrie, et le facteur temps est impératif pour mener des travaux d'analyses efficaces avec cette méthode.

1.2.3.3 Analyse des Modes de Défaillances, de leurs Effets et de leurs Criticités (AMDE-AMDEC)

L'AMDE est une méthode qui exige la connaissance et compréhension de tous les modes de défaillances du système complexe. En effet, la performance de l'AMDE repose sur la typicité des modes de défaillances du système et l'exhaustivité de cette liste. Cette méthode permet :

- D'évaluer les effets de chaque mode de défaillance des composants d'un système sur les différentes fonctions du système.
- De définir l'importance de chaque mode de défaillance sur le fonctionnement habituel du système et d'en estimer l'impact sur la fiabilité du système considéré.
- De hiérarchiser les modes de défaillances connus suivant la facilité de détection et de résolution

Afin d'évaluer la criticité d'un mode de défaillance, l'AMDEC est définie comme la continuité logique de l'AMDE. En effet, la criticité estime le mode des défaillances ; la probabilité d'occurrence, l'ampleur (mineurs, majeurs, critiques ou catastrophiques) et le risque de la non détection. L'AMDEC est basé sur les démarches de l'AMDE et ajoute une étude quantitative de la criticité de la défaillance du système.

Les avantages de ces méthodes sont nombreux tels que :

- L'aptitude pour détecter les défaillances des éléments conduisant à la défaillance globale du système.
- L'outil efficace pour l'identification de défaillances potentielles et les moyens d'en limiter les effets. Cette méthode contribue à prévenir les risques et améliorer les rendements du système.
- La capacité pour donner des informations, des indications pertinentes à gérer au niveau des analyses de sureté de fonctionnement et des opérations à entreprendre.

Cette méthode permet une autre vision du système, en apportant des supports, des idées de décisions et des améliorations au système.

Toutefois, les méthodes AMDE-AMDEC sont très difficiles à maîtriser pour un système complexe avec un grand nombre de composants et d'interactions. Avec ces méthodes, il est impossible de décrire les défaillances multiples. Ces méthodes sont lourdes, impliquent beaucoup d'incertitudes, et nécessitent des travaux importants et laborieux au niveau de la sûreté de fonctionnement des systèmes dans certains domaines industriels où il y a un développement continu des technologies. Ces méthodes sont insuffisantes pour une bonne analyse prévisionnelle inductive et qualitative d'un système industriel complexe. Ce sont des méthodes non adaptées aux systèmes industriels en temps réel.

1.2.3.4 Diagramme de Fiabilité (DF)

La méthode du diagramme de fiabilité est utilisée pour représenter le modèle d'un système à partir de la fiabilité des composants/pièces, des sous-systèmes ou des états de fonctions. En effet, la modélisation permet de chercher et d'évaluer les liens entre les composants et les fonctions du système considéré.

L'avantage du DF permet une analyse qualitative pour la réussite du procédé en identifiant les bons composants du système. Cette méthode permet de reconnaitre les scénarios qui mènent à l'échec du système afin d'éviter les pannes/accidents. Elle permet une analyse quantitative qui vise en particulier à définir la probabilité du bon fonctionnement du système. Les calculs de probabilités reposent sur le bon fonctionnement du système.

Dans la pratique, l'inconvénient du diagramme de fiabilité est qu'il faut s'assurer de l'indépendance entre les scénarios d'évènements du système. Ainsi, le diagramme de fiabilité ne permet pas de modéliser des systèmes dynamiques.

1.2.3.5 Arbre de défaillance dynamique (AdD)

La méthode d'analyse par l'Arbre de défaillance dynamique (AdD) permet d'évaluer les scénarios d'un évènement dangereux ou risqué du système.

L'avantage de cette méthode permet une analyse quantitative pour déterminer les enchaînements et combinaisons des scénarios pouvant conduire à l'événement dangereux du système. Elle permet aussi d'évaluer la probabilité d'occurrence d'événement dangereux afin de disposer des critères pour déterminer les priorités pour la prévention d'accidents potentiels.

Cependant, l'inconvénient principal de l'AdD est qu'il est difficile de prendre en compte les aspects temporels pour les systèmes complexes.

1.2.3.6 Processus de Markov

Le processus de Markov ou la méthode de l'espace des états du système permet de modéliser un système dynamique comportant un grand nombre de variables d'états. Il vise à repérer et identifier le passage par les états de défaillance/panne sur le fonctionnement du système.

Le principal avantage de cette méthode est de permettre la modélisation des systèmes ou sous-systèmes réparables. Le processus de Markov est à la fois performant et flexible pour analyser les systèmes dynamiques tant au niveau de la fiabilité et la disponibilité du système.
Cette méthode permet la modélisation des systèmes dynamiques présentant plusieurs cycles de fonctionnement (fonctionnement et panne).

Autre avantage, le processus de Markov permet de réduire le coût des opérations considérablement. Cependant, le processus a une grande lacune pour son application : la construction et la modélisation pour les systèmes industriels de grande envergure définis par un grand nombre d'état. Ainsi, les inconvénients du processus sont les taux de transitions constants entre états, autrement dit les évènements aléatoires sont établis par les lois de probabilités exponentielles.

1.2.3.7 Réseau de Pétri

Le réseau de Pétri est utilisé pour représenter un modèle dynamique du système industriel comportant un grand nombre de variables d'état (état statique et dynamique). Les avantages du réseau de Pétri sont nombreux. Il permet la possibilité d'analyser le comportement d'un système en présence de défaillances. En effet, la modélisation dynamique permet d'obtenir des mesures en termes de fiabilité du système en assignant des valeurs numériques aux paramètres du modèle. Il permet aussi de modéliser d'une part le fonctionnement normal d'un système et d'autre part les occurrences de défaillances. Un autre avantage est la diminution du coût des opérations du système dans la pratique. Toutefois, le réseau de Pétri est inaccessible à un grand nombre de variables d'état du système industriel compliqué.

1.2.3.8 Simulation Monte-Carlo

La simulation Monte-Carlo permet de modéliser de façon plus délicate le comportement d'un système complexe. Il s'agit d'une méthode développée pour traiter les systèmes dynamiques ; ces systèmes passent d'états stables en états aléatoires régis par les divers phénomènes (défaillances de composants, réparations) auxquels le système est soumis.
Les avantages sont variés. La méthode donne accès à plusieurs paramètres inaccessibles par les autres méthodes et conduit à des analyses très détaillées des systèmes étudiés. Également, la méthode n'est pas restreinte par le nombre d'états du système étudié, même si seuls les états prépondérants du système se manifestent lors de la simulation. C'est une méthode qui permet l'utilisation de toutes les lois de probabilité. L'implémentation informatique de la simulation Monte Carlo n'est pas difficile.

L'inconvénient de la simulation Monte-Carlo est généralement insensible à un grand nombre d'états des systèmes industriels complexes. La dépendance stochastique des variables est problématique.

Autre inconvénient de cette méthode est qu'elle demande beaucoup de temps de calcul exorbitant pour arriver à des systèmes industriels fiables. Par conséquent, les résultats obtenus sont imprécis.

1.2.3.9 Réseau Bayésien

L'analyse par le réseau bayésien expose à la fois les méthodes des arbres de défaillances et de décision des systèmes industriels.

Les avantages du réseau bayésien sont au niveau de la performance et de la flexibilité pouvant combiner plusieurs aspects, dont les statistiques, les probabilités, de l'aide à la décision et la gestion des connaissances des systèmes statiques et dynamiques. L'approche bayésienne offre une modélisation prometteuse et qualitative pour des systèmes industriels complexes où il y a des dépendances et non dépendances entre les variables aléatoires multi-états.

Autre avantage, l'approche bayésienne dynamique offre un outil assez puissant et à la fois novateur : tant au niveau de l'évaluation de fiabilité des systèmes, à la maintenance et à l'analyse intégrée des risques des systèmes industriels.

Cependant, les réseaux bayésiens ne sont pas suffisants pour étudier et interpréter de manière appropriée et efficace les incertitudes reliées à des contraintes opérationnelles des systèmes industriels complexes. Les études approfondies et privilégiées des problèmes de modélisation, en utilisant des réseaux bayésiens, ne sont pas encore assez avancées pour traiter les contraintes de sûreté de fonctionnement des systèmes complexes et de maîtrise des risques industriels.

Le tableau 2 illustre les méthodes d'analyse des risques qui peuvent être utilisées selon les caractéristiques des systèmes industriels.

En résumé, le choix d’une méthode d’analyse des risques des systèmes industriels est une étape à la fois délicate et compliquée où beaucoup de facteurs doivent être considérés. La méthode choisie doit être adaptée à la taille du système étudié et selon la nature des risques encourus à priori. Le choix d’une méthode d’analyse des risques des systèmes industriels peut être orienté par la nature et le type des informations ou des données des systèmes disponibles, la caractéristique du problème du système à analyser, la perception des risques du système, les expertises et expériences de l’équipe responsable pour la gestion du risque.

Dans la section suivante, nous proposons l'utilisation des réseaux bayésiens pour un système industriel complexe. Nous avons choisi les réseaux bayésiens pour plusieurs raisons. Les ingénieurs disposent des outils bien connus comme les arbres de défaillance, les diagrammes de fiabilité etc., pour connaître l'état du système avant de prendre une décision. Pourtant, les outils de l'intelligence artificielle, tels les réseaux bayésiens, peuvent apporter une aide efficace dans la prise de décision de fonctionnement, de maintenance ou de réduction des dangers/risques pour les systèmes industriels.

Tableau 2 : Comparaison des méthodes selon le type du système industriel

Méthode / **Type**	AMDEC	AdD	DF	Processus de Markov	Réseau de Pétri	Simulation Monte-Carlo	Réseau Bayésien
Système Statique	oui	oui	oui	oui	oui	oui	oui
Système dynamique	non	non	non	non	oui	oui	oui
Système réparable	oui	oui	oui	oui	oui	oui	oui
Système non réparable	oui	oui	oui	oui	oui	oui	oui
Comportement fonctionnel du système	non	non	non	oui	oui	oui	oui
Comportement non fonctionnel du système	oui	oui	oui	oui	oui	oui	oui

1.3 Proposition d'un système SCADA.

Dans cette section, nous proposons l'application d’une nouvelle méthodologie utilisant les réseaux bayésiens dynamiques pour démontrer l'analyse de la structure causale des risques d'un système industriel de surveillance de contrôle et d'acquisition de données.

On débute par un survol du concept SCADA (Supervisory Control And Data Acquisition). La revue de littérature sur les problématiques concrètes dans les pratiques et le développement des systèmes SCADA dans les processus industriels. Par la suite, nous présentons les méthodes d'analyses des risques dans ce système. Dans plusieurs domaines d'applications industriels, différentes revues de littérature mettent en lumière comment émergent et évoluent les risques de ce système, et comment les experts, les spécialistes et ingénieurs des systèmes peuvent intervenir et surmonter par des actes ou des tâches précises. En dernier, nous identifions les variables pertinentes de ce système qui fait partie de la recherche.

1.3.1 Problématique étudiée sur le développement du système SCADA

Au plan des fonctionnalités, un système SCADA est un système de contrôle et d'acquisition de données pour l'automatisation industrielle. Il permet la télégestion à grande échelle et en temps réel via un grand nombre de télémesures et de contrôle à distance des installations techniques. Au plan des infrastructures, c'est une technologie de l'information dans le domaine de l'instrumentation, dont l'implémentation peut requérir différents *frameworks* d'instrumentation, ainsi qu'une couche de type *middleware*.

Les applications du système sont très larges et versatiles. Le système peut être appliqué à de nombreux domaines de l'acquisition de données, de surveillance et de contrôle de processus comme : les systèmes manufacturiers, les systèmes d'alimentations, les systèmes d'approvisionnement en eau, énergie, pétrole, produits chimiques et d'autres domaines, les systèmes à distance.

En général, le système est basé sur le contrôle des processus par ordinateurs pour les systèmes à distance. Il peut être utilisé dans l'équipement de surveillance et de contrôle pour effectuer les fonctions d'acquisition de données, le contrôle des périphériques, de mesure, de réglage de paramètres et un modèle d'alarme de signal ou des alertes/ avertissements.

Le développement du système se distingue selon les différents domaines d'applications. Ils dépendent de besoins spécifiques dans la pratique. Avec l'évolution technologique des différents domaines industriels, le développement de ces systèmes fait face à de grands défis dans plusieurs domaines, entre autres dans l'industrie des semi-conducteurs.

Les problématiques étudiées sur le développement de ces systèmes dans les processus industriels complexes sont assez nombreux et souvent critiques. Nous avons abordé ces problématiques en nous basant sur la revue de littérature la plus récente sur les systèmes industriels complexes.

Par exemple, un domaine où la recherche sur les SCADA est la plus active présentement est l'industrie de l'énergie éolienne. Les problèmes récurrents dans ce secteur sur le développement du système sont au niveau de l'amélioration, la fiabilité en temps réel des systèmes de raccordement des réseaux éoliens, de l'optimisation de stratégie d'entretien des éoliens ainsi que les problèmes de contrôle et d'acquisition des données des systèmes d'alertes éoliennes [17]. Les auteurs ont analysé et évalué les lacunes au niveau du contrôle d'acquisition des données des systèmes d'alerte éoliens, et proposent un système plus robuste basé sur une modélisation globale de la prédiction et du système d'asservissement.

Selon Yang et al. [18] la haute fiabilité et la disponibilité de déploiement des éoliens doivent répondre à des stratégies axées sur la maintenance à l'aide d'un système de contrôle et de surveillance d'état et de la prédiction de la fiabilité de pointe des systèmes éoliens. Les auteurs décrivent l'idée de base principale de la façon dont les systèmes d'acquisition des données de surveillance des éoliens contribuent à l'établissement d'une telle stratégie.

Les problématiques étudiées sur le développement du SCADA reposent sur l'analyse du cycle de vie des éoliens [19] à travers les traitements des données opérationnelles des installations. Les chercheurs ont réalisé une étude sur les opérations des turbines éoliennes suivie par des analyses statistiques des données d'étude. Ces chercheurs ont proposé une méthode adaptée pour accroitre la fiabilité des éoliennes selon les résultats obtenus grâce aux analyses statistiques aux conditions environnementales spécifiques de l'endroit.

D'autres problématiques étudiées sur le développement des systèmes SCADA sont aussi au niveau des modèles de diagnostic des pannes d'un système d'alimentation électrique en utilisant des informations incomplètes ou conflictuelles dans la prise de décision [20]. Ces problèmes peuvent générer des dangers et des risques potentiels. Les chercheurs ont étudié la pertinence des informations et les contraintes temporelles du système ; les séquences d'évènements provenant de la surveillance et du contrôle d'acquisition des paramètres électriques ainsi que des caractéristiques de mesures temporelles du système. Par la suite ces chercheurs ont proposé un modèle de diagnostic des pannes/défauts qui intègrent, analysent et traitent des informations provenant de ces multiples sources.

Les recherches sur l'application de la méthode bayésienne temporelle pour le développement du système est devenu populaire et novatrice pour les chercheurs dans plusieurs domaines industriels. Quelques travaux de recherches sur l'utilisation des réseaux bayésiens statiques et temporels dans les systèmes SCADA basés sur des publications sont abordés et détaillés dans le chapitre 3.

1.3.2 Importance des interfaces dans les systèmes SCADA

Dans différents secteurs industriels, les systèmes SCADA évoluent grandement et se dirigent vers une solution entièrement optimisée pour plusieurs entreprises de production et surtout de services privés ou publics. Ceci permettra de prédire la demande ou le besoin de l'entreprise en fonction des données historiques, des modèles météorologiques, de l’heure et de la minute, et même d’une seconde dans la journée. De plus, les opérateurs et les ingénieurs auront ainsi accès à des données sur le terrain grâce auxquelles ils pourront prendre de meilleures décisions suivies des actes nécessaires pour diagnostiquer et ou prédire des défaillances ou pannes du système. Par ailleurs, les activités seront plus efficaces, la qualité du besoin du système sera accrue et les arrêts inutiles seront évités.

Le système est sur le point de devenir le réseau majeur ou l'interface centrale de toutes les données des entreprises industrielles. Les données sont intégrées directement dans le système, puis distribuées aux fins de préparation des procédures de prédictions et plans d’amélioration des processus et des installations, des budgets et de rapports de conformité de la qualité des besoins ou des produits.

Dans le chapitre 3, nous évaluons les raisons pour lesquelles les systèmes SCADA deviendront des interfaces nécessaires et pertinentes de toutes les données des processus industriels. Le choix et le développement d'interfaces ou réseaux sont considérés pour un système industriel de contrôle et d'acquisition des données plus complexes. Nous allons développer les éléments pertinents des interfaces pour répondre aux exigences du système, dans notre cas, c'est un jeu de données d'un processus d'une usine de fabrication des semi-conducteurs.

Conclusion

Pour conclure ce chapitre, nous avons abordé sur la détection de la structure causale des risques dans les systèmes industriels. Nous avons mis l'emphase en particulier sur la priorisation des risques du système sous forme des séquences d'évènements corrélés. L'utilisation des réseaux bayésiens dynamiques a été proposé pour améliorer les méthodes de priorisation des risques dans les systèmes industriels. Ces méthodes ont été étudiés pour détecter et prédire les causalités de risques d'un système industriel. Le développement d'une nouvelle interface utilisateur pour un système de contrôle et d'acquisition des données a été démontré en utilisant des représentations graphiques et des résultats des réseaux bayésiens dynamiques. Le chapitre suivant décrit les concepts, les principes, et les domaines d'applications des réseaux bayésiens dynamiques ainsi que ces particularités comme étant une méthode appropriée et efficace pour la priorisation des risques industriels.

CHAPITRE 2 - Priorisation des risques par les réseaux bayésiens dynamiques

Introduction

Aujourd'hui, les industriels, les chercheurs et la communauté scientifique explorent de vraies solutions plus concurrentielles, plus performantes face à l'évolution des technologies et à la complexité de plus en plus croissante de différents systèmes tels que : industriels, médicaux, biologiques, économiques et financiers. Dans notre société, les lois environnementales et les défis liés à la compétitivité technologique nous obligent également à chercher des approches ou des modèles plus innovants et adaptés afin de répondre à des problèmes spécifiques dans différents domaines respectifs. Les modélisations statistiques (stochastiques) sont devenues les approches les plus utilisées.

Dans ce chapitre, nous présentons l'état de l'art sur les applications des modèles de réseaux bayésiens dynamiques à partir d'une revue de la littérature récente de différents contextes et systèmes complexes. On débute par un résumé des fondements des chaînes de Markov cachées, pour ensuite introduire les réseaux bayésiens dynamiques. Par la suite, nous révisons les fondements de la littérature autour des années 2000, qui ont permis de formaliser les algorithmes d'estimation ou d'apprentissage de ces réseaux. Nous élaborons sur cinq grands domaines d'applications où les réseaux bayésiens dynamiques ont été largement appliqués ces dernières années :

1. Surveillance et diagnostic médical
2. Fiabilité et diagnostic en génie
3. Reconnaissance des activités humaines
4. Reconnaissance d'évènements dans les vidéos
5. Modélisation de systèmes complexes et discrets

Nous présentons également des plus récentes avancées théoriques dans l'estimation des réseaux bayésiens dynamiques. Nous concluons par l'identification de quelques pistes de recherches intéressantes pour de nouvelles applications sur des données en haute fréquence ou en temps réel.

2.1 Fondements des réseaux bayésiens dynamiques

2.1.1 Principes et modèles de Markov cachés

Nous présentons les fondements des chaines de Markov cachées (premier ordre), qui sont des cas particuliers de réseaux bayésiens dynamiques. Par définition, les modèles de Markov cachés sont des outils statistiques permettant de modéliser des phénomènes stochastiques complexes.

Les modèles basés sur les chaines de Markov cachées sont définis par :

- Un ensemble d'états éventuels : $\{X_1, X_2, \ldots, X_k\}$
- Un processus passant d'un état à l'autre, générant ainsi une séquence d'état : $\{X_{i1}, X_{i2}, , \ldots, X_{ik}\}$
- Le principe de chaines de Markov, soit la probabilité d'occurrence d'un état dépend uniquement de l'état précédent :

$$P(X_{ik} / X_{i1}, X_{i2}, , \ldots, X_{ik-1}) = P(X_{ik} / X_{ik-1}) \quad (1)$$

- Un modèle de Markov, qui est défini par les probabilités de transition ;

$$\boldsymbol{a_{ij}} = \boldsymbol{P}(X_i / X_j) \quad (2)$$

- et les probabilités à priori (initiales), définies par :

$$Л_i = \boldsymbol{P}(X_i) \quad (3)$$

Pour un modèle de Markov caché, on doit définir les éléments suivants :

- Une matrice de transition :

$$A = (\boldsymbol{a_{ij}}),\ \boldsymbol{a_{ij}} = \boldsymbol{P}(X_i / X_j) \quad (4)$$

- Les états invisibles, qui génèrent chacun un état observable parmi un nombre

M : $\{V_1, V_2, \ldots, V_M\}$

- Une matrice de transition contenant des probabilités des états observables :

$$B = (b_i(V_M)), b_i(V_M)) = P(V_M / X_i) \quad (5)$$

et

- Un vecteur de probabilité à priori ou dite initiale :

$Л = (Л_i), Л_i = \boldsymbol{P}(X_i)$ *(6)*

De manière synthétique, le modèle de Markov caché est donné par MMC = (A, B, Л).

2.1.2 Réseaux bayésiens dynamiques

La théorie et les définitions des réseaux bayésiens ont été largement expliqués et discutés [21]. Un réseau bayésien comprend trois notions fondamentales qui sont : la théorie de la probabilité, la règle de Bayes et l'indépendance conditionnelle [21]. On peut donc comprendre qu'un réseau bayésien (BN) est la fusion entre la théorie des graphes acycliques et la théorie de probabilité. En fait, ce sont des modèles graphiques probabilistes permettant de représenter de façon intuitive la loi d'une suite de variable aléatoire (v.a.) $X = (X_1, X_2, \dots, X_n)$

Explicitement, un BN est noté μ est définie par un couple $(\sigma, (£_n) 1 \leq n \leq N)$. $\sigma = (X, \varepsilon)$ est un graphe orienté sans circuit où chaque nœud i est associé à la v.a. X_i prenant ses valeurs dans X_i , et où chaque arc est orienté $(i, j) \in \varepsilon$ indique une relation de dépendance entre les v.a. X_i et X_j . $(£_n) 1 \leq n \leq N$ est une suite des lois de probabilité conditionnelle tel que $(£_n)$ représente la loi de probabilité de la v.a. X_n conditionnellement à ses parents X_{pan} ; *pan* désignant les indices des v.a. parents de X_n dans σ . Les relations d'indépendances s introduites par les arcs du graphe permettent de factoriser la distribution jointe de la suite de v.a. *X* de la manière suivante :

$$\boldsymbol{P}(\boldsymbol{X}) = \boldsymbol{P}(X_1, \dots, X_N) = \pi_{n=1}^{N} P(X_n / X_{pan}) \quad (7)$$

Un modèle de réseau bayésien est dit dynamique dans le fait qu'il permet de modéliser des contextes ou des systèmes dynamiques au même état que les modèles de Markov cachés vus dans la section 2.1.1 En général, on note que :

1) si les changements dynamiques sont causés par un processus stationnaire, alors les probabilités ne changent pas dans le temps. L'apprentissage de réseaux bayésiens dynamiques dont la structure n'évolue pas au cours du temps est relativement bien maitrisé [22] [23].

2) si les changements dynamiques sont causés par un processus Markovien, l'état courant dépend seulement de l'état précédent.

On a alors le modèle de Markov du premier ordre :

$$P(X_{ik} / X_{i1}, X_{i2}, , \ldots, X_{ik-1}) = P(X_{ik} / X_{ik-1}) \quad (8)$$

Il se traduit comme modèle de transition inter-temporelle.

3) Si l'observation est générée par l'état courant, alors on définit un modèle d'observation (probabilité d'état observable).

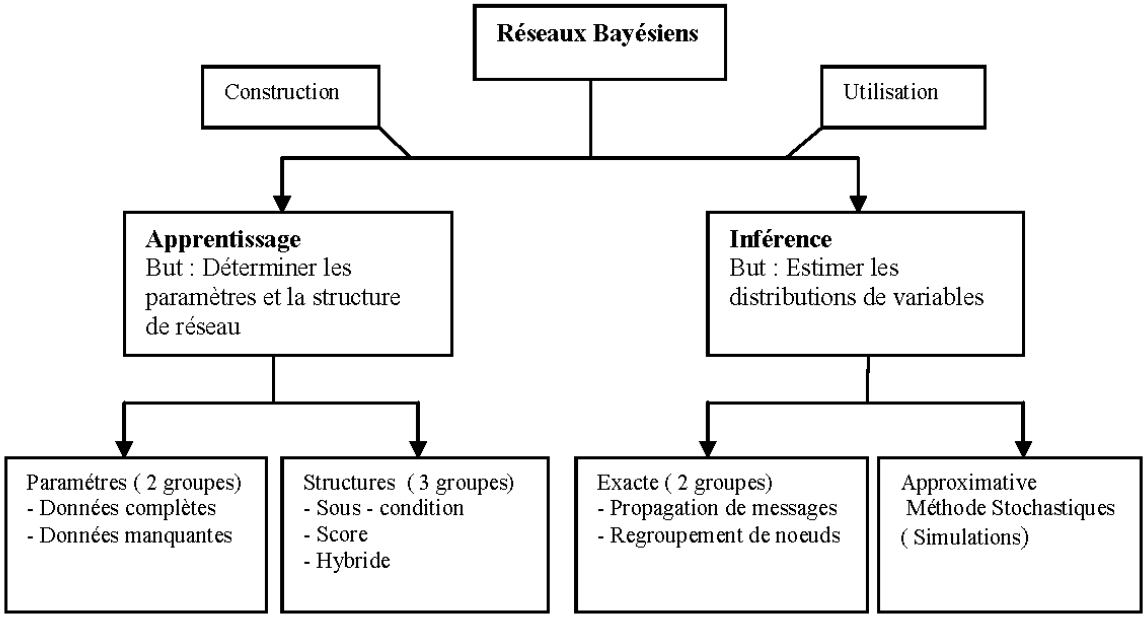
Inspiré des connaissances formelles des réseaux bayésiens classiques, le cadre des réseaux de bayésiens dynamiques (dBN) [24] a permis de fusionner de nombreuses approches issues de la modélisation des séries temporelles, e.g., les modèles de Markov cachés, le filtrage de Kalman, etc.

2.2 Algorithmes d'apprentissage des réseaux bayésiens

Tout d'abord, nous introduisons les concepts d'apprentissage et d'inférence pour construire et utiliser un réseau bayésien. Ces deux concepts sont les principaux défis dans le raisonnement probabiliste bayésien. L'apprentissage est une approche inductive dans un réseau bayésien. Il y a deux types d'apprentissages : paramétrique et structurel. Pour sa part, l'inférence est une approche déductive dont le but est de calculer n'importe quelle probabilité conditionnelle d'une variable du modèle à partir de la structure causale et les distributions de probabilités. Il y a deux types d'inférences de réseaux bayésiens. Les inférences exactes sont les calculs de probabilités à posteriori étant donné qu'un événement est observé. Les inférences approximatives sont la méthode de calcul par élimination des variables, la méthode de rejet, ainsi que le calcul de probabilité conjointe ou marginale (non-conditionnelle).

La figure 1 résume les groupes d'apprentissage et d'inférence dans les réseaux bayésiens.

Figure 1 : Apprentissage et inférence dans les réseaux bayésiens



Les principaux critères pour le choix d'un algorithme d'apprentissage et d'inférence sont basés sur trois éléments : le type de problème (apprentissage ou inférence); la nature des informations/ données disponibles (complètes ou manquantes) et le jugement de l'expert (avec ou sans les connaissances de l'expert).

La revue de littérature suivante récapitule les algorithmes d'inférence dans les réseaux bayésiens dynamiques.

- Les algorithmes de recherches d'indépendance conditionnelle [21] utilisent les tests de l'indépendance. Ceux-ci sont les plus simples et basés sur la théorie de Bayes, mais d'autres algorithmes plus évolués permettent une inférence plus robuste.
- Les algorithmes basés sur des scores sont illustrés et utilisés dans les travaux de recherches de plusieurs auteurs [21], [22], [23], [25], [24]. Ces algorithmes vont chercher soit la structure des données qui maximise un certain score, soit les meilleures structures pour ensuite combiner leurs résultats.

- L'algorithme d'apprentissage de recherche heuristique [21] est une méthode simple pour bâtir la structure de réseaux bayésiens. Toutefois, son utilisation est limitée pour un ensemble d'exemples restreints ou de données plus petites de recherche.
- L'algorithme Structurel-EM [22] a été appliqué pour l'apprentissage des réseaux bayésiens dynamiques. Il est basé sur le principe de l'algorithme EM (espérance maximisation) et permet de traiter des bases d'exemples incomplètes ou insuffisantes sans avoir à ajouter une nouvelle modalité (variable non mesurée) à chaque nœud. C'est une méthode itérative dont la convergence a été prouvée partant d'une structure initiale pour estimer la distribution de probabilité des variables cachées ou manquantes grâce à l'algorithme EM classique. L'espérance d'un score par rapport à ces variables cachées est ensuite calculée pour tous les réseaux bayésiens du voisinage afin de choisir la structure précédente.
- La programmation dynamique [21] a permis l'apprentissage exacte d'une structure bayésienne pour un nombre de variables modéré ou petit (30 variables au maximum). Une méthode de choix pour générer une structure standard permettant de comparer des nouvelles techniques.

La section suivante décrit plus en détails les méthodes algorithmiques des réseaux bayésiens dynamiques réalisés parmi les grands chercheurs dans ce domaine.

2.2.1 Étude bibliographique sur les réseaux bayésiens dynamiques

Plusieurs auteurs et chercheurs se sont intéressés à l'étude des réseaux bayésiens dynamiques.
Les revues de littérature et travaux de recherches réalisés par les auteurs : Murphy, Friedman, Russell, et Ghahramanni, décrivent et expliquent en profondeur les inférences et l'apprentissage dans les réseaux bayésiens dynamiques (dBN).

2.2.1.1 Travaux de recherche de Kevin Murphy

Dans sa thèse, Murphy [24] a proposé la représentation des réseaux bayésiens dynamiques, la manière pour effectuer les inférences exactes et approximatives ainsi que les apprentissages des modèles bayésiens dynamiques à partir des données séquentielles.

L'étude concerne la méthode pour représenter la chaîne de Markov cachée hiérarchique, et l'algorithme de l'arbre de jonction pour l'inférence dans les réseaux bayésiens dynamiques. L'auteur vise le développement d'un nouvel algorithme d'inférence approximative déterministe dans les réseaux bayésiens dynamiques. L'analyse effectuée sur la relation entre l'algorithme de Boyen-Koller et Loopy Belief propagation fait partie de l'étude.

Une nouvelle contribution s'agit aussi d'une application de filtrage Rao-Blackwellised à des réseaux bayésiens dynamiques, ainsi que l'extension de l'algorithme Espérance Maximisation (EM) structurel dans les réseaux bayésiens dynamiques avec différentes applications.

2.2.1.2 Travaux de recherche de Friedman, Murphy et Russell

Friedman et al. [22] ont exploré les méthodes d'inférences Espérance Maximisation (EM) structurel et les méthodes de gradient dans les réseaux bayésiens dynamiques.
Leur étude concerne différents algorithmes d'évaluations des scores en utilisant des réseaux plus explicites. Le développement dans l'inférence approximative de Boyen-Koller a permis d'accélérer les calculs statistiques (d'éviter le goulot d'étranglement) et la simulation stochastique dans les réseaux bayésiens dynamiques.

2.2.1.3 Travaux de recherche de Ghahramanni

Ghahramanni [23] a fourni un tutoriel sur les méthodes d'inférences et d'apprentissages des réseaux bayésiens dynamiques. L'auteur a proposé une approche sur les inférences stochastiques dans les réseaux bayésiens dynamiques. Les inférences abordées sont : la chaîne de Markov cachée, le Monte Carlo par chaîne de Markov (MCMC) selon l'échantillonnage de Gibbs, l'algorithme de l'Espérance Maximisation (EM) et la méthode variationnelle.
Cette étude contribue à l'apprentissage du réseau bayésien dynamique par l'estimation des paramètres en utilisant l'EM, et à l'apprentissage des modèles temporelles selon les données pour faire de la prédiction.
Le tableau suivant résume les différents algorithmes d'inférences et d'apprentissages étudiés des réseaux bayésiens dynamiques par les auteurs mentionnés ci haut.

Tableau 3 : Algorithmes d'inférences et apprentissages des réseaux bayésiens dynamiques

Algorithmes d'inférences de réseau	**Apprentissages de réseau**
Kevin Murphy, 2002 • Inférences exactes. • Inférences approximatives : - Loopy Believe propagation - Méthode variationnelle - Méthode d'échantillonnage. • Inférences graphiques : - Arbre de jonction - Séparateurs - Graphiques arborescents - Passage des variables aléatoires continues (Gaussiens) - Discret exact	L'apprentissage du réseau dans dBN se fait par l'ensemble de calculs de famille marginaux [24] L'apprentissage dans DBN se résume à travers - Estimation des paramètres. - Structure des données (complète, incomplète) - Structure graphique (cartographie/topologie du graphe)
Friedman, Murphy & Russell 1998 • EM structurelle et la méthode de gradient. • Algorithme d'évaluation par des scores • Inférences graphiques de probabilités. • Inférences approximatives pour accélérer les calculs statistiques (éviter le goulot d'étranglement) et de la simulation stochastique dans le réseau bayésien dynamique.	- Structure de réseau bayésien dynamique par extension et combinaison des modifications structurelles et paramétriques apportées à la méthode Espérance-Maximisation (EM). - Structure linéaire du réseau et dans les modèles gaussiens.
Ghahramanni, 1998 • Inférences stochastiques : - Chaîne de Markov cachée ou HMM - MCMC selon l'échantillonnage de Gibbs. - Algorithme EM et la méthode variationnelle.	Estimation des paramètres en utilisant la méthode EM. Apprentissage des modèles temporels selon les données pour faire de la prédiction.

2.2.2 Classification des inférences et apprentissages des modèles probabilistes

Cette section concerne un résumé de la revue de littérature et de recherche sur des inférences et apprentissages des modèles probabilistes.

2.2.2.1 Travaux de recherche de Du et Swamy

Du et al. [25] ont étudié sur la classification des inférences et apprentissages des modèles probabilistes bayésiens. Cette étude concerne deux familles d'inférence approximative pour les modèles probabilistes : l'inférence par MCMC et l'inférence variationnelle. Différents algorithmes de classifications des modèles probabilistes sont illustrés dans le tableau 4.
Les travaux de recherche contribuent à l'apprentissage de structure des réseaux bayésiens à partir de plusieurs inférences probabilistes basés sur l'analyse d'indépendance ou conditionnelle et les scores des paramètres de modèle étudié.

Tableau 4 : Classification des inférences et apprentissages des modèles probabilistes

Inférences des modèles	**Apprentissages des modèles**
Algorithmes de classifications des modèles : • Propagation de croyance : Inférence approximative des modèles graphiques arbitraires, algorithme de passage de message générique et fonction dans un facteur graphique (calcul complexe). • Inférence variationnelle bayésienne : approximation et estimation des paramètres de modèle. Procédure itérative d'EM. La complexité de calcul. • Inférence MCMC selon l'échantillonnage de Gibbs. • Inférence de modèles graphiques probabilistes : - Modèles probabilistes hybrides - Modèles combinés et analyse factorielle des modèles gaussienne et ML) • Inférence variationnelle bayésienne. • Machine Boltzmann par une MCMC de Gibbs (algorithme analytique) ; basé sur les états des occurrences et les corrélations des variables (visibles et cachées). • Stochastique de Hopfield: modèle de réseau de neurones récurrents à temps discret. • Inférence approximative des modèles probabilistes pour les réseaux profonds (Training Deep Network).	• Structure des modèles du réseau bayésien. • Procédure de calculs des paramètres de modèles à grande échelle et des simulations physiques et en statistiques. Estimation et localisation des paramètres des modèles de probabilité (la densité des lois probabilistes) et des modèles statistiques. • Optimisation des paramètres de données incomplètes. • Traitement de lot de variables (valeurs : manquantes, latentes) dans l'espace et temps des modèles. • Apprentissage dans modèles de réseaux neuronaux avec de nombreuses couches cachées. • Apprentissage du réseau d'Hopfield dans la mémorisation associative de la stabilité de réseau.

2.3 Domaines d'applications des réseaux bayésiens dynamiques.

Dans cette partie, nous présentons au Tableau 5 une synthèse des cinq grands domaines d'applications où les approches selon les différents modèles de réseaux ont été utilisées et ce en fonction des quelques critères pertinents de recherche. Ces domaines sont :

1. Diagnostic médical : modélisation des causes de maladies par un réseau de facteurs;
2. Fiabilité et génie : analyse des causes de défaillances sous forme de réseau de facteurs;
3. Reconnaissance d'activité humaine : détection des séquences d'évènements;
4. Reconnaissance vidéo : détection de patterns préalables entre les images d'un vidéo;
5. Systèmes complexes et discrets : réseau d'interdépendances entre facteurs d'un système.

Nous élaborons ici sur ces 5 grands domaines d'application. Le tableau 5 montre et compare les approches étudiées.

Tableau 5 : Synthèse des grands domaines d'applications des réseaux bayésiens dynamiques

Critères de recherche	**Diagnostic médical**	**Fiabilité/Génie**	**Reconnaissance d'activité humaine**	**Reconnaissance vidéo**	**Systèmes complexes et discrets**
Défis/ Problématique de recherche	- Méthodes complexes à cause des données fournies par des experts ou des données incomplètes. - Contexte et situation difficile pour une décision	- États des systèmes complexes. - Méthodes correctives et préventives très coûteuses.	-États de systèmes impliquant des interactions entre les personnes, interfaces homme- machine -Événement incluant des contextes complexes simultanés	- États des mouvements temporels - Contexte d'images complexes ; floues, incomplètes, imprécises.	Système multi-états complexes comportant des dépendances, interactions contextuelles.
Objectifs poursuivis	- Progression et observation des maladies. - Prévision des risques et planification de traitements médicaux. - Aide dans la prise de décision pour les médecins ou spécialistes.	- Prévision et planification des opérations (entretiens, bonne performance des systèmes). - Assurance et augmentation de la fiabilité/sécurité des systèmes. - Réduire les coûts des opérations.	- Identification des activités pour unifier des contextes, des situations complexes -Mécanisme de prédiction d'un geste humain. - Capacité dans la prise de décision de contexte compliqué	- Identification des images pour réduire la redondance des données. - Caractérisations des images	- Prévision des risques (sécurité) - Augmentation de la performance des systèmes. - Prise de décision dans un contexte conflictuel (dépendance/ interaction des systèmes.

Types de données	- Données cliniques incluant des variables mesurées répétées et continues. - Données des soins aléatoires	- Données aléatoires sur les pannes / défaillances/usures des systèmes. - Données de fiabilité des systèmes complexes.	- Données contextuelles aléatoires. - Données réelles des activités humaines.	- Données incertaines/floues d'images ou séquences d'images de vidéo. - Données temporelles sur le mouvement d'un objet.	Données réelles multivariées du système
Types de réseaux	-Modèles graphiques. -Modèles dBN, HMM,	- Modèles dBN, et HMM. - Arbres de défaillance dynamiques.	-Modèles probabilistes graphiques - Modèle dBN unifié	-HMM, Filtre de Kalman -Modèles dBN. -Modèle non linéaire gaussien.	- HMM - Une copule de réseaux bayésiens dynamiques (DCBNs)
Interprétations	- Modèles efficaces pour faire des prévisions, des déductions et estimations des données des soins etc. - Modèles de décision ou d'aide à la décision pour les médecins. Pronostic adéquat dans la pratique.	- Modèles fiables pour des prédictions des risques et des planifications des opérations des systèmes ou sous-systèmes complexes. - Modèle efficace pour une allocation maximale des systèmes de sécurité pour réduire les risques.	- Modèles réalistes pour bien identifier et évaluer les vrais défis dans l'activité (objet ou action de l'événement) - Modèle de précision. - Modèles originaux pour un système d'aide à la décision rapide.	- Modèles efficients et vraisemblables de reconnaissances vocales, des mouvements, des gestes continus d'un objet. - Modèles exacts pour la prise de décision dans un évènement ou un scénario complexes.	Modèle DCBN complexe, nécessitant une analyse/évaluation sur les aspects (paramètres, algorithme etc.) du modèle avant d'utiliser.

2.3.1 Surveillance et diagnostic médical

Selon la revue de la littérature, les réseaux bayésiens dynamiques (dBN) ont été largement utilisés en médecine. Marshall et al. [26] ont présenté une approche plus réaliste en utilisant un modèle dynamique de réseaux bayésiens pour représenter la survie des patients atteints d'une maladie cardiovasculaire. L'article a démontré également les avantages des modèles de réseaux bayésiens dynamiques comme outils de représentations des connaissances dans la prévision de la maladie chez les patients.

Peelen et al. [27] ont étudié les domaines médicaux plus complexes tel que les soins intensifs. Les modèles de Markov discrets basés sur différentes données cliniques de patients sont développés pour faire des prédictions, des déductions et des évaluations médicales.

Sandri et al. [28] ont abordé aussi le domaine de soins intensifs. Les auteurs exposent une nouvelle approche de recherche sur la défaillance d'un organe de divers patients en utilisant les réseaux bayésiens dynamiques afin de modéliser le niveau de la gravité de maladies en suivant leur progression dans le temps. Les études réalisées sur les applications des réseaux bayésiens dynamiques comme outils dans les modélisations des systèmes de soins intensifs ont permis et aidé effectivement les médecins dans leur prise de décisions sur les patients.

Kleinberg et al. [29] ont effectué une autre recherche sur l'application des réseaux bayésiens pour le système de santé. Les modèles graphiques probabilistes et le concept de causalité de Granger sont utilisés afin d'explorer le développement et l'évolution de la maladie chez des patients. Ces modèles ont permis d’identifier, d'évaluer les symptômes et trouver des séquences d'événements indésirables ou des facteurs de risques de maladie substantiels chez les patients.

Comme approche dans le système de santé, les modèles de réseaux bayésiens dynamiques sont largement utilisés pour représenter les différents problèmes spécifiques de santé, plus particulièrement dans le diagnostic et pronostic de différentes maladies. Les applications de ces modèles probabilistes ne cessent pas de se développer dans ce système.

2.3.2 Fiabilité et diagnostic en génie

Selon la revue de la littérature, les réseaux bayésiens dynamiques ont été appliqués dans les systèmes industriels et manufacturiers complexes. Nous allons aborder quelques recherches dans l'ingénierie et le domaine de la fiabilité des systèmes.

Premièrement dans l'industrie aéronautique, Ferreiro et al. [30] ont présenté une approche en utilisant les modèles de réseaux bayésiens dynamiques dans la prévision des opérations d'entretien des aéronefs. Ces modèles ont permis d'évaluer et d'améliorer adéquatement les plans d'entretiens prévus sur les avions et de réduire les coûts des opérations.

Deuxièmement, Hu et al. [9] ont utilisé les réseaux bayésiens dynamiques dans les systèmes industriels assez complexes concernant la sécurité et la fiabilité des systèmes ou des sous-systèmes; les équipements et technologies. Ces réseaux sont largement utilisés pour en fournir des modèles de prévisions de sécurité compte tenu du caractère aléatoire des pannes/arrêts et ou les défectuosités des systèmes. En fait, l'outil développé est basé sur les réseaux bayésiens dynamiques servant à modéliser l'expansion des défauts, les problèmes sur les interactions et les relations entre les composants des systèmes ou les sous-systèmes. Dans la pratique (système énergétique), ces outils présentent un énorme avantage permettant entre autres d'éviter des accidents humains ou des dommages environnementaux.

Troisièmement, Portinale et al. [31] ont étudié les réseaux bayésiens dynamiques pour fournir un outil permettant de modéliser et d'analyser la fiabilité en ingénierie [31]. Grâce à cet outil, les ingénieurs en fiabilité peuvent faire des calculs, des analyses de fiabilité pour des systèmes plus spécifiques (entretiens, sécurité etc.)

Finalement, Khakzad et al. [32] ont montré que les réseaux bayésiens dynamiques sont réputés comme étant une méthode (algorithme) robuste pour modéliser des systèmes dynamiques dans lesquels les éléments ont une dépendance fonctionnelle et consécutive [32]. De plus, les auteurs ont démontré que les modèles de réseaux dynamiques sont appliqués efficacement dans la prévision et l'analyse des risques dans des systèmes complexes qui possèdent de multiples fonctions et tâches.

2.3.3 Reconnaissance des activités humaines

La reconnaissance de l'activité humaine est un champ de recherches très actif dans le groupement de reconnaissances et de traitements d'images. Nous proposons une revue de la littérature dans ce domaine d'application qui utilise de nombreuses méthodes ; les modèles de probabilités et, plus particulièrement les chaînes de Markov cachés (HMM) et les réseaux bayésiens dynamiques (dBN).

Premièrement, selon Aggarwal et al. [33], l'analyse du mouvement humain dans les vidéos est un défi important dans le domaine de la vision par ordinateur. L'article a fourni un aperçu détaillé sur les différentes recherches effectuées dans la reconnaissance des actions humaines. Ces auteurs ont évalué et analysé différentes approches sur les méthodologies déjà développées en comparant les performances dans différentes applications. Les applications sont nombreuses notamment pour les systèmes de la surveillance de lieux publics ou privés, les surveillances des patients, le diagnostic médical par imagerie, la recherche et l'archivage de vidéos etc.

Autre application, Wang et al. [34] ont proposé un nouveau modèle graphique probabiliste dans la reconnaissance d'un événement dans les vidéos de surveillance. Ce nouveau modèle permet d'intégrer divers contextes de l'événement simultanément: la scène, l'interaction de l'objet de l'évènement, et les contextes temporels de l'événement. Avec ce nouveau modèle, les expériences sur les données réelles de surveillance ont été réalisées dans les endroits ou milieux complexes. Les résultats obtenus démontrent que le nouveau modèle en question peut améliorer les performances de la reconnaissance de l'évènement.

Lee et al. [35] ont étudié un nouveau concept d'algorithme de prévision d'aide à la décision plus performant pour une situation ou un contexte compliqué. Ces auteurs ont proposé un algorithme dans lequel les systèmes omniprésents d'aide à la décision sont capables de prédire à l'avance les contextes et situations futures plus rapidement, précisément et pro activement [35] . Ce nouveau concept d'algorithme a été évalué sur diverses données contextuelles.

2.3.4 Reconnaissance d'évènements dans les vidéos

Plusieurs travaux de recherches concernant le suivi d'objets ou d'images dans les vidéos sont dorénavant développés. De nos jours, il est possible d'obtenir des données plus précises et fiables sur un objet en mouvement dans des situations variées. Nous proposons une revue de quelques littératures dans ce domaine d'application qui utilise différentes méthodes pour reconnaitre des événements, des actions ou réellement des interactions entre objets dans les vidéos. Les chaînes de Markov cachées (HMM) et les réseaux bayésiens dynamiques (dBN) sont les modèles les plus souvent utilisés.

Luo et al. [36] ont proposé un nouveau schéma pour l'analyse de vidéos basée sur l'objet, l'interprétation basée sur l'extraction automatique de l'objet vidéo, et la modélisation de l'évènement sémantique. Dans ce cas, le réseau bayésien dynamique est utilisé pour définir, expliquer et déterminer la nature espace-temporel des objets sémantiques. La comparaison réalisée à un modèle de Markov caché (HMM) sur le modèle bayésien dynamique permet d'obtenir une description plus détaillée des caractéristiques des objets dans la vidéo.

Autre cas plus intéressant, Suk et al. [37] ont présenté une nouvelle approche de calcul des mouvements; les gestes de la main dans un flux de vidéo continu en utilisant un modèle de réseau bayésien dynamique. Dans la même étude, les chercheurs ont également développé un algorithme de décodage en temps réel basé sur la théorie de programmation dynamique pour la reconnaissance de gestes continus. Le modèle proposé sur les gestes de la main et la conception de l'algorithme ont eu un succès potentiel dans les résultats des expériences réalisées. De plus, ces approches peuvent être utilisables à d'autres applications connexes telles que la reconnaissance du langage des signes de la main.

Yao et al. [38] ont proposé un algorithme basé sur les réseaux bayésiens dynamiques combinés avec des calculs de probabilité *« Gaussien »* concernant l'étude et le suivi en temps réel de petits d'objets visuels. La méthode proposée est formelle, l'algorithme de calcul conçu ont permis de suivre simultanément de multiples petits objets en présence d'occlusions et des interruptions selon la dynamique de l'objet. Selon les résultats obtenus des expériences réalisés, l'algorithme proposé est concurrentiel et très efficace dans les applications équivalentes.

2.3.5 Modélisation de systèmes complexes et discrets

Dans cette partie, nous allons proposer une revue de quelques littératures sur la modélisation de systèmes complexes et discrets. Les domaines d'applications présentés sont plutôt variés : la médecine, la biologie, l'économie, la finance et la fiabilité.

Dans le premier article, Donat et al. [39] ont développé un modèle général permettant de modéliser et de représenter la dynamique d'un système de fiabilité multi-états. Pour ce faire, ces auteurs ont proposé un réseau bayésien dynamique particulier, nommé modèle graphique de durée (MDG). Ce modèle a un fort potentiel de s'adapter à des systèmes multi-états évoluant au cours du temps. Dans cet article, l'application de cette approche de modélisation graphique de durée est bien présentée dans un problème de fiabilité industrielle.

Le deuxième article, Eban et al. [40] ont étudié l'apprentissage des modèles temporels pour des données d'un système à valeurs réelles multivariées. D'abord, on définit le mot " Copule" (en statistique), un objet mathématique venant de la théorie de probabilité (distribution de probabilité multivariée). Une copule permet de caractériser la dépendance entre les coordonnées d'une variable aléatoire dans R sans se préoccuper de ses lois marginales. Ces auteurs ont présenté un modèle de copule de réseaux bayésiens dynamiques (DCBNs), une modélisation visant à saisir la distribution probabiliste optimale des séquences temporelles d'un contexte d'événement complexe. Ce modèle est appliqué dans différents domaines entre autres, la biologie (neuroscience), la reconnaissance de la parole et l'économie. Ces expériences d'apprentissage ont permis de dégager les avantages qualitatifs (détecter et identifier les variables influentes) et quantitatifs (gain des données ou des informations importantes) du modèle.

Le troisième article, Enright et al. [41] ont exposé l'intégration des connaissances d'experts dans différents domaines d'applications sous une forme de modèles de calculs. Les auteurs proposent une méthode pour rassembler les connaissances (données ou informations d'expertises) sous une forme d'équations différentielles (EDO) dans les réseaux bayésiens dynamiques. Dans le système médical, cette méthode est appliquée dans un contexte réel de soins intensifs et a permis des prédictions efficaces et infaillibles dans la plupart des diagnostics d'un patient.

Enfin, dans ce dernier article, Nicholson et al. [42] ont étudié une combinaison d'approches dans le système médical. Ces auteurs ont proposé une extension d'algorithme de transition d'états (STM) avec des réseaux bayésiens dynamiques pour les outils d'aide à la décision où l'accent est mis sur la modélisation de la dynamique du système [42]. Cette approche présente des avantages explicites de l'état du système à modéliser; cependant, selon ces chercheurs, les modélisateurs doivent analyser les aspects et l'état de problèmes du système avant de l'utiliser.

2.4 Avancées théoriques récentes

Dans cette section, nous présentons une revue de littératures des avancées théoriques plus récents dans l'estimation des réseaux bayésiens dynamiques.

Dans ce premier article, Dondelinger et al. [43] ont fait une étude portée sur un modèle en biologie (cellule vivante) et les avancements portés dans son application. Dans plusieurs travaux de modélisation, on constate que les réseaux bayésiens dynamiques classiques n'ont pas la flexibilité nécessaire pour représenter des systèmes complexes permettant d'obtenir des résultats recherchés ou voulus (précis et fiables). Les chercheurs explorent, dans des réseaux complexes de la régulation des gènes biologiques, un modèle semi-flexible basé sur un réseau dynamique classique quant à une répartition des informations de gènes. Le modèle est évalué dans une série de données de gènes biologiques simulés assujetties à un environnement changeant. Dans cette étude, les avancés théoriques reposent sur les connaissances acquises d'un modèle semi-flexible sur des réseaux complexes de la régulation des gènes.

Dans le deuxième article, Doshi et al. [44] ont présenté un modèle bayésien dynamique infini (idBN), non paramétrique, prenant en compte les modèles d'état-espace qui généralisent les réseaux bayésiens dynamiques. En fait, c'est un modèle non paramétrique qui réalise un processus aléatoire indicé par le temps avec un nombre variable de facteurs (déterministes et aléatoires). Les objectifs de ce modèle sont d'étudier et d'expliquer les changements et variations possibles des valeurs de facteurs et de les prédire dans le futur. L'avancé théorique du modèle conduit à l'inférence d'une structure de données d'un processus aléatoire temporel. Les expériences réalisées sont dans la modélisation des signaux neurologiques dans le traitement de langage ainsi que dans la prévision des conditions météorologiques.

Dans le troisième article, Fenz et al. [45] ont proposé une approche fondée sur l'ontologie des domaines existants pour la construction des réseaux bayésiens. En s'appuyant sur l'ontologie, les avancés théoriques résident dans le développement des méthodes et techniques appropriées pour bâtir une structure de réseau bayésien graphique (nœuds et relations), ainsi pour préserver les contraintes sémantiques de l'ontologie et concevoir les faits de connaissances déjà existantes.

Autre revue de littérature, Gao et al. [46] ont proposé pour les systèmes nommés "intelligents", l'utilisation des réseaux bayésiens dynamiques pour faire l'inférence (calcul) exacte ou approximative dans les contextes difficiles et incertains est complètement lourd et très coûteux. Ces auteurs ont étudié une approche *"sliding window"*, un algorithme pour l'inférence approximative dans les réseaux bayésiens dynamiques pour réduire la charge de calcul. Les connaissances acquises ont permis d'adapter et d'intégrer cet algorithme d'inférence dans les réseaux bayésiens dynamiques, c'est l'essentiel de l'avancé théorique dans cette approche.

On sait que les réseaux bayésiens dynamiques classiques sont fondés sur l'hypothèse de la théorie de la chaîne de Markov homogène et ne peuvent pas faire face à des processus temporels non-homogènes. Depuis quelques années, plusieurs approches ont été proposées et étudiées pour abandonner cette hypothèse d'homogénéité. Dans ce dernier article, Grzegorczyk et al. [47] ont présenté une combinaison de deux types de réseaux bayésiens dynamiques dans un contexte complexe. Le premier est un réseau bayésien avec des probabilités conditionnelles dans la famille linéaire gaussienne. Le deuxième est un réseau multiple bayésien dynamique utilisant la distribution de probabilité a posteriori avec la chaîne de Markov Monte Carlo (MCMC). Les avancées théoriques de cette combinaison sont les améliorations et les modifications apportées au niveau de la méthodologie et des algorithmes de calculs essentiels qui permettent l'évolution des structures de réseaux.

Conclusion

Selon la revue de la littérature visée, la théorie de Bayes (axiomes et règles) a démontrée être un outil très puissant de prévision, de décision ou d'aide à la décision avec la révolution des systèmes industriels et des technologies d'informations. De ces littératures, nous révisons les diverses approches statistiques et probabilistes des données, les réseaux (BN et dBN) qui ont permis d'établir des connaissances incertaines et éventuelles plus formelles. Également, par ces diverses approches nous nous situons au cœur du domaine de l'intelligence artificielle [48].
Les nombreuses recherches et travaux réalisés dans plusieurs domaines d'applications visés ont fait preuve que les industriels, les chercheurs et la communauté scientifique sont vraiment intéressés à ces types de réseaux. Ceux-ci ont des points communs importants entre les domaines respectifs.
Nous évaluons les grands défis des avancés théoriques récents dans l'estimation de ces réseaux, plus particulièrement les réseaux bayésiens dynamiques. Le projet de recherche porte sur la détection de la structure causale des risques dans les systèmes industriels complexes par la méthode des réseaux bayésiens dynamiques.

Afin de démontrer cette méthode, le développement d'une nouvelle interface utilisateur pour les systèmes industriels de contrôle et d'acquisition de données (SCADA) est étudié. Il s'agit d'un système industriel dit complexe par le nombre de tâches et acteurs (ingénieurs, opérateurs, gestionnaires, experts etc..) en interactions, l'importance de la gestion des risques et de la surveillance en temps réel pour l'assurance des politiques de fiabilité industrielle.

Le chapitre suivant décrit la méthodologie adoptée sur l'utilisation des réseaux bayésiens dynamiques d'un système SCADA. L’objectif de recherche sera la priorisation des risques d'échec du système à l'aide de ces réseaux.

CHAPITRE 3 - Proposition d'étude

Introduction

Ce chapitre présente deux sections. Dans la première, nous présentons notre contribution scientifique attendue. Le développement des interfaces dans les systèmes SCADA est étudié. A travers la revue de littérature et des travaux de recherches réalisés par des chercheurs, nous focalisons sur des pratiques multidisciplinaires de processus d'ingénierie moderne comme des méthodes stochastiques assimilées par les systèmes SCADA. L'utilisation des réseaux bayésiens dynamiques pour les systèmes SCADA comme outil de modélisation de prédiction des défaillances/pannes des systèmes est abordé.

Dans la deuxième section, nous décrivons notre contribution scientifique de la thèse. Nous présentons la méthodologie retenue et la démarche proposée en fonction du plan de recherche. Nous détaillons et analysons les sources des données qui ont été compilées par des auteurs du Royaume-Uni et obtenues d'une usine de fabrication des semi-conducteurs [49]. Nous démontrons comment la méthode pourrait aider à mieux orienter l'analyse des risques pour le développement des interfaces intelligentes. Aussi, cette section comprend les hypothèses de recherche élaborées. Nous expliquons le choix, le développement et la configuration des outils de traitement et d'analyse ainsi que les livrables attendus de la recherche.

3.1 Contribution scientifique attendue

Cette thèse s’oriente vers une nouvelle méthode d'analyse de la structure causale des risques dans les systèmes industriels complexes par les réseaux bayésiens dynamiques. L'objectif visé ultimement est de développer une nouvelle interface utilisateurs pour les systèmes industriels de contrôle et d'acquisition de données (SCADA).

La nouvelle interface utilisateur est un outil d'aide à la décision, de prévision des risques et de proposition des solutions originales plus efficaces et plausibles face aux scénarios éventuels des systèmes de production assez complexes. Ces modèles permettent aux ingénieurs de production de réagir très rapidement et de prendre des décisions spontanées liées aux facteurs de risques et fiabilité des systèmes.

Les nouveaux modèles seront expliqués et évalués en fonction des critères pertinents suivants : les défis à relever du système industriel, les objectifs poursuivis et ciblés, le nombre et volume de données traitées, les types de réseaux de modélisation ainsi que les explications et interprétations des résultats et des actions quant à la tendance de l'évolution des besoins des systèmes industriels. C'est une avenue de recherche très favorable et prometteuse dans les systèmes industriels concurrentiels et dynamiques, de nos jours, comme le domaine de l'industrie des semi-conducteurs.

Nous voulons soulever les problématiques de modélisation des prévisions d'un processus industriel qui permettent l'intégration des systèmes SCADA. Plus précisément, nous voulons apporter des solutions pour prédire les défaillances/pannes en utilisant des informations ou des données SCADA de l'industrie, de façon à améliorer la politique de maintenance des systèmes ou sous-systèmes de production en temps réel.

3.1.1 Développement des interfaces des systèmes SCADA

Dans cette thèse, le développement d'une nouvelle interface utilisateur pour les systèmes SCADA est étudié dans le domaine des semi-conducteurs. Il s'agit d'un système dit complexe suite au nombre de tâches et acteurs en interactions, et à l'importance de la gestion des risques [50] et de la surveillance en temps réel pour l'assurance des politiques de fiabilité industrielle.

L’une des exigences les plus importantes et un défi opérationnel pour l’équipe d’ingénieurs de production, est de facilement et rapidement repérer l’émergence spontanée des facteurs de risque et fiabilité, et ce, souvent, en équipes distribuées. Bien que beaucoup de systèmes SCADA possèdent des interfaces dynamiques permettant ce type d’analyse, il reste qu’ils sont pour la plupart basés sur des analyses historiques et sur des modèles de corrélation statiques entre les variables de fiabilité.

3.1.2 Étude bibliographique : utilisation des réseaux bayésiens dynamiques dans les systèmes industriels.

Il existe plusieurs représentations pour extraire des connaissances à partir des données (arbres de décision, réseaux de neurones, etc..). Il en est de même pour les techniques de fouille de données (classification, régression, filtrage, lissage etc.).

Dans le domaine industriel, l'utilisation des réseaux bayésiens dynamiques a attiré beaucoup l’attention de plusieurs chercheurs en raison de leurs propriétés temporelles et relationnelles pour mieux modéliser les systèmes industriels.

Nous décrivons les différents travaux de recherches réalisés par des scientifiques et des chercheurs sur l'utilisation des réseaux bayésiens dynamiques dans les systèmes industriels.

3.1.2.1 Travaux de Ferreiro, Arnaiz, Sierra et Irigoien

L'objectif visé des travaux de recherche de ces auteurs [30] est de démontrer l'utilité des modèles bayésiens dans la stratégie de maintenance des aéronefs comme étant une méthode de pronostic.
Cette étude concerne les mesures de maintenance des systèmes aéronefs, plus précisément les freins, les trains d'atterrissages et autres composants de ces deux systèmes. Les nouveaux modèles de réseau bayésien dynamique au niveau des modules et fonction ont été développés pour la maintenance prédictive. Les modèles sont nommés : proactifs du système.
Les modèles mettent en œuvre différents modules et fonctionnalités de probabilités nécessaires pour permettre le passage au diagnostic; la détection de l'état vers le pronostic.

Les modèles bayésiens dynamiques estiment et établissent divers paramètres de dégradation et de durée de vie résiduelle des freins et des trains d'atterrissages, ainsi que des autres composants à travers le plan opérationnel de l'aéronef.
Les résultats obtenus des travaux de recherches sont nombreux et favorables pour l'aéronautique et les domaines connexes. Le modèle prédit d'environ 0,11 mm d'usure de freins par vol, ce qui est la moyenne de la dégradation lors d'un atterrissage normal. Le nouveau modèle associe la prédiction de l'usure des freins avec un niveau de confiance de 95%. La marge d'erreur du modèle lors de l'exécution est négligeable. Le modèle offre une approche prospective pour l'état de l'aéronef.

Cependant, les études réalisées démontrent certaines limitations. Dans l'ensemble de processus, les modèles de réseau bayésien ne tiennent pas comptes d'autres éléments technologiques émergents. Ainsi, d'autres développements doivent être intégrés et considérés lors du développement des modèles de réseaux bayésiens dynamiques. De plus, la précision des modèles pour son utilisation et efficacité est basée uniquement sur les paramètres du plan d'exploitation de l'aéronef. On constate également l'absence d'informations connexes qui pourraient affecter les résultats des études.

3.1.2.2 Travaux de Hu, Zhang, Ma, et Liang

Les travaux de recherche de ces auteurs [9] proposent une approche pour résoudre les problèmes des propagations des défauts dans des systèmes industriels complexes. L'objectif visé est d'estimer l'état de sécurité et les risques potentiels du système considéré au moyen de la modélisation.

L'étude concerne de cas réel fait sur une turbine à gaz des systèmes de compresseurs. Elle est centrée sur une évaluation rigoureuse de l'état de sécurité et les risques potentiels de l'ensemble du système. Un modèle de prévision de sécurité intégrée (ISPM) utilisant les réseaux bayésiens et l'algorithme de colonie de fourmis ont été développé. Le modèle en question comprend l'intégration des différents éléments tel que : la méthode HAZOP, les modèles de dégradation (défaut), de surveillance et de l'évaluation des risques de prédiction, les causes dans les réseaux bayésiens dynamiques. L'interaction et la dépendance entre les entités dans les systèmes de compresseurs sont considérées et aussi modélisées par des réseaux bayésiens dynamiques.

Cependant, cette étude présente une certaine limitation dans son exécution, entre autres, le temps de calcul pour les fonctions de probabilités du modèle est énorme. De plus, l'application du modèle intégré dans un autre système industriel complexe équivalent pourrait être inapproprié et avoir des résultats non concluants.

3.1.2.3 Travaux de Portinale et Raiteri

Les travaux de recherche des auteurs [31] proposent une approche de modélisation sur la fiabilité des systèmes basée sur une conversion automatique d'un modèle particulier des arbres de défaillances dynamiques en réseaux bayésiens dynamiques.

L'objectif visé de la recherche est de fournir une interface familière aux ingénieurs de fiabilité en leur permettant de modéliser et analyser les systèmes. Le développement des algorithmes simples et modulaires est mis en œuvre pour compiler automatiquement les arbres de défaillances dynamiques dans les réseaux bayésiens dynamiques. Ces auteurs ont démontré que l'utilisation des réseaux bayésiens dynamiques permet à l'utilisateur d'être en mesure de calculer les valeurs des paramètres qui ne sont pas calculables à partir des arbres de défaillance dynamiques. Pourtant ces valeurs sont obtenues correctement à partir de l'inférence de réseaux bayésiens dynamiques.

L'approche de modélisation par les réseaux bayésiens dynamiques est testée sur des cas précis. Les avantages d'avoir un moteur d'inférence complet et performant basé sur les réseaux bayésiens dynamiques sont démontrés pour les tâches d'analyse demandées.

Hormis, la discrétisation et l'effort sur le calcul, les temps nécessaires pour l'analyse de fiabilité des systèmes sont énormes.

3.1.2.4 Travaux de Khakzad, Khan et Amyotte

Les travaux de recherche des auteurs [32] proposent une technique robuste de modélisation pour analyser la fiabilité des systèmes dynamiques Cette étude concerne des systèmes industriels complexes dans lesquels la dépendance séquentielle des composants et leur dépendance fonctionnelle doivent être considérées.
Les réseaux bayésiens dynamiques à temps discret ont été proposés comme solution pour résoudre le problème des arbres de défaillances dynamiques sans avoir recours à des chaînes de Markov. L'utilisation des réseaux bayésiens dynamiques permet de surmonter les problèmes de l'explosion d'espaces d'états du système et la procédure de conversion des erreurs d'arbre de défaillance dynamique.

Pour cette étude, un algorithme de dépendance neutre est développé permettant de modéliser les entrées dynamiques du système. Cet algorithme permet d'éviter la grandeur de la table de probabilité conditionnelle et d'obtenir les distributions de probabilité d'échec du système.
Les modèles des entrées dynamiques contribuent dans la conception et le suivi de processus à temps réel du système étudié. Ces modèles sont grandement utilisés comme un outil inductif pour analyser les défaillances à l'aide des nouvelles observations du système. Ils sont aussi un outil d'évaluation efficace pour une allocation optimale de sécurité afin de réduire les risques ou dangers encourus du système.

Pourtant, le seul inconvénient est l'impossibilité d'appliquer les modèles à temps continu dans l'évaluation de la sécurité et l'analyses des risques des systèmes complexes.

3.1.2.5 Travaux de Takesiha Khoda et Cui

Les travaux de recherche des auteurs [50] consistent à reconfigurer la logique d'un système de surveillance de sécurité pour des systèmes dynamiques. Cette reconfiguration est basée sur des critères de risque du système étudié en utilisant la modélisation par des réseaux bayésiens temporels.

Cette étude concerne des capteurs pour des systèmes de sécurité industriels. Les auteurs ont développé des modèles de réseaux bayésiens statiques et dynamiques pour diagnostiquer les défaillances/défauts, ainsi les causes des évènements anormaux et ses effets dans les systèmes considérés. Ces modèles vérifient la relation et le comportement dynamique des états de surveillance de la sécurité du système.

Les résultats attendus des nouveaux modèles sont focalisés pour le système étudié. Les informations et les fonctions de probabilités sont claires et précises.

Par contre, les réseaux bayésiens statiques sont dédiés seulement au diagnostic ou à l'identification de la cause des défaillances ou des défauts des capteurs. Ces modèles statiques ne fonctionnent pas pour examiner l'effet de diagnostic sur le fonctionnement des capteurs.

La reconfiguration des modèles par des réseaux bayésiens dynamiques en termes d'analyse des risques n'est pas assurée sur d'autres systèmes dynamiques comparables.

3.1.2.6 Travaux de Medjaher, Mechraoui et Noureddine

Les travaux de recherche de ces auteurs [10] proposent une approche de modélisation qui permet de localiser les défaillances d'un système complexe et de prédire les éventuelles dégradations qui pourraient l'affecter. Cette étude concerne un système électrique ; un moteur électrique à aimants dans l'automobile. Les chercheurs ont développé des modèles de réseaux bayésiens statiques pour diagnostiquer les défaillances du moteur et par la suite l'extension par des modèles dynamique pour prédire les défaillances.

Les résultats obtenus de cette étude sont concluants. Pourtant, la limitation de cette recherche est sur l'anticipation des défaillances du moteur. Cette anticipation n'a pas été faite, ce qui aurait complété l'étude.

3.1.2.7 Travaux de Delcroix, Maalej et Piechowiak

Les travaux de recherche des auteurs [51] sont focalisés sur le développement des modèles graphiques probabilistes pour les diagnostics des pannes simples et multiples des systèmes complexes de grande taille. Cette étude est réalisée sur des grands composants et des dispositifs d'une automobile. Les diagnostics des pannes sont à la fois uniques et multiples.

Les auteurs ont développé un algorithme d'inférence probabiliste de diagnostic utilisant des réseaux bayésiens. En fait, l'algorithme regroupe toutes les informations d'un composant ou dispositif sur un même modèle. De plus, cet algorithme sépare aussi distinctement l'état (normal ou panne) ainsi que les conséquences engendrées par ces états sur les variables de sorties d'un composant ou dispositif. Le nouveau modèle permet de calculer les probabilités des défaillances à posteriori de chaque composant ou dispositif.

Les limites de cette étude sont les suivantes :

- Le nouveau modèle ne permet pas de calculer directement les probabilités conjointes des diagnostics (simples ou multiples) d'un composant ou dispositif.
- L'utilisation des réseaux bayésiens se trouve limitée par l'impossibilité d'utiliser l'algorithme d'inférence dans un temps raisonnable.
- Le pourcentage d'erreurs de l'algorithme doit être considéré dans son application.

3.1.2.8 Travaux de Parisot, Boussemart et Simon

Les travaux de recherche des auteurs [52] visent à optimiser la politique de maintenance d'un système de production à travers un modèle utilisant le réseau bayésien dynamique.

Cette étude concerne des cas concrets d'un système de production dans une industrie. Les auteurs ont développé un modèle de maintenance par simulation. Le nouveau modèle permet de déterminer la probabilité du bon fonctionnement des équipements à l'issue d'un certain nombre de cycles accomplis pour chaque configuration du système de production. Le modèle vérifie également l'effet des modes de dégradations du système permettant d'analyser leur disponibilité. Ce nouveau modèle demeure un outil efficace et incontournable d'aide à la décision. En effet, l'application du modèle emmène à la décision de réparation ou de non réparation d'un équipement de production en incluant d'autres critères pertinents du système.

Hormis, pour un système de production complexe, le modèle est difficilement applicable à cause de l'explosion combinatoire des états ou le comportement du système.

3.1.3 Utilisation des réseaux de bayésiens dynamiques dans un SCADA

Dans les années 70, les systèmes SCADA étaient à l'origine des systèmes ou sous-systèmes autonomes, isolés et de source propriétaire. Cependant, les systèmes SCADA ont énormément évolué et sont devenus assez complexes depuis les vingt dernières années dans plusieurs domaines industriels.

Nous abordons à travers de la revue de littérature la construction des réseaux bayésiens dynamiques comme des réseaux d'interfaces sur des modèles de prévisions de défaillances et de fiabilité à partir des données SCADA pour des systèmes et sous-systèmes complexes.

Plusieurs revues et des travaux de recherche démontrent l'utilisation de réseaux dBN comme une méthode pour l'analyse des risques [53] d'échec/panne dans des processus industriels complexes. Ces autres chercheurs utilisent les réseaux dBN pour faire un diagnostic des défaillances [54] et la détection des défauts [55] pour les systèmes industriels SCADA. Les arbres de défaillances dynamiques (DFT) en réseaux bayésiens dynamiques (dBN) ont été présentés. Les chercheurs ont démontré comment l'approche fonctionne sur des données précises en décrivant les avantages d'avoir à la disposition un moteur d'inférence complet basé sur les réseaux bayésiens dynamiques pour les tâches demandées. Cependant, le vrai défi repose sur la question suivante : est-ce-que les algorithmes utilisés suivent justement la réalité dans un système complexe dont les incertitudes technologiques des systèmes d'informations sont présentes ?

D'autres chercheurs utilisent les réseaux dBN pour fournir les modèles réels des systèmes en termes de diagnostic et pronostic des pannes [55; 54; 10; 56] dans divers systèmes industriels modernes et à grande échelle.

Pour plusieurs de ces chercheurs, les réseaux bayésiens temporels ont gagné en popularité comme étant une technique robuste pour modéliser des systèmes dynamiques dans lesquels la dépendance séquentielle des composants et leur dépendance fonctionnelle ne peuvent pas être ignorées. Ces chercheurs ont proposé les dBN discrets comme une alternative viable pour résoudre les arbres de défaillances dynamiques sans avoir recours à des chaînes de Markov. Selon les chercheurs, cette approche permet de surmonter les inconvénients des chaînes de Markov tels que : l'explosion de l'espace d'état et la procédure de conversion des erreurs des arbres de défaillances dynamiques.

Toutefois, la représentation graphique des arbres de défaillances dynamiques dans les réseaux bayésiens présente des limitations et un certain niveau de risques contextuels. Par conséquent, l'utilisation des réseaux bayésiens dynamiques (dBN) fait encore un sujet de recherche à explorer en profondeur concernant les modèles de prédiction des pannes et la disponibilité des systèmes industriels très complexes.

Parmi les travaux de recherche mentionnés ci-haut, les réseaux dBN ont donné des bons résultats, entre autres, dans la gestion des politiques de maintenance et de fiabilité des systèmes compliqués. En fonction des données compilées des événements des défaillances, les calculs de mesure de la fiabilité spécifiques des systèmes et l'utilisation des algorithmes de l'inférence par réseaux bayésiens dynamiques (dBN), les chercheurs en sont arrivés à calculer divers variables et paramètres du système demandé. En résumé, des réseaux innovants de dépendance neutre sont introduits pour modéliser des séquences d'événements dynamiques du système considéré en évitant ainsi la dimension des tables de probabilités conditionnelles.

Pour certains chercheurs l'utilisation des réseaux bayésiens dynamiques (à temps discret) ont permis une bonne analyse et une évaluation des risques de la sécurité des systèmes industriels complexes. Ces chercheurs ont démontré comment la méthode bayésienne dynamique peut être efficacement appliquée pour une allocation optimale des systèmes de sécurité pour obtenir la réduction maximale des risques.
Dans un site de production d'énergie, en s'inspirant de la prévision des probabilités d'échec des systèmes [57], un modèle a été développé qui intègre les données de surveillance d'état. Ce nouveau modèle a conduit à des prévisions de la demande de pièces plus précises et efficaces dans les systèmes. En plus, le modèle a permis de réduire et de quasi éliminer les risques encourus du système en augmentant davantage la performance dans son ensemble.

Dans certains domaines industriels spécifiques, l'utilisation des réseaux dBN pour les systèmes SCADA est encore limité et contraignant. Les problèmes sont concrets dans la gestion des risques des installations complexes, plus précisément dans la gestion des données où la sûreté de fonctionnement est cruciale pour la sécurité des personnes et de l'environnement, comme dans les industries de l'aéronautique ou du nucléaire. En effet, les redondances matérielles [11; 10] sont répandues et problématiques.

L'analyse de la priorisation des risques potentiels industriels pour les systèmes SCADA doit considérer et comprendre l'utilisation des données assez qualitatives et précises dans les domaines des industries modernes.

Bref, l'exploration de la capacité de la modélisation par dBN pour détecter la causalité des risques de défaillances ou pannes d'un système assez complexe est étudiée. Les hypothèses de recherche sont vérifiées et validées.

3.1.4 Hypothèses de recherche

Pour résoudre les problèmes de défaillances, de contrôle et de surveillance dans les systèmes industriels assez complexes, une approche inclusive pour le système SCADA en utilisant les réseaux bayésiens dynamiques est proposée.

Cette thèse met l'accent sur le développement des interfaces dans un système d'information SCADA des semi-conducteurs. Nous voulons prioriser les risques d'échec du système par des représentations graphiques. Les méthodes d'analyse des risques d'échec dépendent d'un petit nombre d'experts de systèmes ou d'ingénieurs de fiabilité. Ces ressources peuvent identifier et calculer les tendances ainsi que les comportements des systèmes en sélectionnant les données de diagnostic et les informations dans les procédures de prévisions de maintenances. De manière plus précise, les interfaces tiennent compte des hypothèses restrictives permettant de mieux refléter la réalité. Ces hypothèses sont :

H1 : Le réseau de variables avant et après la défaillance est représenté par un nombre limité et distinct de facteurs.

H2 : Le réseau de variables avant et après la défaillance peut être représenté graphiquement de manière dynamique dans une interface utilisateur pour aider la prévention et le diagnostic des pannes.

H3 : Les variables liées à la séquence d'évènements au moment de la défaillance peuvent être utilisées comme modèle pour prévoir son occurrence (dont la qualité de la prévision est évaluée par la mesure F1), et trouver la principale cause de celle-ci, permettant ainsi de prioriser les exigences du système de production sur les bonnes variables à surveiller et gérer en cas de panne.

3.2 Méthodologie

Nous adoptons une méthodologie plus quantitative que qualitative. Les points clés de notre méthodologie de recherche reposent sur le concept de traitement des informations et d'assimilation des données SCADA provenant d'un processus de fabrication des semi-conducteurs en utilisant des outils statistiques.

Notre méthodologie démontre comment notre approche peut aider à mieux orienter l'opération du système et la gestion des risques industriels. La méthodologie comprend deux phases à savoir : la préparation des données et les tests des hypothèses de recherche. Ainsi, notre démarche en laboratoire utilise le logiciel R comme plateforme de développement.

3.2.1 Sources des données

Les sources de données de notre étude proviennent d'une usine de fabrication des semi-conducteurs, obtenues par des chercheurs du Royaume-Uni. Le jeu de données d'un SCADA est publié depuis 2010 sur le site web de *"University of California at Irvine"*.

À ce jour, deux articles de publications par différents chercheurs ont été présenté**s**. Le premier article, Mc Cann et al. [49] effectuent une analyse comparative des composants pertinents pour la surveillance et le contrôle efficace de processus dans la fabrication de semi-conducteurs. L'auteur a étudié une gamme de techniques de sélection sur les caractéristiques des effets de causalité rencontrés par les ingénieurs de surveillance et contrôle de processus de production. Les modèles d'analyse de réduction des composants pertinents ont été développés en utilisant la technique d'apprentissage d'arbre de décision simple.

Le deuxième article, Munirathinam et al. [1] proposent des méthodes d'apprentissages automatiques pour générer un modèle prédictif des défaillances du matériel au cours du processus de fabrication des semi-conducteurs. Le modèle de prédiction a été développé à partir des états d'évènements de défaillances unitaires du système. L'évaluation de la performance du nouveau modèle prédictif est basée sur les scores de prédiction.

Le travail de thèse concerne les mêmes sources de données avec des nouvelles contributions de recherche très distinctives. Les travaux de recherche proposent une méthode de sélection et de réduction des composants pertinents avec l'utilisation de la technique d'apprentissage automatique de Forêts Aléatoires qui combine le concept des sous-espaces aléatoires et la technique d'inférence statistique *"Bootstrap"*. L'étude sur les modes défaillances et les modèles de prédiction ont été fait sur deux (2) états du système, suivi par quatre (4) états étendus des séquences d'évènements, ainsi que six (6) différents scénarios combinant les quatre états étendus du système. L'évaluation de la performance des nouveaux modèles de prédiction est basée sur les scores de prédiction et surtout sur les degrés des réseaux bayésiens dynamiques.

3.2.2 Sélection et préparation des données

La sélection et préparation du jeu de données se révèlent difficile selon les méthodes utilisées et la quantité de données du système SCADA étudié.

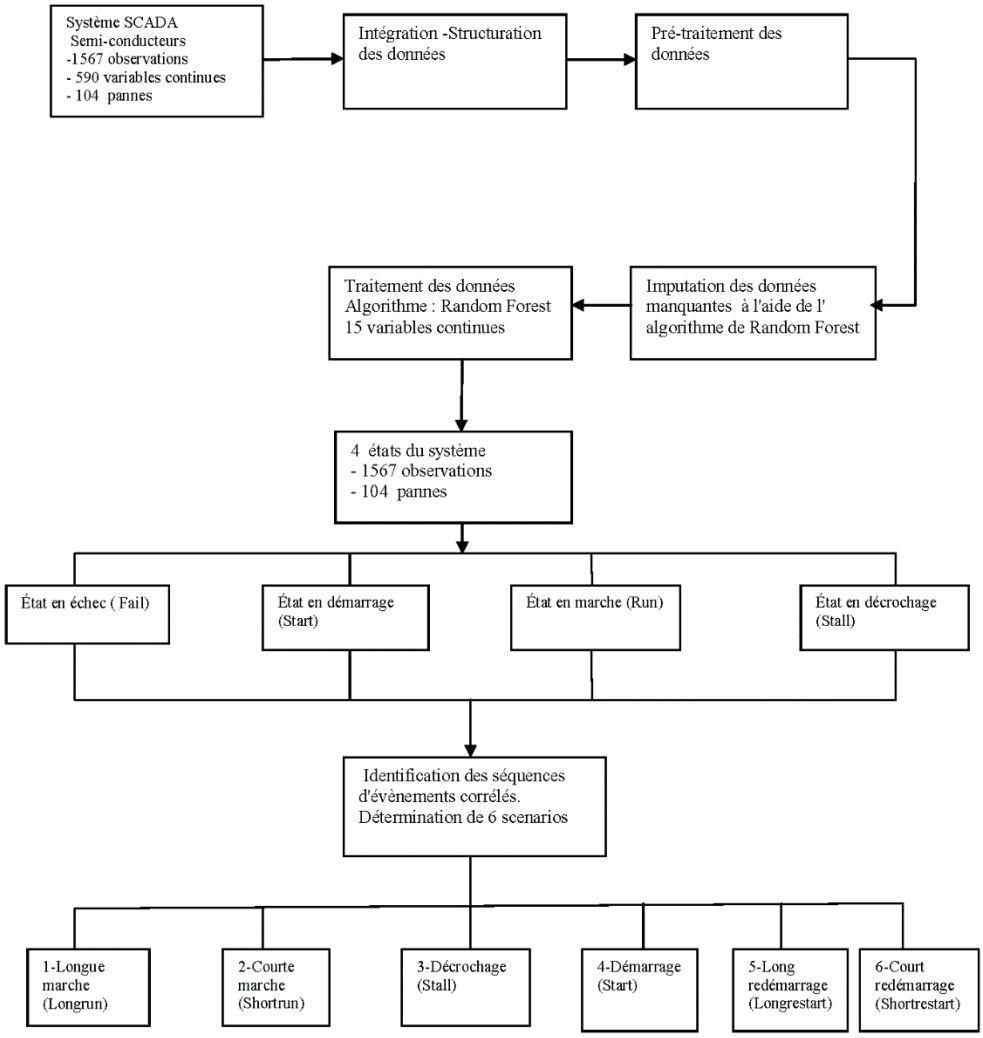
Le tableau 6 illustre un extrait de données du système de fabrication de semi-conducteurs dont 1567 observations, 590 variables de V001 à V590 qui indiquent chacune de variable une activité du système de production et 104 défaillances/pannes.

Tableau 6 : Extrait de données du système de fabrication de semi-conducteurs

Rang	Temps Début	Semaine	Échec	Durée	Statut	État	Stabilité	ObsAvDerÉ	SecDurDerÉ	SDDepDÉ	V001	V002	V003	V004
1	0	1	0	0	3-Marche	4-Décrochage	2-Average	0	0	0	3030,93	2564	2187,7333	1411,1265
2	2220	1	0	2219	3-Marche	4-Décrochage	2-Average	0	0	2220	3095,78	2465,14	2230,4222	1463,6606
3	4920	1	1	2700	1-Échec	1-Échec	2-Average	2	4920	4920	2932,61	2559,94	2186,4111	1698,0172
4	10080	1	0	5159	1-Échec	2-Démarrage	2-Average	2	4920	5160	2988,72	2479,9	2199,0333	909,7926
5	12420	1	0	2340	3-Marche	2-Démarrage	2-Average	2	4920	7500	3032,24	2502,87	2233,3667	1326,52
6	21480	1	0	9060	3-Marche	2-Démarrage	2-Average	2	4920	16560	2946,25	2432,84	2233,3667	1326,52
7	28140	1	0	6660	3-Marche	3-Marche	2-Average	2	4920	23220	3030,27	2430,12	2230,4222	1463,6606
8	28200	1	0	60	3-Marche	4-Décrochage	2-Average	2	4920	23280	3058,88	2690,15	2248,9	1004,4692
9	30540	1	0	2339	3-Marche	4-Stall	2-Average	2	4920	25620	2967,68	2600,47	2248,9	1004,4692
10	34800	1	0	4260	3-Marche	4-Stall	2-Average	2	4920	29880	3016,11	2428,37	2248,9	1004,4692
11	36120	1	1	1320	1-Échec	1-Échec	2-Average	7	31200	31200	2994,05	2548,21	2195,1222	1046,1468
12	39420	1	1	3300	1-Échec	1-Échec	2-Average	0	3300	3300	2928,84	2479,4	2196,2111	1605,7578
13	56400	1	0	16979	3-Marche	2-Démarrage	2-Average	0	3300	16980	2920,07	2507,4	2195,1222	1046,1468
14	159960	1	0	103560	3-Marche	4-Décrochage	2-Average	0	3300	120540	3051,44	2529,27	2184,4333	877,6266
15	172680	1	1	12720	1-Échec	1-Échec	2-Average	2	133260	133260	2963,97	2629,48	2224,6222	947,7739

ObsAvDerÉ : Observation AvantDernierÉchec
SDDerÉ : SecondeDuréeDernierÉchec
SDDepDÉ : SecondeDuréeDepuisDernierÉchec

Figure 2 : Préparation des données



Les étapes de la préparation de données sont illustrées à la figure 2, et s'interprètent ainsi :

1. Intégration et structuration des 2 sections de données originales, lesquelles sont les 590 variables des activités du système, suivi par l'identification de l'état du système en marche ou panne du système.
2. Calcul de différentes variables de temps, en secondes, identifiant la durée entre les observations et depuis la dernière panne.

3. Imputation de toutes les données manquantes par le biais d'un algorithme de forêts aléatoires *(Random Forest)* tel qu'implémenté dans le logiciel R : le *" MissForest "* et le VSURF (Variable Selection Using Random Forest) suivi par la vérification manuelle de la qualité et cohérence de l'imputation. Nous utilisons cette approche donnant un plus petit nombre de variables pour un taux de performance équivalent aux autres méthodes [58].
4. Première réduction du nombre de variables avec VSURF de 590 à 7 variables, utilisant les deux (2) états du système; Échec /Fail et en Marche/Run. Deuxième réduction du nombre de variables avec VSURF de 590 variables à 20 variables, utilisant cette fois-ci les quatre (4) états du système: Échec (Fail), Démarrage (Start), en Marche (Run) et en Décrochage (Stall).
5. Annotation manuelle de 4 états du système, autour des 104 pannes, pour chacune des 1567 observations :

    1-État en Échec (Fail) : 104 observations ;

    2-État en Démarrage (Start) : 208 observations, les 3 suivantes après un état 1-échec ;

    3-État en Marche (Run) : 1067 observations, suite consécutive d'états 3-en marche suivant les états 2-en démarrage ;

    4-État en Décrochage (Stall) :188 observations, les 3 précédentes avant un état 1-échec.
6. Annotation manuelle de 6 scénarios combinant les 4 états et constitution de jeux de données secondaires, chacun contenant un nombre de séquences d'évènements se répétant à plusieurs endroits dans le jeu de données :

    1. Longue marche (Longrun) : 88 cas comportant des séquences de 10 états en marche ;

    2. Courte marche (Shortrun) : 21 cas comportant des séquences de 3 états en marche ;

    3. Décrochage (Stall) : 54 cas comportant des séquences de 3 états en décrochage;

    4. Démarrage (Start) : 55 cas comportant des séquences de 3 états en démarrage ;

    5. Long redémarrage (Longrestart) : 31 cas comportant des séquences de 7 états dont 3 états en décrochage, 1 état en échec et 3 états en démarrage ;

6. Court redémarrage (Shortrestart): 63 cas comportant des séquences de 3 états dont 1 état en décrochage, 1 état en échec, et 2 états en démarrage.

3.2.3 Vérification et validation des hypothèses

La vérification et validation des trois hypothèses découlent à partir de la construction des réseaux bayésiens statiques et dynamiques. Nous avons mis à œuvre un algorithme nommé : EBDBN (Empirical Bayes estimation of Dynamic Bayesian Networks) proposé et développé par des chercheurs [59]. Il s'agit d'un algorithme itératif qui utilise le filtre de Kalman pour estimer les distributions à posteriori d'un modèle d'espace d'état ou un système dynamique linéaire, complémenté d'un modèle non-paramétrique dont l'estimation empirique permet de laisser varier les paramètres en tout temps, donc requérant l'estimation d'hyper-paramètres.

En résumé, nous avons introduit cet algorithme dans notre contribution de recherche des modèles du système étudié. La construction des réseaux sont faits avec le logiciel R comme étant une plateforme de développement. Ce logiciel est adapté pour la modélisation et l'apprentissage des différents modèles d'états statiques ou dynamiques du système.

3.2.4 Limites de la méthode des réseaux bayésiens statiques

La méthode des réseaux bayésiens est considérée comme une représentation de causes à effet statique des relations entre les composants du système. Les réseaux bayésiens statiques permettraient de vérifier notre hypothèse H1.

Cependant, les réseaux bayésiens statiques ne permettent pas de modéliser les liens entre les variables de façon intuitive et ou intentionnelle des composants de système afin d'améliorer les méthodes de priorisation des risques du système étudié.

Les limites de l'utilisation de la méthode bayésienne statique reposent uniquement sur la modélisation des liens entre les variables à temps donné ou établi. En effet, les réseaux bayésiens statiques sont des méthodes pour détecter et localiser les défaillances ou les pannes.

Par contre, il est possible d'utiliser une méthode graphique temporelle, les réseaux bayésiens dynamiques [60], pour les systèmes industriels complexes où des connaissances expertes et des informations importantes tel que : des observations, des tests périodiques , des vérifications etc. sont disponibles. Ces réseaux permettent de représenter graphiquement cette connaissance en tenant compte des incertitudes technologiques et les risques potentiels du système.

Le chapitre suivant présente la méthodologie d'analyses des réseaux et les critères d'évaluations des résultats obtenus.

Conclusion

Les contributions de cette thèse sont récapitulées ainsi : la sélection des variables pertinentes du système SCADA considéré par classification avec l'algorithme des forêts aléatoires. Suivi par une analyse de prévention des risques de défaillances ou pannes des états du système utilisant des représentations graphiques statiques et particulièrement dynamiques où des liens inter temporels entre les variables du système.

L'utilisation de l'approche bayésienne statique ou classique est limitée pour un système où le nombre des opérations et des acteurs en interactions sont variés et conflictuels. Par contre, les avantages sur l'utilisation des réseaux bayésiens dynamiques sont nombreux et pourrait aider à mieux orienter les opérations du système et la gestion des risques industriels.

Ainsi, nous présentons dans le chapitre suivant la méthodologie d'analyses et évaluations des résultats attendus des réseaux bayésiens statiques et dynamiques pour les modèles d'états du système étudié.

CHAPITRE 4 - Analyse des modèles d'états du système

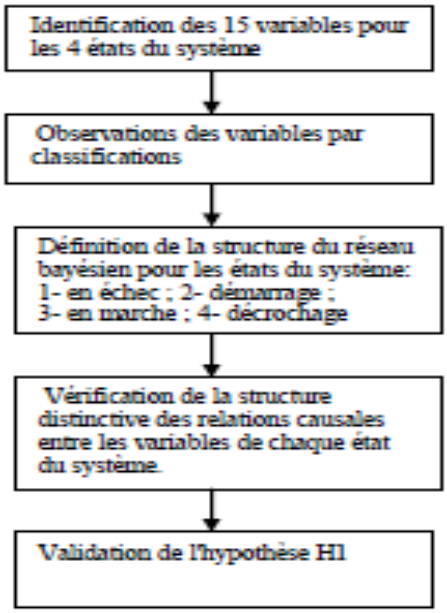
Introduction

Dans ce chapitre, nous présentons la méthodologie d'analyse des modèles et leurs résultats. L'objectif ultime de notre contribution de recherche est le développement des modèles graphiques dynamiques ou réseaux bayésiens dynamiques pour détecter la causalité des risques encourus ou des échecs du système, et spécifiquement de prédire l'état futur du système.

4.1 Réseaux bayésiens statiques des états du système

Notre première l'hypothèse H1 est évaluée à partir de la construction des réseaux bayésiens statiques qui permettent d'analyser les états du système. On fait le diagnostic des défaillances ou pannes du système pour, par la suite, extraire les relations causales existantes entre les variables (composants) du système. La Figure 3 montre les étapes pour la construction de réseaux bayésiens statiques (classiques) du système.

Figure 3 : Construction des réseaux bayésiens statiques des états du système

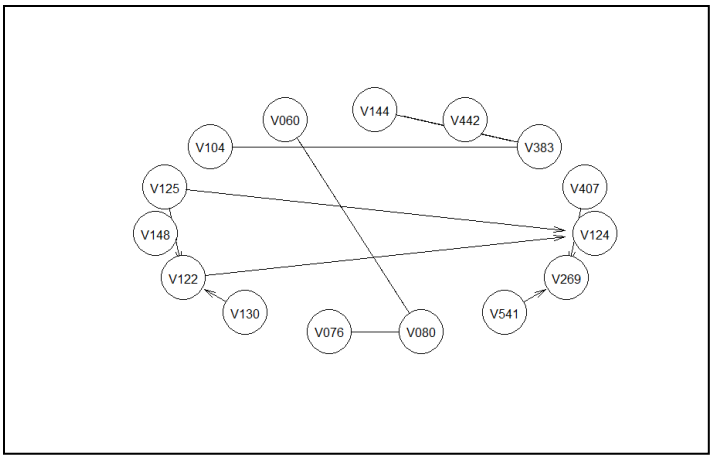


La construction des réseaux bayésiens statiques des états du système sont analysés autour des éventuelles 104 défaillances ou pannes pour les 15 variables continues du système. Les résultats des réseaux statiques permettent de vérifier la première hypothèse H1 en utilisant les représentations graphiques des états de systèmes. Les réseaux pour chacun des quatre (4) états du système sont illustrés pour démontrer l'existence ou non de la structure distinctive des relations causales entre les nœuds du réseau.

4.1.1 États du système en échec (Fail)

Ce premier graphique est une représentation qui capture l'état du système en échec (Fail). Le réseau montré à la Figure 4 présente une structure distinctive des relations causales entres les nœuds des variables.

Figure 4 : État du système en échec

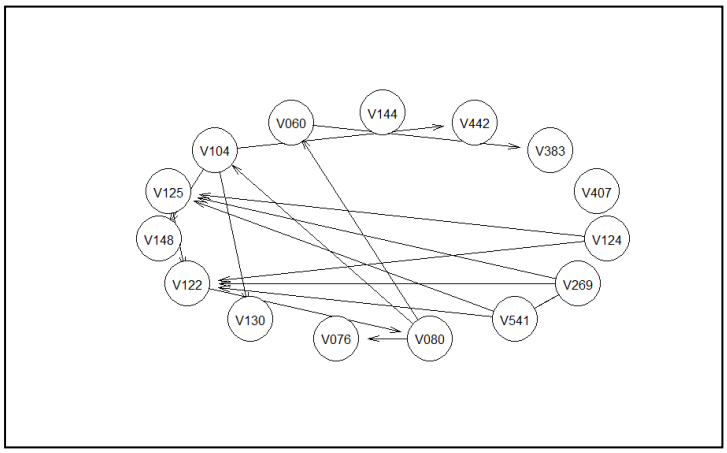


Pour l'état de système en échec, le résultat de la modélisation est basé sur 104 observations pour l'ensemble de 15 variables du système. Chaque annotation de la variable V indique une activité de système de production ayant un rôle important. Prenons l'activité V060, à chaque temps de T1 à T10, elle reçoit de manière répéter les relations de causalité d'au moins 7 variables.

4.1.2 États du système en démarrage (Start)

Le deuxième modèle graphique représente l'état du système en démarrage (Start). Le réseau de la Figure 5 montre une structure différente et explicite des relations causales entre les nœuds des variables. Ces relations sont bien exposées à travers le modèle graphique.

Figure 5 : État du système en démarrage

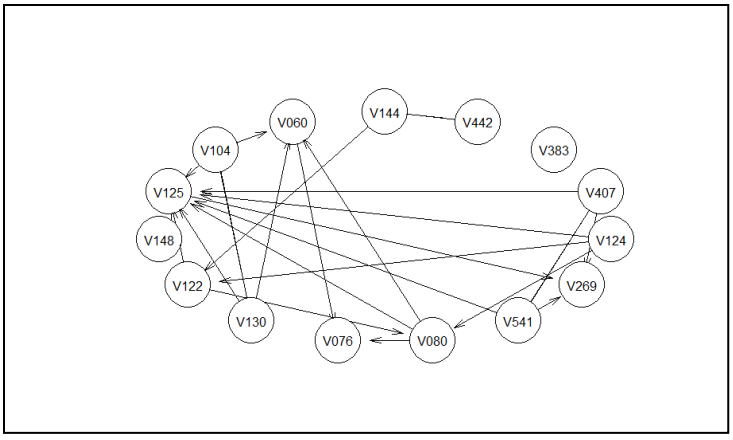


Pour l'état du système en démarrage, le résultat de la modélisation est basé sur 208 observations, sous forme de trois états successifs du système qui suivent après un état en échec et pour l'ensemble des 15 variables du système. Prenons l'activité V122 et V125, à chaque intervalle de temps de T1 à T7 elles reçoivent de manière répétée les relations de causalité d'au moins 7 variables.

4.1.3 États du système en marche (Run)

Le troisième modèle graphique représente l'état du système en fonctionnement ou en marche (Run). Le réseau montré à la Figure 6, donne une structure distinctive plus explicite des relations causales entre les nœuds des variables. Dans ce modèle, les relations causales entre les variables du système sont très significatives.

Figure 6 : État du système en marche

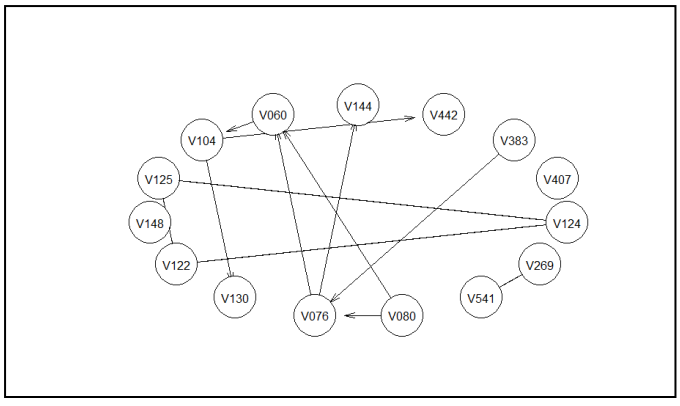


Pour l'état du système en marche, le résultat de la modélisation est basé sur 1067 observations, la suite consécutive des états de système en marche (Run) suivant les états du système en démarrage. Prenons l'activité V060 ou V125, à chaque temps de T1 à T10, elle reçoit de manière répétée les relations de causalité d'au moins 7 variables.

4.1.4 États du système en décrochage (Stall)

Le quatrième modèle graphique représente l'état du système en décrochage (Stall). Le réseau, montré à la Figure 7, illustre une structure distinctive des relations causales entre les nœuds des variables. Dans ce modèle, les relations causales entre les variables du système sont significatives.

Figure 7 : État du système en décrochage



Pour l'état de système en décrochage le résultat de la modélisation est basé sur 188 observations, soit les trois états du système qui précédent après un état d’échec, pour l'ensemble de 15 variables du système.

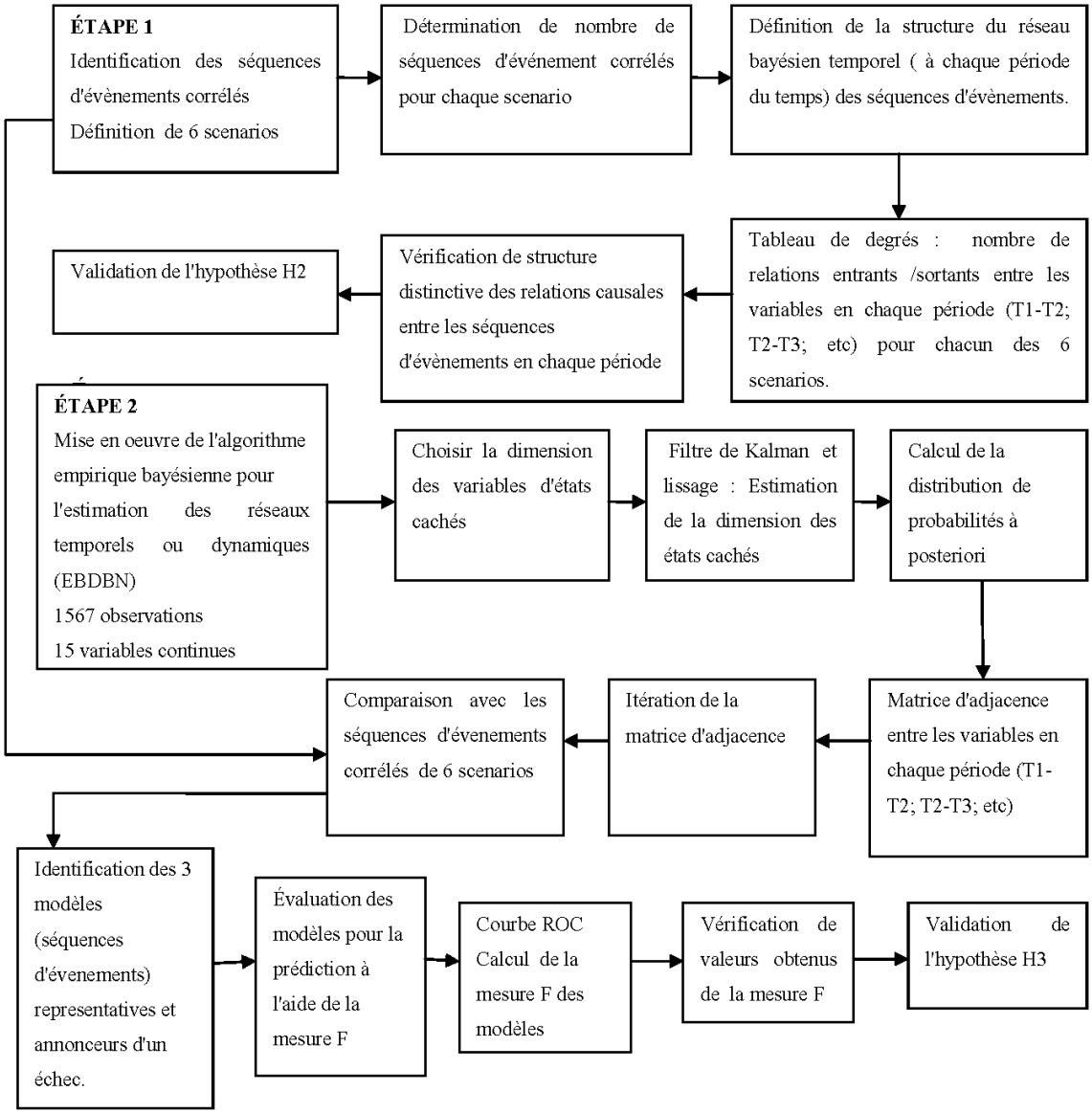
Les quatre réseaux graphiques d'états du système nous démontrent l'existence de la structure distinctive des relations causales entre les variables d'activités V. Nous n'avons pas le dictionnaire des données, cependant, nous pouvons conclure un lien de causalité fréquent entre deux étapes d'un système de production comme l'activité V060 mentionnée.

L'énoncé de notre première hypothèse H1 est ainsi vérifié et validé dans ces réseaux bayésiens statiques.

4.2 Réseaux bayésiens dynamiques des états du système

Les réseaux bayésiens dynamiques sont l'extension de réseaux bayésiens statiques combinant les 4 modèles d'états du système dans le temps. La construction des réseaux bayésiens dynamiques permet de tester la deuxième (H2 - étape 1) et la troisième (H3 - Étape 2) hypothèses de recherche. La Figure 8 illustre les étapes de la construction de ces réseaux suivi par les évaluations de celles-ci.

Figure 8 : Construction des réseaux bayésiens dynamiques des états du système



4.2.1 Évaluation des réseaux bayésiens dynamiques

Nous utilisons la méthode empirique bayésienne pour l'estimation des réseaux bayésiens temporels (EBDBN). Il s'agit d'un algorithme itératif qui utilise le filtre de Kalman pour estimer les distributions à posteriori du modèle d'espace d'évènement du système. Pour les 1565 observations, les calculs itératifs d'une matrice d'adjacence à un niveau de confiance de 95% ; 99% et 99,9% ont été testés afin de d'évaluer les interactions entre les 15 variables continues et ceux entre chaque période T1-T2 et T2-T3, etc. Nous avons retenu l'itération de la matrice d'adjacence avec un niveau de confiance 99,9% (seuil plus grand) pour évaluer les corrélations entre les variables des modèles. Les résultats obtenus de chaque itération de la matrice d'adjacence sont comparés pour prioriser les 23 séquences d'évènements corrélés résultant des 6 scénarios. Ainsi nous voulons déterminer si le modèle le plus représentatif et annonceurs d'un échec peut être identifié.

L'analyse et les interprétations des résultats des réseaux bayésiens dynamiques de six scenarios sont semblables comme celles mentionnées dans les réseaux statiques. Cependant, dans les réseaux dynamiques, l'association causale entre les variables de temps T1 et une autre T2, se produit entre des activités ayant une dépendance logique dans le contexte de Longue Marche (LongRun). Ainsi, comme l'algorithme des réseaux indiquent que V104 pourrait être une activité dont la description est de façonner et couper le petit lingot de silicium-calcium en plaque très mince dit " wafer " et V130 est une activité nécessaire après V104 comme le polissage pour la pureté du wafer.

V060 est également la variable principale et pertinente de l'arbre de décision étudié par les auteurs Munirathinam et Ramadoss. Pourtant ces auteurs ne pouvaient pas indiquer son importance dans un cycle dynamique de temps T1 à T10 ; cependant, dans notre analyse, nos réseaux dynamiques ajoutent, à ce niveau, une formation et des connaissances pertinentes. Nous démontrons que V060 a une influence aussi grande à travers le temps T1 à T10 qu'à l'intérieur d'un temps donné.

4.2.2 Évaluation des tableaux de degrés.

Le tableau de l'Annexe VII à XII indique le nombre des relations entrants et sortants entre les 15 variables en chaque couple de période temps pour les 6 scenarios.

Le scénario de longue marche est caractérisé par un couple (T1-T2 et T2-T3, etc.) de périodes de temps. Il est enveloppé par trois évènements de démarrage et suivi par trois évènements de décrochage. Les T1-T2 et T2-T3 sont différents de neuf couples de longue marche parce que les deux couples de courte marche ont plus de ressemblance avec la structure des réseaux de démarrage qu'ils précèdent et de décrochage qui leurs succèdent.

Le scénario de court redémarrage a plus de similitude avec la structure de réseau d'un décrochage suivi par un échec et après un démarrage. Tandis que le scénario de long redémarrage est plus similaire à trois couples de temps de réseau en décrochage suivi par un échec et après 3 couples de temps de démarrage.

Bref, pour chaque scénario, les réseaux de variables avant et après la défaillance sont représentés de manière dynamique pour détecter et localiser les échecs ou pannes du système.

La partie suivante présente les différentes méthodes d'apprentissage du modèle de prévision de la séquence d'évènement avant la défaillance du système. La performance de notre modèle de prévision sont évalués et testés à partir des algorithmes de classification supervisée.

4.3 Prévision de la défaillance

Dans cette section, nous présentons notre modèle de prévision de la séquence d'évènement au moment de la défaillance le plus représentatif du système. Nous utilisons différentes méthodes d'algorithmes de classification supervisée pour évaluer la mesure de performance et la capacité prédictive du modèle.

4.3.1 Algorithmes de classification

Dans cette section, en premier lieu, nous présentons une aperçue théorique sur les six (6) différents algorithmes d'apprentissages utilisés pour fin d'analyses et évaluations de la mesure de performance du modèle de prévision.

Par la suite, nous voulons expérimenter ces six algorithmes de classification dont l'objectif ultime est d'estimer l'erreur de prévision de notre modèle.

Enfin, les séries d'expérimentations réalisées ont permis d'évaluer la capacité de prévision. De plus ces expériences conduisent à une mesure de qualité et de confiance accordée à notre modèle et de comparer notre contribution avec les travaux de recherches publiés.

4.3.1.1 Arbre de décision

L'arbre de décision est un algorithme de classification supervisée. C'est une méthode statistique non-paramétrique qui permet de classer un ensemble d'individus décrits par des variables qualitatives ou quantitatives. Cette méthode donne aussi l'opportunité de produire des classes les plus homogènes possibles.

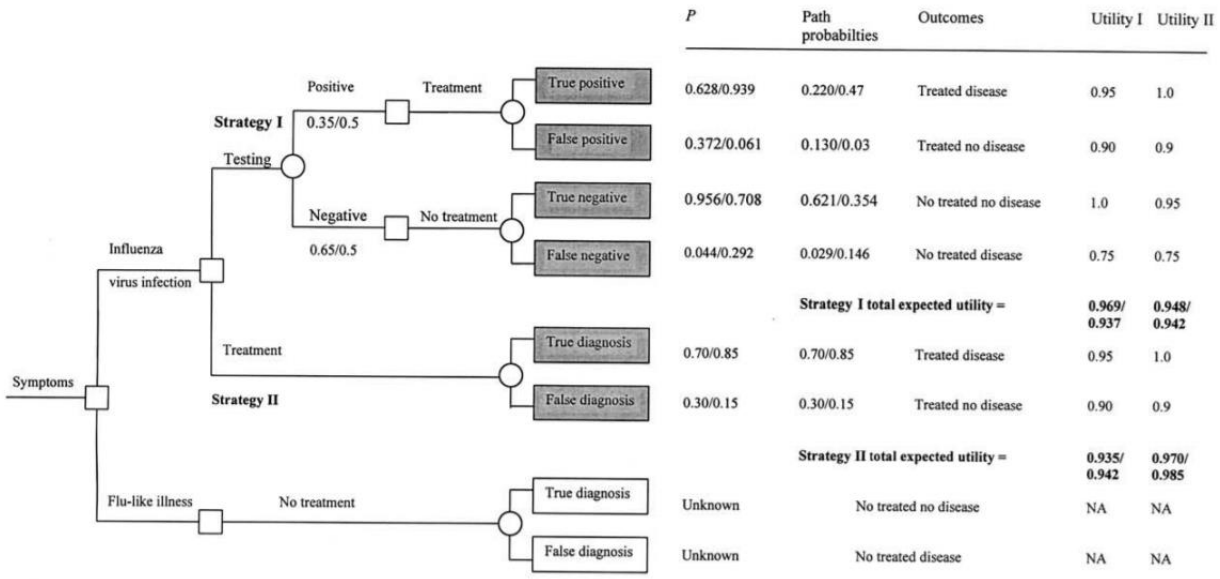
Selon la définition de Cornuejols et Miclet [61], l'apprentissage par arbre de décision est une approche classique en apprentissage automatique. Le but est de créer un modèle qui prédit la valeur d'une variable cible depuis la valeur de plusieurs variables d'entrée. En fait, l'apprentissage désigne une méthode basée sur l'utilisation d'un arbre de décision comme " modèle prédictif " permettant d'évaluer la valeur d'une caractéristique d'un système depuis l'observation d'autres caractéristiques du même système. On l'utilise spécialement en fouilles de données et en apprentissage automatique.

Il existe deux principaux types d'arbre de décisions:

- Les arbres de classification *(Classification Tree)* permettent de prédire à quelle classe la variable-cible appartient, dans ce cas la prédiction est une classe.
- Les arbres de régression *(Regression Tree)* permettent de prédire une quantité réelle (par exemple: la durée de séjour d'un patient dans un hôpital), dans ce cas la prédiction est une valeur numérique.

Un exemple d'un arbre de décision (de type classifiction) est présenté à la Figure 9. Il faut décider si un patient doit être testé pour l'influenza, ou si on le considère malade sur la base des symptômes. Un arbre de décision comprend les 2 classes « testing » et « treatment », et les résultats de probabilité selon les observations concluantes de ces 2 stratégies de soins. On constate que la stratégie du traitement direct donne une fonction d'utilité (Utility II) supérieure à toutes les autres.

Figure 9 : Exemple d'un arbre de décision



Source : [62]

Pour le calcul des probabilités d'un arbre, l'algorithme d'apprentissage « analyse d'arbre de classification et de régression » *(Classification And Regression Tree, CART)* est décrit par [63] :

- Les entrées comprennent : n individus, les variables j continues ou discrètes et une variable supplémentaire contenant la classe de chaque individu (c classes).
- La sortie : l'arbre de décision T

  Ainsi:

$N(j)$= nombre d'individus associés à la position (nœud) j

$N(k|j)$ = nombre d'individus de la classe k associés à la position j

La proportion des individus de la classe k parmi ceux de la position j est :

$$P(k|j) = \frac{N(k|j)}{N(j)} \qquad (9)$$

Les arbres de décision présentent plusieurs avantages en comparaison à d'autres méthodes de fouille de données entre autres :

- La simplicité au niveau de compréhension et d'interprétation. L'arbre de décision est un modèle facilement expliqué à l'aide de la logique booléenne, au contraire d'autres modèles comme les réseaux neuronaux, dont l'explication des résultats est difficile à comprendre.
- L'arbre de décision nécessite peu de préparation des données (pas de normalisation, de valeurs vides à supprimer, ou de variable muette).
- Le modèle peut gérer à la fois des valeurs numériques et des catégories. D'autres techniques sont souvent spécialisées sur un certain type de variables comme les réseaux neuronaux ne sont utilisables que sur des variables numériques.
- Il est possible de valider un modèle à l'aide de tests statistiques, et ainsi de rendre compte de la fiabilité du modèle.
- Performant sur de grands jeux de données: la méthode est relativement économique en termes de ressources de calcul.

Cependant, l'apprentissage par l'arbre de décision présente certains inconvénients en basant sur la revue de littérature et de publications:

- L'apprentissage par l'arbre de décision ne s'affirme pas concernant plusieurs aspects de l'optimal. En fait, les algorithmes d'apprentissages par arbre de décision sont basés sur des heuristiques telles que les algorithmes gloutons cherchant à optimiser le partage à chaque nœud, et de tels algorithmes ne garantissent pas d'obtenir l'optimum global. Certaines méthodes visent à diminuer l'effet de la recherche gloutonne.
- L'apprentissage par arbre de décision peut mener des arbres de décision très complexes, qui s'étendent mal l'ensemble d'apprentissage, dans ce cas, il s'agit du problème de surapprentissage. Les procédures d'élagage sont utilisées pour contourner ce problème, certaines approches comme l'inférence conditionnelle permettent de solutionner le surapprentissage.

- Certains concepts sont difficiles à exprimer à l'aide d'arbre de décision. Dans ces cas, les arbres de décision deviennent très étendus. Pour résoudre ce problème, plusieurs moyens existent, tels que la méthode de proportion, ou l'utilisation d'algorithmes d'apprentissage utilisant des représentations plus expressives : par exemple la programmation logique inductive.
- Lorsque les données incluent des attributs ayant plusieurs niveaux, le résultat ou le rapport d'information dans l'arbre est biaisé en faveur de ces attributs. Pourtant, le problème de la sélection de prédicteurs biaisés peut être contourné par des méthodes telles que l'inférence conditionnelle.

Les évaluations et les analyses du modèle de prédiction utilisant l'arbre de décision sont présentées dans la section 4.3.3.

4.3.1.2 K plus proches voisins

L'algorithme K plus proches voisins, noté K-ppv ou *K-nearest neighbours*, est simpliste et directe [63]. Elle n'a pas besoin d'apprentissage mais simplement le classement des données d'apprentissage. Le principe de cet algorithme est de comparer une donnée de classe inconnue à toutes les données classées. On choisit pour la nouvelle donnée, la classe majoritaire parmi ses k plus proches voisins; on s'entend que la classe peut être lourde pour des grandes bases de données; au sens d'une distance choisie.

Plusieurs mesures de distances peuvent être utilisées : euclidienne, cosinus, Chi carré, Mahalanobis, Minkowsky, Manhattan, Kullback-Leibler, Hamming, etc. Il a été démontré que Mahalanobis performe mieux pour tous types de données [64], que Chi carré est meilleur pour des jeux de données mixtes [65], et que Manhattan est utile en particulier pour les cas à haute dimensionnalité [66]. Cependant, dans notre cas le nombre de 15 variables est limité et elles sont toutes du même type. Puisque les mesures euclidienne et cosinus sont similaires [67], nous choisissons la distance euclidienne qui demeure fiable pour notre modèle.

Soient deux données représentées par deux vecteurs $x_i \; et \; x_j$ la distance entre ces deux données est calculée par :

$$d(x_i , x_j) = \sqrt{\sum_{k=1}^{d} (x_{ik} - x_{jk})^2} \qquad (10)$$

Un modèle prédictif utilisant la méthode de K plus proches voisins est proposé. Nous voulons développer cet algorithme parce que les caractéristiques des données sont non-linéaires, et en plus la règle de cette méthode est un concept intuitif.

Les évaluations et les analyses du modèle de prédiction utilisant l'algorithme K-ppv sont présentées dans la section 4.3.3.

4.3.1.3 Bayésien naïf

La première variété de réseau bayésien est appelée réseau bayésien naïf. Cette variété est largement utilisée pour les problèmes de classification et elle a donné de très bons résultats dans plusieurs publications et travaux de recherche. Le terme classifieur Bayésien est souvent employé. Un réseau bayésien naïf est composé d'un graphe à deux niveaux, un nœud parent discret pour le premier niveau (nommé nœud de classe *C*) et des nœuds enfants (ou feuilles notés X_i) pour le second. Les *n* modalités du nœud de classe *C* représentent le nombre de classe du problème (C_1,C_2,…,C_n). Notons, pour cette structure simple, l'hypothèse très forte d'indépendance des enfants conditionnellement au parent.

En tenant compte de cette hypothèse « naïve » d'indépendance entre les variables X_i , la probabilité jointe peut être exprimée par la formule [61] :

$$P(C, x_1 , \dots , x_n) = P(C) \prod_{i=1}^{n} P(x_i /C) \qquad (11)$$

où *C* est la variable de classe recherchée.

Le modèle de conception bayésiens « naïf » et ses hypothèses de base sont considérés très simplistes. Pourtant, ces classifieurs naïfs ont fait preuve d'une efficacité plus que suffisante dans beaucoup de situations et d'applications réelles complexes.

Selon la revue de littérature et dans plusieurs pratiques de différents domaines d'applications, l'estimation des paramètres de données pour les modèles bayésiens naïfs repose sur le maximum de vraisemblance. Autrement dit, on peut étudier et travailler seulement avec des modèles bayésiens naïf sans tenir compte de la probabilité ou d'utiliser l'approche bayésienne.

Les avantages du classifieur bayésien naïf sont:

- Il demande relativement peu de données d'entraînement pour estimer les paramètres nécessaires à la classification, à savoir moyennes et variances des différentes variables. En effet, l'hypothèse d'indépendance des variables permet de se contenter de la variance de chacune d'entre elle pour chaque classe, sans avoir à calculer de matrice de covariance.
- Il est un type de classification Bayésienne probabiliste simple basée sur le théorème de Bayes avec une forte indépendance (dite naïve) des hypothèses établies.
- Il met en œuvre un classifieur bayésien naïf appartenant à la famille des classifieurs linéaires. Un classifieur bayésien naïf estime que l'existence d'une caractéristique pour une classe, est indépendante de l'existence d'autres caractéristiques.
- Selon la nature de chaque modèle probabiliste, les classifieurs bayésiens naïfs peuvent être entraînés efficacement dans un contexte d'apprentissage supervisé.

Le modèle bayésien naïf est décrit avec le modèle probabiliste pour un classifieur est le modèle conditionnel :

$$p(C|F_1, \dots, F_n)$$

où C est une variable de classe dépendante dont les instances ou *classes* sont peu nombreuses, conditionnées par plusieurs variables caractéristiques $(F_1, \dots, F_n)$

Lorsque le nombre de caractéristiques n est grand, ou lorsque ces caractéristiques peuvent prendre un grand nombre de valeurs, baser ce modèle sur des tableaux de probabilités devient impossible et irréalisable. Par conséquent, nous le dérivons pour qu'il soit plus facilement accessible.

À l'aide du théorème de Bayes nous écrivons :

$$\boldsymbol{p(C|F_1, \dots, F_n) = \frac{p(C)p(F_1,\dots,F_n|C)}{p(F_1,\dots,F_n)}} \quad (12)$$

Cela indique :

$$\boldsymbol{Post\acute{e}rieure = \frac{ant\acute{e}rieure\ X\ vraisemblance}{\acute{e}vidence}}$$

En pratique, seul le numérateur nous intéresse, puisque le dénominateur ne dépend pas de C et les valeurs des caractéristiques sont données. Le dénominateur est donc en réalité constant. Le numérateur est soumis à la loi de probabilité à plusieurs variables

$p(C, F_1, \dots, F_n)$

et peut être factorisé de la façon suivante, en utilisant plusieurs fois la définition de la probabilité conditionnelle.

$$\begin{aligned} & p(C, F_1, \dots, F_n) \\ & = p(C)p(F_1, \dots, F_n/C) \\ & = p(C)p(F_1/C)p(F_2, \dots, F_n|C, F_1) \\ & = p(C)p(F_1/C)\, p(F_2/C, F_1)\, p\,(F_3, \dots, F_n/C, F_1, F_2) \\ & = p(C)p(F_1/C)(F_2/C, F_1)\, p\,(F_3/C, F_1, F_2)\, p(F_4, \dots, F_n/C, F_1, F_2, F_3) \\ & = (C)p(F_1/C)(F_2/C, F_1)\, p\,(F_3/C, F_1, F_2), \dots, p(F_n/C, F_1, F_2, F_3, \dots, F_n) \end{aligned} \quad (13)$$

Nous faisons intervenir l'hypothèse naïve, si chaque F_i est indépendant des autres caractéristiques $F_{i\neq j}$, alors :

$$p(F_i/C, F_j) = p(F_i|C) \quad (14)$$

pour tout $i \neq j$, par conséquent la probabilité conditionnelle peut s'écrire

$$\begin{aligned} p(C|F_1, \dots, F_n) &= p(C)p(F_1|C), p(F_2|C), p(F_3|C) \\ &= p(C) \prod_{i=1}^{n} (p(F_i|C) \end{aligned} \quad (15)$$

En tenant compte de l'hypothèse d'indépendance ci-dessus, la probabilité conditionnelle de la variable de classe C est formulée par :

$$p(C|F_1, \dots, F_n) = \frac{1}{Z} p(C) \prod_{i=1}^{n} p\,(F_i/C) \quad (16)$$

où Z (appelé « évidence ») est un facteur d'échelle qui dépend uniquement de $F_1, \ldots, F_n$, à savoir une constante dans la mesure où les valeurs des variables caractéristiques sont connues.

Les modèles probabilistes ainsi décrits sont plus faciles à manipuler, puisqu'ils peuvent être factorisés par *l'antérieure* $P(C)$ (probabilité *a priori* de C) et les lois de probabilité indépendantes $p(F_i|C)$. S'il existe k classes pour C et si le modèle pour chaque fonction $p(F_i|C) = c$ peut être exprimé selon r paramètres, alors le modèle bayésien naïf correspondant *dépend de (k − 1) + n r k paramètres.*

Dans la pratique, on observe souvent des modèles où k=2 (classification binaire) et r =1 (les caractéristiques sont des variables de Bernoulli). Dans ce cas, le nombre total de paramètres du modèle bayésien naïf ainsi décrit est de 2n+1, ou n est le nombre de caractéristiques binaires utilisées pour la classification.

Pour construire un classifieur à partir du modèle de probabilités, il est nécessaire d'établir le modèle à caractéristiques indépendantes, à savoir le modèle de probabilités bayésien naïf. C'est le classifieur bayésien naïf couple du modèle avec une règle de décision. Il s'agit la règle du maximum à posteriori ou MAP. Le classifieur correspondant à cette règle est la fonction suivante :

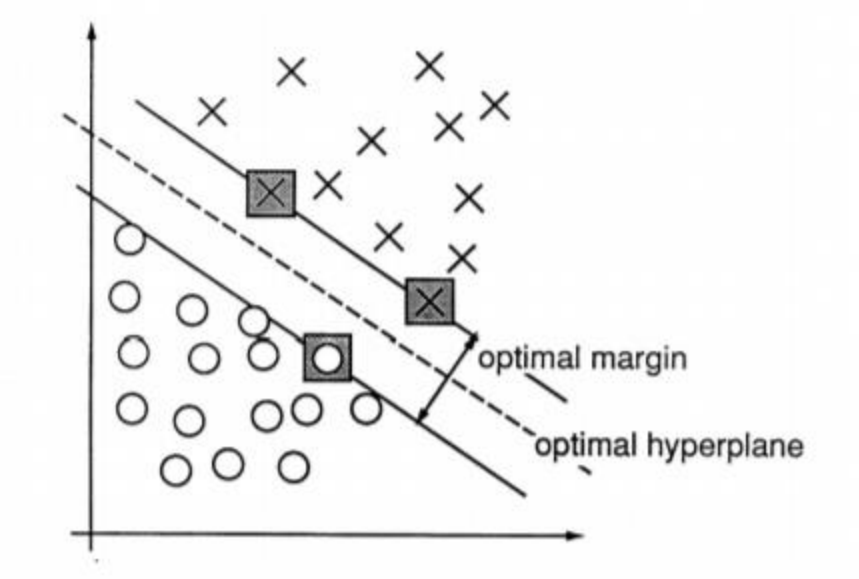
$$\boldsymbol{classifieur}\,(\boldsymbol{f_1}, \ldots, \boldsymbol{f_n}) = \boldsymbol{argmax}\ \prod_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{p}(\boldsymbol{F_i} = \boldsymbol{f_i}\,|\boldsymbol{C} = \boldsymbol{c})\,\boldsymbol{p}(\boldsymbol{C} = \boldsymbol{c}) \qquad (17)$$

Nous avons expérimenté ce classifieur Bayésien naïf pour notre modèle de prédiction. Les évaluations et les analyses des expériences sont présentés dans la section 4.3.3.

4.3.1.4 Séparateurs à vaste marge

Les séparateurs à vaste marge (Support Vector Machines) sont une famille de classifieurs supervisés qui ont été introduits par Valdimir Vapnik [68]. Elles sont évoluées parmi les modèles les plus utilisés pour la classification et la régression. La figure 10 montre un SVM pour la classification.

Figure 10 : Exemple d'un SVM



Source : [69]

Le principe des séparateurs à vastes marges est utilisé pour résoudre des problèmes de discrimination, c'est-à-dire décider à quelle classe appartient un échantillon de données ou de régression, prédire la valeur numérique d'une variable. La résolution passe par la construction d'une fonction :

$$\boldsymbol{y = h(x)} \qquad \textit{(18)}$$

La discrimination à deux classes (binaire), $y \in \{-1, 1\}$ le vecteur d'entrée x étant dans un espace X muni d'un produit scalaire. On peut prendre $X = R^N$

Pour un cas simple d'une fonction de discrimination linéaire obtenue par une combinaison linéaire du vecteur d'entrée $x = (x_{1,} \dots x_N)^T$, avec un vecteur poid $w = (w_1 , \dots , w_N)$:

$$\boldsymbol{h(x) = w^T x + w_0} \qquad \textit{(19)}$$

Il est alors décidé que x est de classe 1 si $h(x) \geq 0$ et de classe -1 sinon, c'est un classifieur linéaire

La frontière de décision $h(x)$= 0 est un hyperplan, appelé hyperplan séparateur ou séparatrice. Le but d'un algorithme d'apprentissage supervisé est d'apprendre la fonction $h(x)$ par le biais d'un ensemble d'apprentissage :

$$\{(x_1, l_1), (x_2, l_2), \dots, (x_p, l_p)\} \subset R^{N\,x}\{-1\,,1\} \qquad (20)$$

où les l_k sont les labels, p est la taille de l'ensemble d'apprentissage, N la dimension des vecteurs d'entrée. Si le problème est linéairement séparable on doit avoir :

$$l_k h(x_k) \geq 0 \quad 1 \leq \mathrm{k} \leq \mathrm{p} \quad \text{autrement dit} \quad l_k = w^T x_k + w_0 \geq 0 \qquad 1 \leq \mathrm{k} \leq \mathrm{p} \qquad (21)$$

La marge est la distance entre l'hyperplan et les échantillons les plus proches, ce sont les vecteurs supports. L'hyperplan qui maximise la marge est donnée par :

$$Arg\max\min\{||x - x_k||: x \in R^{N\,x}, w^{T\,x} + w_0 = 0\} \qquad (22)$$

Il s'agit de trouver w et w_0 remplissant ces conditions, afin de déterminer l'équation de l'hyperplan séparateur :

$$\boldsymbol{h(x) = w^T x + w_0 = 0} \qquad (23)$$

Marge Maximale : il existe un unique hyperplan optimal, défini comme hyperplan qui maximise la marge entre les échantillons et l'hyperplan séparateur. La capacité des classes d'hyperplans séparateurs diminue lorsque leur marge augmente.

La marge est la plus petite distance entre les échantillons d'apprentissage et l'hyperplan séparateur qui satisfasse la condition de séparabilité. La distance d'un échantillon x_k à l'hyperplan par sa projection orthogonale sur l'hyperplan est :

$$\frac{\boldsymbol{l_k(w^{T\,x} + w_0)}}{\boldsymbol{||w||}}$$

L'hyperplan séparateur (w , w_0) de marge maximale est donnée par :

$$\boldsymbol{argmax_{w,w_0}}\{\frac{\mathbf{1}}{||\boldsymbol{w}||}\mathbf{min}[\boldsymbol{l_k}(\boldsymbol{w^T x}+\boldsymbol{w_0})]\} \quad (24)$$

Nous avons choisi cet algorithme pour évaluer la performance de notre modèle de prévision. Les évaluations et les analyses des séries d'expérimentations sont présentées dans la section 4.3.3.

4.3.1.5 Régression Logistique

La régression logistique ou modèle logit est une méthode prédictive [63]. Cette méthode vise à créer un modèle permettant de prédire et expliquer les valeurs prises par une variable visée qualitativement. En général on parle fréquemment de la régression logistique binaire car la variable à prédire prend deux modalités.

Dans le cadre de la régression logistique binaire, on note Y la variable à prédire (variable expliquée) qui prend deux modalités {1,0} et les variables $X = (X_1, X_2, \dots, X_j)$ sont les variables prédictives, et exclusivement continues ou binaires.

Pour faire une estimation, on dispose de :

- Un échantillon Ω d'effectif n, notons n_1 $(resp. n_0)$ les correspondants à la modalité 1(resp.0) de Y.
- $P(Y = 1)$ $(resp. P(Y = 0))$ est la probabilité à priori pour que $Y = 1$ $(resp. Y = 0)$. Pour simplifier, nous écrivons $p(1)$ $(resp. p(0))$
- $p(X|1)$ $(resp. p(X|0))$ est la distribution conditionnelle des X sachant la valeur prise par Y.
- La probabilité a postériori d'obtenir la modalité 1 de $Y(resp. 0)$ sachant la valeur prise par X est représentée par p $(1|X)$ $resp. p$ $(0|X))$

La régression logistique repose sur l'hypothèse fondamentale de la maximisation de la vraisemblance et les concepts produits sur le rapport de vraisemblance. Cette hypothèse est la suivante, où l'on reconnait la mesure nommée " Évidence " $Ev(p) = ln\frac{p}{1-p}$ pour les besoins de l'inférence bayésienne en évitant des renormalisations continuelles sur [0,1].

$$\boldsymbol{ln\frac{p(X|1)}{p(X|0)} = a_0 + a_1x_1 + \cdots + a_jx_j} \qquad (25)$$

Une vaste classe de distribution répond à cette spécification. Comme la distribution multinormale décrite en analyse discriminante linéaire, mais également d'autres distributions notamment celle ou les variables explicatives sont booléennes (0,1).
Notons que par rapport à l'analyse discriminante, ce ne sont pas les densités conditionnelles $p(X|1)$ et $p(X|0)$ qui sont modélisés mais le rapport de ces densités. La restriction introduite par l'hypothèse est moins forte.

La régression logistique ou le modèle Logit peut être écrite de manière différente :

$$\boldsymbol{ln\frac{p(1/X)}{1-p(1/X)} = b_0 + b_1x_1 + \cdots + b_jx_j} \qquad (26)$$

- Il s'agit d'une régression, montrant une relation de dépendance entre une variable à expliquer et une série de variables explicatives.
- Il s'agit d'une régression" logistique" car la loi de probabilité est modélisée à partir d'une loi logistique.

En effet, après la transformation de l'équation (26), nous obtenons :

$$\boldsymbol{p(1|X) = \frac{e^{b_0 + b_1x_1 + \cdots + b_jx_j}}{1+e^{b_0 + b_1x_1 + \cdots + b_jx_j}}} \qquad (27)$$

Pour notre modèle de prédiction, les évaluations et les analyses de la mesure de performance utilisant la régression logistique sont présentés dans la section 4.3.3.

4.3.1.6 Réseau neuronal artificiel

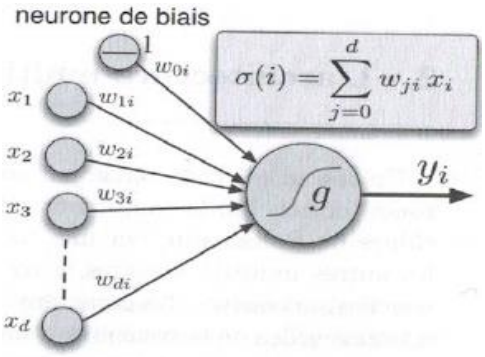
Le réseau de neurone artificiel est une méthode d'apprentissage conçue promptement de l'observation du fonctionnement des systèmes cérébraux humains. La capacité d'apprentissage de ces systèmes est reliée à la modification des poids des connexions. D'abord, nous présentons les définitions nécessaires à la compréhension des réseaux de neurones artificiels ou connexionnistes [70].

Un réseau artificiel ou connexionniste comprend [61] :

- L'espace de représentation : dans un réseau, les données d'entrées sont des vecteurs $\mathbb{R}^d$ notés (en transposition) $x^T = (x_1, \dots, x_d)$. Nous présentons des réseaux de neurones artificiels qui sont des règles de classifications de données numériques.
- Le neurone formel : c'est l'unité de traitement élémentaire dans un réseau de neuronal. Nous considérons notamment les modèles de réseaux multicouches qui classent les unités selon qu'elles sont des neurones d'entrée, cachés, ou de sortie. Un neurone d'entrée est une unité de chargée de transmettre une composante du vecteur x des données d'apprentissage. Un neurone de sortie est une unité qui fournit une hypothèse d'apprentissage, par exemple dans un problème de classification, une décision sur la classe à laquelle est attribué x. Un neurone caché est un neurone qui n'est ni un neurone d'entré, ni de sortie. C'est l'unité qui fait les traitements intermédiaires.

Comme le montre la figure 11, l'état de neurone formel est caractérisé par un ensemble de valeurs σ_i , un pour chaque neurone formel i.

Figure 11 : Modèle d'un neurone formel



Source : [61]

Lorsque le neurone i est un neurone d'entrée on a : $\sigma_i = x_i$ ou x_i est la composante de rang *i* du vecteur x. Le neurone formel est caractérisé par une fonction de sortie g ou fonction d'activation qui permet de calculer pour chaque neurone i une valeur de sortie y_i en fonction de son état d'activation σ_i : $y_i = g(\sigma_i)$

Réseau neuronal multicouches : L'architecture de ce type de réseau est définie par les neurones formels ou unités d'une couche sont reliées uniquement à toutes celle de la couche supérieure. On associe le poids $w(i,j)$ de la connexion à chaque lien entre les deux unités $i\ et\ j$. La couche d'entrée ne sert qu'à la lecture des données. Elle comporte d + 1 éléments. Elle est activée par un vecteur x de R. Ainsi, la première couche cachée effectue le calcul pour chacun de ses neurones formels, puis la seconde couche fait ses calculs, et ainsi de suite. La couche de sortie sert à traduire la décision. Pour une classification, le neurone de la couche finale ayant la valeur de sortie la plus grande indique la classe calculée pour l'entrée.

Le fonctionnement d'un réseau multicouche :

- L'ensemble de neurones d'entrées

y_i avec $i \in source\ (j)$

- Le poids des connexions :

$w(i,j)$ avec $i \in source\ (j)$

- Son état d'activation :

$$\sigma = w(0,j) + \sum_{i \in source\ j} w(i,j)y_i$$

- Sa fonction de sortie :

$$y_i = g(\sigma_j) \qquad (28)$$

Pour calculer la décision, chaque neurone se met dans l'état :

$\sigma = \sum_{i \in source\ j} w(i,j)\mathrm{y_i}$ puis transmet vers les autres neurones formels de {dest (j)} :

y_i = f (σ_i) avec la fonction non-linéaire : fonction signe ou sigmoïde

Les expériences effectuées sur notre modèle de prédiction en utilisant le réseau neuronal artificiel sont présentés dans la section 4.3.3. Ces tests permettent d'évaluer la mesure de performance du modèle et sa capacité de prédiction.

4.3.2 Critères d'analyse et évaluation du modèle de prévision

Dans cette section, nous abordons l'utilisation de six algorithmes de classification supervisée pour fin de critères d'évaluations et d'analyses de la mesure de performance de notre modèle. Il s'agit d'un ensemble de tests pour une estimation des erreurs de prévisions. Les critères de mesure de performance du modèle s'estiment par plusieurs indicateurs connus tel que les scores, les taux d'erreur de prévision etc. et aussi en fonction de la taille de l'échantillon de données.

Tout d’abord, nous abordons quelques définitions des mesures statistiques. Selon [61], la sensibilité et la spécificité sont des mesures statistiques de la performance d'un test de classification binaire, également connu en statistique comme fonction de classification. La sensibilité (aussi appelé le taux positif réel, le rappel ou la probabilité de détection dans certains domaines) mesure la proportion de positifs correctement identifiés comme tels autrement dit les individus (e.g., succès, ou possession d’un actif ou présence d’une maladie) qui sont correctement identifiées comme ayant la condition.

La spécificité ou le taux négatif réel mesure la proportion de négatifs qui sont correctement identifiés comme tels autrement dit les individus qui sont correctement identifiées comme n'ayant pas la condition.

On évalue les résultats d’une classification par le calcul des taux positif = identifié et négatif = rejeté. De ce fait :

- Vrai positif (TP) = correctement identifié
- Faux positif (FP) = identifié incorrectement
- Vrai négatif (TN)= rejeté correctement
- Faux négatif (FN) = rejeté de manière incorrecte

Considérons un groupe de population avec P instances positives et N instances négatives de certaines conditions [61; 63]. Les quatre résultats peuvent être formulés dans une matrice de confusion, présentée au Tableau 7.

Tableau 7 : Matrice de confusion

		Condition prédictive	
	Total de population	Prédiction positive +1	Prédiction négative -1
Vrai condition	Condition positive +1	TP	FN
	Condition négative -1	FP	TN

L'inférence de la matrice a donné :

- Taux de vrais positifs ou Rappel (Sensibilité): la proportion de résultats positifs réels qui sont positifs prédits.

$$TP\ (taux\ de\ vrais\ postifs) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

-Taux de faux positifs :

$$FP\ (taux\ de\ faux\ positifs) = 1 - Spécificité = \frac{\mathrm{TP}}{\mathrm{FP} + \mathrm{TN}}$$

- Précision (valeur prédictive positive): c'est la proportion de positifs prédits qui sont en fait positifs

$$Précision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$Sensibilité = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$Spécificité = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

$$Exactitude = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}$$

$$F - mesure \ = 2 \ \frac{(\text{précision.rappel})}{(\text{précision} + \text{rappel})} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

La valeur de MCC peut être directement calculée à partir de la matrice de confusion en utilisant la formule suivant :

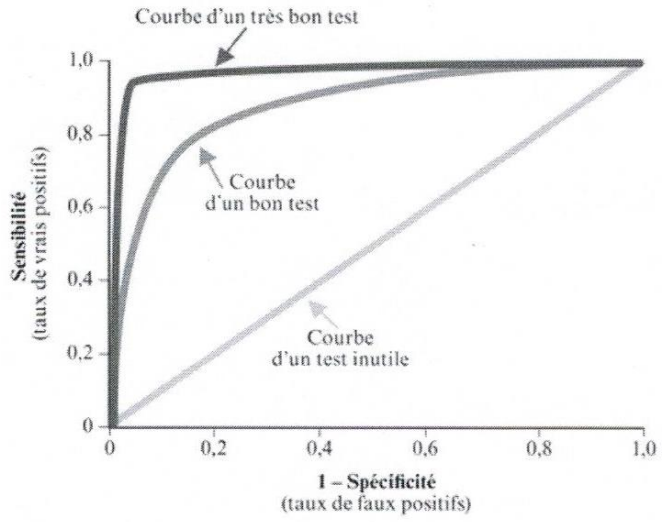
$$MCC = \frac{\text{TP x TN} - \text{FP x FN}}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Le score de Précision et TFP (Taux de faux positif) compris entre la valeur -1 et 1 évaluent la qualité d'un modèle de prédiction. Si ce score est supérieur à 0, le taux de bonne prédiction est supérieur à celui des faux positifs. Et plus il est proche de 1, meilleur le modèle de prédiction.

Les notions de spécificité et sensibilité proviennent de la théorie du signal ; leurs valeurs dépendent directement du seuil s fixée à priori (en général 0,5). En augmentant s, la sensibilité diminue tandis que la spécificité augmente car la règle de décision pour la détection devient plus exigeante ; un bon modèle associe une grande sensibilité et grande spécificité pour la détection d'un signal. Ce lien est représenté graphiquement par la courbe ROC *(Receiver Operating Characteristic)* ou courbe de sensibilité (probabilité de détecter le vrai signal) en fonction de 1 moins la spécificité (probabilité de détecter le vrai signal) pour chaque valeur s du seuil. Le ROC est appelé aussi caractéristique de performance d'un test. C'est un indicateur de mesure de performance d'un classificateur binaire, c'est-à-dire d'un système qui a pour objectif de catégoriser des éléments en deux groupes distincts sur la base d'une ou plusieurs des caractéristiques de chacun de ces éléments.

Nous illustrons la figure 12, un exemple général des courbes ROC résultant de trois tests : Courbe ROC optimale (très bon test); Courbe ROC significative (bon test); Courbe ROC inutile (classificateur aléatoire).

Figure 12 : Exemple de courbes ROC sur un échantillon de tests



L'aire sous la courbe mesure la qualité de la discrimination du modèle tandis qu'une analyse de la courbe aide au choix du seuil s.

Notre contribution comprend le modèle le plus représentatif et annonceur d'un échec, étant identifié avec les 15 variables continues.

Pour ce faire, nous testons relativement six (6) algorithmes de classification d'apprentissage supervisée avec ce modèle avec des ensembles de données classifiées. Ces six algorithmes permettent d'évaluer la pertinence et la performance des modèles de prédiction de la séquence d'évènements avant la défaillance du système. Les indicateurs de performances sont utilisés pour examiner la capacité prédictive du modèle.

Nous utilisons le logiciel ***Tanagra*** comme plateforme d'expérimentation avec les six algorithmes pour effectuer des séries d'expériences des modèles de prédiction. Ces séries d'expérimentations jouent un rôle important dans l'apprentissage des modèles et leur mesure de performances en prédiction.

Notre modèle de prévision porte sur les deux états du système ; Échec (Fail) et En Marche (Run). Les attributs en *Target (output)* ce sont les variables à prédire l'état du système (variable discrète) et en *Input* les 15 variables prédicteurs qui sont tous continues. Les courbes *ROC* et *PR* sont illustrés et analysés pour chacun des six algorithmes évalués.

4.3.3 Modèle de prévision

L'Annexe XIII montre les résultats obtenus des tests de simulation dans Tanagra pour chacun des six algorithmes de classification, et le Tableau 6 résume les mesures de performance.
Ces résultats conduisent à l'apprentissage du modèle en tant sur l'estimation du modèle de prévision et sur l'estimation des erreurs de prévision. De plus, ces résultats d'expérimentations contribuent à la précision du modèle et en même temps donnent, et évaluent la qualité du modèle et leur capacité en prévision.

Nous présentons un tableau de synthèse pour les valeurs des indicateurs de performance du modèle. L'objectif visé de ce tableau est de démontrer que parmi les six algorithmes de classification utilisés les taux d'erreurs estimés restent assez faibles. Ces valeurs d'estimations de d'erreurs de prévision ont mené un gain de compréhension apportés par les différents algorithmes, tant en termes de représentation des connaissances qu'en terme de sélection de variables pertinentes du modèle.
Également, ces indicateurs contribuent aussi à une aide à la décision pour l'amélioration continue du processus de production dans un système complexe tel étudié.

Dans le tableau 8, nous pouvons qualifier que les estimations des erreurs de prévision sont vraies pour notre modèle. Dans le tableau 9, nous comparons la performance du modèle de prédiction de l'état du système aux résultats de recherche de Munirithinam et Ramadoss (2016). Les valeurs supérieures, en comparaison entre nos travaux, sont indiquées en caractères italiques et gras. La meilleure valeur parmi les algorithmes est identifiée par une cellule grise.

Cette analyse comparative permet de voir la performance et la qualité des deux modèles de prédiction dans son ensemble. D'après ce tableau ci-dessous, la mesure performance de notre modèle s'explique et se traduit ainsi :

- Taux d'erreurs estimés sont minimes dans tous les algorithmes de classification utilisée.
- Meilleure performance globale des estimateurs du modèle.
- Plus représentatif pour la gestion et prévision des risques du système de production.

Tableau 8 : Indicateurs de performance du modèle de prévision

Indicateurs	**Bayes**	**Kppv**	**ArD**	**RLog**	**RN**	**SVM**
TP - Sensibilité	0,294	0,500	0,616	0,333	0,667	0
TN - Spécificité	0,949	0,947	***0,976***	0.947	0,949	0,946
FP	0,050	0,052	***0,023***	0,052	0,050	0
FN	0,705	0,500	0,383	0,667	0,333	0
Précision	0,854	0,905	***0,964***	0,865	0,930	0
F-Mesure	0,440	0,644	0,751	0,490	***0,776***	0
MCC	0,324	0,492	***0,643***	0,358	***0,643***	0
PR	0,344	0,552	0,639	0,384	***0,717***	0
ROC	0,764	0,942	***0,945***	0,751	0,670	0,644

Tableau 9 : Indicateurs de performance du modèle de Munirithinam et Ramadoss (2016)

Indicateurs	**Bayes**	**Kppv**	**ArD**	**RLog**	**RN**	**SVM**
TP - Sensibilité	0,746	0,98	***1,0***	***1,0***	-	0,2
TN - Spécificité	-	-	-	-	-	-
FP	0,352	0,24	0,161	0,335	-	0,6
FN	-	-	-	-	-	-
Précision	0,234	0,37	0,472	0,301	-	0,5
F-Mesure	0,356	0,54	0,641	0,463	-	0
MCC	0,3	0,2	0,41	0,27	-	0
PR	0,2	0,1	0,42	0,32	-	0
ROC	0,4	0,3	0,51	0,31	-	0

4.3.4 Fonctions d'estimations de la robustesse du modèle de prévision

L'Annexe XIV montre les résultats obtenus des tests et fonctions de robustesse du modèle de prévision. Nous avons rapporté ici les résultats de nos analyses dans Tanagra utilisant les fonctions *"Train-Test"*, validation croisée ou *"Cross-Validation"*, ainsi que *"Bootstrap"*.

4.3.4.1 Évaluation du modèle avec *"Train-Test"*

La fonction *"Train-Test"* est nécessaire afin de séparer les échantillons et par la suite évaluer la capacité prédictive du modèle.

Pour les six algorithmes, en utilisant la fonction de *"Train-Test"*, nous constatons que les résultats taux d'erreur évalués sont tous assez faibles avec les proportions des échantillons testés. On peut conclure que la capacité prédictive du modèle de prédiction est bonne.

4.3.4.2 Évaluation du modèle avec la Validation croisée et *"Bootstrap"*

Les fonctions de validation croisée et le *"Bootstrap"* qui sont utilisés pour tester la précision prédictive du modèle dans un échantillon du test par rapport à la précision prédictive de l'échantillon d'apprentissage à partir duquel le modèle a été développé.

Pour ce faire, la validation croisée est exécutée aussi bien sur l'échantillon test que sur l'échantillon d'apprentissage. Le *Bootstrap* est également vérifié avec les échantillons de variables.

La validation croisée est un principe simple et largement utilisée pour estimer une erreur contre un surplus d'inférence ou de calcul. L'idée est d'itérer l'estimation de l'erreur sur plusieurs échantillons de validation puis d'en calculer la moyenne. Pour ce faire, on subdivise les données en K bloc et on répète K fois le processus suivant l'apprentissage (K-1 bloc) et le test sur le reste K-ième bloc. Plusieurs études empiriques montrent que K=10 paraît un bon compromis. Dans Tanagra, pour nos séries de tests, nous spécifions bien K=10 *(number of folds = 10)* et réitérons une fois *(number of repetition = 1)*

Le principe de la méthode *Bootstrap* de ré-échantillonnage est de substituer, à la distribution de probabilité inconnue F, dont est issu l'échantillon d'apprentissage, la distribution empirique F_n qui donne un poids 1/n à chaque réalisation. Ainsi on obtient un échantillon de taille n dit échantillon Bootstrap selon la distribution empirique F_n par n tirages aléatoires avec remise parmi leur n observations initiales.

Dans Tanagra, deux estimateurs "*Bootstrap*" sont disponibles. Le standard 0.632 Bootstrap et l'indicateur modifié 0.632 Bootstrap +, tiennent compte des spécificités de la technique d'apprentissage. Le seul paramétrage possible est le nombre de répétitions. On remarque que plus qu'on augmente le nombre de réplications, plus que la variance de l'estimateur est réduit, toutefois, une vingtaine de réplications en Bootstrap est largement suffisant pour obtenir une estimation appropriée. Nous avons choisi une fois de répétition *(number of repetition = 1)* dans nos séries de tests.

En conclusion, la valeur K=10 est un bon compromis pour la validation croisée, et l'augmentation du nombre de répétitions permet aussi d'améliorer la précision du modèle. La fonction Boostrap+ a généralement une variance plus faible que la validation croisée dans chacun des algorithmes.

La différence des taux d'erreurs estimés entre les six différentes techniques reste toujours assez faible. Les taux d'erreurs évalués sont très proches du véritable taux d'erreurs obtenus du calcul. La précision des estimateurs est bonne, ainsi l'estimation prédictive du modèle semble acceptable. On peut confirmer que ces valeurs sont des vraies erreurs estimées du modèle de la prédiction.

Conclusion

La proposition d'une nouvelle méthodologie utilisant les réseaux bayésiens statiques et dynamiques est bien développée pour analyser la structure causale des risques d'échec du système SCADA. Nous avons bien démontré que ces réseaux peuvent contribuer largement au diagnostic et à la prédiction des échecs ou des pannes du système.

Nous avons illustré les réseaux bayésiens statiques et dynamiques pour chaque état du système, représenté par un nombre limité et distinct des facteurs.

Nous avons mené une étude approfondie d'un modèle de prédiction deux états du système à partir différentes méthodes d'apprentissages supervisés. L'étude du modèle permet de démontrer leur performance et la capacité en prédiction.

Les résultats attendus des représentations graphiques des réseaux bayésiens statiques et dynamiques sont concluants pour valider les hypothèses H1 et H2.

En effet, pour l'hypothèse H1, les quatre réseaux graphiques d'états du système nous démontrent l'existence de la structure distinctive des relations causales entre les variables d'activités (composant) du système. Ces 4 états ont des signatures significatives.

Pour l'hypothèse H2, les réseaux bayésiens dynamiques pour les 6 scénarios sont assez déterminants tel que démontrés dans les annexes I à VI. De plus, les tableaux de degrés ont également des signatures assez spécifiques tel que démontré dans les Annexes VII à XII pour le nombre des relations entrants et sortants entre les 15 variables en chaque couple de période temps pour les 6 scenarios.

Pour conclure, les réseaux bayésiens dynamiques ne donnent pas des résultats significatifs pour vérifier l'hypothèse H3, le score de prévision demeure faible. Cependant, les résultats de séries d'expérimentations avec des six algorithmes de classification de quinze (15) variables pour notre modèle sont meilleurs par rapport aux résultats des études antérieures de prévision des auteurs de références, Mc Cann et al. (2010) et Munirathinam et Ramadoss (2016).

Dans le chapitre suivant, nous allons aborder les avantages et les valeurs ajoutées de notre contribution en utilisant les réseaux bayésiens dynamiques et le modèle de classification. Nous présentons les limites de la thèse et les recommandations pour la recherche à venir.

CHAPITRE 5 - Discussion et Conclusion

Introduction

Dans ce chapitre, nous présentons les valeurs ajoutées de nos contributions, en particulier par l'utilisation des réseaux bayésiens dynamiques. Nous discutons aussi les atouts du modèle de prédiction face à la complexité d'un système de production.

Nous abordons également les limites de notre recherche au niveau du jeu de données du système SCADA. S'ajoute à cela le modèle qui contribue à la gestion et priorisation des risques du système de production étudié.

Nous concluons ce travail par l'identification de quelques pistes de recherche intéressantes pour la méthodologie proposée sur le jeu de données du système SCADA.

5.1 Contributions de la thèse

Le Tableau 9 résume les contributions de la thèse. Nous avons tout d'abord réduit la dimensionnalité des variables avec la méthode de Forêt aléatoires. L'utilisation de cette méthode est concluante car elle permet de sélectionner les bonnes variables du modèle de classification.

À l'aide des variables identifiées, nous avons développé des modèles utilisant les réseaux bayésiens dynamiques. Ils nous ont permis d'apprendre les dépendances causales en modélisant des phénomènes de défaillances aléatoires.

Les représentations graphiques bayésiennes développées sont très compactes avec une facilité d'acquisition et d'utilisation de diverses connaissances liées aux risques et leurs interactions.

Pour son utilisation, les réseaux bayésiens dynamiques montrent une grande flexibilité permettant de demander le même modèle de réseau graphique pour des objectifs différents. Dans notre cas, nous l'avons utilisé pour faire le diagnostic et la prédiction.

Tableau 10 : Contributions de la thèse

Contributions de la thèse	**Références et travaux connexes : Mc Cann et al. 2010 et Munirathinam et Ramadoss 2016.**
- Réduction et sélection des variables par méthode de Random Forest. - Détection des relations causales par les réseaux de bayes statiques. - Détection des séquences d'évènements et relations causales intertemporelles par les réseaux de bayes statiques. - Sélection des variables clés et pertinentes. Ce sont des variables plus précises. - Simulations avec six algorithmes de classifications + des fonctions d'estimations des erreurs des classifieurs. - Classification parcimonieuse, plus représentative et pragmatique - Génère explicitement bien les techniques de validation des combinaisons des variables. - Apprentissage, validation, tests plus détaillés/complets du modèle - Présence des bons résultats des indicateurs de performances du modèle avec 6 algorithmes de classification.	- Réduction des variables avec arbre de décision simple. - Plus grande dimension de la taille des variables étudiés : Division et catégorie des variables. - Simulations avec des algorithmes de classifications. - Création et tests du modèle de classification et prédiction - Absence des résultats des indicateurs de performances avec le réseau neuronal artificiel du modèle.

Les réseaux bayésiens dynamiques permettent de faire une modélisation même si les données sont de nature incertaine. Les algorithmes dédiés au calcul offrent un outil puissant pour la fusion des données manquantes ou incomplètes.

La valeur ajoutée à utilisateurs de SCADA avec des réseaux bayésiens dynamiques comme méthode de priorisation des risques sous forme de séquences d'évènements corrélés ont été évalué sur différents facteurs de succès. Ces facteurs sont basés sur la méthodologie adoptée : le jeu de données du système industriel choisi, la construction des réseaux ainsi que la performance de l'algorithme (EBDBN) emprunté pour la simulation.

Nous avons enfin évalué notre modèle avec les six algorithmes de classification. Notre évaluation est assez décisive et plus représentatif. Ainsi, pour chaque algorithme de classification emprunté dans les séries d'expérimentations réalisées, les différentes valeurs des indicateurs de performance obtenus sont tous assez significatifs, et les taux d'erreurs estimées sont très faibles pour l'ensemble de la prédiction par période de temps donnée.

5.2 Limites de la thèse

Nous avons étudié le développement d'un réseau d'interface pour un système SCADA ou système de contrôle et d'acquisition de données afin de démontrer notre méthode d'analyse de la structure causale des risques. Le jeu de données provient d'une usine de fabrication des semi-conducteurs. Ces données comportent 1567 observations datées à la seconde près et sur une durée de 3 mois, dont 104 évènements de défaillances/pannes et rapportant 590 variables continues.

Nous constatons que les mesures des données métrologiques sont problématiques. En effet, l'intervalle de la prise de mesure entre les variables est trop large, autrement dit, les distances de prise de mesure entre les variables sont très étendues. Par conséquent, ces données ne sont pas très cohérentes et probantes pour effectuer des tests pour un modèle de prédiction plus poussés.
En utilisant de l'algorithme de *Random Forest*, les variables continues ont été réduites de 590 à 15. Ceci conduit à une limitation de cette recherche. En effet, nous avons expérimenté un modèle de prédiction avec les 15 variables continues entre T1-T2, T2-T3, ..., T9-T10.

5.3 Discussion et recommandation

Dans l'industrie des semi-conducteurs, le processus de fabrication est très complexe. Les ingénieurs, les analystes et les experts ont souvent des données de production très volumineuses et parfois conflictuelles. Ces données rendent onéreuses les analyses et les processus de maintien de la productivité et l'efficacité des opérations ainsi que l'application des stratégies de maintenance et la fiabilité du système.

Les réseaux bayésiens représentent une valeur ajoutée très significative dans un processus de fabrication des semi-conducteurs qui requièrent une mesure automatique des données de production volumineuse, une technique d'analyse de données rigoureuse, et une extraction de données pertinente, donc on devrait les intégrer dans les systèmes SCADA.

Les réseaux bayésiens dynamiques donnent une interface plus intuitive du point de vue de l'ingénieur du système de production. Ainsi, il serait recommandable d'utiliser ces réseaux dynamiques dans les systèmes SCADA ou des liens intertemporels entres les variables sont présents. Ces réseaux permettraient d'identifier les points de surveillance et d'intervention pour prévenir les pannes ou défaillances du système.

Nous recommandons une autre recherche basée sur une bonne analyse et évaluation de ce jeu de données pour performer d'avantage les modèles de prédictions par les réseaux bayésiens dynamiques. Le nombre d'observations doit également permettre un grand nombre d'occurrences dans les chaînes d'évènements, de l'ordre des centaines préférablement.

Conclusion

La complexité de détecter des phénomènes de défaillances dans un système industriel nous amène à rechercher des moyens pour améliorer la gestion de risques et d'assurer les stratégies de maintenance et la fiabilité du système.

En fait, les différents travaux publiés et la revue de littérature réalisée dans les approches statistiques et probabilistes pour les systèmes industriels complexes nous ont permis de comprendre la particularité et la performance des réseaux bayésiens dynamiques et ainsi choisir les différents cas étudiés. Ces travaux publiés nous ont également permis d'évaluer la pertinence de l'application des réseaux bayésiens dynamiques pour un système SCADA de données sélectionnées. De plus, ces publications nous ont conduit à élaborer les méthodologies empruntées pour prioriser les risques de défaillances du système telles que la construction des réseaux bayésiens et la simulation.

La construction des réseaux bayésiens statiques et dynamiques ont permis de vérifier et valider les résultats de relations entre des séquences d'états aléatoires du système. Ces réseaux ont montré la structure distinctive des relations causales des risques dans les divers cas étudiés. Les réseaux bayésiens statiques sont considérés comme étant un outil d'aide pour la prise de décision. Tandis que, les réseaux bayésiens temporels ou dynamiques sont utilisés comme outil de prévision pour les différents états du système. La qualité de l'offre en matière d'outils rend les réseaux bayésiens dynamiques de plus en plus séduisants pour les applications industrielles complexes.

Les simulations que nous avons conduites ont permis de vérifier la performance et l'efficacité des réseaux bayésiens statiques et dynamiques. Les données des simulations ont apporté de bons résultats pour fins de diagnostic, afin de détecter les risques des défaillances du système.

En conclusion, cette thèse réunit les avantages des réseaux bayésiens dynamiques pour prioriser les risques de défaillances aléatoires du système industriel. En effet, le développement des réseaux bayésiens dynamiques pour un système SCADA favorise une perception spontanée et réaliste de la situation dans les processus industriels complexes. En conséquence, l'équipe de production (Ingénieurs, Opérateurs, Experts, etc.) sont en meilleure position pour interagir et prendre rapidement les décisions nécessaires.

Plusieurs avenues utilisant notre méthodologie de vérification et validation des résultats de simulation au moyen des réseaux bayésiens dynamiques peuvent être explorées.

ANNEXE I - Réseau bayésien dynamique du système de longue marche (LongRun)

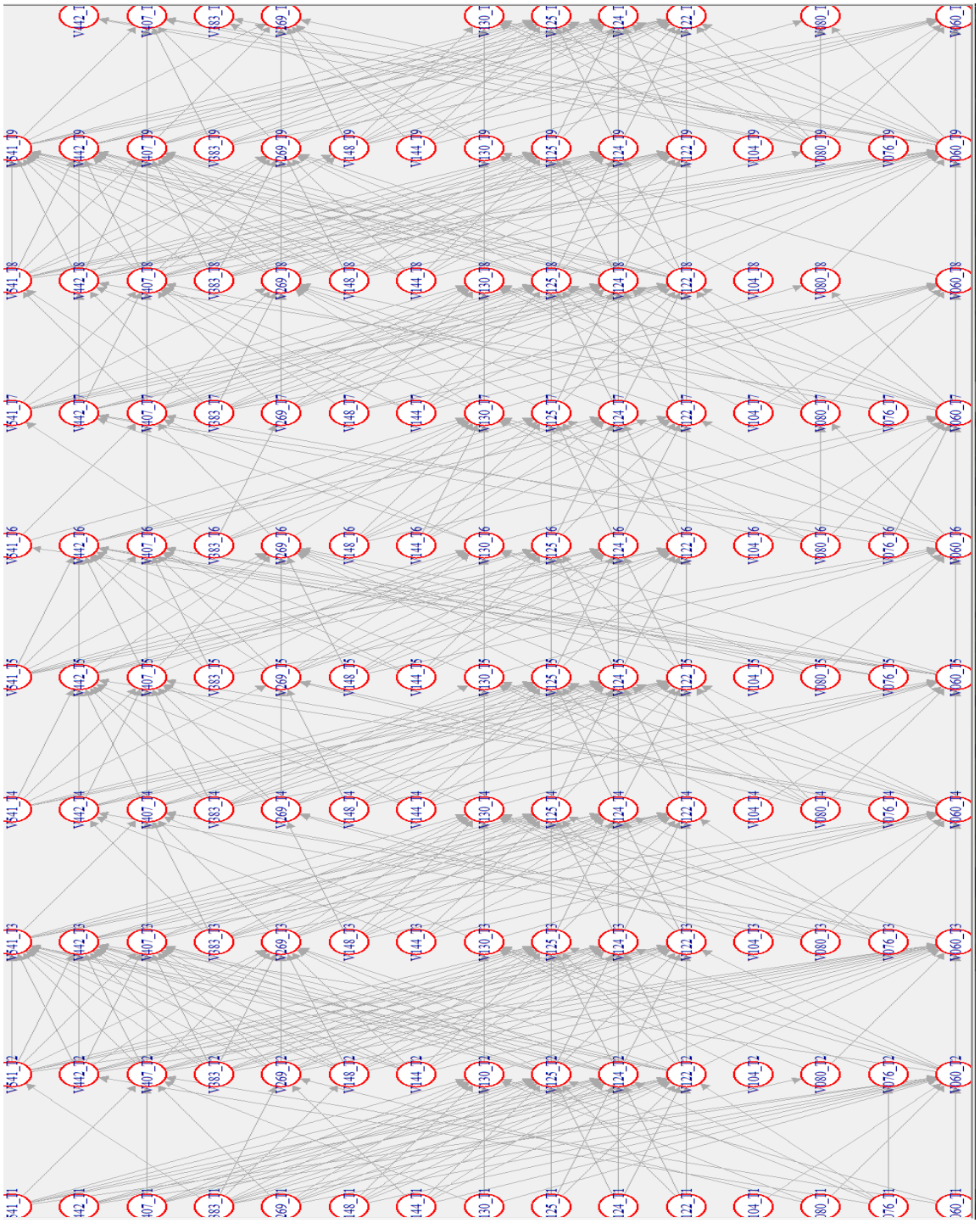


Figure 13 : Réseau bayésien dynamique du système de longue marche (LongRun)

ANNEXE II - Réseau bayésien dynamique du système de courte marche (ShortRun)

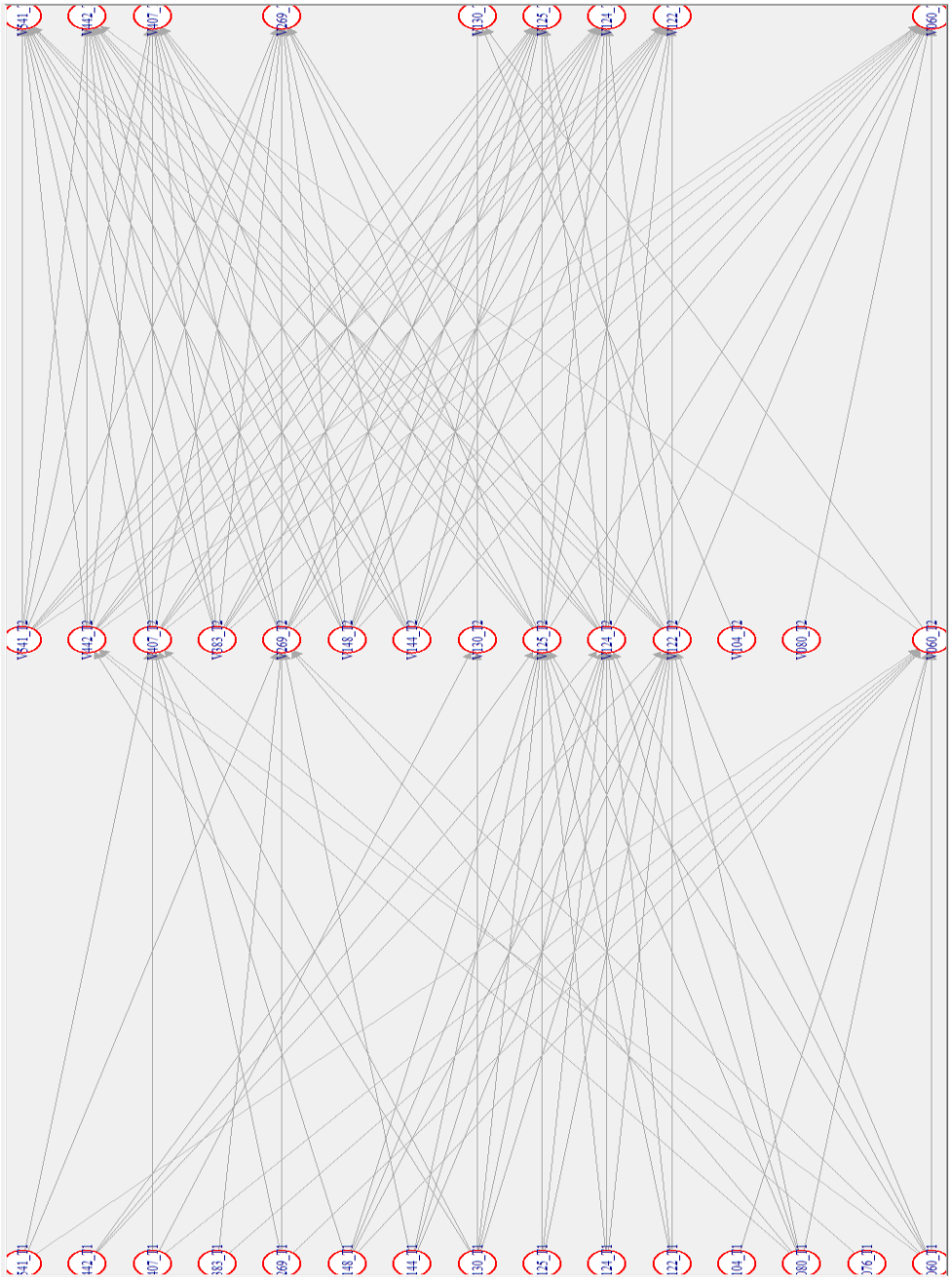


Figure 14 : Réseau bayésien dynamique du système de courte marche (ShortRun)

ANNEXE III - Réseau bayésien dynamique du système en décrochage (Stall)

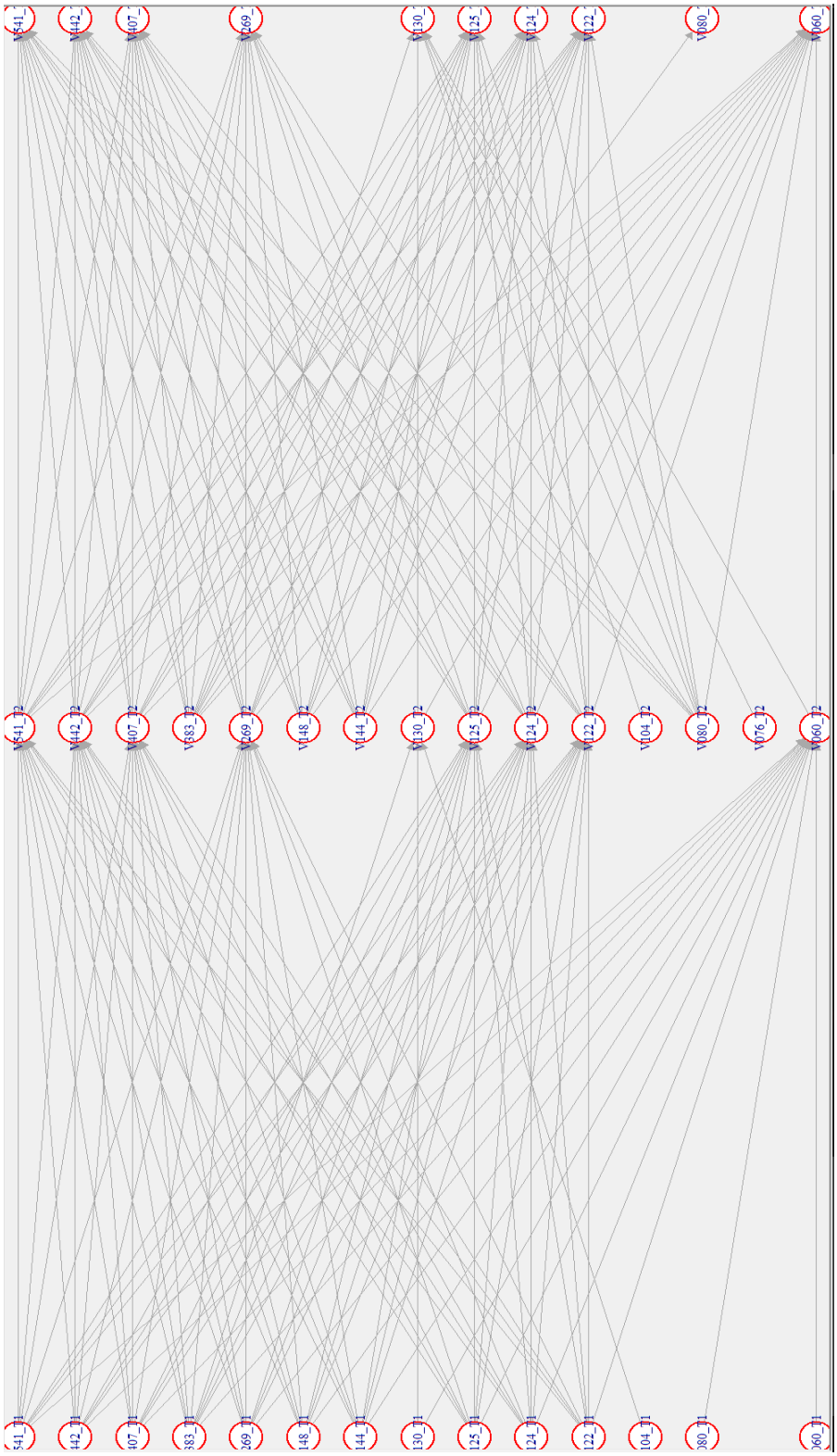


Figure 15 : Réseau bayésien dynamique du système en décrochage (Stall)

ANNEXE IV - Réseau bayésien dynamique du système en démarrage (Start)

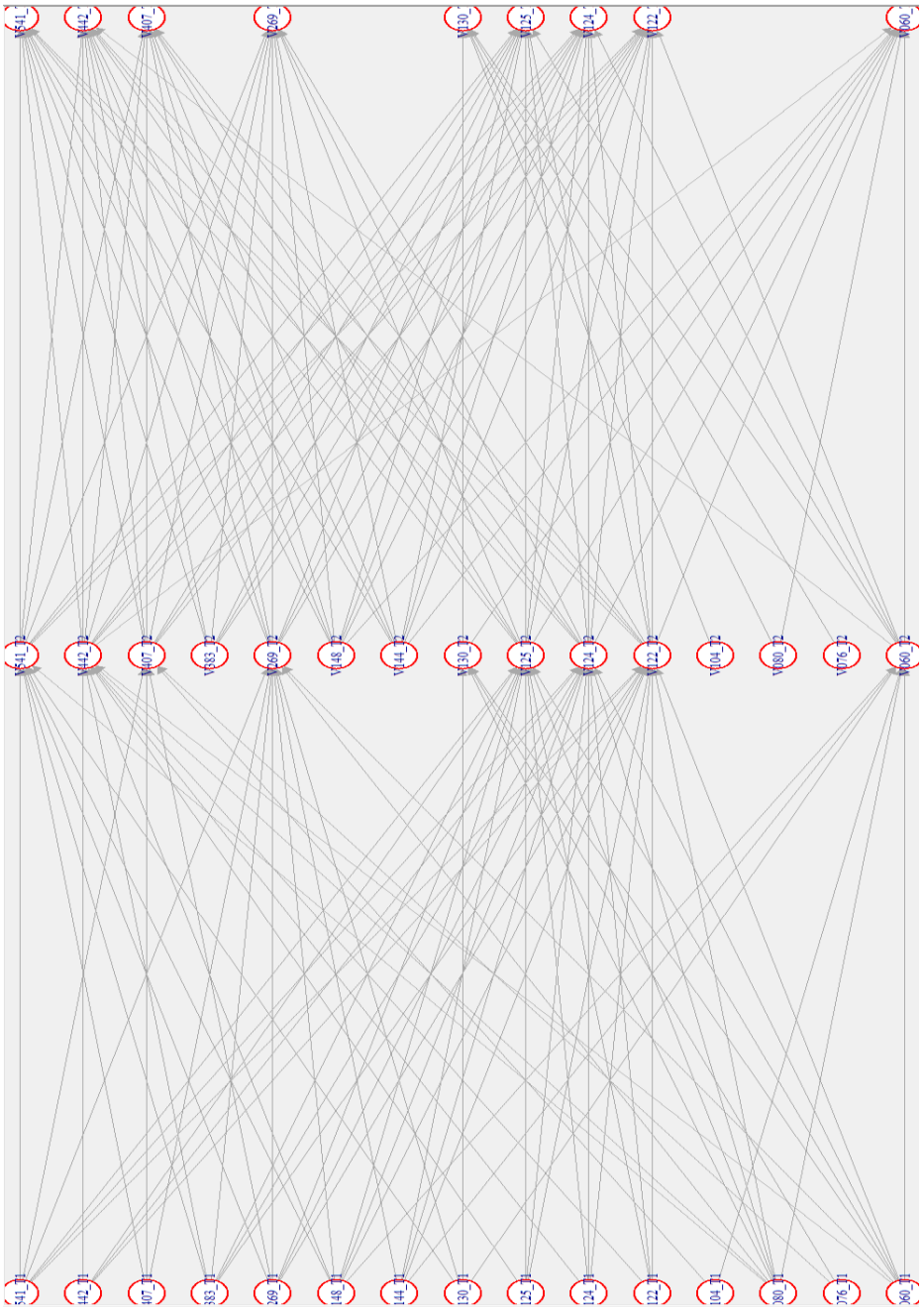


Figure 16 : Réseau bayésien dynamique du système en démarrage (Start)

ANNEXE V - Réseau bayésien dynamique du système de long redémarrage (Longrestart)

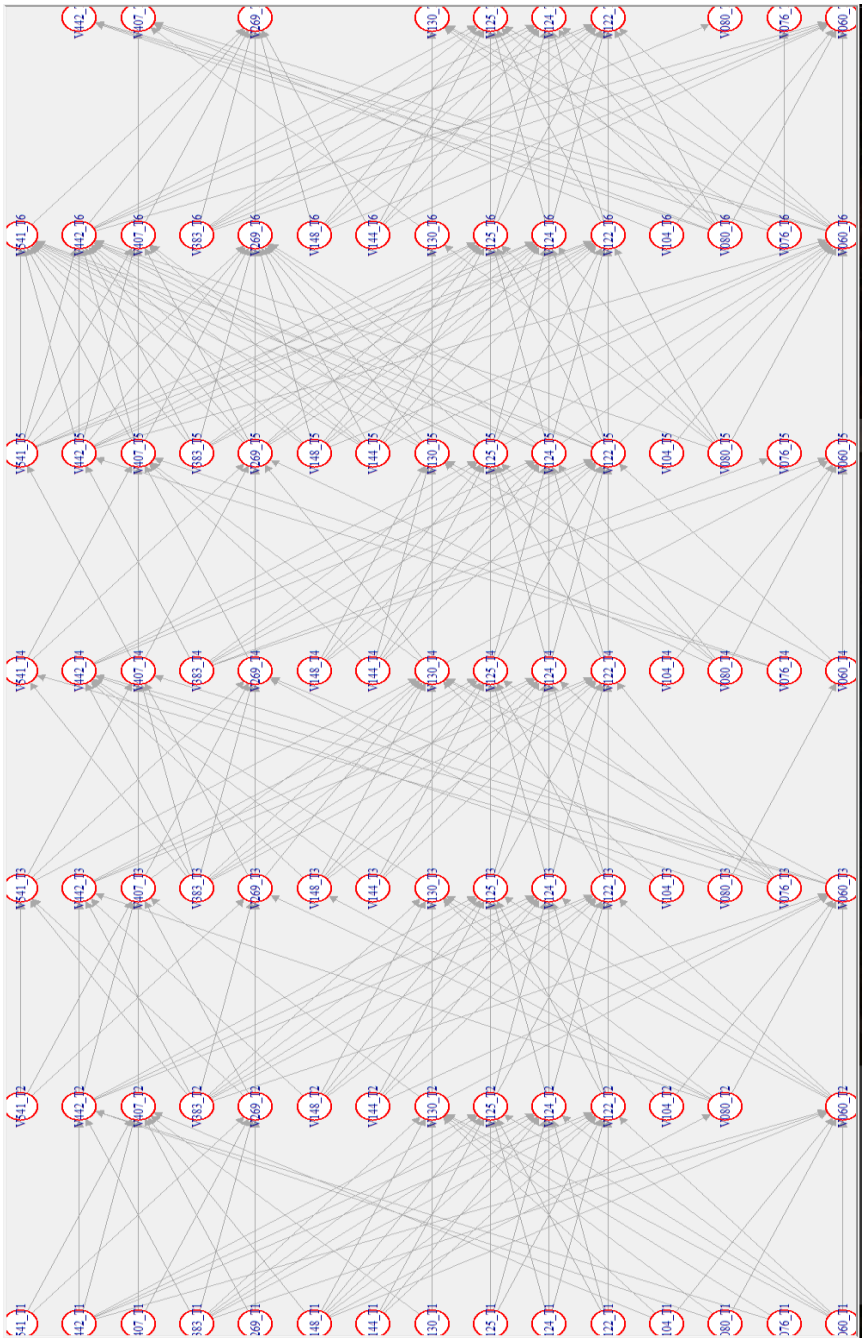


Figure 17 : Réseau bayésien du système de Long redémarrage (Long restart)

ANNEXE VI - Réseau bayésien dynamique du système de court redémarrage (Shortrestart)

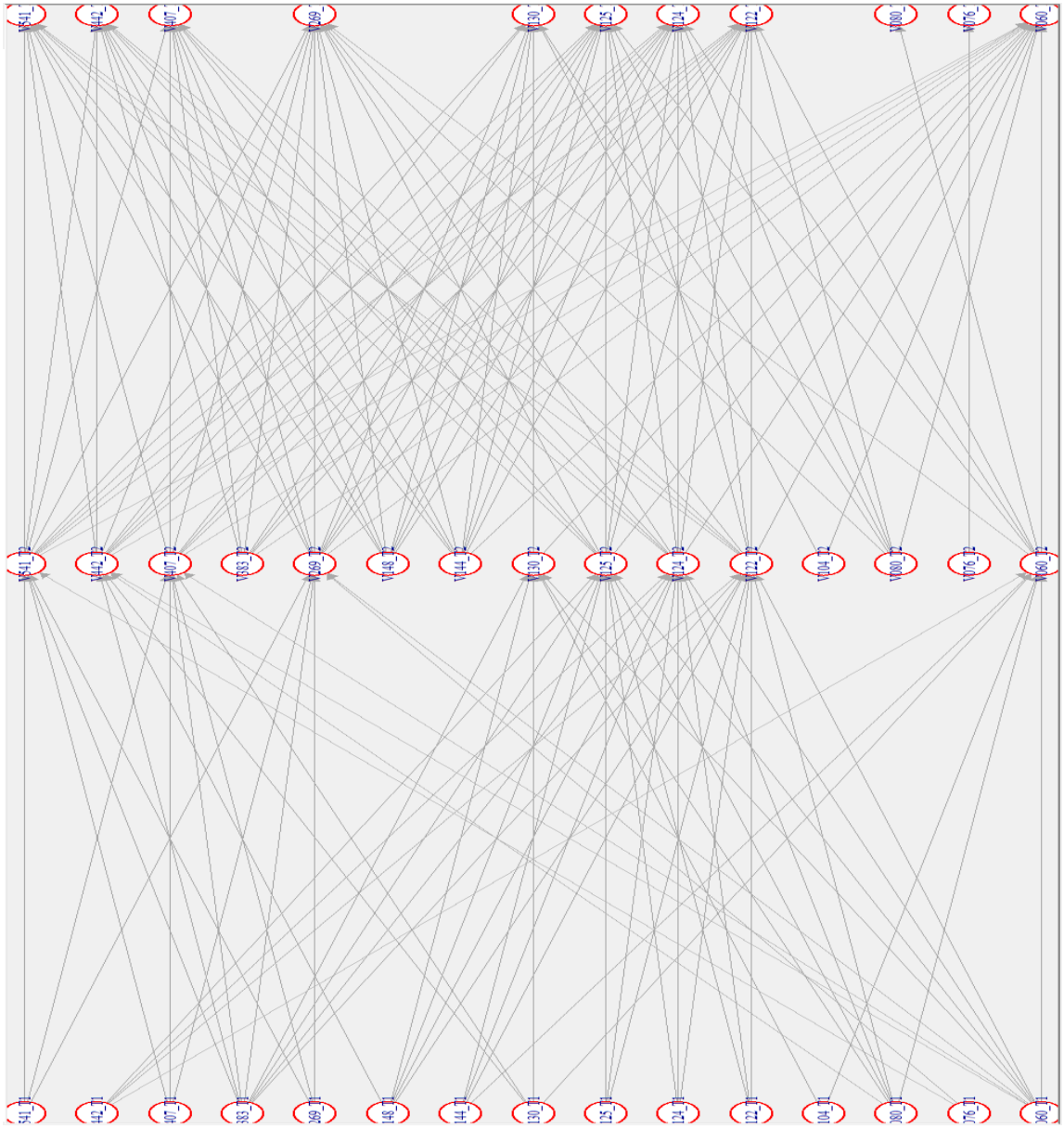


Figure 18 : Réseau bayésien dynamique du système court redémarrage (Shortrestart)

ANNEXE VII - Tableau des degrés du scénario longue marche (Longrun)

Tableau 11 : Nombre des relations entrants et sortants entre les variables en chaque période de temps (Scénario de longue marche)

	T1	T2	T2	T3	T3	T4	T4	T5	T5	T6	T6	T7	T7	T8	T8	T9	T9	T10
Variables	Sortant	Entrant	Sortant	Entrant	Sortant	Entrant	Sortant	Entrant	Sortant	Entrant	Sortant	Entrant	Sortant	Entrant	Sortant	Entrant	Sortant	Entrant
V060_T1	3	10	6	13	3	8	7	7	6	6	6	6	7	7	2	11	9	7
V076_T1	5	1	1	0	5	0	1	0	2	0	2	0	1	0	0	0	1	0
V080_T1	3	1	2	0	4	0	6	0	4	0	5	2	6	2	1	1	8	3
V104_T1	1	0	1	0	3	0	2	0	1	0	2	0	2	0	1	0	2	0
V122_T1	3	11	8	11	3	11	3	12	4	8	3	7	4	11	8	10	3	12
V124_T1	3	11	8	11	3	10	3	12	4	8	3	8	4	11	8	10	3	12
V125_T1	3	11	8	11	3	11	3	12	4	8	3	8	4	11	8	10	3	12
V130_T1	3	8	2	5	3	10	3	5	3	7	4	10	4	9	1	3	4	5
V144_T1	5	0	8	0	5	0	6	0	5	0	4	0	6	0	8	0	5	0
V148_T1	5	1	8	0	6	0	5	0	5	0	5	0	4	0	8	2	7	0
V269_T1	6	3	8	10	5	3	6	5	7	10	2	1	6	9	8	10	5	7
V383_T1	9	0	8	0	7	0	7	0	4	0	5	0	8	0	7	0	5	2
V407_T1	5	5	8	10	4	7	5	8	4	8	2	5	7	8	9	10	5	6
V442_T1	4	1	8	11	4	3	5	7	5	9	3	3	5	4	8	9	3	2
V541_T1	6	1	8	10	5	0	6	0	7	1	2	1	6	2	9	10	5	0

ANNEXE VIII - Tableau des degrés du scénario courte marche (Shortrun)

Tableau 12 : Nombre des relations entrants et sortants entre les variables en chaque période de temps (Scénario de courte marche)

	T1		T2	
Variables	Sortants	Entrants	Sortants	Entrants
V060_T1	6	8	3	12
V076_T1	1	0	0	0
V080_T1	6	0	1	0
V104_T1	1	0	1	0
V122_T1	3	9	8	10
V124_T1	3	9	8	10
V125_T1	3	9	8	10
V130_T1	6	2	1	3
V144_T1	4	0	8	0
V148_T1	5	0	8	0
V269_T1	3	6	8	10
V383_T1	2	0	8	0
V407_T1	3	6	8	10
V442_T1	3	3	8	11
V541_T1	3	0	8	10

ANNEXE IX - Tableau des degrés du scénario en décrochage (Stall)

Tableau 13 : Nombre des relations entrants et sortants entre les variables en chaque période de temps (Scénario en décrochage)

	T1	T2	T2	T3
Variables	Sortants	Entrants	Sortants	Entrants
V060_T1	1	13	2	13
V076_T1	0	0	1	0
V080_T1	1	0	9	1
V104_T1	1	0	1	0
V122_T1	8	10	8	11
V124_T1	8	10	8	11
V125_T1	8	10	8	11
V130_T1	2	3	2	6
V144_T1	8	0	8	0
V148_T1	8	0	8	0
V269_T1	8	10	8	11
V383_T1	9	0	10	0
V407_T1	8	10	8	11
V442_T1	8	10	8	11
V541_T1	8	10	8	11

ANNEXE X - Tableau des degrés du scénario en démarrage (Start)

Tableau 14 : Nombre des relations entrants et sortants entre les variables en chaque période de temps (Scénario en démarrage)

	T1		T2	
Variables	Sortants	Entrants	Sortants	Entrants
V060_T1	7	6	6	9
V076_T1	2	0	1	0
V080_T1	8	0	2	0
V104_T1	2	0	1	0
V122_T1	4	11	8	11
V124_T1	4	11	8	11
V125_T1	4	11	8	11
V130_T1	4	4	2	5
V144_T1	5	0	8	0
V148_T1	7	0	8	0
V269_T1	6	9	7	10
V383_T1	6	0	6	0
V407_T1	3	5	7	9
V442_T1	4	8	8	11
V541_T1	6	7	7	10

ANNEXE XI - Tableau des degrés du scénario de long redémarrage (Longrestart)

Tableau 15 : Nombre des relations entrants et sortants entre les variables en chaque période de temps (Scénario long redémarrage)

	T1	T2	T2	T3	T3	T4	T4	T5	T5	T6	T6	T7
Variables	Sortants	Entrants	Sortants	Entrants	Sortants	Entrants	Sortants	Entrants	Sortants	Entrants	Sortants	Entrants
V060_T1	5	6	5	6	3	1	3	5	1	9	7	7
V076_T1	2	0	0	0	6	0	2	1	0	0	3	1
V080_T1	3	1	5	0	2	0	5	0	6	0	7	1
V104_T1	3	0	2	0	2	0	1	0	1	0	2	0
V122_T1	3	8	3	8	3	8	3	8	7	10	3	9
V124_T1	3	8	3	8	3	8	3	8	7	10	3	9
V125_T1	3	8	3	8	3	8	3	8	7	10	3	9
V130_T1	2	7	3	6	2	9	4	5	4	2	2	4
V144_T1	3	0	4	0	4	0	5	0	8	0	5	0
V148_T1	7	0	5	1	5	0	4	0	7	0	5	0
V269_T1	2	4	3	4	3	6	2	5	7	11	1	7
V383_T1	7	0	8	0	8	0	6	0	6	0	6	0
V407_T1	2	7	1	5	3	5	3	6	4	6	2	5
V442_T1	6	4	5	4	3	6	3	2	8	11	5	3
V541_T1	2	0	3	3	3	2	2	1	7	11	1	0

ANNEXE XII - Tableau des degrés du scénario de court redémarrage (Shortrestart)

Tableau 16 : Nombre des relations entrants et sortants entre les variables en chaque période de temps (Scénario court redémarrage)

	T1	T2	T2	T3
Variables	Sortants	Entrants	Sortants	Entrants
V060_T1	8	6	6	13
V076_T1	2	0	1	1
V080_T1	6	0	6	1
V104_T1	2	0	2	0
V122_T1	3	9	8	12
V124_T1	3	9	8	12
V125_T1	3	9	8	12
V130_T1	4	6	2	8
V144_T1	4	0	9	0
V148_T1	5	0	8	0
V269_T1	3	6	9	12
V383_T1	8	0	7	0
V407_T1	3	6	7	9
V442_T1	4	5	8	9
V541_T1	3	5	9	9

ANNEXE XIII – Résultats du modèle de prévision

1. Algorithme d'arbre de décision du modèle de prédiction

Le modèle avec les deux états (échec et marche) du système est testé avec la classification d'arbre de décision. Les résultats obtenus avec l'arbre de décision (C4.5 dans Tanagra) sont les suivants :

Performances des classifieurs :

- Taux d'erreur = 0,0417

Tableau 17 : Valeur de prédiction avec l'algorithme de classification d'arbre de décision

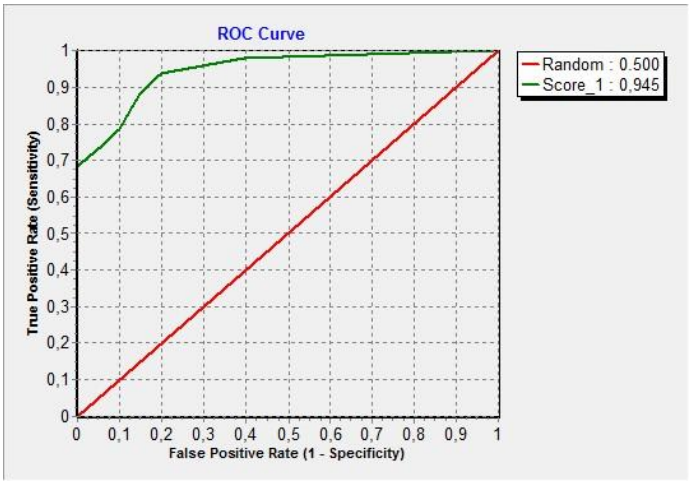
État	Rappel	1-Précision
Échec	0,5873	0,3833
Marche	0,9793	0,0233

Tableau 18 : Matrice de confusion avec l'algorithme de classification d'arbre de décision

État	Échec	Marche	Somme
Échec	37	26	63
Marche	23	1088	1111
Somme	60	1114	1174

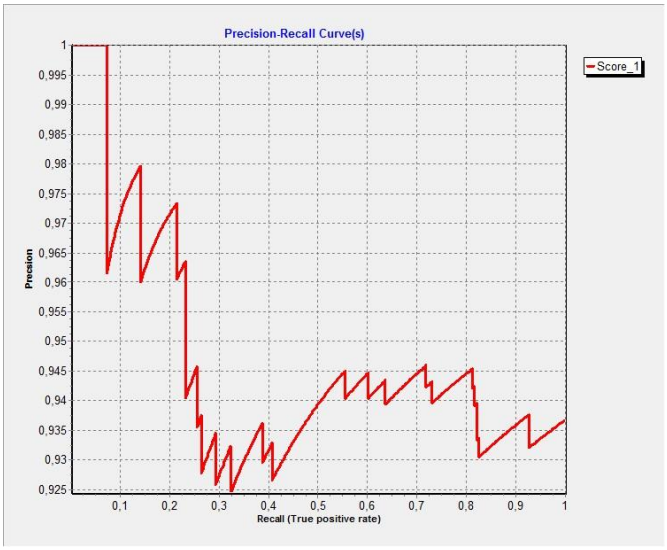
Courbe de sensibilité et spécificité *(ROC curve)*

Figure 19 : Courbe ROC de l'algorithme de classification d'arbre de décision



Courbe de rappel de précision *(Precision Recall curve)*

Figure 20 : Courbe de rappel de précision avec l'algorithme d'arbre de décision



2. Algorithme K plus proches voisins du modèle de prédiction

Le modèle avec les deux états du système est testé avec l'algorithme de classification K plus proches voisins ou k-ppv (K-NN dans Tanagra). Les résultats obtenus sont les suivants :

Performances des classifieurs

- Taux d'erreur = 0,0537

Tableau 19 : Valeur de prédiction avec l'algorithme de classification K-ppv

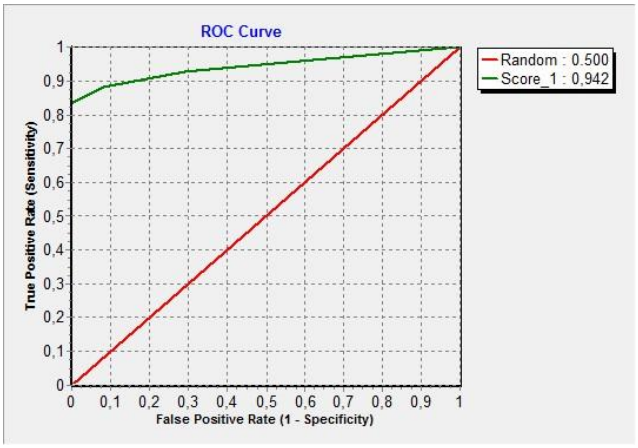
État	Rappel	1-Précision
Échec	0,0317	0,5000
En Marche	0,9982	0,0521

Tableau 20 : Matrice de confusion avec l'algorithme de classification K-ppv

État	Échec	Marche	Somme
Échec	2	61	63
Marche	2	1109	1111
Somme	4	1170	1174

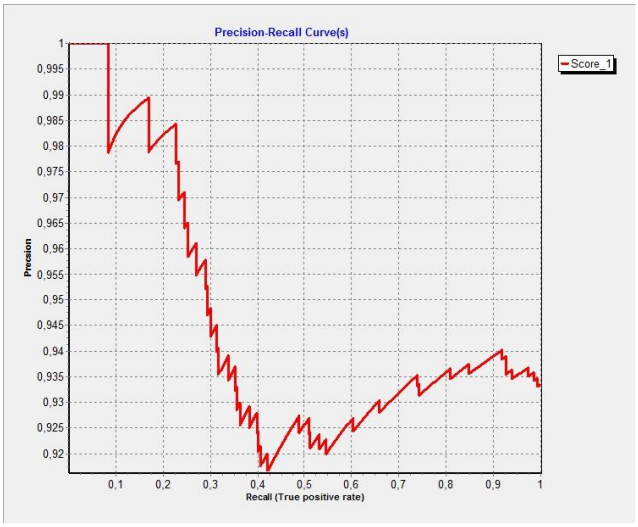
Courbe de sensibilité et spécificité *(ROC curve)*

Figure 21 : Courbe ROC avec l'algorithme de classification K-ppv



Courbe de Rappel de précision *(Precision rappel curve)*

Figure 22 : Courbe de Rappel de précision avec l'algorithme de classification K-ppv



3. Algorithme bayésien naïf du modèle de prédiction

Le modèle avec les deux états du système est testé avec l'algorithme de classification bayésienne naïve (Naïves Bayes continue dans Tanagra). Les résultats obtenus sont les suivants :

Performances des classifieurs

- Taux d'erreur = 0,0596

Tableau 21 : Valeur de prédiction avec l'algorithme de classification Bayésienne naïve

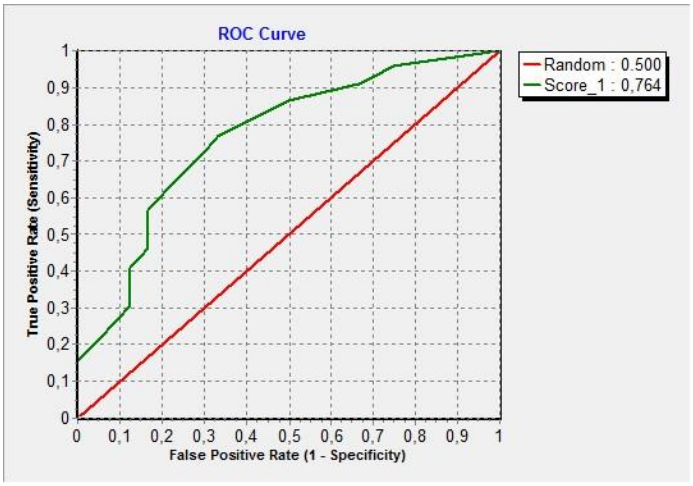
État	Rappel	1-Précision
Échec	0,0794	0,7059
En Marche	0,9892	0,0501

Tableau 22 : Matrice de confusion avec l'algorithme de classification Bayésienne naïve

État	Échec	Marche	Somme
Échec	7	58	63
Marche	12	1099	1111
Somme	17	1157	1174

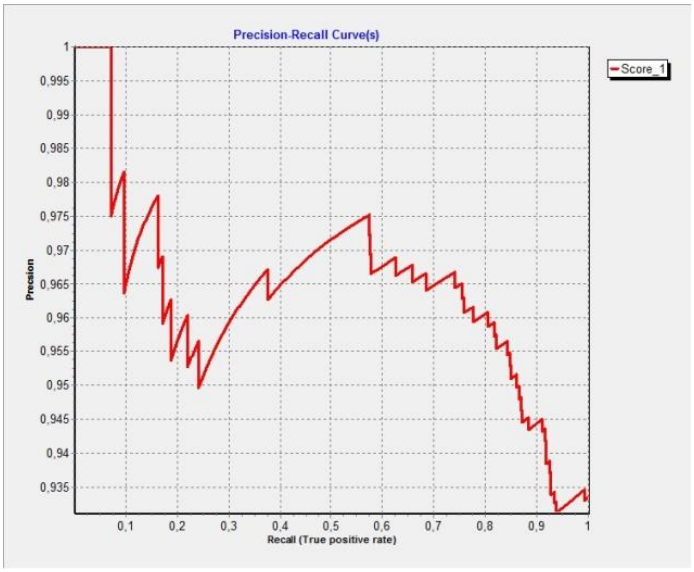
Courbe de sensibilité et spécificité (ROC curve)

Figure 23 : Courbe ROC avec l'algorithme de classification Bayésienne naïve



Courbe de Rappel de précision *(Precision rappel curve)*

Figure 24 : Courbe de Rappel de précision avec la classification Bayésienne naïve



4. Algorithme de séparateurs à vaste marge du modèle de prédiction

Le modèle avec les deux états du système est testé avec l'algorithme des séparateurs à vaste marge (SVM dans Tanagra). Les résultats obtenus sont les suivants :

Performances des classifieurs

-Taux d'erreur = 0,0545

Tableau 23 : Valeur de prédiction avec l'algorithme des séparateurs à vaste marge

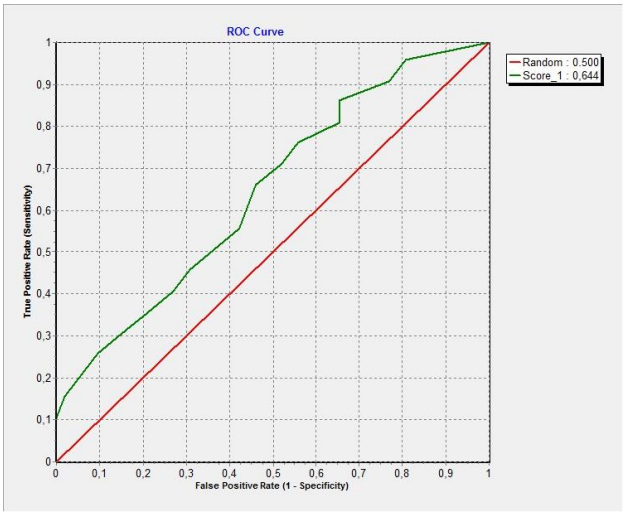
État	Rappel	1-Précision
Échec	0,0011	1,0000
Marche	1,0000	0,0537

Tableau 24 : Matrice de confusion avec l'algorithme des séparateurs à vaste marge

État	Échec	Marche	Somme
Échec	0	63	63
Marche	1	1110	1111
Somme	1	1174	1174

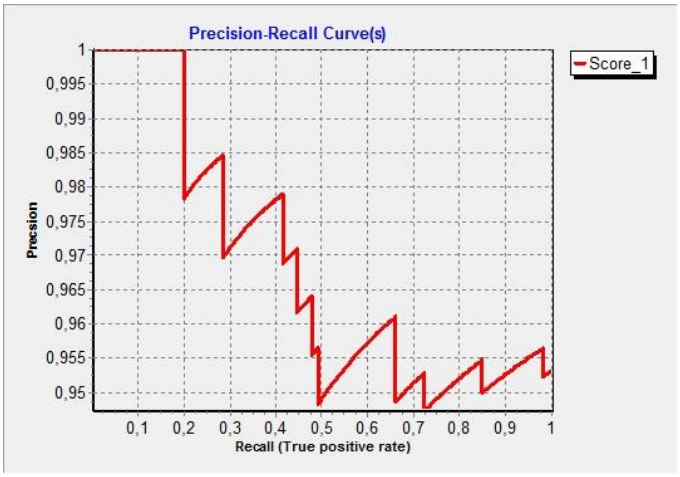
Courbe de sensibilité et spécificité *(ROC curve)*

Figure 25 : Courbe ROC avec l'algorithme des séparateurs à vaste marge



Courbe de rappel de précision *(Precision Recall curve)*

Figure 26 : Courbe de Rappel de précision avec l'algorithme des séparateurs à vaste marge



5. Algorithme de régression logistique du modèle de prédiction

Le modèle avec les deux états du système est testé avec l'algorithme de régression logistique (Binary Logistic Regresion dans Tanagra). Les résultats obtenus sont les suivants :

Performances des classifieurs

-Taux d'erreur = 0,0545

Tableau 25 : Valeur de prédiction avec l'algorithme de régression logistique

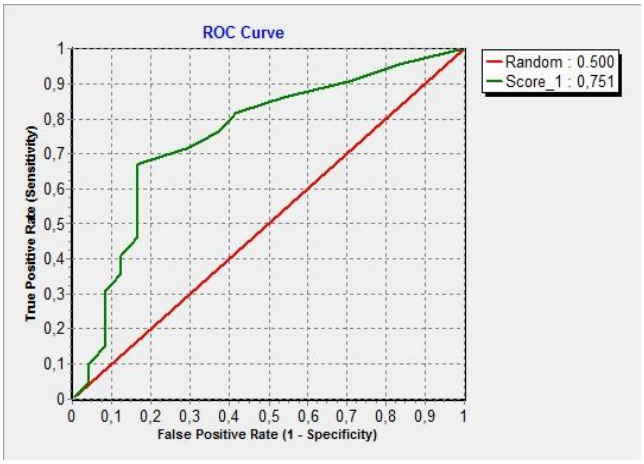
État	Rappel	1-Précision
Échec	0,0159	0,6667
En Marche	0,9982	0,0529

Tableau 26 : Matrice de confusion avec l'algorithme de régression logistique

État	Échec	Marche	Somme
Échec	1	62	63
Marche	2	1109	1111
Somme	3	1171	1174

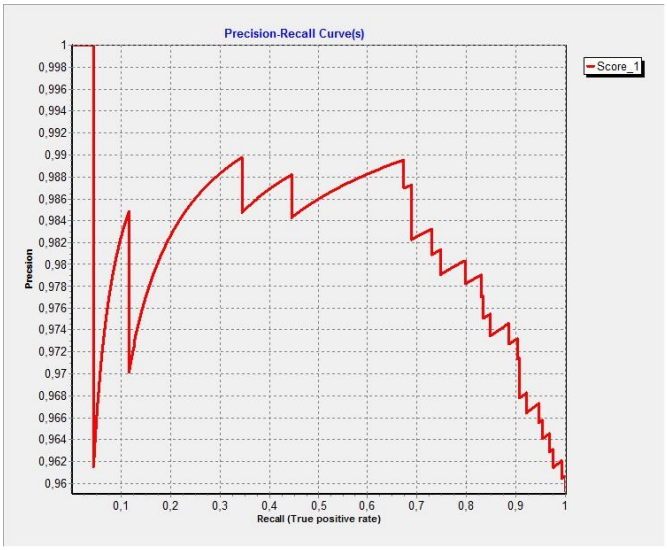
Courbe de sensibilité et spécificité *(ROC curve)*

Figure 27 : Courbe ROC avec l'algorithme de régression logistique



Courbe de rappel de précision *(Precision Recall curve)*

Figure 28 : Courbe de Rappel de précision avec l'algorithme de régression logistique



6. Algorithme de réseau neuronal artificiel du modèle de prédiction

Le modèle avec les deux états du système est testé avec l'algorithme de réseau neuronal artificiel multicouches (Multi layer Perceptron dans Tanagra) sont les suivants :

Performances des classifieurs

- Taux d'erreur = 0,0520

Tableau 27 : Valeur de prédiction avec l'algorithme de réseau neuronal artificiel

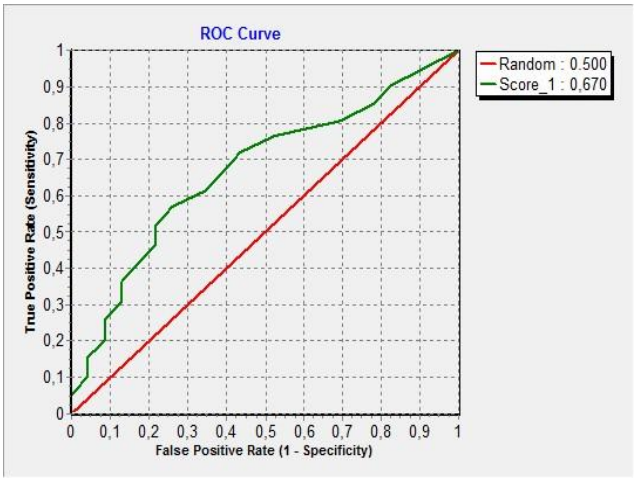
État	Rappel	1-Précision
Échec	0,0635	0,3333
En Marche	0,9982	0,0505

Tableau 28 : Matrice de confusion avec l'algorithme de réseau neuronal artificiel

État	Échec	Marche	Somme
Échec	4	59	63
Marche	2	1109	1111
Somme	6	1168	1174

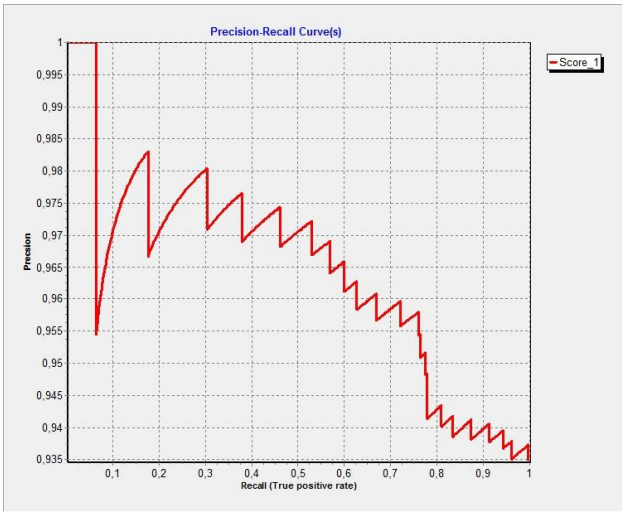
Courbe de sensibilité et spécificité (ROC curve)

Figure 29 : Courbe ROC avec l'algorithme de réseau neuronal artificiel



Courbe de rappel de précision *(Precision Recall curve)*

Figure 30 : Courbe de Rappel de précision avec l'algorithme de réseau neuronal artificiel



ANNEXE XIV – Résultats des fonctions "Train-Test", "Cross-Validation", et "Bootstrap"

1. Évaluation du modèle avec la fonction *"Train-Test"*

Pour chaque algorithme de classification, les séries de tests sont effectués à une proportion de 70% (valeur acceptable) des échantillons.

1.1 Arbre de décision

- Taux d'erreur = 0,0567

Les tableaux 29 et 30 montrent les résultats obtenus de la fonction Train-Test.

Tableau 29 : Arbre de décision - Valeur de prédiction

État	Rappel	1-Précision
Échec	0,0556	0,7500
Marche	0,9910	0,0487

Tableau 30 : Arbre de décision - Matrice de confusion

État	Échec	Marche	Somme
Échec	1	17	18
Marche	3	332	335
Somme	4	349	353

1.2 K plus proches voisins (K-ppv)

- Taux d'erreur = 0,0562

Les tableaux 31 et 32 montrent les résultats obtenus de la fonction Train-Test.

Tableau 31 : K-ppv - Valeur de prédiction

État		Rappel	1-Précision
Échec		0,0000	1,0000
Marche		0,9997	0,0566

Tableau 32 : K-ppv - Matrice de confusion

État	Échec	Marche	Somme
Échec	0	20	20
Marche	0	333	333
Somme	0	353	353

1.3 Bayésien naïf

- Taux d'erreur : 0,0652

Les tableaux 33 et 34 montrent les résultats obtenus de la fonction Train-Test.

Tableau 33 : Bayésien naïf - Valeur de prédiction

État	Rappel	1-Précision
Échec	0,0714	0,9091
En Marche	0,9705	0,0380

Tableau 34 : Bayésien naïf - Matrice de confusion

État	Échec	Marche	Somme
Échec	1	13	14
Marche	10	329	339
Somme	11	342	353

1.4 Régression Logistique

- Taux d'erreur : 0,0482

Les tableaux 35 et 36 montrent les résultats obtenus de la fonction Train-Test.

Tableau 35 : Régression logistique - Valeur de prédiction

État	Rappel	1-Précision
Échec	0,0006	1,0000
En Marche	0,9989	0,0482

Tableau 36 : Régression logistique - Matrice de confusion

État	Échec	Marche	Somme
Échec	0	17	17
Marche	0	336	336
Somme	0	353	353

1.5 Séparateurs à vaste marge

- Taux d'erreur : 0,0538

Les tableaux 37 et 38 présentent les résultats obtenus de la fonction Train-Test.

Tableau 37 : Séparateurs à vaste marge - Valeur de prédiction

État	Rappel	1-Précision
Échec	0,0000	1,0000
Marche	1,0000	0,0538

Tableau 38 : Séparateurs à vaste marge - Matrice de confusion

État	Échec	Marche	Somme
Échec	1	17	18
Marche	0	334	334
Somme	1	353	353

1.6 Réseau neuronal artificiel

- Taux d'erreur : 0,0425

Les tableaux 39 et 40 montrent les résultats obtenus de la fonction Train-Test.

Tableau 39 : Réseau neuronal artificiel - Valeur de prédiction

État	Rappel	1-Précision
Échec	0,0012	1,0000
Marche	0,9941	0,0370

Tableau 40 : Réseau neuronal artificiel - Matrice de confusion

État	Échec	Marche	Somme
Échec	0	13	13
Marche	2	338	340
Somme	2	351	353

2. Évaluation du modèle avec les fonctions « validation croisée » (*Cross-Validation)* et *« Bootstrap »*

Pour chaque algorithme de classification, les séries de tests sont effectués pour les deux fonctions ; la validation croisée *(Cross-Validation)* et *« Bootstrap »*.

2.1 Arbre de décision

Les tableaux 41, 42 et 43 présentent les résultats obtenus de ces deux fonctions.

Fonction validation croisée

- Taux d'erreur = 0,0880

Tableau 41 : Arbre de décision - Valeur de prédiction

État	Rappel	1-Précision
Échec	0,0968	0,8868
Marche	0,9576	0,0501

Tableau 42 : Arbre de décision - Matrice de confusion

État	Échec	Marche	Somme
Échec	6	56	62
Marche	47	1061	1108
Somme	53	1117	1170

Fonction *"Bootstrap"*

Tableau 43 : Arbre de décision - Valeur de taux d'erreur

Répétition	Test set Erreur	Bootstrap	Bootstrap +
1	0,0910	0,0729	0,0872

Nous constatons que l'écart de taux d'erreur estimé de la validation croisée et du Bootstrap est très minime (moins de 0,8% ; écart : 0,0880 -0,0872 =0,008).

2.2 K plus proche de voisin (K-ppv)

Les tableaux 44, 45 et 46 montrent les résultats obtenus de ces deux fonctions.

Fonction validation croisée

- Taux d'erreur = 0,0547

Tableau 44 : K-ppv - Valeur de prédiction

État	Rappel	1-Précision
Échec	0,0011	1,0000
Marche	0,9982	0,0531

Tableau 45 : K-ppv - Matrice de confusion

État	Échec	Marche	Somme
Échec	0	62	62
Marche	0	1106	1108
Somme	0	1168	1170

Fonction *"Bootstrap"*

Tableau 46 : K-ppv - Valeur de taux d'erreur

Répétition	Test set Erreur	Bootstrap	Bootstrap +
1	0,0700	0,0640	0,0601

Nous constatons que l'écart de taux d'erreur estimé de la validation croisée et du Bootstrap est assez faible (moins de 0,93% ; écart : 0,0640 -0,0547 =0,0093).

2.3 Bayésien naïf

Les tableaux 47, 48 et 49 montrent les résultats obtenus de ces deux fonctions.

Fonction validation croisée

- Taux d'erreur = 0,0624

Tableau 47 : Bayésien naïf - Valeur de prédiction

État	Rappel	1-Précision
Échec	0,0645	0,7895
Marche	0,9865	0,0504

Tableau 48 : Bayésien naïf - Matrice de confusion

État	Échec	Marche	Somme
Échec	4	58	62
Marche	15	1093	1108
Somme	19	1151	1170

Fonction *"Bootstrap"*

Tableau 49 : Bayésien naïf - Valeur de taux d'erreur

Répétition	Test Erreur	Bootstrap	Bootstrap +
1	0,0782	0,0714	0,3376

Nous constatons que l'écart du taux d'erreur estimé de la validation croisée et du Bootstrap est assez faible (moins de 1%, écart : 0,0714 -0,0624 =0,009).

2.4 Séparateurs à vaste marge (SVM)

Les tableaux 50, 51 et 52 montrent les résultats obtenus de ces deux fonctions.

Fonction validation croisée

- Taux d'erreur = 0,0530

Tableau 50 : Séparateurs à vaste marge - Valeur de prédiction

État	Rappel	1-Précision
Échec	0,0007	1,0000
En Marche	1,0000	0,0530

Tableau 51 : Séparateurs à vaste marge - Matrice de confusion

État	Échec	Marche	Somme
Échec	0	62	62
Marche	0	1108	1108
Somme	0	1170	1170

Fonction *"Bootstrap"*

Tableau 52 : Séparateurs à vaste marge - Valeur de taux d'erreur

Répétition	Test Erreur	Bootstrap	Bootstrap +
1	0,0540	0,0539	0,0539

Nous constatons que l'écart du taux d'erreur estimé de la validation croisée et du Bootstrap est non significatif (moins de 0,9% - écart : 0,0539 -0,0530 =0,009).

2.5 Régression Logistique

Les tableaux 53, 54 et 55 montrent les résultats obtenus de ces deux fonctions.

Fonction validation croisée

- Taux d'erreur = 0,0538

Tableau 53 : Régression logistique - Valeur de prédiction

État	Rappel	1-Précision
Échec	0,0004	1,0000
Marche	0,9991	0,0530

Tableau 54 : Régression logistique - Matrice de confusion

État	Échec	Marche	Somme
Échec	0	62	17
Marche	1	1107	1108
Somme	1	1169	1170

Fonction *"Bootstrap"*

Tableau 55 : Régression logistique - Valeur de taux d'erreur

Répétition	Test Erreur	Bootstrap	Bootstrap +
1	0,0555	0,0551	0,0553

Nous constatons que l'écart d'erreur évalué de la validation croisée et de Bootstrap est non significatif (moins de 0,2% écart : 0,0551 -0,0530 =0,0021).

2.6 Réseau neuronal artificiel

Les tableaux 56, 57 et 58 présentent les résultats obtenus de ces deux fonctions.

Fonction validation croisée

- Taux d'erreur = 0,0556

Tableau 56 : Réseau neuronal artificiel - Valeur de prédiction

État	Rappel	1-Précision
Échec	0,0002	1,0000
En Marche	0,9973	0,0531

Tableau 57 : Réseau neuronal artificiel - Matrice de confusion

État	Échec	Marche	Somme
Échec	0	62	62
Marche	3	1105	1108
Somme	3	1167	1170

Fonction *"Bootstrap"*

Tableau 58 : Réseau neuronal artificiel - Valeur de taux d'erreur

Répétition	Test set Erreur	Bootstrap	Bootstrap +
1	0,0560	0,0552	0,0552

Nous constatons que l'écart de taux d'erreur estimé de la validation croisée et du Bootstrap est non significatif (moins de 0,5% écart : 0,0556 -0,0552 = 0,004).

BIBLIOGRAPHIE

[1] Munirathinam, S., & Ramadoss, B., (2016), "Predictive Models for Equipment Fault Detection in the Semiconductor Manufacturing Process", ***International Journal of Engineering and Technology***, 8 (4): 273.

[2] Vidal, L.-A., & Marle, F., (2008), "Understanding project complexity: implications on project management", ***Kybernetes***, 37 (8): 1094-1110.

[3] Xia, W., & Lee, G., (2005), "Complexity of Information Systems Development Projects: Conceptualization and Measurement Development", ***Journal of Management Information Systems***, 22 (1): 45-83.

[4] Bosch-Rekveldt, M., Jongkind, Y., Mooi, H., Bakker, H., & Verbraeck, A., (2011), "Grasping project complexity in large engineering projects: The TOE (Technical, Organizational and Environmental) framework", ***International Journal of Project Management***, 29 (6): 728-739.

[5] Villemeur, A., (1988), ***Sûreté de fonctionnement des systèmes industriels***, Paris, Eyrolles.

[6] Thamhain, H., (2013), "Managing risks in complex projects", ***Project Management Journal***, 44 (2): 20-35.

[7] Vidal, L.-A., Marle, F., & Bocquet, J.-C., (2011), "Measuring project complexity using the Analytic Hierarchy Process", ***International Journal of Project Management***, 29 (6): 718-727.

[8] Dubé, L., (2014), "Exploring how it professionals experience role transitions at the end of successful projects", ***Journal of Management Information Systems***, 31 (1): 17-45.

[9] Hu, J., Zhang, L., Ma, L., & Liang, W., (2011), "An integrated safety prognosis model for complex system based on dynamic Bayesian network and ant colony algorithm", ***Expert Systems with Applications***, 38 (3): 1431-1446.

[10] Medjaher, K., Mechraoui, A., & Zerhouni, N., (2008), "Diagnostic et pronostic de défaillances par réseaux bayésiens", Paper presented at the ***4èmes Journées Francophones sur les Réseaux Bayésiens, JFRB'2008.***: 80-93.

[11] Mazouni, M. H., & Aubry, J.-F., (2009), "De l'analyse préliminaire de risque au système d'aide à la décision pour le management des risques", Paper presented at the ***8ème Congrès international pluridisciplinaire en Qualité et Sûreté de Fonctionnement, Qualita 2009***: CDROM.

[12] Faucher, J., (2009), ***Pratique de l'AMDEC-2e édition: Assurez la qualité et la sûreté de fonctionnement de vos produits, équipements et procédés***, Dunod.

[13] Tixier, J., Dusserre, G., Salvi, O., & Gaston, D., (2002), "Review of 62 risk analysis methodologies of industrial plants", ***Journal of Loss Prevention in the process industries***, 15 (4): 291-303.

[14] Mili, A., (2009), ***Vers des méthodes fiables de contrôle des procédés par la maîtrise du risque: Contribution à la fiabilisation des méthodes de process control d'une unité de Recherche et de Production de circuits semi-conducteurs***, Institut National Polytechnique de Grenoble-INPG.

[15] Mili, A., Bassetto, S., Siadat, A., & Tollenaere, M., (2009), "Dynamic risk management unveil productivity improvements", ***Journal of Loss Prevention in the Process Industries***, 22 (1): 25-34.

[16] Leroy, A., & Signoret, J.-P., (1992), ***Le risque technologique***, Presses universitaires de France.

[17] LIANG, Y., & Ruiming, F., (2013), "An Online Wind Turbine Condition Assessment Method Based on SCADA and Support Vector Regression", ***Automation of Electric Power Systems***, 14: 003.

[18] Yang, W., Court, R., & Jiang, J., (2013), "Wind turbine condition monitoring by the approach of SCADA data analysis", ***Renewable Energy***, 53: 365-376.

[19] Kaidis, C., Uzunoglu, B., & Amoiralis, F., (2015), "Wind turbine reliability estimation for different assemblies and failure severity categories", ***Renewable Power Generation, IET***, 9 (8): 892-899.

[20] Zhang, Y., Chung, C. Y., Wen, F., & Zhong, J., (2016), "An analytic model for fault diagnosis in power systems utilizing redundancy and temporal information of alarm messages", ***IEEE Transactions on Power Systems***, 31 (6): 4877-4886.

[21] Daly, R., Shen, Q., & Aitken, S., (2011), "Learning Bayesian networks: approaches and issues", ***The Knowledge Engineering Review***, 26 (02): 99-157.

[22] Friedman, N., Murphy, K., & Russell, S., (1998), "Learning the structure of dynamic probabilistic networks", ***Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence***, Madison, Wisconsin, Morgan Kaufmann Publishers Inc., 139-147.

[23] Ghahramani, Z., (1998), "Learning dynamic Bayesian networks", In C. L. Giles, & M. Gori (Eds.), ***Adaptive Processing of Sequences and Data Structures***, Vol. 1387, Springer Berlin Heidelberg, 168-197.

[24] Murphy, K. P., (2002), ***Dynamic bayesian networks: representation, inference and learning***, Unpublished Ph.D., University of California, Berkeley, Berkeley, CA.

[25] Du, K.-L., & Swamy, M. N. S., (2014), "Probabilistic and Bayesian Networks", ***Neural Networks and Statistical Learning***, Springer London, 563-619.

[26] Marshall, A. H., Hill, L. A., & Kee, F., (2010), "Continuous Dynamic Bayesian networks for predicting survival of ischaemic heart disease patients", Paper presented at the ***Computer-Based Medical Systems (CBMS), 2010 IEEE 23rd International Symposium on***, 12-15 Oct. 2010: 178-183.

[27] Peelen, L., de Keizer, N. F., Jonge, E. d., Bosman, R.-J., Abu-Hanna, A., & Peek, N., (2010), "Using hierarchical dynamic Bayesian networks to investigate dynamics of organ failure in patients in the Intensive Care Unit", ***Journal of Biomedical Informatics***, 43 (2): 273-286.

[28] Sandri, M., Berchialla, P., Baldi, I., Gregori, D., & De Blasi, R. A., (2014), "Dynamic Bayesian Networks to predict sequences of organ failures in patients admitted to ICU", ***Journal of biomedical informatics***, 48: 106-113.

[29] Kleinberg, S., & Hripcsak, G., (2011), "A review of causal inference for biomedical informatics", ***Journal of Biomedical Informatics***, 44 (6): 1102-1112.

[30] Ferreiro, S., Arnaiz, A., Sierra, B., & Irigoien, I., (2012), "Application of Bayesian networks in prognostics for a new Integrated Vehicle Health Management concept", ***Expert Systems with Applications***, 39 (7): 6402-6418.

[31] Portinale, L., Raiteri, D. C., & Montani, S., (2010), "Supporting reliability engineers in exploiting the power of Dynamic Bayesian Networks", ***International Journal of Approximate Reasoning***, 51 (2): 179-195.

[32] Khakzad, N., Khan, F., & Amyotte, P., (2013), "Risk-based design of process systems using discrete-time Bayesian networks", ***Reliability Engineering & System Safety***, 109: 5-17.

[33] Aggarwal, J. K., & Ryoo, M. S., (2011), "Human activity analysis: A review", ***ACM Comput. Surv.***, 43 (3): 1-43.

[34] Wang, X., & Ji, Q., (2014), "Context augmented Dynamic Bayesian Networks for event recognition", ***Pattern Recognition Letters***, 43: 62-70.

[35] Lee, S., & Lee, K. C., (2012), "Context-prediction performance by a dynamic Bayesian network: Emphasis on location prediction in ubiquitous decision support environment", ***Expert Systems with Applications***, 39 (5): 4908-4914.

[36] Luo, Y., Wu, T.-D., & Hwang, J.-N., (2003), "Object-based analysis and interpretation of human motion in sports video sequences by dynamic bayesian networks", ***Computer Vision and Image Understanding***, 92 (2–3): 196-216.

[37] Suk, H.-I., Sin, B.-K., & Lee, S.-W., (2010), "Hand gesture recognition based on dynamic Bayesian network framework", ***Pattern Recognition***, 43 (9): 3059-3072.

[38] Yao, R., Zhang, Y., Zhou, Y., & Xia, S., (2014), "Multiple small objects tracking based on dynamic Bayesian networks with spatial prior", ***Optik - International Journal for Light and Electron Optics***, 125 (10): 2243-2247.

[39] Donat, R., Leray, P., Bouillaut, L., & Aknin, P., (2010), "A dynamic Bayesian network to represent discrete duration models", ***Neurocomputing***, 73 (4–6): 570-577.

[40] Eban, E., Nelken, I., Rothschild, G., Mizrahi, A., & Elidan, G., (2013), "Dynamic Copula Networks for Modeling Real-valued Time Series", Paper presented at the ***Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS 2013)***.

[41] Enright, C. G., Madden, M. G., & Madden, N., (2013), "Bayesian networks for mathematical models: Techniques for automatic construction and efficient inference", ***International Journal of Approximate Reasoning***, 54 (2): 323-342.

[42] Nicholson, A. E., & Flores, M. J., (2011), "Combining state and transition models with dynamic Bayesian networks", ***Ecological Modelling***, 222 (3): 555-566.

[43] Dondelinger, F., Lèbre, S., & Husmeier, D., (2013), "Non-homogeneous dynamic Bayesian networks with Bayesian regularization for inferring gene regulatory networks with gradually time-varying structure", ***Mach. Learn.***, 90 (2): 191-230.

[44] Doshi, F., Wingate, D., Tenenbaum, J., & Roy, N., (2011), "Infinite dynamic bayesian networks", Paper presented at the ***Proceedings of the 28th International Conference on Machine Learning (ICML-11)***: 913-920.

[45] Fenz, S., (2012), "An ontology-based approach for constructing Bayesian networks", ***Data & Knowledge Engineering***, 73 (0): 73-88.

[46] Gao, X.-G., Mei, J.-F., Chen, H.-Y., & Chen, D.-Q., (2013), "Approximate inference for dynamic Bayesian networks: sliding window approach", ***Applied Intelligence***: 1-17.

[47] Grzegorczyk, M., & Husmeier, D., (2011), "Non-homogeneous dynamic Bayesian networks for continuous data", ***Machine Learning***, 83 (3): 355-419.

[48] PIECHOWIAK, S., (2003), ***Intelligence artificielle et diagnostic***, Ed. Techniques Ingénieur.

[49] McCann, M., Li, Y., Maguire, L. P., & Johnston, A., (2010), "Causality Challenge: Benchmarking relevant signal components for effective monitoring and process control", Paper presented at the ***NIPS Causality: Objectives and Assessment***: 277-288.

[50] Kohda, T., & Cui, W., (2007), "Risk-based reconfiguration of safety monitoring system using dynamic Bayesian network", ***Reliability Engineering & System Safety***, 92 (12): 1716-1723.

[51] Delcroix, V., Piechowiak, S., & Maalej, M.-A., (2003), "Calcul des diagnostics les plus probables a posteriori", ***Revue d'intelligence artificielle***, 17 (4): 627-654.

[52] Parisot, J., Boussemart, M., & Simon, C., (2011), "Aide à la décision en maintenance par simulation d'un réseau bayésien dynamique", Paper presented at the ***9ème Congrès International Pluridisciplinaire Qualité et Sûreté de Fonctionnement, Qualita'2011***: CDROM.

[53] Weber, P., Medina-Oliva, G., Simon, C., & Iung, B., (2012), "Overview on Bayesian networks applications for dependability, risk analysis and maintenance areas", ***Engineering Applications of Artificial Intelligence***, 25 (4): 671-682.

[54] Lerner, U., Parr, R., Koller, D., & Biswas, G., (2000), "Bayesian fault detection and diagnosis in dynamic systems", Paper presented at the ***AAAI/IAAI***: 531-537.

[55] Isermann, R., (2005), "Model-based fault-detection and diagnosis–status and applications", ***Annual Reviews in control***, 29 (1): 71-85.

[56] Sheppard, J. W., & Kaufman, M., (2005), "A Bayesian approach to diagnosis and prognosis using built-in test", ***Instrumentation and Measurement, IEEE Transactions on***, 54 (3): 1003-1018.

[57] Tracht, K., Goch, G., Schuh, P., Sorg, M., & Westerkamp, J. F., (2013), "Failure probability prediction based on condition monitoring data of wind energy systems for spare parts supply", ***CIRP Annals-Manufacturing Technology***, 62 (1): 127-130.

[58] Díaz-Uriarte, R., & De Andres, S. A., (2006), "Gene selection and classification of microarray data using random forest", ***BMC bioinformatics***, 7 (1): 3.

[59] Rau, A., Rau, M. A., & GeneNet, S., (2012), "Package 'ebdbNet'", https://cran.r-project.org/web/packages/ebdbNet.

[60] Tchangani, A., & Noyes, D., (2006), "Modeling dynamic reliability using dynamic Bayesian networks", ***Journal Européen des systèmes automatisés***, 40 (8): 911-935.

[61] Cornuejols, A., & Miclet, L., (2010), ***Apprentissage artificiel : concepts et algorithmes*** (2e édition. ed.), Paris, Eyrolles.

[62] Sintchenko, V., Gilbert, G., Coiera, E., & Dwyer, D., (2002), "Treat or test first? Decision analysis of empirical antiviral treatment of influenza virus infection versus treatment based on rapid test results", ***Journal of clinical virology***, 25 (1): 15-21.

[63] Hastie, T., Tibshirani, R., & Friedman, J. H., (2009), ***The elements of statistical learning : data mining, inference, and prediction*** (2nd ed. ed.), New York, Springer.

[64] Walters-Williams, J., & Li, Y., (2010), "Comparative study of distance functions for nearest neighbors", ***Advanced Techniques in Computing Sciences and Software Engineering***: 79-84.

[65] Hu, L.-Y., Huang, M.-W., Ke, S.-W., & Tsai, C.-F., (2016), "The distance function effect on k-nearest neighbor classification for medical datasets", ***SpringerPlus***, 5 (1): 1304.

[66] Aggarwal, C. C., Hinneburg, A., & Keim, D. A., (2001), "On the surprising behavior of distance metrics in high dimensional space", Paper presented at the ***International Conference on Database Theory***, Springer: 420-434.

[67] Qian, G., Sural, S., Gu, Y., & Pramanik, S., (2004), "Similarity between Euclidean and cosine angle distance for nearest neighbor queries", Paper presented at the ***Proceedings of the 2004 ACM symposium on Applied computing***, ACM: 1232-1237.

[68] Vapnik, V., (1995), ***The Nature of Statistical Learning Theory***, Berlin, Springer.

[69] Cortes, C., & Vapnik, V., (1995), "Support-vector networks", ***Machine learning***, 20 (3): 273-297.

[70] Tang, H., Tan, K. C., & Zhang, Y., (2007), ***Neural networks : computational models and applications***, New York, Springer.