

UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

**ESTIMATION DE LA DURÉE DE VIE RÉSIDUELLE D'UN ACTIF PAR L'IA -
APPLICATION SUR DES DONNÉES NON INTÈGRES DANS UNE CENTRALE
HYDROÉLECTRIQUE**

**MÉMOIRE PRÉSENTÉ COMME EXIGENCE PARTIELLE DE LA
MAÎTRISE EN GÉNIE INDUSTRIEL**

PAR

DURAND CHANCIER KAMGAING KEMI

AVRIL 2022

Université du Québec à Trois-Rivières

Service de la bibliothèque

Avertissement

L'auteur de ce mémoire ou de cette thèse a autorisé l'Université du Québec à Trois-Rivières à diffuser, à des fins non lucratives, une copie de son mémoire ou de sa thèse.

Cette diffusion n'entraîne pas une renonciation de la part de l'auteur à ses droits de propriété intellectuelle, incluant le droit d'auteur, sur ce mémoire ou cette thèse. Notamment, la reproduction ou la publication de la totalité ou d'une partie importante de ce mémoire ou de cette thèse requiert son autorisation.

UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

Ce mémoire a été dirigé par :

Georges Abdul-Nour, directeur de recherche, Ph.D.	UQTR
---	------

François Gauthier, codirecteur de recherche, Ph.D.	UQTR
--	------

Jury d'évaluation du mémoire :

François Gauthier	UQTR
-------------------	------

Dragan Komljenovic	IREQ/HQ
--------------------	---------

Toualith Jean-Marc MEANGO	IREQ/HQ
---------------------------	---------

REMERCIEMENTS

Je tiens à remercier toutes les personnes ayant contribué au succès de ce projet de recherche :

- Monsieur Georges Abdul-Nour, mon directeur de recherche, qui m’a guidé tout au long de cette aventure, m’a prodigué de précieux conseils, m’a fait confiance et surtout, m’a ouvert la porte à de nombreuses opportunités.
- Monsieur François Gauthier, mon codirecteur de recherche, qui m’a invité à rejoindre la chaire de recherche Hydro-Québec en gestion des actifs, et m’a encadré de façon irréprochable tout au long de cette maîtrise.
- Mes superviseurs d’Hydro-Québec Production, notamment Monsieur Patrice Lapointe, Monsieur Russel Larouche et Madame Geneviève Pleau pour leur excellente collaboration et leur disponibilité.
- Monsieur Marc-Antoine Roy, étudiant de maîtrise en génie industriel qui m’a beaucoup aidé pour mon intégration et m’a passé dignement le flambeau sur ce projet de recherche.

Je tiens également à remercier toutes les personnes, qui de près comme de loin, ont su me donner la motivation, le courage et la force d’avancer malgré les circonstances. Un merci particulier à mes parents et à ma famille pour leur soutien et leur optimisme. Un merci à tous mes enseignants durant cette maîtrise pour l’excellente formation. Un merci à mon ami Michel Kirumba Kimuha avec qui nous nous motivions et serrions les coudes tout au long de notre parcours.

RÉSUMÉ

Compte tenu du durcissement du cadre réglementaire et d'une compétitivité sévère du marché industriel, les grandes entreprises se tournent davantage vers la maintenance prédictive. Ce type de maintenance permet d'anticiper de nombreuses défaillances à l'avance ou de prédire la fin de vie des actifs physiques, permettant ainsi de sauver des coûts importants et d'assurer au mieux la sécurité des personnes. Cette recherche s'inscrit dans ce cadre pour le développement d'une approche méthodologique et l'estimation des durées de vie résiduelles des équipements d'Hydro-Québec Production. L'objectif principal consiste au développement d'un algorithme de traitement des données de défaillance en faisant usage de l'intelligence artificielle en vue d'estimer la durée de vie résiduelle d'un actif physique. Une revue de la littérature a été effectuée pour se renseigner sur les avancées significatives dans le domaine. La revue de littérature a permis de faire ressortir quatre catégories de variables indépendantes qui agissent sur la variable dépendante qu'est la durée de vie résiduelle. Il s'agit de la décision économique, de la prédiction, de la disponibilité et de l'intégrité des données. Un accent particulier a été porté aux catégories prédiction et intégrité des données. La prédiction englobe toutes les solutions de prédiction de durée de vie résiduelle parmi lesquelles les plus répandues s'avèrent les méthodes basées sur les données et les méthodes basées sur un modèle. La catégorie intégrité des données englobe quelques solutions proposées dans la littérature pour pallier au problème d'intégrité des données qui constitue un frein au domaine du pronostic. Les principales solutions se réfèrent à l'utilisation des méthodes bayésiennes, de l'intelligence artificielle et d'autres méthodes telles que l'utilisation de sources externes. La méthodologie de cette recherche consiste au développement d'un algorithme de pairage des bases de données. Ensuite, le traitement du langage naturel, une branche de l'intelligence artificielle, est exploité pour analyser les rapports de maintenance et ainsi montrer les apports de l'intelligence artificielle dans l'amélioration de l'intégrité des données.

Puis, la combinaison des modèles de traitement du langage naturel et de l'algorithme de pairage constitue l'algorithme final permettant de traiter des données historiques brutes non intégrées et d'en extraire des caractéristiques précises et pertinentes. La méthode statistique a été choisie pour sa simplicité, sa pertinence lors de l'absence des données de maintenance conditionnelle, et surtout de sa réputation reconnue par de nombreux experts du domaine de l'ingénierie de la fiabilité pour l'estimation de la durée de vie résiduelle d'un équipement. Le contexte concerne l'estimation de la fin de vie par vieillissement dû à la dégradation de l'actif physique. Trois hypothèses ont été posées comme prémisses à cette étude. La première hypothèse concerne la considération seule des arrêts non planifiés pour cause de dégradation. La seconde suppose qu'un arrêt non planifié de durée importante équivaut à la maintenance à la suite d'une défaillance critique. La troisième hypothèse stipule que la maintenance du composant générique à la suite d'une défaillance critique est considérée comme un remplacement. Une équipe d'experts a relevé la pertinence du stator qui a été choisi comme composant générique pour tester les résultats de l'étude. L'algorithme a permis d'extraire les défaillances critiques par court-circuit du stator sur plusieurs centrales. Un test d'homogénéité des données de centrales a été effectué et une estimation globale de la durée de vie des stators a été évaluée à 25 ans grâce à une analyse Weibull. L'originalité de cette recherche s'observe surtout par la façon inédite de croiser deux bases de données sans clé directe de jointure et d'en extraire automatiquement des caractéristiques spécifiques épurées.

TABLE DES MATIÈRES

REMERCIEMENTS	iv
RÉSUMÉ	v
TABLE DES MATIÈRES	vii
LISTE DES TABLEAUX.....	x
LISTE DES FIGURES.....	xii
LISTE DES ABRÉVIATIONS.....	xiii
INTRODUCTION	1
Chapitre 1 – MISE EN CONTEXTE	3
1.1. Problématique	3
1.1.1 Problématique générale.....	3
1.1.2 Problématique industrielle	4
1.2. Questions de recherche	5
1.3. Objectifs de la recherche	7
1.4. Concepts et mots-clefs	7
1.4.1 Concepts.....	7
1.4.2 Mots-clefs.....	9
1.5. Conclusion	9
Chapitre 2 – REVUE DE LA LITTÉRATURE	10
2.1 Pronostic et gestion de l'état de santé	10
2.1.1 Concepts et définition	10
2.1.2 Cycle de vie d'un actif physique.....	13
2.1.3 Types de maintenance	16
2.2 Méthodologies de prédiction de la durée de vie résiduelle	19
2.2.1 Méthodes basées sur un modèle.....	20
2.2.2 Méthodes basées sur les données	22
2.2.3 Autres méthodes.....	26
2.3 Le pronostic dans l'industrie électrique	30
2.4 Intégrité des données	31

2.4.1	Avis d'experts et méthodes bayésiennes.....	31
2.4.2	Intelligence artificielle	32
2.4.3	Autres méthodes.....	35
2.5	La place de l'IA dans le domaine du PGES	37
2.6	Traitement du langage naturel.....	40
2.6.1	Origines, évolution et applications.....	40
2.6.2	Classification de texte	41
2.7	Matrice de la revue de littérature	44
2.8	Synthèse de la revue de littérature.....	55
2.9	Conclusion	57
Chapitre 3 – MÉTHODOLOGIE DE LA RECHERCHE.....		58
3.1	Hydro-Québec Production.....	58
3.2	Bases de données.....	59
3.3	Intégrité des données chez Hydro-Québec Production.....	62
3.4	Démarche	64
3.4.1	Pairage des bases de données.....	64
3.4.2	Classification automatique de texte	66
3.4.3	Combinaison du pairage et de la classification automatique	67
3.4.4	Estimation des DVR.....	69
3.5	Conclusion.....	70
Chapitre 4 - RÉSULTATS.....		71
4.1	Algorithme de traitement des données	71
4.1.1	Pairage des bases de données.....	71
4.1.2	Classification de texte	79
4.1.3	Combinaison du pairage et de la classification de texte:	84
4.2	Estimation de la durée de vie résiduelle du stator.....	93
4.2.1	Choix d'une méthode de prédiction des durées de vie.....	93
4.2.2	Estimation des durées de vie pour chaque centrale.....	93
4.2.3	Test d'homogénéité des centrales	98
4.3	Extension de l'algorithme à d'autres systèmes et sous-systèmes	101

4.4 Conclusion	102
CONCLUSION	103
RÉFÉRENCES	105
ANNEXE I - Graphiques comparatifs des données de SIPROD et CEGTA	111
ANNEXE II - Code Python de l'algorithme de traitement des données	113
ANNEXE III - Code python de la classification automatique des sous-systèmes de l'alternateur	129
ANNEXE IV - Code Python de la classification automatique des modes de défaillance du stator	135

LISTE DES TABLEAUX

TABLEAU 2.1: QUELQUES MÉTHODOLOGIES D'ÉVALUATION DES COÛTS DU CYCLE DE VIE adapté de: de Almeida Pais et al. (2021)	15
TABLEAU 2.2: LES TYPES DE MAINTENANCE - Prajapati et al. (2012).....	18
TABLEAU 2.3: RÉSUMÉ DES DIVERSES APPROCHES D'ESTIMATION DE DVR PRÉSENTÉES PAR DIFFÉRENTS AUTEURS.....	20
TABLEAU 2.4: AVANTAGES ET INCONVÉNIENTS DES APPROCHES BASÉES SUR UN MODÈLE ET BASÉES SUR LES DONNÉES - Y. Wang et al. (2020)	25
TABLEAU 2.5: VUE D'ENSEMBLE DES DIFFÉRENTES MÉTHODOLOGIES DE PRÉDICTION DE DVR.....	29
TABLEAU 2.6: DIVERSES SOLUTIONS AU PROBLÈME D'INTÉGRITÉ DES DONNÉES.....	36
TABLEAU 2.7: MATRICE DES APPROCHES DE PRÉDICTION DE DVR	45
TABLEAU 2.8: MATRICE DES APPROCHES DE PRÉDICTION DE DVR (SUITE 1) .	46
TABLEAU 2.9: MATRICE DES APPROCHES DE PRÉDICTION DE DVR (SUITE 2) .	47
TABLEAU 2.10: MATRICE DES APPROCHES DE PRÉDICTION DE DVR (SUITE 3)	48
TABLEAU 2.11: MATRICE DES APPROCHES DE PRÉDICTION DE DVR (SUITE 4)	49
TABLEAU 2.12: MATRICE DES APPROCHES DE PRÉDICTION DE DVR (SUITE 5)	50
TABLEAU 2.13: MATRICE DES APPROCHES DE PRÉDICTION DE DVR (SUITE 6)	51
TABLEAU 2.14: MATRICE DE L'INTÉGRITÉ DES DONNÉES	52
TABLEAU 2.15: MATRICE DE L'INTÉGRITÉ DES DONNÉES (SUITE 1)	53
TABLEAU 2.16: MATRICE DE L'INTÉGRITÉ DES DONNÉES (SUITE 2)	54
TABLEAU 3.1: INFORMATIONS PERTINENTES CONTENUES DANS LES BASES DE DONNÉES	60
TABLEAU 3.2: DESCRIPTION DES INFORMATIONS PERTINENTES CONTENUES DANS LES BASES DE DONNÉES.....	61
TABLEAU 3.3: EXTRAIT DE SIPROD PRÉSENTANT UN PROBLÈME D'INTÉGRITÉ DES DONNÉES	63
TABLEAU 4.1: RÉSULTATS DU PAIRAGE SIPROD-CEGTA DE LG2-LG2A.....	78

TABLEAU 4.2: PARAMÈTRES OPTIMISÉS DU BOW POUR CHAQUE ALGORITHME.....	82
TABLEAU 4.3: RÉSULTATS DE LA CLASSIFICATION DES SOUS-SYSTÈMES	82
TABLEAU 4.4: RÉSULTATS DE LA CLASSIFICATION DES MODES DE DÉFAILLANCE.....	83
TABLEAU 4.5: TABLE RÉSULTANTE DE LA COMBINAISON DU PAIRAGE ET DE LA CLASSIFICATION DE TEXTE	84
TABLEAU 4.6: RÉSULTATS DU PAIRAGE SUR LES AUTRES CENTRALES	86
TABLEAU 4.7: RÉSULTATS DU PAIRAGE SUR LES AUTRES CENTRALES (SUITE 1).....	87
TABLEAU 4.8: RÉSULTATS DU PAIRAGE SUR LES AUTRES CENTRALES (SUITE 2).....	88
TABLEAU 4.9: CAPACITÉ DE GÉNÉRALISATION DU MODÈLE DE CLASSIFICATION SUR D'AUTRES CENTRALES	89
TABLEAU 4.10: RÉSULTATS DE L'EXTRACTION DES DONNÉES DE CC STATOR	91
TABLEAU 4.11: RÉSULTATS DE L'ALGORITHME PAR RAPPORT AU FICHIER DE CC STATOR DE L'IREQ.....	92
TABLEAU 4.12: NOMBRE DE DÉFAILLANCES PAR CC STATOR.....	94
TABLEAU 4.13: COEFFICIENTS R^2 DES DISTRIBUTIONS ET TEST DE KS	95
TABLEAU 4.14: CRITÈRE D'ÉLIMINATION DE LA LOI NORMALE	96
TABLEAU 4.15: PARAMÈTRES WEIBULL LG2-LG2A & BEAUHARNOIS	96
TABLEAU 4.16: DURÉE DE VIE ESTIMÉE DU STATOR LG2-LG2A & BEAUHARNOIS.....	96
TABLEAU 4.17: PARAMÈTRES WEIBULL DES DONNÉES GLOBALES	98
TABLEAU 4.18: DURÉE DE VIE ESTIMÉE GLOBALE DU STATOR	98
TABLEAU 4.19: DVR DES STATORS DE LA CENTRALE BEAUHARNOIS.....	100

LISTE DES FIGURES

Figure 1.1: Récapitulatif des questions de recherche.....	7
Figure 2.1: Vue d'ensemble du PGES de Li et al. (2020)	12
Figure 2.2: Cycle de vie d'un actif physique adapté de: de Almeida Pais et al. (2021) ...	14
Figure 2.3: Les philosophies de maintenance de Ayu et Yunusa-Kaltungo (2020).....	17
Figure 2.4: Complexité et applicabilité des approches basées sur un modèle et basées sur les données	26
Figure 2.5: Les différents types d'apprentissages	38
Figure 2.6: Diagramme de flux d'un projet de classification de texte adapté de Kang et al. (2020)	43
Figure 3.1 : Vue en coupe d'un GTA.....	58
Figure 3.2: Méthodologie de la recherche	65
Figure 4.1 : Représentativité des données par système.....	73
Figure 4.2: Principe d'association temporelle de SIPROD et CEGTA	74
Figure 4.3: Diagramme de flux de l'algorithme de pairage	76
Figure 4.4: Outil de visualisation du pairage SIPROD-CEGTA	77
Figure 4.5: Répartition des classes de sous-systèmes alternateur	80
Figure 4.6: Nuage de mots illustrant la fréquence des mots de la classe stator	81
Figure 4.7: Logigramme de filtrage des données stator	85
Figure 4.8: Logigramme d'extraction des données de court-circuit stator	90
Figure 4.9: Courbe de vieillissement stator – centrale LG2-LG2A	97
Figure 4.10: Courbe de vieillissement stator – centrale Beauharnois	97
Figure 4.11: Courbe de vieillissement stator – Global.....	99
Figure A1.1: Nombre de données Alternateur - SIPROD – Beauharnois.....	111
Figure A1.2: Nombre de données Alternateur - CEGTA – Beauharnois.....	111

LISTE DES ABRÉVIATIONS

BOW : *Bag Of Words*
BT : Bon de travail
CC : Court-circuit
CNN : *Convolutionnal Neural Network*
DL : *Deep Learning*
DTM : *Document Term Matrix*
DVR : Durée de vie résiduelle
GAN : *Generative Adversarial Network*
GTA : Groupe turbine-alternateur
HQ : Hydro-Québec
HQP : Hydro-Québec Production
IA : Intelligence artificielle
IREQ : Institut de Recherche sur l'Électricité du Québec
KS : Kolmogorov-Smirnov
LTSM : *Long Short Term Memory*
MC : Maintenance corrective
MCON : Maintenance conditionnelle
ML : *Machine Learning*
MLE : *Maximum Likelihood Estimation*
MP: Maintenance préventive
MPd : Maintenance prédictive
NA : Non attribué
PGES : Pronostic et gestion de l'état de santé
RNN: *Recurrent Neural Network*
SGDC: *Stochastic Gradient Descent Classifier*
SMOTE : *Synthetic Minority Over-sampling Technique*
SVM : *Support Vector Machine*

TF : *Term Frequency*

TF-IDF : *Term Frequency Inverse Document Frequency*

INTRODUCTION

Toute organisation possédant une grande quantité d'actifs fait éventuellement face à une décision cruciale : les réparer ou les remplacer, et quand? Cette décision peut entraîner des conséquences significatives. En effet, remplacer trop tôt peut signifier un gaspillage de ressources, et remplacer trop tard peut entraîner des conséquences onéreuses. Compte tenu du dynamisme actuel du marché, de telles situations sont de plus en plus inacceptables. La durée de vie résiduelle (DVR) d'un système est un facteur important pour la prise de décisions des actifs majeurs. Si une compagnie peut prédire correctement la DVR d'un système réparable, alors estimer les coûts d'entretien et de réparation du système jusqu'à ce point devient possible. Il en est de même pour l'évaluation des bénéfices potentiels du remplacement du système en entier à un moment donné (Wong, Jardine, & Banjevic, 2011).

À cet effet, les auteurs exploitent de nombreuses méthodes à travers la littérature pour estimer la DVR des équipements réparables, dont les plus répandues sont les méthodes basées sur un modèle et les méthodes basées sur les données.

Cependant, d'après Wong et al. (2011), dans le cas des systèmes réparables, trois problèmes principaux compliquent l'estimation de la durée de vie des composants ou des systèmes : l'intégrité des données, la dépendance/indépendance des composants et les effets des réparations. À ce propos, ce mémoire porte sur le problème d'intégrité des données et sur quelques solutions proposées par de nombreux auteurs pour y remédier : l'intégration des avis d'experts à travers des approches bayésiennes, l'intelligence artificielle (IA) et d'autres propositions telles que le *crowdsourcing*, et l'usage de bases de données externes.

Dans le cadre de l'amélioration de la gestion de ses actifs, Hydro-Québec (HQ) a lancé de nombreux projets dans chacun de ses groupes. Cette recherche s'applique au groupe Hydro-Québec production (HQP).

L'entreprise soutient une vision de maintenance prédictive axée sur l'estimation des fins de vie des actifs associée à des prises de décisions économiques optimales en termes de remplacement. Dès lors, HQP s'active dans ce sens pour déterminer les DVR des équipements majeurs des groupes turbines alternateur (GTA) des centrales hydroélectriques. Toutefois, HQP se heurte à un problème d'intégrité des données à résoudre en première intention. Ce projet de recherche s'inscrit donc dans les volets de gestion du cycle de vie et de prise de décision du cadre global de la gestion des actifs.

La suite de ce mémoire est organisée comme suit : le chapitre 1 présente le contexte de cette étude en y ressortant la problématique, les questions de recherche, le but visé, les objectifs secondaires, les concepts et les mots-clefs. Les chapitres 2 et 3 explicitent respectivement la revue de la littérature et la méthodologie de recherche. Ensuite, le chapitre 4 discute des résultats obtenus et enfin un récapitulatif des points abordés suivi de la mise en évidence de l'originalité et des limites de la recherche concluent le document.

CHAPITRE 1 – MISE EN CONTEXTE

Ce chapitre est une mise en contexte général de la recherche. De la problématique découle des questions de recherche et des objectifs. Cette démarche permet de ressortir les concepts et mots-clefs.

1.1. Problématique

La problématique est divisée en deux sous-sections. La première présente la problématique générale de la compagnie HQ, tandis que la seconde présente la problématique industrielle du groupe HQP.

1.1.1 Problématique générale

Au sens strictement économique, une entreprise a pour but de générer de la richesse. Ainsi, les grandes entreprises, quel que soit le secteur d'activité, doivent assurer leur pérennité en veillant à toujours satisfaire la demande, peu importe les fluctuations du marché. En effet, la concurrence est rude, car de nombreuses compagnies proposent des services similaires et il s'avère nécessaire de se démarquer.

C'est à cet égard que le leader du secteur de la production, transport et distribution d'énergie électrique HQ fait face à une pléthore de perturbations. La mondialisation occasionnant un besoin de disponibilité plus élevée, les changements climatiques à l'origine de verglas, les inondations et la surabondance de végétation, le vieillissement des actifs entraînant des arrêts de ligne, des pannes et par conséquent une diminution du service à la clientèle. La pénurie de main-d'œuvre se révèle également comme un facteur nuisible que les chercheurs essaient de pallier au moyen de nouvelles technologies.

Il est important de noter aussi les restrictions budgétaires courantes, le durcissement du cadre réglementaire associé aux considérations socio-environnementales et de sécurité. Dès lors, dans la quête d'agilité et de résilience, HQ se doit de relever tous ces défis (Hydro-Québec, 2022).

Les données sont le fondement, la base de toute quête d'amélioration, et donc constituent un pilier important dans la relève des défis cités précédemment. Cependant, un sérieux problème d'intégrité des données mine la réalité des industries qu'il s'avère nécessaire de solutionner en première intention.

1.1.2 Problématique industrielle

HQP est la division de HQ chargée de la production d'électricité au moyen de nombreuses centrales hydro-électriques constituées de systèmes GTA. Parmi les nombreux défis à relever notés dans la section 1.2.1, la nécessité d'une disponibilité élevée est incontournable. C'est à cet effet que les GTA se doivent d'être disponibles en tout temps. Il se pose donc la question à savoir : comment anticiper au mieux les pannes ou la fin de vie des actifs majeurs constituant les systèmes GTA.

Pour anticiper, il faudrait avant tout pouvoir prédire. Dès lors, prédire les durées de vie. Étant donné que ces actifs sont déjà en exploitation, l'expression DVR est plus appropriée. Cependant, la prédiction de la DVR nécessite l'usage de données, qui peuvent être de deux principales sortes : les données historiques de défaillances permettant d'analyser le comportement passé des équipements et s'y fier pour prédire le futur et des données de maintenance conditionnelle au travers de capteurs de surveillance en temps réel. Seules les données historiques des GTA ont été rendues disponibles par HQP pour cette étude.

Toutefois, il se trouve qu'il y a de nombreuses bases de données des événements historiques comportant des informations parfois incomplètes, imprécises, mêlées ou erronées.

La question de savoir comment choisir les bonnes informations, les assembler et les compléter au mieux dans ce labyrinthe complexe se montre dès lors pertinente. En effet, seules des données épurées et fiables permettent de réaliser des analyses fiables, et par conséquent une prédiction de DVR non biaisée.

1.2. Questions de recherche

La problématique fait ressortir l'intérêt de prédire la DVR des actifs majeurs, tout en notifiant l'obstacle que représente l'intégrité des données. Il se révèle intéressant de se poser les questions principales suivantes :

- Comment déterminer la DVR des équipements réparables en prenant en compte les aspects économiques et de fiabilité des systèmes?
- Quelles sont les approches méthodologiques encadrant le problème d'intégrité des données?

Pour répondre à la première interrogation, la compréhension des fondements de la DVR apparaît comme incontournable. Certains auteurs comme Okoh, Roy, Mehnen et Redding (2014), Ahmadzadeh et Lundberg (2014) et Y. Wang, Zhao et Addepalli (2020) définissent la DVR et font des revues présentant une vue d'ensemble des différentes techniques d'estimation de DVR. Il est évident que les études de DVR vont de pair avec les concepts de fiabilité au travers duquel de nombreuses approches statistiques sont utilisées. C'est par exemple le cas de Navas, Sancho et Carpio (2017) qui passent en revue de nombreux modèles statistiques pour l'analyse de fiabilité d'un système réparables de chemins de fer, avant de finalement démontrer lequel colle le mieux à leur étude. Waghmode et Patil (2016) font également une analyse de fiabilité, qui en revanche démontre l'importance de la fiabilité dans l'optimisation des coûts du cycle de vie des actifs. La considération des risques économiques est donc incontournable au sujet d'études de DVR.

Wong et al. (2011) remarquent que lorsqu'une grande quantité d'actifs vieillissent, il est attendu que les coûts opérationnels, de maintenance et administratifs puissent augmenter. La décision de réparation ou de remplacement est influencée par ces coûts pour l'unité existante contre la dépense d'acquisition et d'installation d'une nouvelle unité. Une bonne comparaison des coûts doit être effectuée dans le but de déterminer la décision économique optimale. Khalifa, Khan et Thorp (2015) proposent quant à eux un modèle représentant un outil efficace pour la prise de décisions optimales concernant la stratégie de maintenance (réparation ou remplacement) et l'évaluation de DVR basée sur une comparaison du risque économique total.

En ce qui concerne la deuxième interrogation, Ye, Wang, Lu et Li (2020) soulignent que durant un processus de collection de données, l'incomplétude est l'un des plus sérieux problèmes de qualité de données. Albert et al. (2012) sont du même avis, car ils soutiennent qu'il est très difficile de recueillir les données pour fournir les informations requises. Tous ces auteurs proposent des méthodes, dont des méthodes bayésiennes jumelées à l'intégration des avis d'experts, ou au *crowdsourcing*. D'autres auteurs suggèrent l'utilisation de bases de données externes ou de l'IA.

Avec l'avènement des nouvelles technologies, les grandes entreprises se tournent davantage vers la quatrième révolution industrielle, dite industrie 4.0. L'IA est l'un des outils clés nécessaires à l'implantation de cette industrie du futur. Dès lors, cette technologie attire beaucoup d'intérêt et de convoitise au vu de ses nombreuses applications. En juxtaposant cet intérêt aux concepts discutés dans les deux questions principales, il s'avère pertinent de se poser la question spécifique suivante :

Quelle est la place de l'IA dans le domaine du pronostic et gestion de l'état de santé (PGES) des actifs physiques, en particulier pour la prédiction de la DVR?

La figure 1.1 présente un récapitulatif des questions principales et la question spécifique énoncée, ainsi que les notions énoncées au travers.

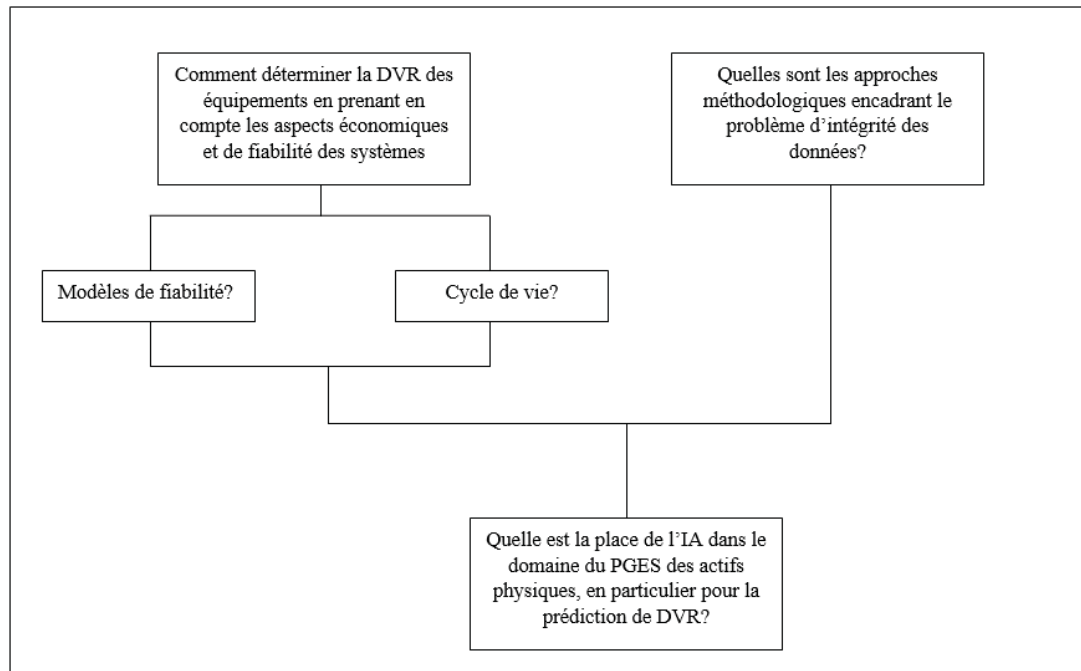


Figure 1.1: Récapitulatif des questions de recherche

1.3. Objectifs de la recherche

L'objectif principal de cette recherche consiste au développement d'un algorithme de traitement des données historiques en faisant usage de l'intelligence artificielle en vue d'estimer la DVR d'un actif physique.

Ce but visé est rattaché aux objectifs secondaires suivants :

- Explorer les méthodes de prédiction de DVR et choisir une pertinente pour l'application aux données de HQP.
- Exploiter les données épurées par l'algorithme pour l'estimation de la DVR d'un ou de certains équipements.

1.4. Concepts et mots-clefs

1.4.1 Concepts

La disponibilité est le principal concept qui englobe cette recherche. En effet, la DVR se révèle une mesure de performance inhérente à cette notion. Ainsi, pour mieux cerner l'étude, une définition claire de ce concept se montre significative.

La norme AFNOR X 60-500 « Terminologie relative à la fiabilité - maintenabilité – disponibilité » définit la disponibilité comme « l'aptitude d'une entité à être en état d'accomplir une fonction requise dans des conditions données [...] en supposant que la fourniture des moyens extérieurs nécessaires soit assurée ». Cette aptitude se trouve impactée par la fiabilité de l'entité, sa maintenabilité, et par conséquent sa logistique de maintenance. La compréhension de la disponibilité montre qu'elle englobe quelques notions importantes, notamment la fiabilité, la maintenabilité.

Selon la norme AFNOR X 60-500, la fiabilité représente l'aptitude d'une entité à accomplir une fonction requise, dans des conditions données, pendant un intervalle de temps donné, dans un environnement donné. Dans un sens plus large, la fiabilité est vaguement associée à l'absence de défaillances. Ainsi, pour des analyses techniques, la fiabilité est représentée par une distribution de probabilité. Un décideur d'entreprise donnerait au paramètre fiabilité une vision de performance, justifiée par le degré de tolérance et d'acceptation des risques dans le but de satisfaire des clients, propriétaires ou actionnaires.

La maintenabilité désigne l'aptitude d'une entité à être maintenue ou rétablie, dans des conditions données d'utilisation, sur un intervalle de temps donné, dans un état dans lequel elle peut accomplir une fonction requise, lorsque la maintenance est accomplie dans des conditions données, avec des procédures et des moyens prescrits. Cette définition a également été proposée par la norme française AFNOR citée précédemment. La notion de maintenabilité est donc directement reliée à la maintenance, dont le terme apparaît dans la définition.

La norme AFNOR NF-X 60 000 « maintenance industrielle – fonction maintenance » définit la maintenance comme « l'ensemble des actions permettant de maintenir ou de rétablir un bien dans un état spécifié ou en mesure d'assurer un service déterminé ». la

norme DIN EN 13306: 2015-09, 2.1 « terminologie de la maintenance » l'explique comme étant « la combinaison de toutes les actions techniques, administratives et de gestion effectuées tout au long du cycle de vie d'un actif, dans le but de le conserver ou de le restaurer dans un état dans lequel il peut remplir la fonction requise ». La maintenance regroupe ainsi les actions de dépannage et de réparation, de réglage, de révision, de contrôle et d'inspection des actifs. Dans ce cas de figure, un décideur d'entreprise donnerait au paramètre maintenance une vision de coûts et ressources, justifiées par le besoin de maintenir la performance ou de respect d'une obligation réglementaire dans le but d'éviter des pannes, des défauts ou une diminution de rendement.

1.4.2 Mots-clefs

Les mots-clefs relatifs à cette recherche sont : **DVR, PGES, systèmes complexes réparables, intégrité des données, IA , classification de texte.**

1.5. Conclusion

Ce premier chapitre a présenté le cadre de la recherche, en y ressortant les principales notions relatives de cette étude. Il s'agit de l'estimation de la DVR, du problème d'intégrité des données associé et de l'intérêt pour l'IA. Ces notions n'ont cependant été qu'introduites et nécessitent un approfondissement. À cet effet, le prochain chapitre relate la revue de littérature.

CHAPITRE 2 – REVUE DE LA LITTÉRATURE

La revue de littérature commence par relater brièvement la notion de pronostic et gestion de l'état de santé (PGES) des actifs physiques, suivie de la DVR et ses principales méthodologies de prédiction. En effet, la DVR est l'idée phare à retenir lorsqu'il s'agit de la partie pronostic du PGES. Dans le même élan, un regard sera porté sur le pronostic dans le secteur électrique. Ensuite, le thème d'intégrité des données est abordé en parcourant les diverses solutions proposées dans la littérature. La revue couvre également la place de l'IA dans le domaine du PGES. Une attention particulière est ensuite portée au traitement du langage naturel (TLN), une branche de l'IA.

2.1 Pronostic et gestion de l'état de santé

2.1.1 Concepts et définition

Le but du PGES est de proposer des méthodes et outils pour la mise en place des politiques de maintenance relatives à des actifs spécifiques compte tenu de leurs conditions d'exploitation et de dégradation, résultant à une haute disponibilité et des moindres coûts. Le PGES peut être considéré comme une approche holistique pour une gestion efficace et efficiente de la santé d'un actif physique. La gestion de la santé représente le processus de céduler propices et optimales des actions de maintenance basées sur les extrants de diagnostics et pronostics, les ressources disponibles et la demande opérationnelle. Le diagnostic consiste en la détection, isolation et identification de défauts lorsqu'ils surgissent. Le pronostic, cependant, consiste à prédire des défauts ou défaillances avant qu'ils ne surviennent, compte tenu des conditions présentes et/ou historiques. En résumé, le PGES se concentre sur l'évaluation et la minimisation des impacts de défaillances et le contrôle des coûts de maintenance (Fink et al., 2020). Les auteurs R. Li, Verhagen et Curran (2020) constatent également que le PGES a émergé comme l'une des solutions clés pour l'amélioration des systèmes, la sécurité, la

maintenabilité, la prise en charge et des coûts abordables pour les actifs industriels majeurs. La littérature reconnaît le diagnostic et le pronostic comme les deux principales disciplines du PGES dont le but est d'optimiser l'exploitation et la maintenance des actifs tout en améliorant la sécurité, la fiabilité et la rentabilité. La surveillance, le diagnostic et le pronostic sont dès lors les principaux acteurs sur lesquelles l'attention est focalisée dans les recherches relatives au PGES. Les applications stratégiques de PGES ont démontré une efficacité dans la réduction des temps d'arrêt des équipements et des coûts de maintenance.

Toutefois, il est important de cerner la mise en place d'un programme de PGES. À cet effet, Jardine, Lin et Banjevic (2006) et Teixeira, Lopes et Braga (2020) ressortent les principales étapes constituant le PGES : l'acquisition de données, le traitement des données et les prises de décisions relatives à la maintenance.

L'acquisition est le processus de collecte et de stockage des données utiles provenant des actifs à l'étude. Ces données peuvent être de deux principaux types, notamment les données historiques et les données de surveillance, encore connues comme données de maintenance conditionnelle (MCON). Les données historiques représentent les événements du passé (défaillances, révisions, installations, entretiens, etc.) et les données de MCON concernent des mesures inhérentes à l'état ou condition de santé de l'actif physique. Ces données de MCON sont très versatiles et peuvent être de toutes sortes : vibrations, analyse d'huile, acoustique, pression de température, humidité, climat, environnement, images, etc. Toutes ces données sont collectées par différents types de capteurs.

Le traitement de données consiste à extraire les informations importantes provenant des données. Cette étape requiert un nettoyage de données, avant d'enchaîner par une analyse des données épurées. Le nettoyage de celles-ci est nécessaire, car les données brutes sont souvent mal structurées et imparfaites. Ce bruit provient généralement des erreurs humaines ou des défauts de capteurs.

En ce qui concerne l'analyse de données, de nombreux modèles, algorithmes et outils sont disponibles dans la littérature, quel que soit le type de la donnée de MCON collectée.

La dernière étape du programme de PGES concerne les prises de décisions relatives à la maintenance. Une prise de décision émane de certains critères. Dans le cas du PGES, les prises de décisions, par rapport à la maintenance, sont principalement dues à deux principales techniques évoquées plus tôt : le diagnostic et le pronostic. Évidemment, on pourrait considérer le pronostic supérieur au diagnostic, car il permet de prévenir et ainsi éviter le défaut ou la défaillance avant qu'il ne survienne, et par conséquent limiter les dépenses. Cependant, le pronostic ne peut pas complètement remplacer le diagnostic, car en réalité il existe toujours des défauts ou défaillances non prédictibles. De plus, il est difficile de concevoir un outil de prédiction avec une précision et une sensibilité égale à 100 % (Jardine et al., 2006). Rui Li et al. (2020) présentent dans la figure 2.1 une vue d'ensemble du PGES.

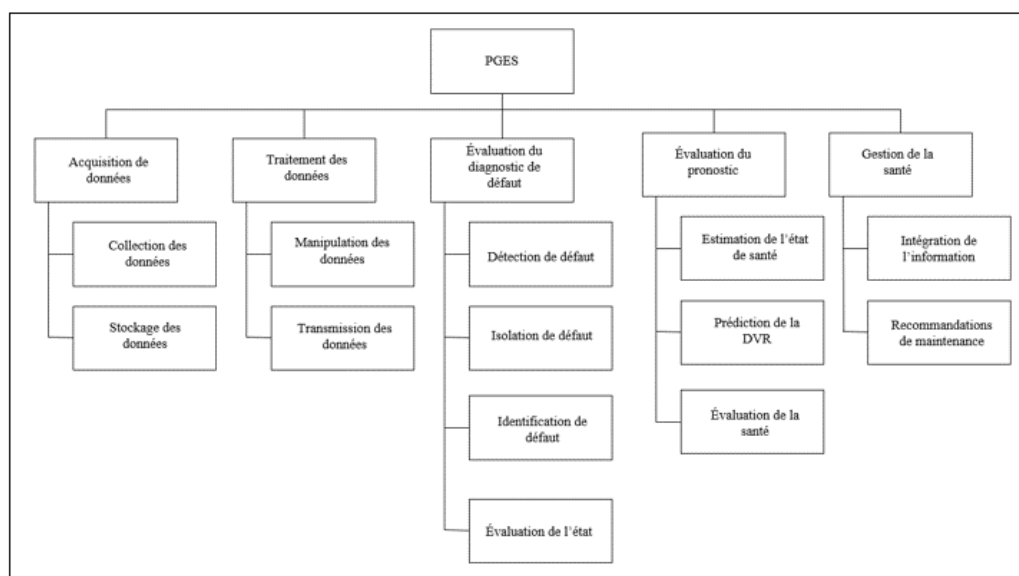


Figure 2.1: Vue d'ensemble du PGES de Li et al. (2020)

2.1.2 Cycle de vie d'un actif physique

Un grand nombre de changements se produisent au cours de la vie d'un bien physique, tant interne qu'externe. Il est donc important d'établir des stratégies pour soutenir la prise de décision, tout en anticipant au mieux certains changements tels que les changements législatifs, les impacts environnementaux et la demande de production (de Almeida Pais, Raposo, Farinha, Cardoso, & Marques, 2021). La figure 2.2 illustre les différents changements et paramètres à prendre en considération tout au long du cycle de vie d'un actif physique. La courbe baignoire repérée dans (Levin, Kalal, & Rodin, 2019) illustre l'aspect fiabilité du cycle de vie, soutenant assez bien le besoin de PGES quand l'équipement vieillit. La courbe baignoire montre que le cycle de vie d'un actif physique est constitué de trois principales périodes : la période de rodage (défaillances prématurées), la période de fonctionnement normal (défaillances aléatoires) et la période de fin de vie (vieillessement). Biard et al. (2021) réalisent une étude pour déterminer le cycle de vie et les coûts de remplacement futurs des équipements d'un réseau de distribution électrique. Les résultats de l'étude permettront aux auteurs de proposer des stratégies de remplacement permettant d'optimiser la performance, minimiser les risques ou optimiser les coûts selon les objectifs de la compagnie : l'usage jusqu'à la défaillance, le remplacement basé sur le degré de détérioration observé, le remplacement basé sur l'âge et le remplacement basé sur une inspection ciblée.

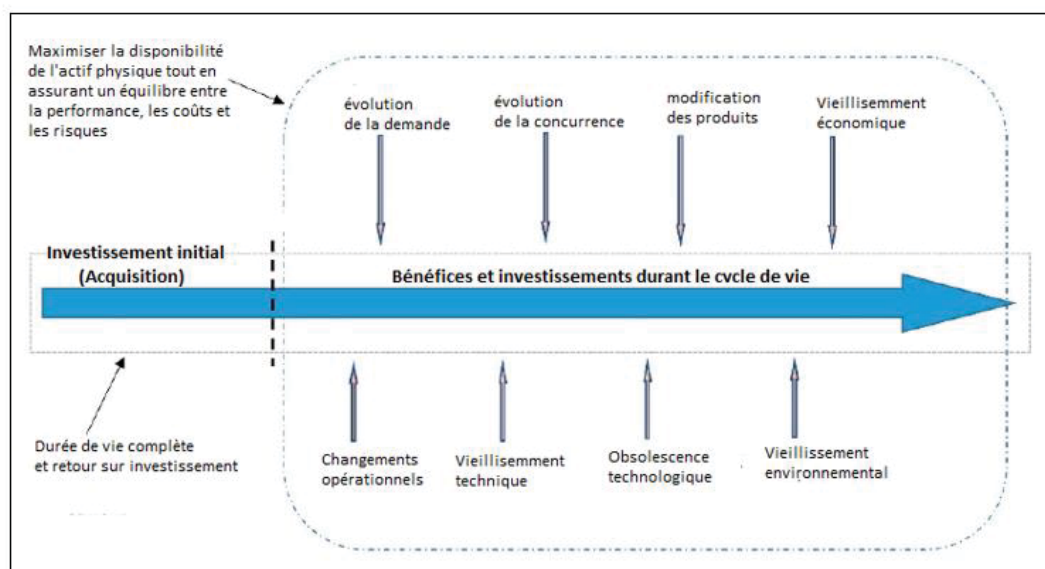


Figure 2.2: Cycle de vie d'un actif physique adapté de: de Almeida Pais et al. (2021)

Le tableau 2.1 résume quelques méthodologies discutées par de Almeida Pais et al. (2021) pour l'évaluation des coûts du cycle de vie d'un actif. Il est nécessaire de choisir la méthode appropriée pour soutenir une bonne décision concernant le remplacement des actifs, et mettre en évidence des variables telles que le coût d'acquisition, la valeur de retrait, les coûts d'exploitation, les coûts de maintenance, les coûts d'exploitation, le taux d'inflation et le taux d'actualisation.

TABLEAU 2.1: QUELQUES MÉTHODOLOGIES D'ÉVALUATION DES COÛTS DU CYCLE DE VIE adapté de: de Almeida Pais et al. (2021)

Méthodologies	Commentaires
Coût total de possession	Prend en considération les coûts d'acquisition, les passations de commande, la recherche et la qualification des fournisseurs, le transport, la réception, l'inspection, le rejet, le remplacement, le temps d'arrêt dû à une panne et les coûts d'élimination.
Coût du cycle de vie	Somme de tous les fonds dépensés pour soutenir l'actif depuis sa conception et sa fabrication jusqu'à la fin de sa vie utile, en passant par son exploitation
Évaluation du cycle de vie	Similaire au coût du cycle de vie, mais évalue la performance environnementale relative.
Analyse des coûts du cycle de vie	Largement utilisée pour soutenir les décisions au niveau des projets
Évaluation de la durabilité du cycle de vie	Utilise principalement des variables quantitatives, telles que les mesures des impacts économiques, environnementaux et sociaux
Valorisation du cycle de vie	Consiste en deux principaux éléments : un cadre de quatre phases (But & portée, inventaire du cycle de vie, évaluation de l'impact et l'interprétation) et une combinaison des calculs.
Investissement dans le cycle de vie	Dans ce cas, le paradigme est différent. Les coûts d'investissement initial et de maintenance généralement considérés comme une perte sont plutôt vus comme des investissements. En effet, ils sont à l'origine d'un rendement qui est indexé sur la qualité de l'investissement initial et au cours de la vie de l'actif, valorisent l'actif tout au long de son cycle de vie.

2.1.3 Types de maintenance

Comme discuté dans la section 2.2.1, le PGES propose des méthodes et des outils pour la mise en place des politiques de maintenance efficaces et efficientes optimisant rentabilité et sécurité. Basri, Abdul Razak, Ab-Samat et Kamaruddin (2017) notent qu'il existe trois principaux types de maintenance : la maintenance corrective (MC), la maintenance préventive (MP) et la maintenance prédictive (MPd). Le tableau 2.2 détaille les types de maintenance et leurs principales caractéristiques (Prajapati, Bechtel, & Ganesan, 2012). Ayu et Yunusa-Kaltungo (2020) parlent plutôt de trois philosophies de maintenance illustrées dans la figure 2.3.

La MC est le type de maintenance le plus simple à mettre en place. Selon Basri et al. (2017), ce type de maintenance consiste à réaliser des actions de réparation ou de remplacement pour remettre un système en état de fonctionnement uniquement après une défaillance. Ding et Kamaruddin (2015) ajoutent que la MC doit être adoptée dans des cas où les marges de profits sont larges. Toutefois, cette politique de maintenance peut causer de grandes pertes de productions et de sérieux dommages sur le système, le personnel et l'environnement à cause des défaillances imprévues.

Compte tenu de cet important inconvénient, Ding et Kamaruddin (2015) soulignent l'apparition de la MP. Ce type de maintenance a été proposé pour prévenir l'occurrence de la défaillance et limiter les conséquences y afférentes. Basri et al. (2017) ajoutent que la MP est planifiée et effectuée après une période spécifique de temps, ou lorsqu'un système spécifique a été utilisé pour réduire la probabilité de la défaillance.

Le coût élevé de l'entretien des équipements complexes rend nécessaire l'amélioration des systèmes d'assistance à la maintenance et de remplacer les concepts traditionnels, comme les stratégies correctives et préventives, par de nouvelles comme les pronostics (Ahmadzadeh & Lundberg, 2014). Ding et Kamaruddin (2015) soutiennent également cette idée et soulignent que la MPd a vu le jour et évolué grâce aux avancées technologiques.

Tout comme la MP, la MPd consiste à prévenir les défaillances et conserver le système opérationnel, mais en faisant une estimation de la DVR à partir de modèles de prédiction. La MPd est principalement basée sur les conditions en temps réel du système. Selon la norme NF EN 13306 X 60-319 « Terminologie de la maintenance », la MPd se définit comme une « maintenance conditionnelle exécutée en suivant les prévisions extrapolées de l'analyse et de l'évaluation de paramètres significatifs de la dégradation du bien ». La surveillance est effectuée au moyen de divers capteurs collectant des informations permettant de refléter des comportements anormaux des équipements. En résumé, la MP et la MPd sont toutes les deux des approches de maintenance proactive avec des objectifs similaires, mais la MP est effectuée uniquement lorsque l'équipement est à l'arrêt et programmée de façon systématique, tandis que la MPd englobe la surveillance de l'équipement et des prévisions qui définissent le moment idéal pour mettre l'équipement à l'arrêt et effectuer l'action de maintenance (Basri et al., 2017).

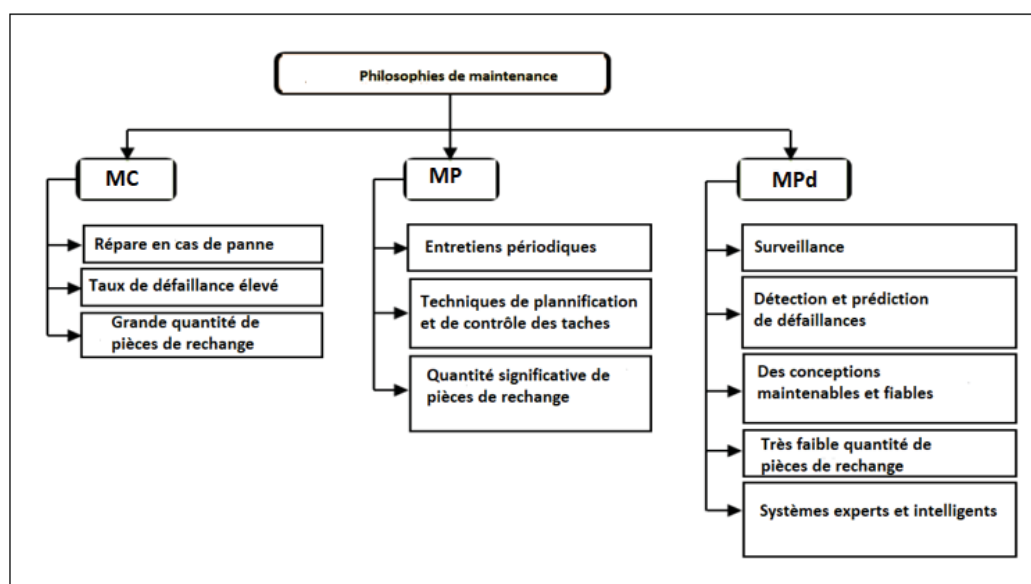


Figure 2.3: Les philosophies de maintenance de Ayu et Yunusa-Kaltungo (2020)

TABLEAU 2.2: LES TYPES DE MAINTENANCE - Prajapati et al. (2012)

Types de maintenance				
Caractéristiques	MC	MP	MPd	
Approche de maintenance	Réactive	Proactive	Proactive	
Catégorie de maintenance	Réparation après défaillance	Périodique	Maintenance basée sur le diagnostic (surveillance conditionnelle)	Maintenance basée sur le pronostic (centrée fiabilité)
Temps d'arrêt	Élevé	Faible	Proche du minimum	Négligeable
Bon pour les défaillances	Aléatoirement basé sur l'âge	Basé sur l'âge	Prévient l'apparition (proche de l'optimum)	Prévient l'apparition
Coûteux (main d'œuvre)	Maximal	Faible	Modéré	Minimal
Coût de déploiement initial	Aucun	Légèrement supérieur	Coûteux	Très coûteux
Coût de calcul	Faible	Légèrement supérieur	Supérieur	Élevé
Planification calendaire requise	Non applicable	Basé sur la durée de vie utile standard du composant ou de l'historique des défaillances	Basé sur les conditions réelles	Basé sur la prédiction de la DVR de l'équipement
Action	Inspecter, réparer ou remplacer après défaillance	Inspecter, réparer ou remplacer à des intervalles prédéfinis à la conception et mis à jour avec de l'expérience	Inspecter, réparer ou remplacer au besoin Collection continue des données de MCON	Prédire la DVR basée sur la charge de contrainte réelle
Type de prédiction	Aucun	Aucun	Système en marche ou à l'arrêt, proche de l'analyse de tendance en temps réel	Système en marche ou à l'arrêt, analyse de tendance en temps réel

2.2 Méthodologies de prédiction de la durée de vie résiduelle

Les chercheurs, dépendamment du contexte des données disponibles et du niveau de complexité ou de tolérance, ont recours à diverses méthodes pour l'estimation de la DVR des actifs physiques. Ces méthodes peuvent toutefois être regroupées en plusieurs catégories, englobant ainsi les différentes techniques et leurs variantes. Selon Sikorska, Hodkiewicz et Ma (2011), il y a un faible consensus entre les chercheurs du domaine du pronostic quant aux classifications les plus appropriées pour regrouper les modèles de prédiction de DVR. En effet, Sikorska et al. (2011) proposent quatre principales catégories : les approches basées sur les connaissances, les approches basées sur l'espérance de vie, les réseaux de neurones artificiels et les approches basées sur la physique (ou basées sur un modèle). Jardine et al. (2006) énoncent quant à eux trois familles : les approches statistiques, les approches d'IA et les approches basées sur un modèle. Okoh et al. (2014) notent à leur tour quatre groupes : les approches basées sur un modèle, les approches basées sur l'analyse, les approches basées sur les connaissances et les approches hybrides. Y. Wang et al. (2020) discutent de trois principales méthodologies : les approches basées sur un modèle, les approches basées sur les données et les approches hybrides. Ahmadzadeh et Lundberg (2014) et Laayouj et Jamouli (2017) soulignent quatre principales méthodologies : les méthodes basées sur un modèle, les méthodes basées sur l'expérience, les méthodes basées sur les données et les méthodes hybrides. Le tableau 2.3 présente un résumé de la revue de littérature faite par les auteurs cités précédemment, portant sur les principales méthodologies de prédiction de DVR.

TABLEAU 2.3: RÉSUMÉ DES DIVERSES APPROCHES D'ESTIMATION DE DVR PRÉSENTÉES PAR DIFFÉRENTS AUTEURS

Auteurs	Approches basées sur un modèle	Approches basées sur les données (incluant IA)	Approches hybrides	Autres approches
Sikorska et al. (2011)	X	X		Approches basées sur l'espérance de vie Approches basées sur la connaissance
Jardine et al. (2006)	X	X		Approches statistiques
Ahmadzadeh et Lundberg (2014)	X	X	X	Approches basées sur les expériences
Okoh et al. (2014)	X		X	Approches analytiques Approches basées sur la connaissance
Y. Wang et al. (2020)	X	X	X	

Ahmadzadeh et Lundberg (2014) remarquent que dans la littérature, les deux méthodologies les plus répandues dans les recherches récentes sont les approches basées sur un modèle et les approches basées sur les données. Cela se confirme à travers le tableau 2.3 où l'on observe que ces deux méthodologies sont les plus citées. Les auteurs rajoutent également que la sélection du meilleur modèle dépend du niveau de précision et de disponibilité des données. Dans le cas d'estimations rapides qui sont moins précises, la méthode basée sur les données est préférée, alors que la méthode basée sur un modèle est appliquée lorsque la précision de l'estimation est importante.

2.2.1 Méthodes basées sur un modèle

Les approches basées sur un modèle sont applicables dans des situations où des modèles théoriques précis peuvent être construits des premiers principes.

Ces modèles doivent être configurés spécifiquement pour les systèmes surveillés et devraient simuler avec précision les réponses aux signaux de commande. Ils impliquent généralement de construire techniquement des modèles théoriques compréhensifs pour décrire la physique du système et les modes de défaillance. Ces approches basées sur la physique dépendent d'une compréhension fondamentale de la physique de défaillance du système (Ahmadzadeh & Lundberg, 2014). En effet, l'idée sous-jacente repose dans la formulation de modèles mathématiques reflétant la mécanique de dégradation du système à l'étude (Jardine et al., 2006; Y. Wang et al., 2020). Sikorska et al. (2011) complètent en notant que ces modèles impliquent une compréhension approfondie du comportement du système en réponse aux contraintes, tant du point de vue macroscopique que microscopique. En outre, cela suppose que ce comportement peut être décrit de manière précise et analytique. Les modèles physiques estiment la durée de vie restante d'un composant ou d'une machine en résolvant une équation déterministe ou un ensemble d'équations dérivées de nombreuses données empiriques.

Robinson, Marzat et Raïssi (2017) utilisent dans leurs travaux une méthode basée sur la physique. Un filtre de Kalman étendu est d'abord appliqué pour estimer l'état de dégradation actuel du système, puis la méthode de fiabilité inverse du premier ordre permet de calculer la fonction de densité de probabilité de la DVR. Les auteurs précisent que la méthode de fiabilité inverse du premier ordre est une méthode de propagation de l'incertitude qui peut être adaptée au calcul de la DVR. Puis, une solution analytique ou numérique de l'équation différentielle décrivant l'évolution de la dégradation du système est nécessaire pour calculer le modèle d'estimation de la DVR. Daigle et Goebel (2013) développent une méthode basée sur un modèle qui consiste en un problème d'estimation conjointe état-paramètre dans lequel l'état du système ainsi que la progression des dommages sont estimés, suivi d'un problème de prédiction dans lequel l'estimation conjointe état-paramètre est propagée dans le temps pour prédire la DVR. Le modèle de ces auteurs est construit et appliqué sur la physique d'une pompe centrifuge.

C. J. Li et Lee (2005) proposent une approche de prédiction de la DVR des engrenages présentant une fissure de fatigue sur la dent, basée sur un modèle d'identification de la rigidité de l'engrènement de l'engrenage, un modèle de dynamique d'engrenage et un modèle de la mécanique de rupture.

2.2.2 Méthodes basées sur les données

Traditionnellement, le processus de dégradation est représenté sous la forme de modèles physiques, qui ne sont pas toujours faisables (Jardine et al., 2006; Mosallam, Medjaher, & Zerhouni, 2016). C'est à cet effet que Ge, Guo et Dou (2019) soutiennent que la prédiction basée sur les données surpasse la limitation de la nécessité de modèles mathématiques précis et compliqués à développer. Y. Wang et al. (2020) rajoutent que grâce à l'essor du big data et de l'internet des objets, les méthodes basées sur les données attirent de plus en plus d'attention dans les industries. Cette énorme quantité de données, collectées grâce à ces nouvelles technologies, devient utile après un traitement et une transformation en informations logiques et exploitables pour l'estimation de la DVR. La précision de l'estimation dépend de la quantité et de la qualité des données. La méthodologie basée sur les données est fondée sur des techniques statistiques et d'apprentissage, la plupart tirant leur origine dans la théorie de reconnaissance des formes. De telles approches englobent les algorithmes conventionnels numériques comme la régression linéaire ou les filtres de Kalman, tout comme les algorithmes communément retrouvés dans les communautés d'apprentissage automatique et d'exploration de données. D'autres algorithmes incluent les réseaux de neurones, les arbres de décision et les machines à vecteurs de support (Ahmadzadeh & Lundberg, 2014).

Y. Wang et al. (2020) proposent un état de l'art sur l'approche basée sur les données, en particulier en utilisant l'apprentissage profond (*Deep Learning* – DL) en termes des quatre principales architectures représentatives profondes, incluant l'auto-encodage, le réseau de croyances profondes, le réseau de neurones convolutifs (*Convolutional Neural*

Network - CNN) et le réseau de neurones récurrents (*Recurrent Neural Network* - RNN). Le DL est un variant de l'apprentissage automatique (*Machine Learning* – ML) qui tire son origine du réseau de neurones artificiels. Il est important de revenir sur les quatre principales architectures d'apprentissage profond énoncées précédemment. L'auto-encodage peut apprendre une nouvelle représentation des données à travers une reconstruction des données entrantes, qui contient deux phases : l'encodage et le décodage. Dans le cadre de la prédiction de durées de vie résiduelles, l'auto-encodage est normalement utilisé pour extraire les caractéristiques de dégradation. Le réseau de croyances profondes est basé sur un ensemble de machines de Boltzmann restreintes empilées les unes sur les autres. Le CNN est un réseau multicouche se concentrant sur les entrées bidimensionnelles telles que l'image, en utilisant des couches convolutives empilées et regroupant des couches pour obtenir l'apprentissage des caractéristiques. Le RNN est une architecture profonde qui contient des connexions de retour masquées ou de couches de sortie aux couches précédentes, ainsi il est capable de traiter l'information.

Ge et al. (2019) proposent dans leur recherche une méthode basée sur les données pour la prédiction de la DVR des machines sous la contrainte d'un petit échantillon de données. Les auteurs utilisent le test de Kolmogorov-Smirnov (KS) pour extraire l'état de santé caractéristique de la machine, qu'ils juxtaposent ensuite avec l'architecture de DL nommée LTSM (*Long Short Term Memory*) reposant sur des données de MCON en temps réel. X. Li et al. (2020) font face à un problème d'observations partielles qu'ils traitent également à partir de la méthode basée sur les données. Les observations partielles sont des informations impossibles à recueillir totalement à chaque étape, elles sont données partiellement. La situation est illustrée et étudiée sur une roue de coupe. Les données de surveillance utilisées dans ce cas sont des images prises de la partie visible de la roue en temps réel. Le DL est appliqué s'appuyant sur ce jeu de données d'images à travers un apprentissage supervisé pour prédire la DVR de la roue de coupe.

Globalement, cette revue de littérature a noté que les approches basées sur un modèle sont précises, mais parfois non faisables ou très difficiles à développer. Les approches basées sur les données quant à elles sont plus simples à développer, mais nécessitent une grande quantité de données. Pour une comparaison plus détaillée, Y. Wang et al. (2020) proposent dans le tableau 2.4 quelques avantages et inconvénients de chacune des deux méthodologies.

TABLEAU 2.4: AVANTAGES ET INCONVÉNIENTS DES APPROCHES BASÉES SUR UN MODÈLE ET BASÉES SUR LES DONNÉES - Y. Wang et al. (2020)

APPROCHE BASÉE SUR UN MODÈLE	APPROCHE BASÉE SUR LES DONNÉES
Avantages	
Efficiente et descriptive	Rapides implémentation et déploiement
Haute précision	Faible coût du développement des algorithmes et faible exigence de connaissance de la physique du système
Facile à valider, certifier et vérifier	Les algorithmes peuvent être adaptés pour être utilisés sur d'autres systèmes
La dynamique des états peut être estimée à chaque intervalle de temps	Aucune exigence sur les hypothèses et les estimations empiriques des paramètres physiques
L'extrapolation hors du jeu d'entraînement permet d'obtenir des prédictions plus robustes	Capacité à transformer des données bruyantes de haute dimension en informations de plus faible dimension pour les décisions de pronostic.
Inconvénients	
Les modèles haute fidélité sont coûteux, prennent du temps et nécessitent des calculs intensifs.	Exigence d'un algorithme robuste
Possibilité de réutilisation limitée	Exigence de la gestion des incertitudes
Difficulté de comprendre la physique des dommages pour les systèmes mécaniques complexes	Précision affectée par la surgénéralisation et le surapprentissage pendant l'entraînement de l'algorithme
Trop stochastique et complexe pour modéliser le défaut	Données inadéquates pour les systèmes nouvellement développés
Requiers l'examen des hypothèses simplificatrices	Résultats contre-intuitifs dus à l'absence de la connaissance de la physique du système
	Requiers une énorme quantité de données
	A un horizon de prédiction court
	Faible précision avec des données de haute dimension

La figure 2.4 illustre la synthèse de cette comparaison en exposant l'applicabilité de ces deux méthodologies en fonction de leur complexité.

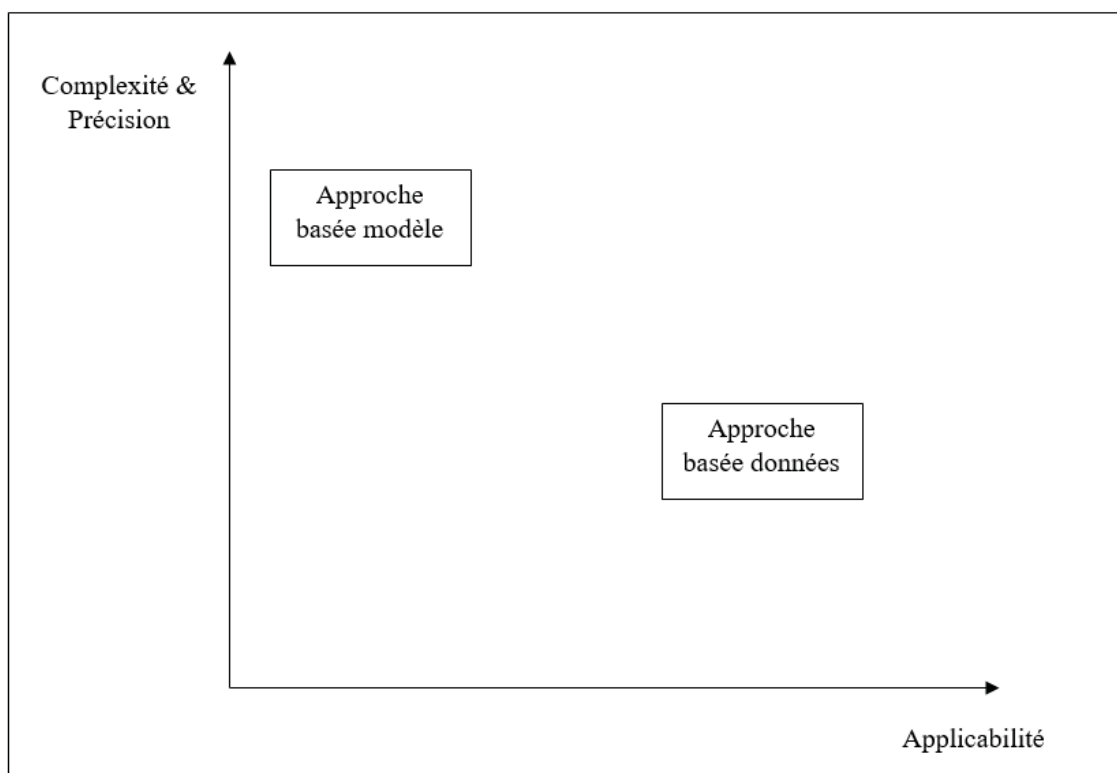


Figure 2.4: Complexité et applicabilité des approches basées sur un modèle et basées sur les données

2.2.3 Autres méthodes

Bien que les approches basées sur un modèle et les approches basées sur les données sont les plus populaires, il existe d'autres méthodologies utilisées par des auteurs. Comme énoncé plus haut dans la section 2.2, il existe un faible consensus entre les différents auteurs du domaine de pronostic quant à la classification des différents modèles de prédiction. Ainsi, de nombreuses terminologies ont été employées dépendamment des auteurs, parmi lesquelles des méthodologies telles que les approches basées sur l'espérance de vie, les approches basées sur la connaissance et les approches hybrides.

Selon Sikorska et al. (2011), les méthodes basées sur l'espérance de vie déterminent la durée de vie des différents composants de la machine par rapport au risque de

détérioration attendu dans des conditions de fonctionnement connues. Ces approches englobent les modèles statistiques et les modèles stochastiques. Sikorska et al. (2011) et Jardine et al. (2006) citent quelques techniques : les fonctions de fiabilité agrégées, les méthodes de probabilité conditionnelle englobant les réseaux bayésiens et les chaînes de Markov, les extrapolations de tendance, les modèles autorégressifs et leurs variants et les modèles à risques proportionnels. Shahin, Simon et Weber (2020) exploitent une forme évoluée de modèle de Markov caché pour représenter le processus d'évolution de l'état de santé d'un système soumis à de multiples conditions opérationnelles. La DVR est ensuite obtenue après la combinaison d'une méthode de rééchantillonnage et une simulation Monte-Carlo à deux niveaux, permettant la considération des incertitudes. Wong et al. (2011) considèrent dans leur étude un modèle à risques proportionnel incluant une covariable basée sur un âge virtuel reflétant les informations valides de réparations et d'entretien.

Les approches basées sur la connaissance concernent la collecte d'informations stockées auprès d'experts et l'interprétation des règles établies. Elles peuvent être considérées comme un système de performance de la prestation de services basée sur les principes du retour d'information sur les services pour l'analyse (Okoh et al., 2014). En d'autres mots, elles évaluent la similitude entre une situation observée et une banque de données de défaillances définies précédemment et déduisent l'espérance de vie à partir d'événements antérieurs (Sikorska et al., 2011). Les approches basées sur la connaissance sont subdivisées en deux principales catégories : les systèmes experts et les systèmes flous. Un système expert est un logiciel qui simule la performance des experts humains dans un domaine particulier. Les règles sont formulées sous la forme de déclarations précises SI-ALORS et sont souvent basées sur des faits heuristiques acquis par un ou plusieurs experts durant des années. Les systèmes flous comprennent une base de connaissance, mais également une logique floue permettant d'assurer un certain degré de vérité comparativement au système expert principalement fondé sur une logique binaire oui-non (Sikorska et al., 2011).

Dans certains cas réels de prédiction, les tendances des paramètres caractéristiques sont diversifiées, rendant difficile l'estimation avec une seule méthode de prédiction. Dès lors, deux ou plusieurs méthodologies sont utilisées en conjonction pour produire des résultats plus précis. La méthode résultante est appelée méthode hybride (Ahmadzadeh & Lundberg, 2014). Y. Wang et al. (2020) remarquent dans leur revue que la plupart des chercheurs se concentrent sur les approches basées sur les données et les approches basées sur un modèle, alors qu'un nombre relativement faible de recherches se concentrent sur les approches hybrides. Les auteurs rajoutent que jusqu'en 2017, on ne trouve que 8 % de la littérature dans ce domaine travaillant sur les approches de pronostic hybrides. Cependant, une approche hybride peut souvent être la meilleure option, car elle tente d'intégrer les avantages de chaque approche considérée. Sur la base d'une certaine quantité de données et des modèles relativement précis, une approche hybride peut généralement atteindre une plus grande précision pour la prédiction de la DVR. Néanmoins, l'inconvénient des approches hybrides est qu'elles présentent également les défauts des deux approches utilisées en conjonction et une complexité accrue dans l'obtention de la solution. Un exemple d'application d'approche hybride se retrouve dans (F.-K. Wang & Mamo, 2019) où les auteurs combinent la méthode de moyenne mobile pondérée exponentielle à des modèles de ML pour la prédiction de la DVR des roulements à billes. Les résultats obtenus ont surpassé les algorithmes de ML utilisés individuellement.

Le tableau 2.5 ci-dessous présente la synthèse de l'état de l'art sur les méthodologies de prédiction de DVR discutées précédemment.

TABLEAU 2.5: VUE D'ENSEMBLE DES DIFFÉRENTES MÉTHODOLOGIES DE PRÉDICTION DE DVR

Méthodologies	Principe sous-jacent	Quelques techniques	Commentaires
Approche basée sur un modèle	Équations mathématiques représentant la physique de dégradation du système	Méthode des éléments finis, équations différentielles ou empiriques, filtres à particules, etc.	Haute précision, mais faible applicabilité
Approche basée sur les données	Exploitation des données de MCON via des algorithmes d'IA	Algorithmes classiques de ML Algorithmes de DL	Moyenne précision et grande applicabilité. Cependant, requiers une énorme quantité de données.
Approches hybrides	Combinaison de deux ou plusieurs approches de natures différentes		Peut améliorer la précision, cependant l'approche est plus complexe compte tenu de la conjonction de plusieurs méthodes
Approches statistiques et stochastiques	Utilisation des techniques de statistique classique et/ou de modèles stochastiques	Les fonctions de fiabilité agrégées, les méthodes de probabilité conditionnelle englobant les réseaux bayésiens et les chaînes de Markov, les extrapolations de tendance, les modèles autorégressifs et leurs variants, les modèles à risques proportionnels, etc.	
Approches basées expériences	Les données récoltées proviennent exclusivement des expérimentations effectuées		Utilise des modèles probabilistes ou stochastiques du phénomène de dégradation ou du cycle de vie des composants en tenant compte des données et des connaissances accumulées par les expériences
Approches basées sur la connaissance	Utilisation des systèmes experts et systèmes flous		

2.3 Le pronostic dans l'industrie électrique

En général, les entreprises du secteur électrique sont constituées de trois principales unités, dont la production, le transport et la distribution. La production est en charge de générer de l'électricité. Le transport est responsable d'acheminer l'énergie sous haute tension. Puis, la distribution a pour mission d'alimenter les clients sous moyenne ou basse tension. Ces trois unités fonctionnent en synergie formant une chaîne appelée réseau électrique. Ce réseau électrique se doit d'être hautement fiable, compte tenu de l'énorme clientèle et de la dépendance des activités mondiales à l'énergie électrique. Dès lors, les entreprises électriques investissent dans l'amélioration de la gestion de leurs actifs. Parmi les différentes stratégies mises au point pour l'amélioration des actifs, l'optimisation de la gestion de la maintenance est incontournable à travers l'implémentation de la maintenance prédictive.

Selon Koziel, Hilber, Westerlund et Shayesteh (2021), trois méthodes sont principalement exploitées par les entreprises électriques pour l'évaluation de la condition d'un composant :

- Les systèmes de notation basés sur les informations disponibles. L'état de santé de l'équipement est représenté par une cote ou score. Ces cotes sont basées sur l'âge du composant et une analyse d'expert appuyée par des mesures effectuées sur le composant.
- Les modélisations markoviennes pour la dégradation des actifs. Une application se retrouve dans (Ruopu Li et al., 2020) où les auteurs utilisent un modèle de Markov caché sur les données d'inspection des fissures de la roue d'une hydroturbine provenant d'opérations réelles pour prédire la DVR et ensuite optimiser les intervalles de maintenance.
- Et enfin, l'utilisation de la distribution des défaillances pour modéliser la probabilité de défaillance, qui est la manière la plus courante de modéliser l'état du composant. La loi de survie est utilisée pour modéliser la fonction de fiabilité

et l'indisponibilité d'un équipement en raison de sa défaillance en fin de vie (c'est-à-dire la probabilité d'une défaillance future compte tenu de son âge). Biard et al. (2021) étudient des stratégies de remplacement des transformateurs d'un réseau électrique de distribution à travers plusieurs scénarios en faisant usage des données historiques et des analyses Weibull.

Cependant, l'énumération des méthodes utilisées dans le secteur électrique pour l'évaluation de l'état de santé des actifs aux fins de pronostic par Koziel et al. (2021) est incomplète. En effet, certains auteurs font usage des approches basées sur les données. Liu et Geng (2016) se servent du SVM (*Support Vector Machine*) pour la prédiction des défaillances d'un moulin à charbon. W. Zhao et al. (2020) exploitent les réseaux de neurones artificiels pour analyser les données de vibrations d'une pompe-turbine d'un groupe hydro-électrique et prédire la DVR sous la contrainte d'une plage de fonctionnement étendue.

2.4 Intégrité des données

Comme discuté précédemment, l'intégrité des données est l'un des principaux problèmes auxquels font face les scientifiques pour pronostiquer la DVR des équipements complexes. Une brève revue relate quelques solutions proposées dans la littérature dans les lignes qui suivent.

2.4.1 Avis d'experts et méthodes bayésiennes

Lorsqu'il s'agit de la complétion des données, les approches bayésiennes sont célèbres. En effet, les méthodes bayésiennes sont généralement utilisées pour intégrer des conditions à priori permettant de corriger les paramètres de distributions déterminés à partir de données incomplètes ou corrompues. Damodaran, Ravikumar et Ramachandran (2014) soulignent la principale différence entre les statistiques classiques et les statistiques bayésiennes en matière de fiabilité. Selon eux, les statisticiens classiques

soutiennent que les paramètres de distribution sont des valeurs fixes, alors que les bayésiens pensent que les paramètres sont des variables aléatoires suivant leurs propres lois de distribution. Les auteurs présentent dans leur étude la combinaison de deux distributions à priori (gamma et uniforme) et d'un modèle bayésien. Sun et al. (2020) établissent un cadre d'inférence bayésienne considérant les avis d'experts et des distributions à priori (bêta et uniforme) pour l'évaluation de la fiabilité d'un produit nécessitant une très haute fiabilité, mais possédant des données de défaillances limitées.

(Albert et al., 2012) ils proposent une méthode pour combiner les avis experts comme un préalable avant de les intégrer à travers une approche bayésienne. Plusieurs scénarios sont proposés quant à l'intégration des avis d'experts. Un consensus peut être réalisé entre les experts afin de déterminer les paramètres de la loi de fiabilité. La moyenne des paramètres estimés peut aussi être incorporée. Une moyenne pondérée des avis d'experts peut aussi être calculée en fonction de l'importance de chacun des experts. Les experts peuvent être également questionnés au niveau de leur incertitude face à leur estimation et de donner un coefficient de pondération à chaque expert en fonction de leur précision estimée. Plus l'incertitude est grande, moins l'impact sera grand sur le résultat final de la décision. Enfin, les avis peuvent être groupés par école de pensée, puis pondérés en considérant le poids de chaque groupe comme égal. Ces méthodes, légèrement différentes les unes des autres, permettent de compléter un lot de données incomplet ou encore d'augmenter la justesse de l'estimation.

2.4.2 Intelligence artificielle

Chollet (2021) définit l'IA comme l'effort d'automatisation des tâches intellectuelles normalement effectuées par les humains. C'est dans ce sens que de nombreuses tâches fastidieuses, qui consomment énormément de temps si réalisées manuellement par des humains, peuvent être exécutées avec une vitesse et une précision époustouflantes par les machines.

Les scientifiques explorent et exploitent les techniques d'intelligence artificielle pour résoudre de nombreux problèmes dans divers domaines. C'est dans cette lancée que l'IA est exploitée par les auteurs pour proposer des solutions au problème d'intégrité des données.

Fink et al. (2020) soutiennent que si les données sensorielles sont peut-être celles auxquelles la plupart des gens pensent en premier lieu lorsqu'il s'agit de données de maintenance, il existe d'autres types d'informations qui peuvent être aussi importantes et qui sont souvent disponibles dans le PGES. Les rapports de défaillance/incident, les dossiers de maintenance et les journaux d'événements peuvent aider à identifier la fréquence, les causes et les remèdes des défaillances, mais nécessitent un traitement différent de celui des données (numériques) des capteurs. Ces données peuvent être utilisées comme sources pour les données d'entrée et les étiquettes/cibles. Cependant, dans la plupart des cas, les enregistrements de pannes et les rapports de maintenance sont saisis manuellement par les techniciens sous forme de texte libre, ce qui introduit un certain degré de variabilité et parfois même d'ambiguïté. Les techniques de TLN de l'IA peuvent aider à automatiser ce processus et à obtenir plus d'informations à partir de ces données, par exemple pour identifier les causes profondes des défaillances. Les modes de défaillance détectés par la machine pourront être ainsi utiles aux fins d'étude de fiabilité. En effet, la classification du système en modes de défaillance ou mécanismes de défaillance est cruciale lors des analyses de fiabilité. Une application de TLN est visible dans (Rajpathak & De, 2016) où les auteurs développent un modèle de fiabilité fusionné en combinant les données de défaillance sur garantie avec les modes de défaillance découverts à partir des données textuelles de réparation non structurées. Les auteurs ont surtout souligné que la prise en compte des modes de défaillance textuels associés aux défaillances des composants a amélioré l'estimation globale de la fiabilité. L'analyse de Thushara et al. (2020) met en évidence la façon dont les techniques de TLN peuvent aider à la préparation et à la conservation des données textuelles pour les analyses de fiabilité, même lorsqu'il existe un manque de standardisation dans la collecte des données opérationnelles et de maintenance et les pratiques de gestion.

En particulier, la mise en œuvre d'algorithmes supervisés et non supervisés permet une approche efficace pour identifier les enregistrements d'intérêt et les classer dans des catégories cohérentes.

Certaines techniques d'apprentissage de l'IA offrent la possibilité d'augmenter des données de toute sorte à partir de modèles génératifs. Behera et Misra (2021) discutent dans leur article de deux techniques utilisées pour produire des échantillons artificiels ressemblant aux données originales : SMOTE (*Synthetic Minority Over-sampling Technique*) et GAN (*Generative Adversarial Network*). SMOTE utilise le concept d'interpolation pour générer une nouvelle fausse instance entre l'instance réelle actuelle et l'une des voisines les plus proches. Cependant, cette technique est surtout utilisée lorsqu'il y a un déséquilibre dans le jeu de données pour une classification binaire. De plus, elle ignore largement les propriétés de distribution des données des échantillons de défauts minoritaires et se base plutôt sur les caractéristiques de leurs données. Par conséquent, cela conduit à un effet d'aveuglement lors du sur-échantillonnage. Le GAN cependant est une technique plus récente proposant une solution possible aux limites de SMOTE, car elle considère la distribution des données avant de générer de nouveaux échantillons artificiels. Les auteurs exploitent dès lors un variant du GAN pour générer des données de défaillance permettant ainsi de compléter le jeu de données utilisé par la suite pour prédire la DVR d'un moteur de turboréacteur. Ces mêmes auteurs reportent que la littérature est très calme au sujet de l'application du GAN dans le domaine de pronostic. Kim, Kim et Choi (2020) font également face à un problème d'incomplétion des données de défaillance aux fins de prédiction de la DVR d'un équipement complexe. Ces auteurs exploitent quant à eux une autre technique d'augmentation basée sur la déformation temporelle dynamique. En fait, ils précisent que cette technique est majoritairement exploitée dans la littérature pour identifier une similarité entre deux jeux de données en trouvant le chemin de correspondance optimal entre eux, alors que leur étude propose la première tentative de générer des données virtuelles en prédisant le chemin de mappage de déformation dynamique temporelle. Une fois les données de

défaillance complétées, les auteurs ont fait usage de l'approche basée sur les données via des réseaux de neurones pour prédire la DVR.

2.4.3 Autres méthodes

D'après Ye et al. (2020), des méthodes récentes recourent à des sources externes comme des bases de données ou internet. Les auteurs adoptent le *crowdsourcing* comme source externe, où des centaines de milliers de travailleurs ordinaires sur la plateforme peuvent fournir des informations de haute qualité basée sur des connaissances contextuelles et l'habileté cognitive humaine. La populaire approche bayésienne est ensuite exploitée par les auteurs en la combinant aux données de *crowdsourcing* pour améliorer la précision de l'imputation.

Le tableau 2.6 ci-dessous présente un sommaire des notions discutées en guise de solutions proposées par différents auteurs au problème d'intégrité des données.

TABLEAU 2.6: DIVERSES SOLUTIONS AU PROBLÈME D'INTÉGRITÉ DES DONNÉES

Méthodes		Commentaires	Avantages	Inconvénients
Méthodes bayésiennes	- Intégration des données de référence ou distributions à priori - Intégration des avis d'experts	Corriger les paramètres des distributions obtenues à partir de données incomplètes ou corrompues	Fournissent un cadre pratique pour un large éventail de modèles	Coût de calcul élevé
Intelligence artificielle	TLN – Classification de texte	Classifier automatiquement les modes de défaillances permettant d'améliorer l'analyse de fiabilité	Transfert de connaissance possible entre les modèles	Coût de calcul élevé
	Augmentation de données : - GAN - SMOTE	Créer des échantillons artificiels ressemblant aux données initiales		
Autres méthodes	Sources externes : - Bases de données - internet	Faire usage des bases de données publiées par certains organismes pour compléter les données à disposition.	Coût de calcul négligeable	Le niveau technologique des équipements, les conditions d'exploitation et les programmes de maintenance associés sont rarement similaires.
	<i>Crowdsourcing</i>	De nombreux travailleurs ordinaires sur la plateforme peuvent fournir des informations de haute qualité basée sur des connaissances contextuelles et l'habileté cognitive humaine		

2.5 La place de l'IA dans le domaine du PGES

Le terme IA est apparu à de nombreuses reprises dans cette revue de littérature, ce qui témoigne son indéniable présence dans le domaine du PGES. En effet, ce document a ressorti l'exploitation de l'IA à travers les approches basées sur les données qui s'avère l'une des méthodologies de pronostic les plus répandues. Outre cela, l'IA s'est montrée utile en proposant quelques solutions au problème d'intégrité des données.

Pour davantage cerner les apports de l'IA dans le domaine du PGES, il serait judicieux de comprendre plus en détail de façon générale ce qu'est l'IA et ses différentes applications. Quand l'on parle d'intelligence, la notion d'apprentissage est incontournable. En effet, l'être humain devient plus intelligent et habile après avoir appris des notions appuyées par quelques exemples (on parle d'entraînement). C'est de la même façon que la machine a besoin d'apprendre par des exemples. Et ces exemples ne représentent rien d'autre que des données compte tenu du monde numérique de l'ordinateur. Toutefois, il est important de noter que contrairement à l'humain qui n'a besoin que de quelques exemples pour capter et reproduire correctement, la machine requiert une grande quantité de données. C'est ainsi que la notion d'apprentissage automatique est une science irréfutable de l'IA, exploitant divers algorithmes classiques alliant mathématiques, statistiques et optimisation. L'apprentissage profond est une forme d'apprentissage plus récente basée sur des réseaux de neurones artificiels permettant à la machine de résoudre des problèmes plus complexes tout en proposant des performances hallucinantes par rapport à celles obtenues grâce aux algorithmes classiques de ML. En intelligence artificielle, il existe trois principales catégories d'apprentissage : l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage par renforcement. Dans l'apprentissage supervisé, le jeu d'entraînement fourni à l'algorithme inclut les solutions désirées (étiquettes). Dans l'apprentissage non supervisé, les données d'entraînement ne comportent pas d'étiquettes. Le système essaie ainsi d'apprendre par lui-même.

Dans l'apprentissage par renforcement, un agent observe l'environnement, sélectionne et exécute des actions, et obtient des récompenses en retour pour finalement apprendre par lui-même la meilleure stratégie (Deng & Liu, 2018). La figure 2.5 illustre les précédentes lignes en représentant les différents types d'apprentissages. Quelques applications de l'IA dans le monde réel sont : classification de produits, analyse de sentiments, détection de tumeurs, création d'assistant personnel virtuel, prédiction de la météo, prévisions de ventes, détection de fraude de carte de crédit, segmentation de clients, création de robots virtuels, vision par ordinateur, etc.

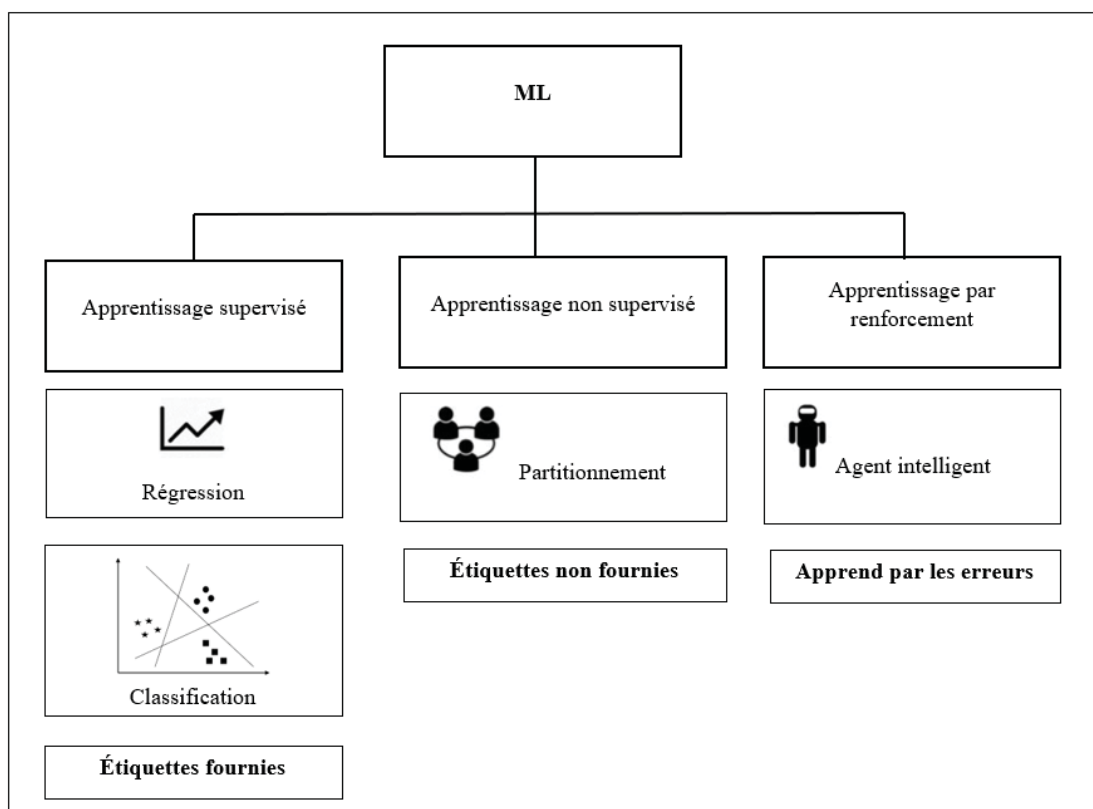


Figure 2.5: Les différents types d'apprentissages

Fink et al. (2020) proposent une brillante revue de littérature discutant du potentiel, des défis et des futures directions de l'IA (DL en particulier), dans les applications de PGES.

Les auteurs présentent l'usage du TLN avec les rapports de maintenance permettant de classer les modes de défaillance. Ensuite, la vision par ordinateur est relatée, permettant les inspections et contrôles de qualité. Dans la vision par ordinateur, celui-ci est nourri par des séquences d'images en temps réel de l'équipement sous surveillance, permettant ainsi de détecter des défauts et de les prévenir. Ensuite, les auteurs ressortent les applications de DL dans le PGES pour les données de MCON collectées par des capteurs. À ce sujet, ils commencent par présenter les algorithmes d'apprentissage supervisé majoritairement exploités dans la littérature, dont les LSTM et CNN pour des détections d'anomalie, diagnostics et pronostics. Ensuite, les algorithmes d'apprentissage non supervisé sont discutés dont l'application la plus courante dans le PGES est la reconstruction de signal. La quantification et la propagation de l'incertitude des données de MCON et l'aide à la décision sont également discutés. Enfin, les auteurs notent quelques directions prometteuses du DL dans le PGES : le transfert de connaissances, le DL pour les approches de flotte, les algorithmes génératifs, l'apprentissage par renforcement et le ML induit par la physique. Khan et Yairi (2018) et Z. Zhao et al. (2021) proposent des travaux similaires discutant des développements et contributions de l'IA dans le domaine du PGES. La particularité de l'article de Z. Zhao et al. (2021) est la mise en lumière de l'importance de la mise en place des jeux de données et codes *open source* pour le développement favorable de la communauté de recherche axée sur l'IA dans le PGES. Le peu de jeux de données et codes *open source* disponible ont été dévoilés dans la foulée.

Dans le secteur électrique plus précisément, outre l'élaboration des prévisions précises de la demande, les algorithmes d'apprentissage automatique sont utilisés pour identifier et anticiper les défaillances (Biard & Nour, 2021). C'est le cas de Jiao et Fang (2019), Liu et Geng (2016) et W. Zhao et al. (2020) qui usent de l'IA pour détecter ou prédire les défaillances des équipements d'une centrale électrique.

2.6 Traitement du langage naturel

Selon Zhou, Duan, Liu et Shum (2020), le TLN est une branche importante de l'IA qui étudie les interactions entre les humains et les ordinateurs via le langage naturel. Cette discipline étudie les technologies fondamentales pour l'expression du sens des mots, des phrases et des documents, ainsi que pour le traitement syntaxique et sémantique, comme la séparation des mots, les analyseurs syntaxiques et sémantiques, et développe des applications telles que la traduction automatique, les questions-réponses, la recherche d'informations, le dialogue, la génération de textes et les systèmes de recommandation.

2.6.1 Origines, évolution et applications

Dans l'article de Kang, Cai, Tan, Huang et Liu (2020), les auteurs dévoilent une vue d'ensemble du TLN. Selon eux, le TLN consiste en deux directions de recherche : la compréhension de langage naturel et la génération de langage naturel. La mission principale de la compréhension de langage naturel est de comprendre le langage humain en déchiffrant les documents et en extrayant des informations précieuses pour les tâches en aval. En revanche, la génération de langage naturel est la production de textes dans des langues naturelles qui sont compréhensibles par les humains à partir de la fourniture de données structurées, de textes, de graphiques, d'audio et de vidéo. Les premières recherches sur le TLN commencent avant les années 1956. Les premiers chercheurs ont exploré des pistes d'application du modèle probabiliste des processus de Markov discret associé au concept d'entropie en thermodynamique. Les premières explorations sur le TLN étaient donc principalement basées sur des règles prédéfinies et des probabilités. Ensuite, des techniques statistiques avancées ont de plus en plus été exploitées, notamment les modèles de Markov cachés. Les modèles de Markov cachés ont apporté une avancée significative dans les recherches sur le TLN, surtout dans le domaine de reconnaissance de la voix. À partir du 21^e siècle, le TLN a connu une révolution hallucinante avec l'exploitation des réseaux de neurones permettant ainsi la prise en

considération du contexte sémantique et par conséquent des résultats spectaculaires dans la compréhension et la génération de langage naturel.

Quelques applications du TLN sont : les moteurs de recherche, les assistants virtuels et chatbots, la segmentation de phrases, l'extraction d'informations, les questions-réponses, la traduction, la récapitulation de texte, l'analyse de sentiments, la détection de spams, le sous-titrage, le commentaire des images, la classification de texte, etc.

2.6.2 Classification de texte

La classification automatique de texte consiste à organiser des documents en classes prédéterminées, généralement à l'aide d'algorithmes d'apprentissage automatique. D'une manière générale, il s'agit de l'une des méthodes les plus importantes pour organiser et exploiter les gigantesques quantités d'informations qui existent sous forme de textes non structurés (Altinel & Ganiz, 2018).

Selon Kang et al. (2020), le processus d'une application TLN peut, en général, être divisé en quatre phases : le prétraitement de texte, la représentation de texte, l'entraînement du modèle et l'évaluation du modèle. Le prétraitement de texte consiste à nettoyer le texte dans le but de limiter le vocabulaire pour améliorer l'efficacité et la précision de l'analyse. Dépendamment du problème à l'étude, une partie ou la totalité des étapes suivantes doivent être effectuées durant la phase de prétraitement: le nettoyage, la tokénisation, la correction des fautes d'orthographe, le marquage des catégories grammaticales, le retrait des mots communs et le *stemming* ou la lemmatisation. Le nettoyage consiste généralement à écarter tous les caractères inutiles aux fins de l'analyse et à normaliser le texte en lettres minuscules. La tokénisation consiste à séparer une phrase en une liste de mot, ou un paragraphe en une liste de phrase. Le marquage des catégories grammaticales permet de faire ressortir le sens d'un mot dans une phrase (par exemple nom, verbe, adjectif, etc.). Le *stemming* ou la lemmatisation transforment les mots sous une forme canonique. La différence entre le

stemming et la lemmatisation est que la lemmatisation essaie de garder l'aspect sémantique du mot, contrairement au *stemming*. Après le prétraitement de texte, il est nécessaire de le représenter sous une forme compréhensible par l'ordinateur. En effet, un ordinateur est incapable de comprendre un texte brut contrairement aux humains, dès lors ces textes doivent être convertis sous forme numérique. Il existe à deux principales méthodes de représentation de texte : le *Bag Of Words* (BOW) et le *Word embeddings*. BOW consiste en une matrice DTM (*Document Term Matrix*) creuse de dimensions $m \times n$ où m représente le nombre de documents présents dans le corpus et n représente le nombre de mots présents dans le vocabulaire. Les trois méthodes les plus communes utilisées dans le BOW sont : *One-Hot encoding*, *Term Frequency* (TF) et TF-IDF (*Term Frequency Inverse Document Frequency*). Dans le cas de One-Hot Encoding, la DTM est constituée uniquement des chiffres binaires 0 et 1. La valeur 1 correspond à la présence d'un mot au moins une fois dans un document. En ce qui concerne TF, la DTM est constituée de la fréquence d'apparition du mot dans le document. La valeur du mot correspond au rapport de la fréquence d'apparition du mot dans le document sur le nombre total de mots présents dans le document. TF-IDF lui en revanche est utilisé lorsqu'on veut attribuer plus de poids à un mot en fonction de sa rareté dans le corpus. En effet, dépendamment des applications, plus un mot est rare, plus il est pertinent pour la discrimination. Altinel et Ganiz (2018) rejoignent Kang et al. (2020) en notant que le BOW montre cependant quelques inconvénients dont le plus notable est le fléau de la dimension connu sous l'expression « *curse of dimensionality* » en anglais. En effet, dans le BOW, les caractéristiques à extraire sont les mots présents dans le vocabulaire. Donc plus le vocabulaire est grand, plus la matrice DTM devient grande et creuse, occasionnant éventuellement des problèmes computationnels. L'autre inconvénient s'avère l'inaptitude du BOW à conserver le contexte des mots, dès lors, l'aspect sémantique est écarté, ou au pire forcé à travers l'utilisation des n-grams (groupes de n mots).

La deuxième méthode *Word embeddings* quant à elle est plus récente et a été créée pour corriger les limites de BOW. En effet, cette méthode conserve l'aspect sémantique en

prenant en compte le contexte des mots dans les documents. De plus, la matrice est dense. Dans le cas de BOW, des algorithmes classiques de ML suffisent, cependant la méthode Word embeddings est employée dans des architectures de DL dont les CNN, RNN et LSTM sont les plus exploités. La figure 2.6 ci-dessous résume les lignes précédentes en illustrant le diagramme de flux d'un projet de classification de texte.

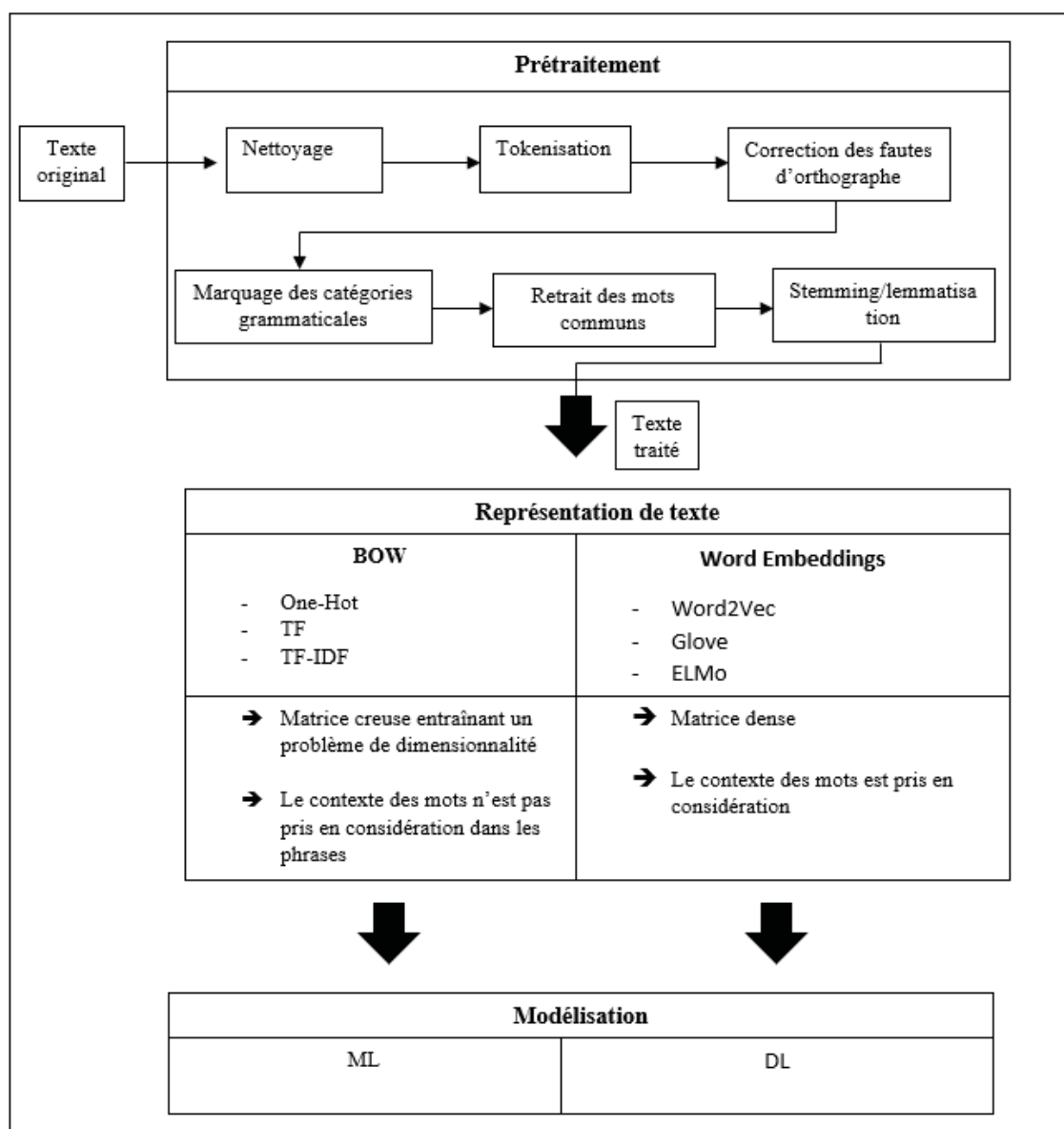


Figure 2.6: Diagramme de flux d'un projet de classification de texte adapté de Kang et al. (2020)

2.7 Matrice de la revue de littérature

La matrice de la revue de littérature permet de donner une vue d'ensemble des différents articles cités précédemment répondant aux deux questions principales de cette recherche, notamment les différentes méthodes d'estimation de DVR et les différentes approches méthodologiques encadrant le problème d'intégrité des données. La problématique, le but, la méthodologie, les variables indépendantes, les variables dépendantes et les résultats de chaque article sont présentés dans cette matrice. La matrice est constituée en deux sous-matrices dont une portant sur les méthodes de prédiction de DVR (tableau 2.7 à 2.13) et l'autre portant sur l'intégrité des données (tableau 2.14 à 2.16).

TABLEAU 2.7: MATRICE DES APPROCHES DE PRÉDICTION DE DVR

Références	Ahmadzadeh et Lundberg (2014)	Wong et al. (2011)	Ge et al. (2019)
Problématique	Améliorer le soutien à l'entretien des systèmes et remplacer les concepts traditionnels comme la prévention et des stratégies correctives avec de nouvelles comme les pronostics.	Déterminer le moment idéal pour remplacer un actif en prenant en compte tous les aspects économiques et techniques	Pour les complexes dispositifs électromécaniques, il est souvent difficile d'établir des modèles mathématiques précis et complets
But	Étudier la littérature sur les méthodes de modélisation des estimations de durée de vie résiduelle.	Incorporer des informations sur les actions de réparation et d'entretien afin d'améliorer les prédictions DVR	Fournir un système simple et méthode efficace pour prédire la durée de vie restante de la machine sous l'état d'un petit échantillon
Méthodologie	Revue de littérature	Étude de cas	Étude de cas
Variables indépendantes	Approches basées sur l'expérience, approche basée sur les données, approches basées sur un modèle, approches hybrides, données de MCON.	Données de MCON, âge virtuel, Modèle à risques proportionnel, coûts d'exploitation, d'entretien et d'administration	Méthode basée sur les données, données de MCON, distance de KS, LSTM
Variables dépendantes	DVR	DVR et décision économique optimale	DVR
Résultat	Le choix du meilleur modèle dépend du niveau d'exactitude et de disponibilité des données. En cas de rapides estimations moins précises, la méthode basée sur les données est préférée en raison de sa facilité de calcul. Cependant lorsque l'exactitude est importante et qu'il y a accès à moins de données, l'approche basée sur un modèle est conseillée.	L'intégration d'informations de réparation et d'entretien via une covariable d'âge virtuel simple a un effet prometteur sur le modèle à risques proportionnel pour l'estimation de la DVR	Comme le modèle s'applique aux petits échantillons, il peut également s'appliquer lorsque les données complètes de la machinerie sont difficiles à obtenir

TABLEAU 2.8: MATRICE DES APPROCHES DE PRÉDICTION DE DVR (SUITE 1)

Références	Y. Wang et al. (2020)	X. Li et al. (2020)	Okoh et al. (2014)
Problématique	Peu d'études se concentrent sur le DL dans le cadre de prédiction de DVR.	Les approches basées sur les données ont besoin d'une grande quantité de données de MCON pour être précises, mais il n'est pas toujours possible de recueillir de telles données dans la vie réelle	Assurer une prestation de service durable
But	Faire une brève introduction sur les méthodes de prédiction de DVR et passer en revue les approches DL utilisées à cet effet.	Proposer une prédiction DVR basée sur l'apprentissage profond méthode pour les composants des machines avec observations partielles	Identifier les mécanismes de défaillance et ressortir la classification des différentes approches de prédiction de DVR
Méthodologie	Revue de littérature	Étude expérimentale	Revue de littérature
Variables indépendantes	Approches basées sur un modèle, approches hybrides, approches basées sur les données, algorithmes de DL	CNN, LSTM, données de MCON (uniquement des images avec observations partielles), mécanisme d'attention supervisée	Méthodes analytiques, méthodes basées sur un modèle, méthodes basées sur la connaissance, méthodes hybrides
Variables dépendantes	DVR	DVR	DVR
Résultat	Les techniques DL sont considérées comme des solutions puissantes en raison de leur capacité à fournir une plus grande agilité pour traiter les données associées. La tendance à la croissance de la littérature montre un intérêt croissant et suggère un avenir prometteur du DL dans la prédiction DVR.	Les résultats de l'expérience suggèrent que la méthode proposée offre une approche prometteuse sur les problèmes de pronostic avec observations partielles	l'importance des services d'ingénierie tout au long du cycle de vie d'un actif a été relevée pour la durabilité des services industriels.

TABLEAU 2.9: MATRICE DES APPROCHES DE PRÉDICTION DE DVR (SUITE 2)

Références	Shahin et al. (2020)	Sikorska et al. (2011)	Jardine et al. (2006)
Problématique	Évaluer l'état de santé d'un élément observable seulement à partir de variables d'observation	De nombreuses recherches ont été effectuées pour le développement de modèles de prédiction. Cependant, l'implémentation dans les industries montre un succès limité.	De nombreux articles sont publiés chaque année pour des recherches appliquées au PGES. Il s'avère utile de proposer une vue d'ensemble des contours du domaine.
But	Exploiter une forme évoluée de modèle de Markov caché pour représenter le processus d'évolution de l'état de santé d'un système	Discuter des questions commerciales qui doivent être prises en compte lors de la sélection d'une approche de modélisation appropriée pour l'essai. Présenter également des tableaux de classification et des diagrammes de processus pour aider l'industrie et le personnel de recherche à sélectionner les modèles de pronostic appropriés pour prédire la durée de vie utile restante des actifs d'ingénierie dans leur environnement commercial spécifique.	Résumer et réviser les recherches et développements en matière de PGES avec un accent sur les modèles, algorithmes et technologies utilisées pour traiter les données et prendre des décisions relatives à la maintenance
Méthodologie	Simulation	Revue de littérature	Revue de littérature
Variables indépendantes	Méthodes stochastiques notamment : les modèles markoviens, les réseaux de Petri, les réseaux Bayésiens statiques et dynamiques, les modèles de Markov (tous les variants), bootstrap, algorithme de Viterbi	Les méthodes basées sur l'espérance de vie (statistiques et stochastiques), les méthodes basées sur la connaissance (systèmes experts, systèmes flous), réseaux de neurones artificiels et les modèles physiques	MCON, données de MCON, programme MCON (acquisition des données, traitement des données et prises de décisions pour recommander les politiques de maintenance efficaces), approches statistiques, intelligence artificielle, les approches basées sur un modèle
Variables dépendantes	DVR	Pronostic et diagnostic	Pronostic et diagnostic
Résultat	La DVR estimé est cohérent et proche de la réalité. Les transitions d'états de la chaîne cachée sont les moments où l'estimation de durée de vie prend plus de sens et offre une meilleure correspondance avec la réalité.	Il existe de nombreuses méthodes employées dans le cadre du PGES, que ce soit pour en diagnostic ou pronostic. Les acteurs doivent s'assurer du choix de la meilleure méthode applicable à leur scénario pour obtenir les meilleurs résultats possible.	Les méthodes avancées de maintenance (MCON) ne sont pas encore suffisamment implantées en industrie comparativement aux méthodes conventionnelles comme la maintenance corrective et la maintenance préventive systématique

TABLEAU 2.10: MATRICE DES APPROCHES DE PRÉDICTION DE DVR (SUITE 3)

Références	Robinson et al. (2017)	Daigle et Goebel (2013)	C. J. Li et Lee (2005)
Problématique	La défaillance de systèmes complexes peut entraîner des désastres humains et industriels	Assurer la sécurité tout en satisfaisant les services	La connaissance de la DVR d'un engrenage fissuré peut permettre à l'équipe de maintenance de planifier adéquatement la maintenance et prévenir des défaillances catastrophiques
But	Proposer une méthode analytique (basé sur la physique) pour le pronostic de la DVR	Développer un cadre pour une approche de pronostic basée sur un modèle qui gère les processus simultanés de progression des dommages	Présenter une méthode basée sur un modèle pour prédire la DVR d'un engrenage avec une fissure de fatigue
Méthodologie	Simulation	Simulation	Simulation
Variables indépendantes	Méthode de fiabilité inverse de premier ordre, filtre de Kalman (variant de filtre à particules), équations différentielles	Filtre à particules, paramètres de dégradation, équations mathématiques, algorithme de contrôle de variance (pour réduire l'incertitude)	MEF (méthode des éléments finis), équations dynamiques, modèle intégré d'estimation de la taille de la fissure, simulateur de la dynamique d'engrenage, algorithme de propagation de fissure
Variables dépendantes	DVR	DVR	DVR
Résultat	La méthode de fiabilité inverse de premier ordre peut être utilisée comme une alternative valable aux modèles basés sur les filtres à particules pour le pronostic de DVR dans le cas où les algorithmes ont un faible coût de calcul	Une bonne performance de l'approche a été observée. Cependant malgré la robustesse des filtres à particules, ils souffrent d'une grosse complexité de puissance de calcul	Erreur inférieure à 7 %, ce qui est très bon.

TABLEAU 2.11: MATRICE DES APPROCHES DE PRÉDICTION DE DVR (SUITE 4)

Références	F.-K. Wang et Mamo (2019)	Ruopu Li et al. (2020)	Biard et al. (2021)
Problématique	Les roulements sont des composants essentiels au fonctionnement des machines rotatives, et pourraient causer des dommages catastrophiques au système complet en cas de défaillance	L'hydro-électricité la source dominante d'électricité dans le monde. Dès lors la disponibilité des équipements doit être irréprochable en vue de satisfaire la grande demande.	Équilibrer la performance, les coûts et les risques
But	Proposer une approche hybride qui combine la moyenne mobile pondérée exponentielle et le ML pour la prédiction de la DVR de roulements à billes	Proposer un nouveau cadre basé sur l'évaluation du cycle de vie pour une planification de maintenance dynamique des turbines hydrauliques	Estimer le volume de remplacement et les coûts associés des actifs majeurs dans les années à venir
Méthodologie	Simulation	Étude de cas	Analyse statistique
Variables indépendantes	Moyenne mobile pondérée exponentielle, ML (SVM, <i>Random Forest</i>), évolution différentielle, données de vibration	Modèle de Markov caché, données d'inspection	Analyse Weibull, données opérationnelles (historiques), vieillissement, causes externes,
Variables dépendantes	DVR	DVR, intervalles de maintenance, planification de maintenance dynamique	Cycle de vie, équilibre du triangle performance-coûts-risques
Résultat	La méthode proposée est un outil fiable pour la prédiction de la DVR des roulements à billes	La méthodologie développée est fiable pour la planification de la maintenance, et permet d'optimiser les coûts par rapport aux méthodes conventionnelles tels que les entretiens périodiques. Le cadre présenté dans l'article a des bénéfices significatifs en termes de MCON pour un composant sans système de surveillance en temps réel.	Les résultats permettent de sélectionner une stratégie de remplacement selon les objectifs de l'entreprise en optimisant soit la performance, les coûts ou en minimisant les risques

TABLEAU 2.12: MATRICE DES APPROCHES DE PRÉDICTION DE DVR (SUITE 5)

Références	Liu et Geng (2016)	W. Zhao et al. (2020)	Fink et al. (2020)
Problématique	Augmentation des exigences sur la sécurité et la fiabilité des machines	Les unités hydroélectriques doivent parfois fonctionner sous des conditions extrêmes hors des limites de conception. Dès lors, les machines, prédisposées à de sérieux dommages et des courtes durées de vie	Bien que les actifs industriels complexes aient fait l'objet d'une surveillance intensive et que de grandes quantités de signaux de surveillance de l'état aient été collectées, l'application des approches d'apprentissage profond pour la détection, le diagnostic et la prévision des défauts des actifs industriels complexes sont limités.
But	Proposer une méthode pour la prédiction de défauts d'un équipement de centrale électrique basée sur la régression SVM	Développer une procédure de surveillance basée sur l'application de réseaux de neurones pour gérer les effets des conditions opérationnelles étendues sur une pompe-turbine	Fournir une évaluation approfondie des développements actuels, des moteurs, des défis, des solutions potentielles et des besoins futurs en matière de recherche dans le domaine de l'apprentissage profond appliqué aux applications de PGES.
Méthodologie	Étude expérimentale	Simulation	Revue de littérature
Variables indépendantes	SVM, données de MCON	Données de vibration, mappage autorégressif, réseaux de neurones artificiels	DL : TLN et systèmes de recommandation, TLN pour les rapports de maintenance, vision par ordinateur, quantification de l'incertitude, transfert de connaissance, algorithmes génératifs, apprentissage (supervisé, non supervisé, semi-supervisé, par renforcement), ML induit par la physique, les approches de flotte
Variables dépendantes	Prédiction de défaut	DVR	Pronostic et diagnostic
Résultat	Le modèle a une bonne précision de prédiction et une habilité de généralisation. Cependant, à cause des impacts à long terme des tendances et des effets saisonniers, la méthode proposée est seulement utile pour la prédiction de la tendance dans des périodes courtes.	Les zones où les opérations ne sont pas recommandées peuvent être identifiées et les effets estimés	Le DL offre de nombreuses orientations prometteuses dans le domaine de la gestion du cycle de vie des produits, avec la possibilité d'améliorer la disponibilité, la sécurité et la rentabilité des actifs industriels complexes. Cependant, plusieurs exigences doivent être satisfaites avant de pouvoir réaliser des progrès significatifs.

TABLEAU 2.13: MATRICE DES APPROCHES DE PRÉDICTION DE DVR (SUITE 6)

Références	Khan et Yairi (2018)	Z. Zhao et al. (2021)
Problématique	Compte tenu des progrès des capacités technologiques modernes, la mise en place d'une stratégie intégrée de gestion de la santé et de diagnostic devient un élément important du cycle de vie opérationnel d'un système	il y a encore une lacune dans la couverture simultanée de la surveillance, du diagnostic et du pronostic basés sur des méthodes fondées sur l'IA, et l'importance d'une communauté de sources ouvertes, y compris les jeux de données et les codes de sources ouvertes, n'a pas été pleinement soulignée.
But	Proposer une revue de littérature sur la contribution du DL dans le PGES, avec un accent sur les défis et les opportunités futures	Fournir un aperçu systématique du développement actuel, des technologies communes, des jeux de données open source, des codes et des défis des méthodes de PGES basées sur l'intelligence artificielle sous trois aspects : surveillance, diagnostic et pronostic
Méthodologie	Revue de littérature	Revue de littérature
Variables indépendantes	Algorithmes de DL	Algorithmes de DL, GAN, codes et jeux de données open source, transfert de connaissance, modélisation de l'incertitude
Variables dépendantes	Pronostic et diagnostic	Pronostic et diagnostic
Résultat	Le DL offre de nombreuses opportunités au domaine du PGES, cependant il y a encore de nombreux défis à relever qui ouvrent la porte à de nombreuses recherches futures	Le DL offre de nombreuses opportunités au domaine du PGES, cependant il fait face à certaines limites résumées par la capacité de généralisation des modèles sur des scénarios réels

TABLEAU 2.14: MATRICE DE L'INTÉGRITÉ DES DONNÉES

Références	Ye et al. (2020)	Albert et al. (2012)	Damodaran et al. (2014)
Problématique	Incomplétude pendant le processus de collecte de données	Il est très difficile de recueillir les données pour fournir les informations requises	La majorité des modèles de fiabilité pour logiciel ont été conçus avec un cadre classique où des paramètres sont considérés comme inconnus et fixes.
But	Proposer une méthode pour l'imputation de valeur manquante en combinant le réseau bayésien et le <i>crowdsourcing</i>	Considérer le problème de combiner les opinions de différents experts d'une manière explicitement basée sur le modèle pour construire un a priori subjectif valide dans une approche statistique bayésienne	Proposer un nouveau modèle bayésien de fiabilité combinant deux distributions à priori pour prédire le nombre total de défaillances et la prochaine défaillance d'un logiciel
Méthodologie	Études expérimentales	Applications	Application
Variables indépendantes	Réseau bayésien, MLE et <i>crowdsourcing</i> , internet, bases de données externes	Approche bayésienne, modèle basé, Avis d'experts (consensus sur les avis d'experts, moyenne sur les avis des experts, moyenne pondérée sur les avis d'experts, incertitude sur les avis d'experts), régression indirecte des probits, répartition sous-jacente des probabilités (Log-Normal), algorithme MMMC	Approche bayésienne, distributions à priori (gamma et uniforme), données de défaillance, MLE
Variables dépendantes	Précision de l'imputation	Distribution à priori	MTBF
Résultat	Les résultats expérimentaux montrent que l'approche proposée arbore une meilleure performance comparée à d'autres méthodes d'imputation	L'approche décrite dans le document est assez générique en ce sens qu'elle ne dépend pas des distributions particulières impliquées dans le processus d'obtention, cela ne dépend pas non plus des questions qui sont posées aux experts. Un aspect critique de la méthode par rapport au modèle de mélange ou aux approches plug-in est qu'il est informatiquement plus exigeant, afin de faciliter la combinaison hiérarchique d'opinions tout en tenant compte de l'erreur au sein de l'expert.	Les temps de défaillance prédits par le modèle sont proches des temps de défaillance réels. Le modèle est simple et facile à adopter, et en plus, même en cas de faible échantillon ce modèle bayésien donnerait de bons résultats.

TABLEAU 2.15: MATRICE DE L'INTÉGRITÉ DES DONNÉES (SUITE 1)

Références	Sun et al. (2020)	Rajpathak et De (2016)	Thushara et al. (2020)
Problématique	Les données de défaillances sont insuffisantes pour une analyse fiable	Traditionnellement, le processus d'estimation de la fiabilité utilise les données de garantie, mais il n'offre qu'un soutien limité pour traiter le problème de confusion des défaillances, par lequel différents modes de défaillance associés à la défaillance d'un composant sont confondus en un seul mode de défaillance. L'estimation de la fiabilité qui en résulte n'a pas le niveau de précision requis.	Les onduleurs sont une source majeure de défaillances matérielles et contribuent à des pertes d'énergie significatives sur les sites photovoltaïques.
But	Établir un cadre d'inférence Bayésienne d'évaluation de fiabilité pour un produit (machine CNC) nécessitant une très haute fiabilité possédant des données de défaillances limitées.	Proposer la fiabilité d'un système réel en fusionnant les données de défaillance sur garantie et les modes de défaillances extraits de données textuelles non structurées	Utiliser le ML pour analyser un jeu de données de rapports de maintenance en vue de ressortir les modes de défaillance prédominants.
Méthodologie	Étude de cas	Étude de cas	Étude de cas
Variables indépendantes	Approche bayésienne, avis d'experts, distributions à priori (uniforme et Bêta)	Modes de défaillance, données textuelles, données de défaillance, distribution Weibull, MLE, estimation par les moindres carrés, classification de texte	Rapports de maintenance, ML, classification de texte, modélisation structurelle des thèmes, partitionnement, distribution Weibull, modes de défaillance
Variables dépendantes	Fiabilité	Fiabilité	Fiabilité
Résultat	Les résultats obtenus sont précis et montrent une réduction de l'incertitude significativement à 95 % d'intervalle de confiance	Cette approche pratique a démontré l'importance de la fusion des modes de défaillances issues des données textuelles et des données de garantie, et a témoigné d'améliorations significatives en fournissant une estimation de fiabilité plus précise des composants	Cette approche permet d'étendre les modes de défaillance en informations exploitables qui peuvent soutenir le développement de la maintenance conditionnelle, ce qui pourrait réduire davantage les défaillances et les pertes d'énergie associées sur les sites photovoltaïques.

TABLEAU 2.16: MATRICE DE L'INTÉGRITÉ DES DONNÉES (SUITE 2)

Références	Behera et Misra (2021)	Kim et al. (2020)
Problématique	La plupart des travaux publiés considèrent la disponibilité des données d'entraînement avec un nombre égal d'échantillons normaux et de défauts concernant différents états de santé de la machine. Cependant, les scénarios pratiques doivent faire face à l'indisponibilité des données de défaut, ce qui entraîne un ensemble de données d'entraînement déséquilibré. Ce problème peut conduire à des inexactitudes avec une prédiction de défaut manquée dans les approches d'estimation de DVR.	Les pronostics de bases de données requièrent une grande quantité de données de défaillance qui n'est pas facile à obtenir en situations réelles
But	Proposer un nouveau cadre de pronostic basé sur le réseau génératif antagoniste conditionnel et un réseau récurrent profond pour augmenter les données et prédire la DVR de systèmes complexes	Présenter l'utilisation de la méthode de déformation temporelle dynamique pour augmenter les données de défaillance.
Méthodologie	Études expérimentales	Application
Variables indépendantes	GAN, SMOTE, données artificielles, internet des objets	Données de défaillance, déformation temporelle dynamique, données virtuelles, réseau de neurones
Variables dépendantes	DVR	DVR
Résultat	Cette approche améliore la précision de la prédiction de la DVR d'au moins 15 % comparé à d'autres travaux publiés sur le jeu de données utilisé.	Des exemples numériques sont donnés pour montrer que l'augmentation des données basée sur la déformation temporelle dynamique proposée peut prédire la DVR avec moins d'incertitude que le modèle de réseau neuronal conventionnel sans mappage des données.

2.8 Synthèse de la revue de littérature

Les différentes notions abordées dans cette revue de littérature peuvent être regroupées en quatre catégories de variables indépendantes agissant sur la réponse qu'est la DVR. Les quatre catégories de variables indépendantes sont : décision économique, disponibilité, prédiction et intégrité des données. Tout au long de la revue de littérature, l'importance d'une décision économique optimale et une optimisation des coûts a clairement été mise en évidence, tout comme la nécessité d'une disponibilité élevée des actifs. De plus, les différentes techniques de pronostic ont été dévoilées, de même que quelques solutions au problème d'intégrité des données. La figure 2.7 ci-dessous illustre le diagramme Ishikawa représentant la synthèse de la revue de littérature à travers les variables indépendantes agissant sur la DVR.

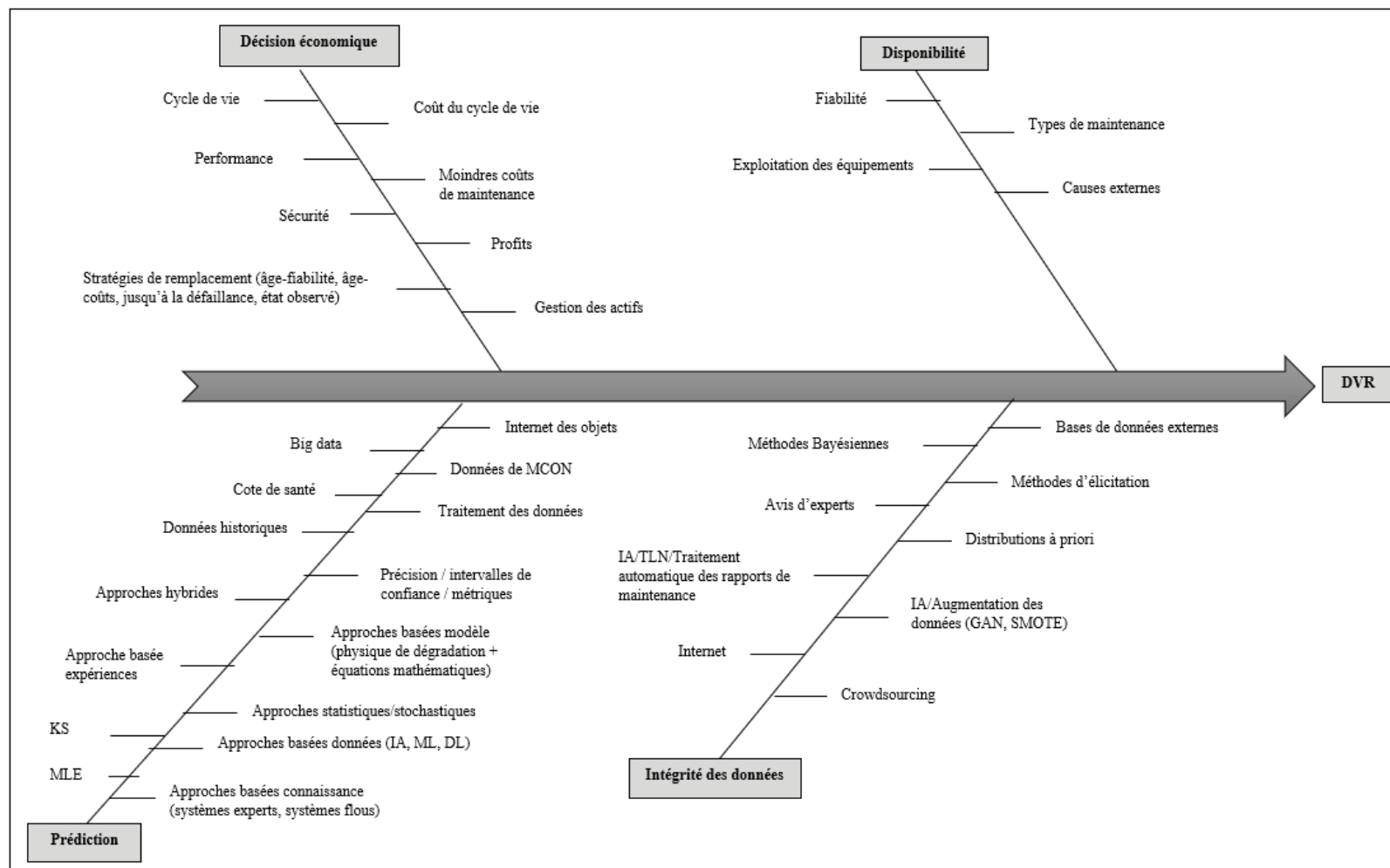


Figure 2.7 : Diagramme Ishikawa

2.9 Conclusion

Ce paragraphe marque la fin du chapitre 2, portant sur la traditionnelle et indispensable revue de littérature dans un projet de recherche. La notion de PGES a tout d'abord été abordée, pour montrer l'importance du pronostic dans la gestion d'un actif, en passant par la gestion du cycle de vie et les politiques de maintenance. Les méthodologies de prédiction ont ensuite été présentées, dont parmi elles les deux les plus répandues dans la littérature sont les approches basées sur un modèle et les approches basées sur les données. Les approches basées sur un modèle sont les plus précises, toutefois ne sont pas toujours réalisables compte tenu de la complexité de certains systèmes dont la physique de dégradation s'avère compliquée pour le développement d'équations mathématiques correspondantes. Les approches basées sur les données gagnent cependant davantage en popularité, car elles sont plus simples à implémenter, mais nécessitent une énorme quantité de données de MCON. Un regard a par la suite été porté au pronostic dans l'industrie électrique. Dans les applications réelles, les données nécessaires pour l'implémentation des techniques de prédiction se montrent parfois incomplètes, voire erronées ou imprécises. À cet effet, certaines approches exploitées dans la littérature ont été discutées parmi lesquelles la réputée méthode bayésienne, l'intelligence artificielle et l'usage des sources externes. La contribution de l'IA dans le PGES a ensuite été relatée, avec un léger regard dans le secteur électrique. Le TLN, en particulier la classification de texte, a été présenté compte tenu de son importance dans ce projet de recherche. La matrice de la revue de la littérature et la synthèse ont servi d'épilogue. En résumé, une intégrité des données soignée, des décisions économiques optimales, un outil de prédiction adéquat et une disponibilité élevée démontrent la robustesse d'une estimation de DVR. L'objectif principal de cette recherche consiste au développement d'un algorithme de traitement des données de défaillance. Le prochain chapitre dévoile la méthodologie appliquée.

CHAPITRE 3 – MÉTHODOLOGIE DE LA RECHERCHE

Ce troisième chapitre présente les différentes étapes parcourues pour atteindre les objectifs de cette recherche, qui se résument à proposer un algorithme de traitement des données historique de HQP afin de les rendre exploitables aux fins d'estimation de DVR.

3.1 Hydro-Québec Production

HQP est le groupe de HQ en charge de générer de l'énergie électrique. Le parc de production compte 61 centrales hydroélectriques et 24 centrales thermiques, ce qui représente une puissance installée de 37,2 GW (Hydro-Québec, 2021b).

Les centrales hydroélectriques sont constituées de nombreux groupes turbine-alternateur (GTA) où la turbine et l'alternateur fonctionnent ensemble pour produire de l'électricité. Le rôle de la turbine est de transformer l'énergie hydraulique en énergie mécanique, de manière à faire tourner un alternateur. L'alternateur, à son tour, transforme l'énergie mécanique en énergie électrique. La figure 3.1 tirée de (Hydro-Québec, 2021a) illustre une vue en coupe d'un GTA. Outre la turbine et l'alternateur, un GTA est constitué d'importants systèmes tels que les automatismes, les liens réseau, les auxiliaires, etc.

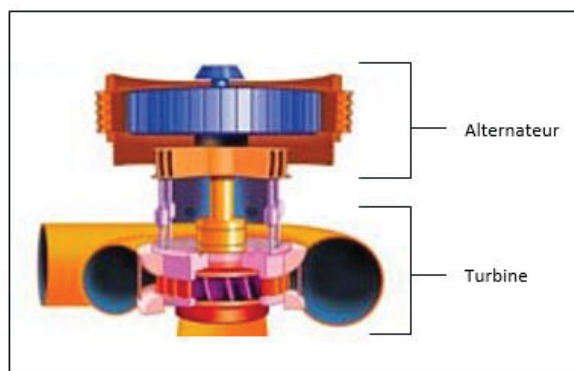


Figure 3.1 : Vue en coupe d'un GTA

3.2 Bases de données

Tout au long de ce mémoire, il a été ressorti deux principaux types de données pour l'estimation de la DVR : les données historiques et les données de MCON. Pour rappel, les données historiques sont les données de défaillance ou de toute action effectuée sur l'équipement dans le temps. Les données de MCON sont des données de surveillance collectées en temps réel par des capteurs.

Les données mises à disposition par HQP pour la présente étude sont des données historiques. Quatre bases de données comportant des données historiques ont été explorées : SIPROD, CEGTA, Diag, travaux majeurs.

- SIPROD contient toutes les informations de défaillances et d'entretiens survenus sur les équipements.
- CEGTA contient toutes les informations des changements d'état des GTA. Elle est la base de donnée la plus fiable, car toutes les informations qu'elle contient sont vérifiées et validées par les autorités.
- Diag contient les informations de dates de mise en service ou réhabilitation des équipements, leur âge et leur cote de santé.
- Travaux majeurs contient les informations de travaux majeurs effectués sur les équipements tels que des réhabilitations ou rénovations.

Le tableau 3.1 dévoile une synthèse de ces quatre bases de données avec les informations pertinentes qu'elles contiennent. Le tableau 3.2 décrit les informations pertinentes présentes dans le tableau 3.1.

TABLEAU 3.1: INFORMATIONS PERTINENTES CONTENUES DANS LES BASES DE DONNÉES

Base de données	SIPROD	CEGTA	Diag	Travaux majeurs
Centrale	✓	✓	✓	✓
Numéro du groupe	✓	✓	✓	✓
Système	✓	✓	✓	✓
Composant générique	✓		✓	✓
Composant ACE		✓		
Composants primaires & secondaires	✓			
Type de travail	✓			
Bon de travail (BT)	✓			✓
Description du bon de travail	✓			
Date	✓	✓		✓
Durée		✓		✓
Coûts	✓			
Faits saillants		✓		
Heures de main d'œuvre	✓			
État		✓		
Année de mise en service			✓	
Âge			✓	
Cote de santé			✓	
Nature du travail				✓

TABLEAU 3.2: DESCRIPTION DES INFORMATIONS PERTINENTES CONTENUES DANS LES BASES DE DONNÉES

Information	Description
Centrale	Nom de la centrale
Numéro du groupe	Numéro du GTA. Une centrale est constituée de plusieurs GTA
Système	L'un des systèmes du GTA
Composant générique	Un sous-système, un des principaux composants de l'un des systèmes
Composant ACE	Composant générique associé au composant primaire (exemple : bobinage du stator. La bobine représente le composant primaire et le stator représente le composant générique du système alternateur)
Composants primaires & secondaires	Composants primaires et secondaires à l'origine de l'intervention sur le composant générique
Type de travail	Conditionnel, correctif, projet, préventif, etc.
Bon de travail (BT)	Numéro du bon de travail
Description du bon de travail	Description de l'intervention
Date	Date
Durée	Durée de l'évènement
Coûts	Coûts de l'intervention
Faits saillants	Faits saillants de l'évènement d'arrêt du GTA
Heures de main d'œuvre	Heures de main d'œuvre travaillées sur l'intervention
État	L'état est un code correspondant au type d'arrêt. Exemple : arrêt non planifié, arrêt planifié long terme, refus de démarrage, etc.
Année de mise en service	Année de mise en service initiale ou après réhabilitation du composant
Âge	Âge du composant
Cote de santé	Cote représentant une appréciation de l'état de santé du composant en fonction de plusieurs critères
Nature du travail	Description du travail majeur réalisé. Cet attribut est présent uniquement dans la base de données des travaux majeurs.

3.3 Intégrité des données chez Hydro-Québec Production

Comme discuté précédemment, HQP dispose de plusieurs bases de données relatives aux activités de maintenance comportant chacune des informations intéressantes. Il est important de savoir comment faire bon usage de ces données brutes.

Certaines bases de données contiennent parfois des informations mêlées, erronées ou incomplètes. C'est par exemple le cas de la base de données SIPROD. Le tableau 3.3 présente un extrait de SIPROD sur des interventions du composant générique (ou sous-système) stator du système alternateur du GTA. On remarque que les enregistrements stator sont parfois bien répertoriés, parfois mal répertoriés et parfois non répertoriés. En effet, le tableau 3.3 montre quelques enregistrements stator répertoriés comme palier ou alternateur hydraulique, et quelques enregistrements stator non attribués (NA). Ces données mêlées et incomplètes peuvent nuire considérablement aux analyses.

TABLEAU 3.3: EXTRAIT DE SIPROD PRÉSENTANT UN PROBLÈME D'INTÉGRITÉ DES DONNÉES

Bon de travail	Description du bon de travail	Date	Composant générique
84071483	resserrer tiges de serrage stator 90 % loose	2014-08-18	ALTERNATEUR HYDRAULIQUE
83643092	A22 Alarme Haute température encoche stator T9 154 degrés	2012-11-23	ALTERNATEUR HYDRAULIQUE
81177044	Alarme de haute température stator (enreg)	2005-11-10	ALTERNATEUR HYDRAULIQUE
84009078	Cales manquantes dans le haut du stator côté Est à 3HR	2014-01-17	ALTERNATEUR HYDRAULIQUE
80982595	MEC RÉPARATION STATOR	2004-10-18	STATOR
81118809	Installer lexan pour permettre de monter le stator	2005-04-27	STATOR
82367015	Faire retouche de peinture cordage stator gr # 1	2007-12-07	STATOR
82740004	A1: alarme de haute temp. stator (A01st1249)	2009-04-16	STATOR
83540522	point temp. stator1 en alarme tandis que les autres points sont normaux	2012-04-12	STATOR
83541578	alt#1:sonde #1 temp.stator indique 20 degrés de plus que les autres sondes	2012-04-16	STATOR
83803337	Poussière de ciment sur bobinage du stator gr-13	2013-02-04	(NA)
H0097022	RP ALARME TEMPÉRATURE STATOR ALTERNATEUR	1993-07-29	(NA)
H0097023	RP DECL. ALT. 15 (EAU STATOR PAR SYST I	1995-03-10	(NA)
80799989	ELE3-Alt.17 al. intermittente sonde temp.stator#10	2003-12-08	(NA)
H0097105	RP BOBINE STATOR ENDOMMAGE	1998-01-30	(NA)
83193770	Fuite d'Eau au stator	2011-02-01	PALIER
84492278	Backwash Radiateur stator	2017-03-09	PALIER

3.4 Démarche

Cette section dévoile la démarche suivie pour le développement de l'algorithme de traitement des données qui fournira des données exploitables pour l'estimation de la durée de vie d'un ou de plusieurs composants. La figure 3.2 présente la méthodologie de la recherche à travers les différentes étapes parcourues allant de l'étape 1 à l'étape 15. La méthodologie peut se scinder en quatre principaux groupes d'étapes : le pairage des bases de données, la classification automatique de texte, la combinaison du pairage et de la classification automatique, puis l'estimation des DVR.

3.4.1 Pairage des bases de données

3.4.1.1 *Étape 1 : Choix des bases de données*

SIPROD, CEGTA, Diag et travaux majeurs comportent toutes des informations historiques des actions de maintenance effectuées sur les équipements du GTA, avec chacune leur particularité comme discuté dans la section 3.2. Le choix des bases de données a été fait avec minutie sur la base de plusieurs critères. Cette étape est fondamentale; elle représente la fondation du projet dont les étapes subséquentes en dépendent.

3.4.1.2 *Étape 2 : Choix d'une centrale*

HQ possède de nombreuses centrales. Cependant, l'étude initiale nécessite un candidat expérimental pour l'essai initial de la méthode. Le choix a été fait en fonction de la quantité et de la qualité des données disponibles. Dès lors, la centrale la plus récente et comportant le plus grand nombre de groupes a été privilégiée.

3.4.2 Classification automatique de texte

3.4.2.1 *Étape 5 : Fractionnement en sous-systèmes*

Chaque système d'un GTA est constitué de plusieurs composants génériques ou sous-systèmes. La répartition d'un système en différentes classes de sous-systèmes qui le constituent est importante pour la classification automatique. Par ailleurs, il est commun de subdiviser un système en sous-systèmes pour des analyses de fiabilité.

3.4.2.2 *Étape 6 : Classification automatique des sous-systèmes*

Comme évoqué plus haut, la base de données SIPROD qui contient la description des BT présente quelques erreurs et manques au niveau de la catégorisation des sous-systèmes affiliés aux BT.

Vérifier manuellement et corriger l'attribut « composant générique » de toutes les lignes d'une base de données comprenant des milliers d'instances s'avère très fastidieux et consomme beaucoup de temps. L'être humain, pour vérifier et corriger la classe attribuée à un BT, se base sur la lecture de la description du BT et des informations qu'elle contient. De la même manière, un algorithme de classification automatique effectue une fouille de texte dans la description du BT pour classer automatiquement les sous-systèmes. Ainsi, un algorithme robuste serait capable de corriger automatiquement les erreurs humaines et de compléter les cases manquantes avec un certain degré de précision.

L'algorithme se base sur un apprentissage supervisé prenant en entrée les textes contenant la description des BT, ainsi que les étiquettes de classes sur lequel l'algorithme se base pour apprendre. Un jeu de données constitué de catégorisations fiables de sous-systèmes associées aux descriptions de BT a été monté sur la centrale choisie initialement.

3.4.2.3 *Étape 7 : Choix d'un sous-système*

Tout comme pour le choix d'un système, l'idée est de restreindre la portée de l'étude. Le choix a été fait selon le niveau d'impact du composant sur le système choisi précédemment, et des avis d'experts.

3.4.2.4 *Étape 8 : Classification automatique des modes de défaillance*

Contrairement aux sous-systèmes, aucune des quatre bases de données mises à disposition pour le projet ne comporte l'attribut « mode de défaillance » directement. Cette information pourtant très importante permet de mettre en exergue l'effet par lequel une défaillance est observée sur le composant générique.

L'algorithme de classification automatique des modes de défaillance se charge de créer l'attribut « mode de défaillance » dans la base de données. Tout comme pour la classification automatique des sous-systèmes, il s'agit d'un apprentissage supervisé prenant en entrées les descriptions des BT et les étiquettes de classes de modes de défaillances. Un jeu de données constitué de catégorisations fiables des modes de défaillance du sous-système choisi associées aux descriptions de BT a été monté sur la centrale choisie initialement.

3.4.3 Combinaison du pairage et de la classification automatique

3.4.3.1 *Étape 9 : Association du pairage des bases de données et de la classification automatique*

Le pairage des bases de données permet d'améliorer la qualité de l'information en permettant aux bases de données choisies de se compléter mutuellement.

La classification automatique permet de corriger les erreurs, compléter les données manquantes et ajouter des informations de valeur (les modes de défaillance) dans la base de données contenant la description des BT. L'association du pairage et de la classification automatique renforce d'autant plus la qualité de l'information.

3.4.3.2 Étape 10 : Choix du mode de défaillance pertinent

La présente étude se focalise sur le mode de défaillance le plus critique et pertinent du composant générique choisi. Le soin de ce choix a été laissé aux experts des équipements du GTA qui se sont focalisés sur le mode de défaillance le plus représentatif du vieillissement de l'actif.

3.4.3.3 Étape 11 : Extension de l'algorithme à d'autres centrales

Une fois l'algorithme bâti sur une centrale, il est nécessaire de l'étendre à d'autres centrales et d'en apprécier les résultats obtenus.

3.4.3.4 Étape 12 : Extraction des données + validation

L'algorithme extrait toutes les instances du composant générique choisi relatives au mode de défaillance pertinent sélectionné.

Ensuite, une validation de l'algorithme est nécessaire :

- La capacité de l'algorithme à généraliser sur les autres centrales a été observée.
- Les données extraites par l'algorithme ont été fournies à une équipe d'experts pour validation en fonction des archives et de leur expertise.

3.4.4 Estimation des DVR

3.4.4.1 *Étape 13 : Choix d'une méthode de prédiction de durée de vie*

La revue de la littérature a relaté de nombreuses méthodologies de prédiction des durées de vie. Cela a permis d'en choisir une simple et applicable au contexte de ce projet.

3.4.4.2 *Étape 14 : Estimation des durées de vie pour chaque centrale*

La méthode choisie à l'étape 13 a été utilisée pour estimer les durées de vie. Les données traitées et extraites par l'algorithme ont permis d'alimenter la méthode choisie.

Il est important de préciser le contexte et d'émettre certaines hypothèses durant cette étape. Ce projet vise à estimer les fins de vie des équipements, et non de prédire les prochaines défaillances. En clair, l'étude se concentre sur le vieillissement par dégradation. Pour ce faire, quelques hypothèses ont été émises :

- Seuls les événements d'arrêt non planifié pour cause de dégradation sont considérés. Les arrêts non planifiés pour cause externe et les arrêts planifiés sont écartés.
- Un arrêt non planifié de durée importante équivaut à une défaillance critique
- La maintenance du composant générique suite à une défaillance critique est considérée comme un remplacement.

3.4.4.3 *Étape 15 : Test d'homogénéité des données des centrales*

L'algorithme extrait des données traitées pour chaque centrale. Ensuite, ces données sont utilisées pour estimer la durée de vie du composant générique selon la centrale. Les GTA de chaque centrale sont soumis à diverses conditions d'exploitation, divers plans de maintenance et fonctionnent dans des environnements différents.

De plus, les fabricants des composants et les technologies installées peuvent varier en fonction des centrales et des GTA. Compte tenu de ces subtilités, les données de chaque centrale sont considérées comme des populations différentes. La statistique propose un test d'homogénéité entre des populations pour savoir avec quel degré de confiance il est possible de les mélanger. Selon le résultat du test, la décision autorise soit à mélanger les données des centrales et faire une estimation globale de la durée de vie, soit à laisser les durées de vie individuellement par centrale.

3.5 Conclusion

Ce chapitre a relaté la méthodologie de cette recherche en détaillant clairement les différentes étapes à parcourir pour l'atteinte des objectifs. Toutes ces étapes peuvent être segmentées en quatre principaux groupes d'étapes, en l'occurrence le pairage des bases de données, la classification automatique de texte, la combinaison du pairage et de la classification automatique, puis l'estimation des DVR. À noter que les trois premiers groupes d'étapes constituent l'algorithme de traitement des données. Cette méthodologie permettra d'améliorer l'intégrité des données de HQP nécessaire pour une estimation robuste des fins de vie des composants des GTA.

CHAPITRE 4 - RÉSULTATS

Ce quatrième chapitre présente en détail le développement de l'algorithme de traitement des données et l'estimation de la DVR d'un composant de GTA.

4.1 Algorithme de traitement des données

L'algorithme est constitué de trois principaux groupes d'étapes : le pairage des bases de données, la classification de texte et la combinaison du pairage et de la classification.

4.1.1 Pairage des bases de données

4.1.1.1 *Étape 1 : Choix des bases de données*

Les bases de données ont été choisies dans le but d'estimer la DVR des composants des GTA. Comme discuté dans la section 3.4, le contexte de fin de vie des équipements a été précisé et des hypothèses ont été émises. Selon ces hypothèses, avoir une idée claire de la durée de chaque arrêt est nécessaire, tout comme le type d'arrêt (arrêts non planifiés uniquement). De plus, la considération unique des arrêts pour dégradation a été précisée. Le tableau 3.1 montre que la seule base de données réunissant les informations de durée et d'état de l'arrêt est la base de données CEGTA. CEGTA est également la base de données possédant des informations entièrement validées. Cependant, selon les hypothèses, il faudrait filtrer les interventions pour cause de dégradation et CEGTA ne dispose pas de cette information, sauf exception dans quelques cas rares où des remarques sur l'intervention ont été ajoutées dans les faits saillants. SIPROD apparaît comme une option convaincante, car elle possède tous les BT effectués sur les GTA et leurs descriptions. Les descriptions de BT dans SIPROD soutenues par les faits saillants (rarement déclarés) de CEGTA permettent de discriminer un effet de dégradation. SIPROD et CEGTA se montrent ainsi les meilleurs candidats pour la réalisation de l'algorithme.

4.1.1.2 Étape 2 : Choix d'une centrale

LG2 est la centrale la plus récente comportant un grand nombre (16) de GTA. La récence joue un rôle important compte tenu de l'évolution des systèmes des technologies de l'information vis-à-vis de la qualité des données. Un grand nombre de GTA témoigne l'envergure de la centrale et ainsi la quantité de données disponibles. LG2-A est une centrale considérée comme une extension de LG2 qui comporte 6 GTA. Ainsi, l'association des centrales LG2 et LG2A a été choisie comme candidat expérimental pour un total de 22 GTA.

4.1.1.3 Étape 3 : Choix d'un système

La figure 4.1 illustre la distribution du nombre d'arrêts et de la durée totale des arrêts de chaque système. Les données de cette analyse proviennent de CEGTA. On remarque que les deux systèmes les plus importants du GTA, notamment la turbine et l'alternateur se démarquent du lot. Le choix final a été porté à l'alternateur après entente collective de l'équipe de recherche HQP.

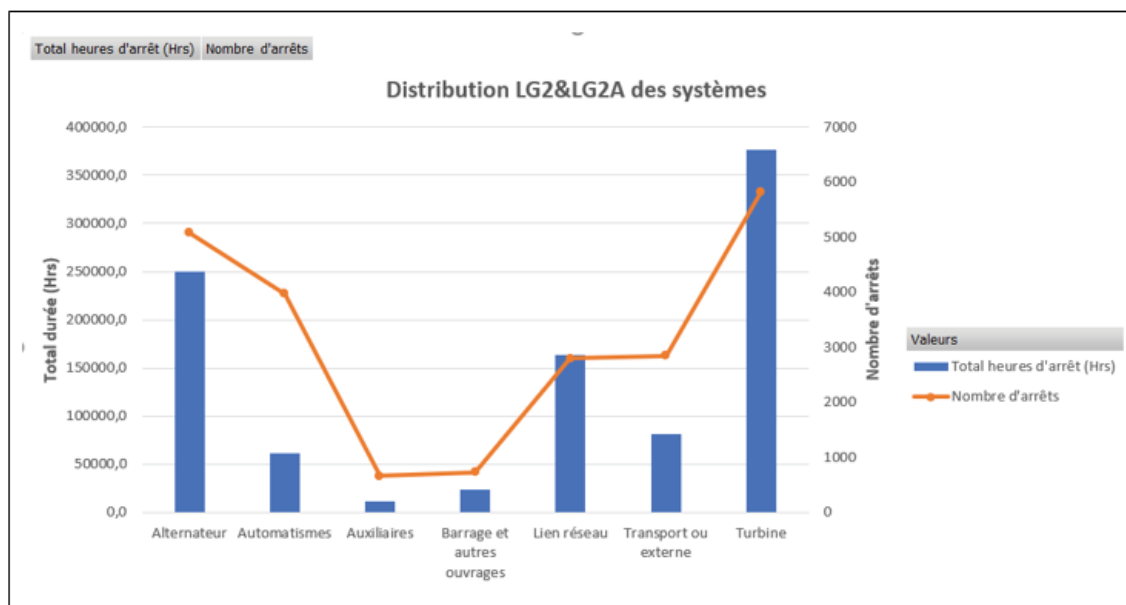


Figure 4.1 : Représentativité des données par système

4.1.1.4 Étape 4 : Pairage de SIPROD et CEGTA

Il est important de réduire la portée des bases de données pour se focaliser sur l'essentiel et limiter le bruit. Le système alternateur est filtré sur les deux bases de données, ainsi que le numéro du GTA (dit groupe) sélectionné. Ensuite dans SIPROD, seuls les types de travail conditionnel et correctif sont filtrés. Dans CEGTA, seuls certains états d'arrêts planifiés précis (entretien planifié long terme et entretien planifié en temps réel) et d'arrêts non planifiés précis (déclenchement en charge, refus de démarrage, arrêt par l'opérateur) sont filtrés.

Le numéro du BT représente la clé primaire de la base de données SIPROD, car cet attribut permet d'identifier de façon unique chaque enregistrement. Dès lors, la façon la plus simple d'associer SIPROD et CEGTA serait sur la base du numéro de BT. Cependant, CEGTA ne possède pas cette colonne. Une solution alternative a été optée sur la base temporelle.

CEGTA contient des informations telles que la date de début et la date de fin de l'évènement, et donc sa durée. SIPROD ne possède que l'information de la date de création du bon de travail. Ainsi, un bon de travail créé durant la plage de temps de l'évènement CEGTA lui est associé. De nombreux BT peuvent être ouverts pour un même évènement d'arrêt de CEGTA. Un évènement CEGTA pourrait donc englober de nombreux enregistrements SIPROD. La figure 4.2 illustre le principe d'association sur une base temporelle de SIPROD et CEGTA.

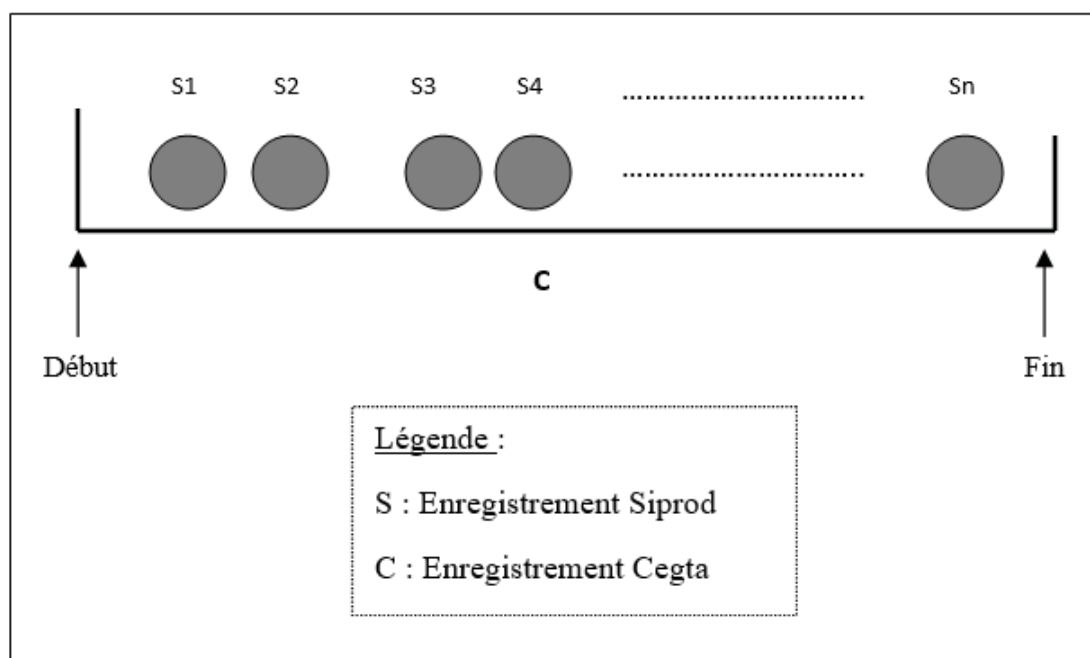


Figure 4.2: Principe d'association temporelle de SIPROD et CEGTA

En situation réelle, il arrive souvent chez HQP qu'un BT soit ouvert quelque temps avant ou après l'évènement CEGTA. Pour gérer ces cas de figure, des ajustements d'intervalles de temps (48 h) ont été faits pour maximiser les chances de pairage.

Parfois, deux évènements CEGTA sont très rapprochés, par exemple, ils ont eu lieu le même jour.

Pour discriminer correctement la meilleure destination pour l'enregistrement SIPROD, une fouille de texte est réalisée dans les champs d'informations de SIPROD (description de BT et composant générique) et de CEGTA (Composant ACE et faits saillants) pour un mappage. La priorité est donnée à l'évènement CEGTA ayant le plus de similarité avec l'enregistrement SIPROD.

Malgré toutes les astuces de correspondance discutées précédemment, il existerait toujours quelques enregistrements SIPROD et CEGTA non pairés. SIPROD dispose d'un attribut très important; les coûts associés au BT. Ainsi, fixer un seuil de dépenses au-dessus duquel chaque enregistrement devrait absolument avoir un équivalent CEGTA se montre intéressant. Ainsi, toutes les instances SIPROD dépassant un certain seuil de coûts (par exemple 10 000 \$) sont pairées à l'évènement CEGTA le plus proche ayant la meilleure similarité de mappage de fouille de texte dans un intervalle de temps fixé (par exemple 2 ou 4 semaines). La figure 4.3 décrit l'algorithme de pairage de SIPROD et CEGTA.

Un outil de graphique dynamique a été développé pour visualiser la juxtaposition temporelle des enregistrements des deux bases de données. Le graphique est dynamique, c'est-à-dire qu'il y a possibilité d'interaction, comme rogner le graphique pour se focaliser sur une plage de temps précise, ou encore placer le curseur sur un point et avoir ses informations pertinentes dans une bulle. La figure 4.4 illustre la juxtaposition temporelle des enregistrements SIPROD et CEGTA du groupe 10 de la centrale LG2-LG2A entre octobre 1997 et octobre 1999. En légende, SIPROD_25 pairé indique qu'il s'agit d'un point SIPROD ayant été pairé à un point CEGTA dont l'état d'arrêt est planifié long terme. Conjointement, CEGTA_25 indique qu'il s'agit d'un point CEGTA dont l'état d'arrêt est planifié long terme. Cette distinction SIPROD_25 et CEGTA_25 a été faite au code couleur pour être en mesure de repérer visuellement les points d'état d'arrêts non planifiés.

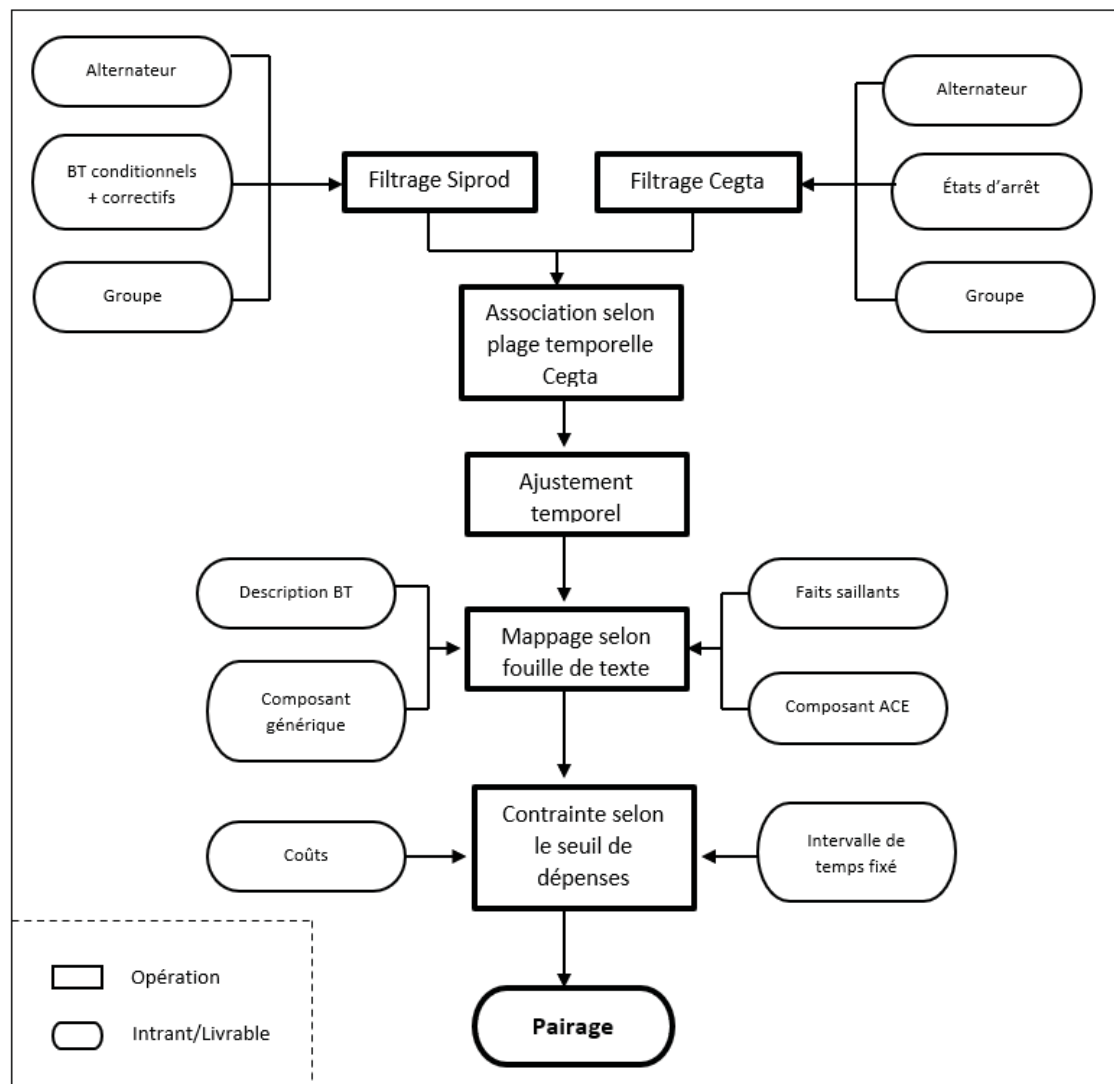


Figure 4.3: Diagramme de flux de l'algorithme de pairage

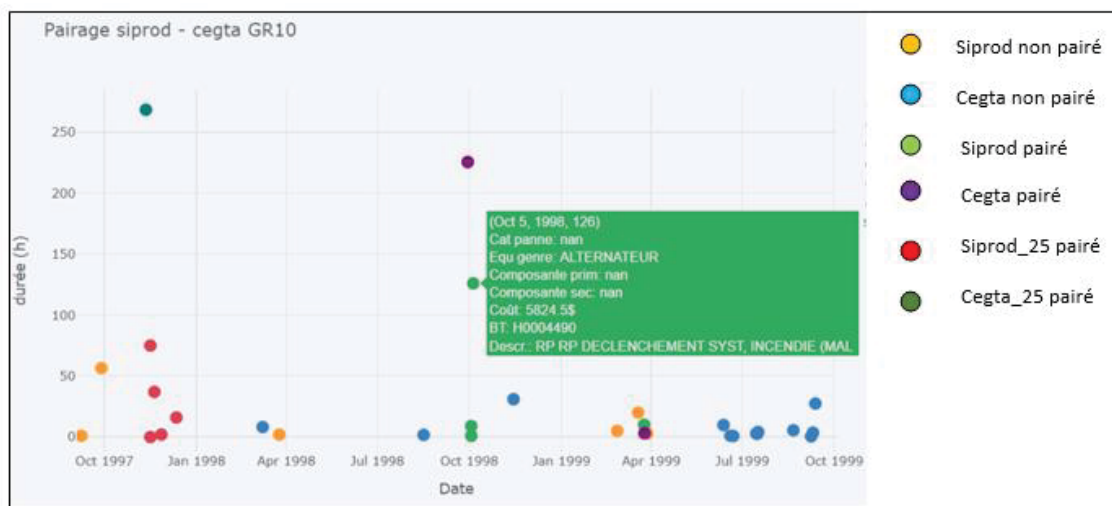


Figure 4.4: Outil de visualisation du pairage SIPROD-CEGTA

Le tableau 4.1 présente les résultats du pairage de SIPROD et CEGTA pour des données allant de 1980 à 2020 de la centrale LG2-LG2A. L'ajustement temporel a été fixé à 2 jours avant et après l'évènement CEGTA. Le seuil de coûts a été fixé à 10 000 \$ et la fenêtre de temps maximum de proximité à 5 semaines. Le seuil de coûts a été recommandé par les experts selon l'hypothèse que les dépenses affectées à un BT à la hauteur de 10 000\$ et plus signifient que le travail effectué est suffisamment important. La fenêtre de temps de 5 semaines a également été recommandée par les experts. Ayant filtré uniquement les entretiens en temps réel, les entretiens planifiés long terme et les défaillances, la fréquence d'occurrence de ces évènements est en moyenne largement supérieure à 1 mois. Dès lors, 5 semaines (environ 1 mois) représentent une plage temporelle permettant de maximiser les chances de pairage des BT aux coûts importants tout en restant suffisamment conservateur.

TABLEAU 4.1: RÉSULTATS DU PAIRAGE SIPROD-CEGTA DE LG2-LG2A

Centrale	Groupe	Total SIPROD	SIPROD pairés	Ratio SIPROD	Total CEGTA	CEGTA pairés	Ratio CEGTA
LG2	1	325	187	58 %	211	82	39 %
LG2	2	336	212	63 %	267	103	39 %
LG2	3	388	271	70 %	285	119	42 %
LG2	4	375	224	60 %	273	103	38 %
LG2	5	348	180	52 %	222	86	39 %
LG2	6	369	208	56 %	277	101	37 %
LG2	7	292	158	54 %	196	82	42 %
LG2	8	374	231	62 %	277	96	35 %
LG2	9	380	245	65 %	248	112	45 %
LG2	10	337	181	54 %	190	77	41 %
LG2	11	383	228	60 %	275	112	41 %
LG2	12	318	191	60 %	214	85	40 %
LG2	13	414	284	69 %	282	126	45 %
LG2	14	300	172	57 %	200	75	38 %
LG2	15	308	189	61 %	238	81	34 %
LG2	16	317	207	65 %	188	94	50 %
LG2-A	21	198	108	55 %	170	67	39 %
LG2-A	22	144	83	58 %	127	46	36 %
LG2-A	23	148	81	55 %	146	52	36 %
LG2-A	24	165	82	50 %	124	45	36 %
LG2-A	25	175	113	65 %	163	68	42 %
LG2-A	26	185	87	47 %	115	41	36 %

Les résultats du pairage montrent que certains enregistrements SIPROD n'ont pas trouvé de correspondance CEGTA, et vice versa. Cela s'explique par le fait que :

- Certains évènements CEGTA sont ouverts puis fermés sans intervention nécessitant l'ouverture d'un BT.
- Certains BT SIPROD sont créés longtemps avant ou après l'évènement CEGTA.
- Certains problèmes inhérents à l'alternateur peuvent avoir des incidences sur d'autres systèmes tels que les automatismes ou la turbine. Ainsi, certains évènements CEGTA classés sous des systèmes autres que l'alternateur correspondent en réalité à certains BT relatifs à l'alternateur. C'est d'autant plus le cas lors des arrêts planifiés

long terme où divers travaux sont effectués sur plusieurs systèmes en parallèle. L'évènement CEGTA créé à cet effet portera le nom d'un seul système (par exemple turbine, automatismes ou alternateur). L'algorithme filtre CEGTA au système alternateur uniquement, dans le but de limiter le bruit.

4.1.2 Classification de texte

4.1.2.1 *Étapes 5 & 6 : Fractionnement en sous-systèmes et classification*

Le système alternateur a été subdivisé en 10 différentes classes : sectionneurs-disjoncteurs, palier & injection, barres blindées, rotor, protection incendie, refroidissement général, excitation, stator, freinage et général. Chaque classe correspond à un sous-système de l'alternateur à l'exception de la classe « refroidissement général » et de la classe « générale ». La classe refroidissement général a été créée pour des descriptions de BT représentant un refroidissement sans précision sur le composant générique affecté. En effet, la plupart des composants génériques possèdent un système de refroidissement. Par exemple, si la description du BT est « radiateur stator brisé », il s'agit d'un problème de refroidissement du stator, car le composant générique stator est précisé. Par contre si la description est « radiateur brisé », il n'y a aucune indication sur le sous-système y afférent, dès lors, elle est catégorisée comme un refroidissement général. Dans la même logique, la classe « générale » comporte des descriptions de BT vagues et générales ne permettant pas de situer le sous-système affecté. Exemple : « Refus de démarrage du groupe ». Dans cet exemple, il est impossible de savoir quel est le sous-système en cause du refus de démarrage du groupe.

Un jeu de données a été constitué sur la base des 10 classes définies précédemment et les descriptions de BT associées. L'attribution des classes à chaque BT a été faite manuellement pour s'assurer de la fiabilité des données nécessaires pour l'entraînement du modèle d'apprentissage automatique.

Le jeu de données a été monté sur LG2-LG2A seulement, comportant 2642 lignes. La figure 4.5 illustre la répartition des classes du système alternateur.

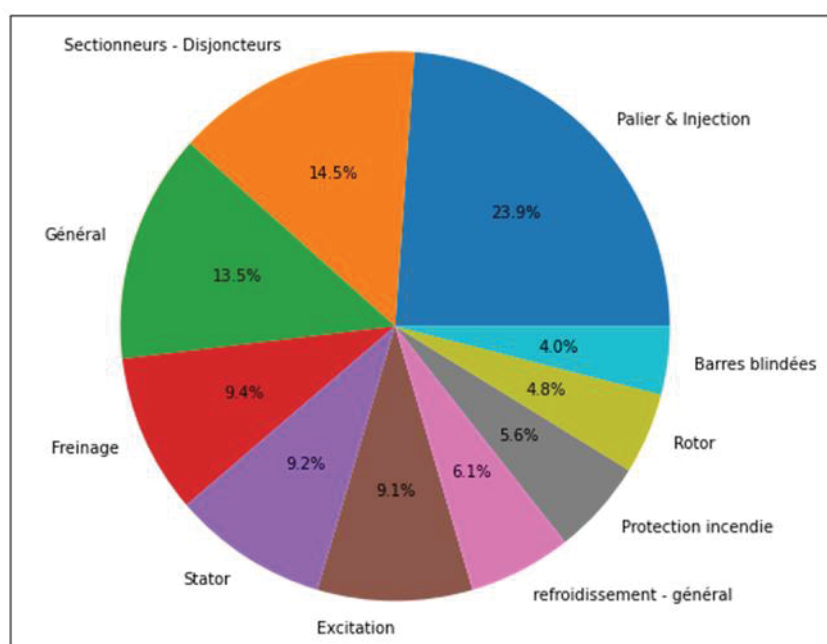


Figure 4.5: Répartition des classes de sous-systèmes alternateur

La classification de texte a été abordée dans la section 2.5.2 du chapitre 2 et illustrée dans la figure 2.6. La première étape d'un projet de classification de texte concerne le prétraitement des données. Dans cette application de classification de sous-systèmes, la phase de prétraitement consiste au nettoyage du texte, suivie d'une correction des fautes d'orthographe puis de la tokénisation, du retrait des mots communs et du *stemming*. Ces opérations ont été réalisées exactement comme décrites dans la section 2.5.2.

La deuxième étape consiste à représenter le texte de manière numérique compréhensible par l'ordinateur. Les deux principales méthodes de représentation de texte sont les BOW et les *Word embeddings*.

Géron (2019) représente une excellente référence pour une introduction solide à l'univers du ML. De nombreux algorithmes y sont notamment présentés tout en expliquant leurs concepts mathématiques et les notions d'optimisation.

TABLEAU 4.2: PARAMÈTRES OPTIMISÉS DU BOW POUR CHAQUE ALGORITHME

Algorithme	SGDC	Random Forest	Régression logistique
Fréquence minimale dans le corpus	1	3	1
N-gram	1	1	1
TF ou TF-IDF	TF	TF	TF

Le tableau 4.2 présente les paramètres optimisés du BOW des modèles proposant les meilleurs scores. Les paramètres sont détaillés comme suit :

- Fréquence minimale dans le corpus : fréquence d'apparition minimale d'un mot dans le corpus (ensemble de tous les textes) pour être pris en considération dans le BOW.
- N-gram : mot unique si $n=1$ ou groupe de mots si $n>1$ considérés dans le BOW.
- TF ou TF-IDF : l'algorithme obtient un meilleur score avec TF ou avec TF-IDF?

TABLEAU 4.3: RÉSULTATS DE LA CLASSIFICATION DES SOUS-SYSTÈMES

Métrique	Précision	Sensibilité	F1-score
SGDC	92 %	92 %	92 %
Random Forest	91 %	90 %	90 %
Régression logistique	93 %	92 %	92 %

Le tableau 4.3 présente les résultats des algorithmes (après optimisation des hyperparamètres) sur le jeu de test. Les métriques reportées représentent la moyenne globale des scores obtenus sur chacune des dix classes. Le modèle de régression logistique arbore les meilleurs résultats.

4.1.2.2 Étape 7 : Choix d'un sous-système

Le sous-système choisi pour la suite de l'étude par l'équipe est le stator, compte tenu de sa pertinence pour le système alternateur. En effet, le stator et le rotor représentent l'induit et l'inducteur directement responsables de la création de l'énergie électrique. Le stator a été priorisé pour l'étude sur avis des experts.

4.1.2.3 Étape 8 : Classification des modes de défaillance

Le sous-système stator a été subdivisé en trois classes de modes de défaillance : court-circuit, refroidissement et général. Tout comme pour la classification des sous-systèmes, un jeu de données a été monté sur le vocabulaire relatif à chaque classe. La procédure suivie pour la classification des sous-systèmes a été répétée. Le tableau 4.4 reporte les résultats obtenus sur le jeu de test des algorithmes (après optimisation des hyperparamètres). Dans ce cas, l'algorithme de *Random Forest* s'en sort le mieux.

TABLEAU 4.4: RÉSULTATS DE LA CLASSIFICATION DES MODES DE DÉFAILLANCE

Métrique	Précision	Sensibilité	F1-score
SGDC	97 %	97 %	97 %
Random Forest	100 %	100 %	100 %
Régression logistique	97 %	97 %	97 %

4.1.3 Combinaison du pairage et de la classification de texte:

4.1.3.1 *Étape 9 : Association du pairage et de la classification automatique*

Il est question ici d'intégrer les modèles de classification des sous-systèmes et des modes de défaillances à la base de données résultante de l'association SIPROD-CEGTA. Le sous-système affilié à la description BT de chaque enregistrement SIPROD pairé à un évènement CEGTA est prédit automatiquement. Ensuite, vu que l'étude se focalise sur le stator, le mode de défaillance est prédit uniquement si le sous-système est prédit comme stator. Ainsi, pour chaque évènement CEGTA pairé, il est possible d'identifier les BT associés, les sous-systèmes à l'origine de l'intervention, et même les modes de défaillance relatifs (uniquement pour le stator dans le cadre de cette étude). Étant donné que l'attribut « faits saillants » comporte parfois une description de l'évènement CEGTA, cette information pourrait être exploitée. Si le composant ACE de l'évènement CEGTA est relatif au stator, alors le modèle de classification des modes de défaillance du stator est passé sur la description des faits saillants. La table résultante est présentée dans le tableau 4.5. Ensuite, un filtrage permet de conserver uniquement tous les évènements CEGTA pairés relatifs au stator. Le processus de filtrage est résumé dans la figure 4.7.

TABLEAU 4.5: TABLE RÉSULTANTE DE LA COMBINAISON DU PAIRAGE ET DE LA CLASSIFICATION DE TEXTE

Attributs SIPROD	SIPROD 1	SIPROD 2		SIPROD n
Attributs CEGTA	CEGTA	CEGTA		CEGTA
Prédiction sous-système (SIPROD)	SIPROD 1	SIPROD 2		SIPROD n
Prédiction mode de défaillance si stator (SIPROD)	SIPROD 1	SIPROD 2		SIPROD n
Prédiction mode de défaillance si composant ACE relatif au stator (CEGTA)	CEGTA	CEGTA		CEGTA

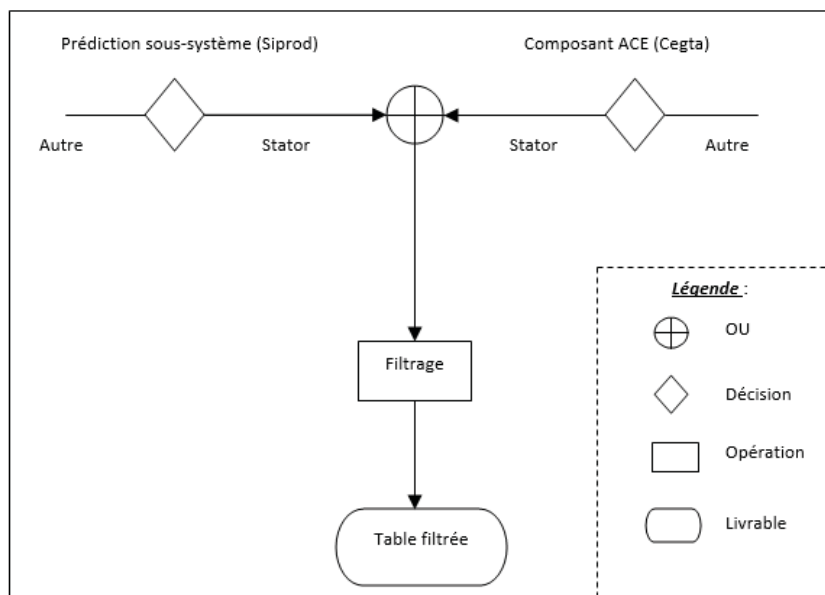


Figure 4.7: Logigramme de filtrage des données stator

4.1.3.2 Étape 10 : Choix du mode de défaillance pertinent

Le mode de défaillance le plus pertinent au sous-système stator est le court-circuit (CC). Le CC est un phénomène qui se produit lorsque deux fils électriques entrent en contact direct, le plus souvent suite à un défaut d'isolation dû à l'usure ou à un facteur externe tel que l'eau. Le CC se traduit par une augmentation brusque du courant pouvant aller jusqu'à provoquer un incendie. La suite de l'étude sera donc consacrée à des données de court-circuit du stator.

4.1.3.3 Étape 11 : Extension de l'algorithme à d'autres centrales

Une fois l'algorithme monté sur la base de la centrale LG2-LG2A, il permet d'apprécier les résultats sur d'autres centrales. Les centrales choisies par l'équipe HQP sont : LG1, LG4, Manic5-Manic5A, Carillon et Beauharnois.

Dans un premier temps la capacité de pairage SIPROD-CEGTA a été observée sur ces différentes centrales. Les tableaux 4.6 à 4.8 reportent les résultats obtenus.

TABLEAU 4.6: RÉSULTATS DU PAIRAGE SUR LES AUTRES CENTRALES

Centrale	Groupe	Total SIPROD	SIPROD païrés	Ratio SIPROD	Total CEGTA	CEGTA païrés	Ratio CEGTA
LG1	1	136	96	71 %	157	55	35 %
LG1	2	106	72	68 %	113	40	35 %
LG1	3	89	59	66 %	113	35	31 %
LG1	4	120	102	85 %	170	67	39 %
LG1	5	85	62	73 %	120	39	33 %
LG1	6	99	71	72 %	144	44	29 %
LG1	7	104	60	58 %	120	36	30 %
LG1	8	101	67	66 %	116	42	36 %
LG1	9	99	72	73 %	136	45	33 %
LG1	10	63	45	71 %	109	30	28 %
LG1	11	84	70	83 %	112	45	40 %
LG1	12	89	66	74 %	82	33	40 %
LG4	1	214	149	70 %	212	75	35 %
LG4	2	176	101	57 %	203	62	31 %
LG4	3	164	115	70 %	220	59	27 %
LG4	4	172	126	73 %	251	63	25 %
LG4	5	134	94	70 %	225	53	24 %
LG4	6	176	106	60 %	208	69	33 %
LG4	7	149	113	76 %	196	66	34 %
LG4	8	177	134	76 %	217	73	34 %
LG4	9	205	161	79 %	243	86	35 %

TABLEAU 4.7: RÉSULTATS DU PAIRAGE SUR LES AUTRES CENTRALES (SUITE 1)

Centrale	Groupe	Total SIPROD	SIPROD païrés	Ratio SIPROD	Total CEGTA	CEGTA païrés	Ratio CEGTA
Manic-5	51	99	65	66 %	160	38	24 %
Manic-5	52	93	58	62 %	122	32	26 %
Manic-5	53	81	32	40 %	139	17	12 %
Manic-5	54	69	37	54 %	161	28	17 %
Manic-5	55	94	62	66 %	168	36	21 %
Manic-5	56	131	68	52 %	164	38	23 %
Manic-5	57	82	54	66 %	157	36	23 %
Manic-5	58	112	68	61 %	178	47	26 %
Manic-5A	41	122	74	61 %	162	30	19 %
Manic-5A	42	133	91	68 %	162	38	24 %
Manic-5A	43	126	66	52 %	163	34	21 %
Manic-5A	44	149	101	68 %	178	51	29 %
Carillon	1	66	30	46 %	135	19	14 %
Carillon	2	29	74	39 %	103	18	18 %
Carillon	3	29	67	43 %	21	120	18 %
Carillon	4	35	62	57 %	21	140	15 %
Carillon	5	64	19	30 %	147	14	10 %
Carillon	6	50	19	38 %	127	10	8 %
Carillon	7	77	32	42 %	132	20	15 %
Carillon	8	76	46	61 %	144	31	22 %
Carillon	9	70	30	43 %	136	19	14 %
Carillon	10	112	74	66 %	135	20	15 %
Carillon	11	66	36	55 %	151	27	18 %
Carillon	12	69	32	46 %	150	21	14 %
Carillon	13	109	44	40 %	158	30	19 %
Carillon	14	57	19	33 %	153	15	10 %

TABLEAU 4.8: RÉSULTATS DU PAIRAGE SUR LES AUTRES CENTRALES (SUITE 2)

Centrale	Groupe	Total SIPROD	SIPROD païrés	Ratio SIPROD	Total CEGTA	CEGTA païrés	Ratio CEGTA
Beauharnois	1	61	144	42 %	98	24	25 %
Beauharnois	2	112	54	48 %	111	26	23 %
Beauharnois	3	73	23	32 %	88	14	16 %
Beauharnois	4	75	29	39 %	86	19	22 %
Beauharnois	5	63	24	38 %	100	16	16 %
Beauharnois	6	55	18	33 %	80	10	13 %
Beauharnois	7	88	33	38 %	117	15	13 %
Beauharnois	8	79	36	46 %	104	24	23 %
Beauharnois	9	111	54	49 %	104	26	25 %
Beauharnois	10	46	20	44 %	104	14	14 %
Beauharnois	11	149	80	54 %	124	33	27 %
Beauharnois	12	61	25	41 %	87	9	10 %
Beauharnois	13	87	34	39 %	95	19	20 %
Beauharnois	14	58	25	43 %	78	14	18 %
Beauharnois	15	133	61	46 %	116	31	27 %
Beauharnois	16	57	25	44 %	86	14	16 %
Beauharnois	17	80	34	43 %	87	21	24 %
Beauharnois	18	76	34	45 %	103	24	23 %
Beauharnois	19	85	38	45 %	92	18	20 %
Beauharnois	20	46	17	37 %	104	10	10 %
Beauharnois	21	88	31	35 %	97	17	18 %
Beauharnois	22	89	45	51 %	100	24	24 %
Beauharnois	23	89	32	36 %	73	18	25 %
Beauharnois	24	131	63	48 %	82	28	34 %
Beauharnois	25	61	26	43 %	62	15	24 %
Beauharnois	26	66	34	52 %	79	16	20 %
Beauharnois	27	88	24	27 %	81	11	14 %
Beauharnois	28	103	41	40 %	74	16	22 %
Beauharnois	29	128	81	63 %	91	26	29 %
Beauharnois	30	104	47	45 %	88	25	28 %
Beauharnois	31	63	28	44 %	78	13	17 %
Beauharnois	32	69	33	48 %	87	18	21 %
Beauharnois	33	111	40	36 %	80	17	21 %
Beauharnois	34	86	55	64 %	112	24	21 %
Beauharnois	35	71	28	39 %	81	19	24 %
Beauharnois	36	79	32	41 %	79	19	24 %

Ensuite, la capacité de généralisation à d'autres centrales des modèles de classification entraînés sur LG2 a été évaluée. Pour ce faire, une comparaison a été faite entre les prédictions et les classes attribuées manuellement après lecture individuelle des BT. L'exercice a été effectué sur l'ensemble des données des centrales LG1, LG4 et Manic-5. Le tableau 4.9 présente les résultats obtenus qui sont très bons et témoignent de la capacité à généraliser des modèles de classification.

TABLEAU 4.9: CAPACITÉ DE GÉNÉRALISATION DU MODÈLE DE CLASSIFICATION SUR D'AUTRES CENTRALES

	Nombre de BT alternateur	Prédictions correctes	Prédictions manquées	Précision
Sous-systèmes	1240	1230	10	99.2 %
Modes de défaillance du stator	180	175	5	97.2 %

4.1.3.4 Étape 12 : Extraction des données et validation

L'algorithme extrait par suite du filtrage des données stator les événements CEGTA pairés dont au moins un des BT associés est inhérent au court-circuit stator. De plus, si la description des faits saillants de CEGTA est présente et que l'attribut « composant ACE » se révèle un composant du stator, alors l'algorithme est capable de prédire directement l'évènement CEGTA comme court-circuit. La figure 4.8 illustre le procédé d'extraction des données de CC stator. Le tableau 4.10 dévoile les résultats des données de CC stator extraites par l'algorithme.

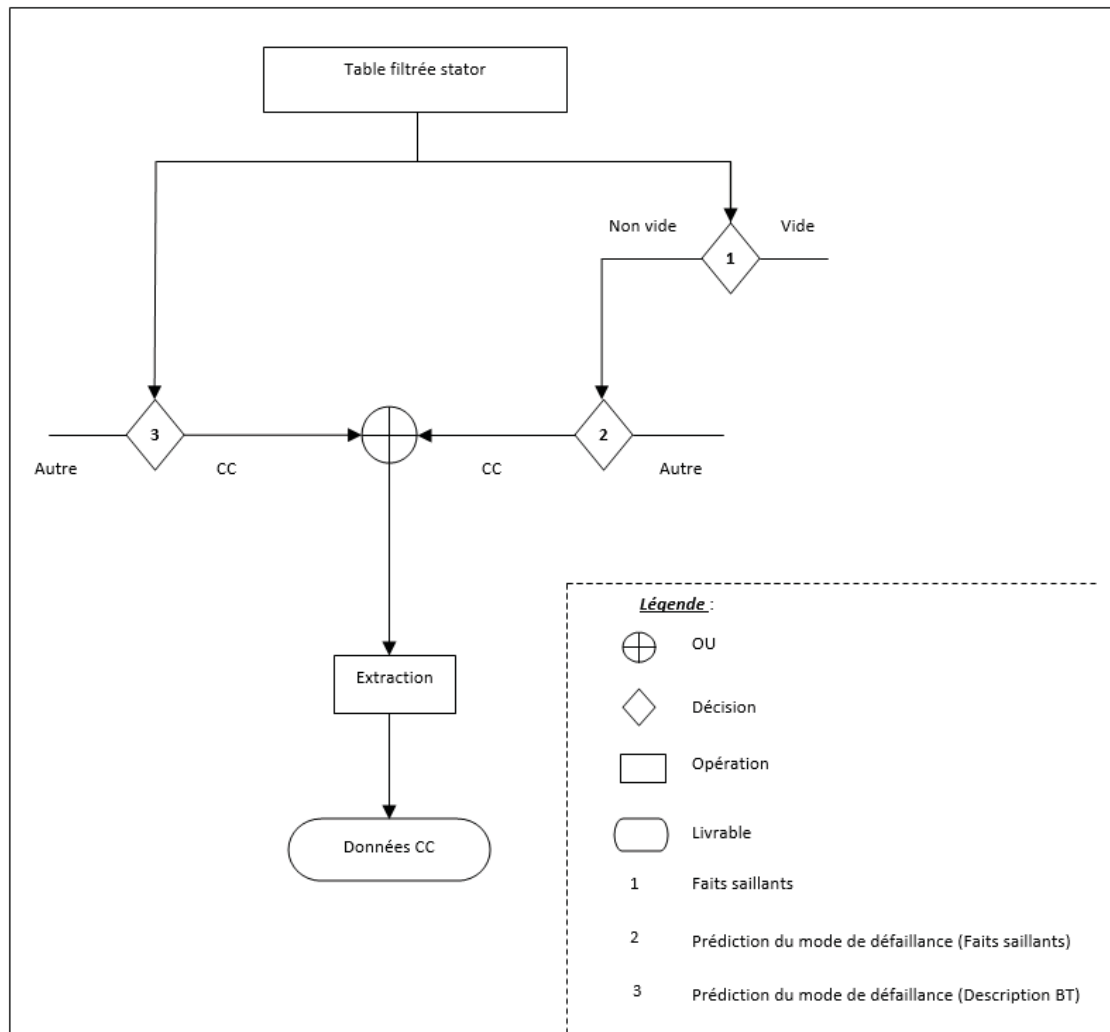


Figure 4.8: Logigramme d'extraction des données de court-circuit stator

TABLEAU 4.10: RÉSULTATS DE L'EXTRACTION DES DONNÉES DE CC STATOR

Centrale	Nombre d'arrêts CC non planifiés (défaillances)	Nombre d'arrêts CC planifiés (entretiens)	Nombre total d'arrêts CC
LG2-LG2A	12	76	88
Beauharnois	21	27	48
Carillon	3	5	8
LG1	1	27	28
LG4	2	58	60
Manic5-Manic5A	3	20	23
TOTAL	42	213	255

La capacité de généralisation de l'algorithme et la concordance des données extraites avec un fichier fourni par des experts ont permis de valider l'algorithme:

- L'algorithme a démontré une capacité à généraliser sur d'autres centrales. En effet, les modèles de classification automatique entraînés sur une partie de LG2 seulement ont montré des résultats satisfaisants sur d'autres centrales, témoignant ainsi la robustesse des modèles d'IA de l'algorithme.
- Les données extraites par l'algorithme ont été fournies à l'équipe IREQ (Institut de recherche sur l'électricité du Québec) en charge des alternateurs pour une validation en fonction des archives et de leur expertise. L'équipe IREQ alternateur a comparé les données extraites par l'algorithme (arrêts non planifiés seulement en prévision à l'étape E14) à un fichier d'archivage de CC stator version bêta. Version bêta, car le fichier n'était pas totalement complet au vu du travail laborieux de fouille manuelle d'archives du passé. Le tableau 4.11 présente les résultats de la comparaison des deux sources de données de CC stator (défaillances seulement).

TABLEAU 4.11: RÉSULTATS DE L'ALGORITHME PAR RAPPORT AU FICHIER DE CC STATOR DE L'IREQ

Ligne	Fichier IREQ	Algorithme	Nombre	Commentaires
A	1	1	23	L'algorithme a repéré 23 sur 45 données du fichier de l'IREQ
B	1	0	22	L'algorithme n'a pas repéré 22 sur 45 données du fichier de l'IREQ
C	0	1	19	19 sur 42 des données extraites par l'algorithme n'existent pas dans le fichier de l'IREQ

Le tableau 4.11 ci-dessus mérite un développement plus explicite. La ligne A indique le nombre de données du fichier détectées par l'algorithme. Le fichier comporte au total 45 données de défaillance CC stator correspondant au total des centrales choisies précédemment. Sur les 42 données extraites par l'algorithme, 23 sont présentes dans le fichier. La ligne B du tableau indique les 22 données présentes dans le fichier et absentes dans l'extraction de l'algorithme. Ce qui signifie que 22 sur 45 données du fichier n'ont pas été détectées par l'algorithme. Après investigation, il s'avère que les informations étaient absentes dans les bases de données. En effet, de nombreuses défaillances du fichier datent d'avant les années 1980, or, CEGTA contient des données de 1980 et plus. D'autre part, la grande majorité des BT du 20^e siècle ne possèdent aucune description de BT dans SIPROD, dès lors il est impossible pour l'algorithme ou même un humain de discriminer le BT. En bref, il ne s'agit pas d'une mauvaise performance de l'algorithme, mais de l'absence de l'information ou de la donnée brute dans les bases de données sources. La ligne C indique que 19 sur 42 données extraites par l'algorithme n'existent pas dans le fichier de l'IREQ. Ces 19 données ont fait l'objet d'une investigation chez l'équipe IREQ alternateur qui ont été validées. En effet, CEGTA est une base de données fiable dont le contenu est entièrement validé, dès lors l'évènement a forcément eu lieu.

De plus, le fichier en question n'était encore qu'en cours de développement, et il s'agissait d'une version bêta comme spécifiée précédemment au vu du travail laborieux de la fouille manuelle des archives de longue date. L'algorithme a montré un réel intérêt à extraire automatiquement des données pertinentes d'évènements précis ayant eu lieu dans le passé.

4.2 Estimation de la durée de vie résiduelle du stator

4.2.1 Étape 13 : Choix d'une méthode de prédiction des durées de vie

La revue de littérature a mis en évidence les différentes méthodologies d'estimation de DVR : les méthodes basées sur les données, les méthodes basées sur un modèle, les méthodes hybrides, les méthodes basées sur l'expérience, les méthodes basées sur la connaissance et les méthodes statistiques. L'article de Sikorska et al. (2011) présente des tableaux comparatifs permettant de choisir une approche appropriée selon le scénario à l'étude et les données disponibles. Selon les auteurs, les fonctions de fiabilité agrégées de la famille des méthodes statistiques sont simples et bien comprises par la communauté des ingénieurs en fiabilité. De plus, des intervalles de confiance sont disponibles. D'autre part, les fonctions de fiabilité agrégées sont applicables lorsque les données de MCON ne sont pas disponibles, comme c'est le cas dans cette étude. Biard et al. (2021) soutiennent également que plusieurs experts s'accordent sur l'efficacité de l'analyse statistique pour l'estimation de la durée de vie des actifs. Compte tenu des raisons énoncées précédemment, l'approche statistique a été retenue pour estimer la DVR des stators des GTA de HQP.

4.2.2 Étape 14 : Estimation des durées de vie pour chaque centrale

Cette étape de la méthodologie a précisé le contexte sur lequel se concentre l'étude, c'est-à-dire le vieillissement par dégradation. Les hypothèses spécifient dans un premier temps la considération unique des arrêts non planifiés (défaillances). Le tableau 4.10

montre que les centrales Carillon, LG1, LG4 et Manic5-Manic5A n'ont pas connu assez de défaillances par CC stator dans leur historique. La quantité de données disponibles pour ces centrales n'est pas significative pour une analyse statistique. Dès lors, seules les centrales LG2-LG2A et Beauharnois ont été sélectionnées pour la suite de l'étude. Le tableau 4.12 reporte le nombre de données de CC stator résultant de la combinaison de l'extraction par l'algorithme et du fichier de l'IREQ.

TABLEAU 4.12: NOMBRE DE DÉFAILLANCES PAR CC STATOR

Centrale	LG2-LG2A	Beauharnois
Dégradation	14	34
Causes externes	2	3
TOTAL	16	37

Les hypothèses spécifient dans un deuxième temps que seules les défaillances par dégradation doivent être considérées. Ainsi, selon le tableau 4.12, on obtient une population d'effectif 14 pour la centrale LG2-LG2A et une population d'effectif 34 pour la centrale Beauharnois. À noter que l'hypothèse selon laquelle un arrêt de durée importante équivaut à une défaillance critique a été prise en considération dans les résultats présentés dans le tableau 4.12. En effet, les arrêts non planifiés de durée insignifiante ont été écartés. Seules les défaillances critiques par dégradation sont intéressantes pour l'étude.

Une analyse statistique des données recueillies des centrales LG2-LG2A et Beauharnois a été effectuée pour déterminer la loi de survie ou distribution du stator. Les dates de mise en service des stators des groupes de chaque centrale ont permis de déterminer les points de départ des âges inter-pannes. Les paramètres des distributions ont été estimés par la méthode de régression linéaire. Le tableau 4.13 dévoile le coefficient R^2 obtenu pour chaque distribution ainsi que les résultats des tests de Kolmogorov-Smirnov.

TABLEAU 4.13: COEFFICIENTS R^2 DES DISTRIBUTIONS ET TEST DE KS

Distribution	Weibull	Exponentielle	Normale	Lognormale
LG2-LG2A	0.958	0.830	0.977	0.881
Test de KS	Oui	Non	Oui	Oui
Beauharnois	0.924	0.836	0.922	0.867
Test de KS	Oui	Oui	Oui	Oui

Dans le cas des deux centrales, le coefficient R^2 des lois Weibull et Normale se démarquent largement du reste. Concernant la loi normale, une règle de base stipule qu'elle ne s'utilise que lorsque la moyenne des durées de vie est supérieure à trois fois l'écart-type. Cette restriction imposée réduit la probabilité théorique de trouver une durée de vie négative à environ 0.1 %. En effet, le temps est toujours positif alors que la variable normale est définie de $-\infty$ à $+\infty$. Le tableau 4.14 présente la vérification de cette règle de base sur les données des deux centrales. Les résultats de cette vérification permettent d'éliminer la loi normale comme candidat. De ce fait, la loi Weibull, la plus populaire des lois permettant de modéliser de nombreuses situations d'usure de matériel s'avère la plus représentative des données des centrales LG2-LG2A et Beauharnois. Le tableau 4.15 reporte les paramètres de la loi Weibull et les intervalles de confiance au seuil $\alpha = 10$ %. Le tableau 4.16 présente l'espérance de la durée de vie des stators déterminée à partir des paramètres de la distribution Weibull. Les figures 4.9 et 4.10 illustrent respectivement les courbes de vieillissement des stators des centrales LG2-LG2A et Beauharnois. Lewis (1996) demeure un excellent ouvrage recommandé en ingénierie de la fiabilité. De nombreux concepts statistiques y sont notamment reportés, dont les principales distributions et leurs fondements mathématiques dans le chapitre 3. On y retrouve donc logiquement la formule mathématique de la loi Weibull et l'explication de ses différents paramètres.

TABLEAU 4.14: CRITÈRE D'ÉLIMINATION DE LA LOI NORMALE

Centrale	Moyenne des durées de vie (en années)	3 fois l'écart-type	Décision
LG2-LG2A	18.43	21.78	Non
Beauharnois	23.56	54.57	Non

TABLEAU 4.15: PARAMÈTRES WEIBULL LG2-LG2A & BEAUHARNOIS

Paramètres ($\alpha=10\%$)	LG2-LG2A	Beauharnois
$m_{\min}$	1.29	0.60
m	2.04	0.81
$m_{\max}$	3.24	1.09
$\theta_{\min}$	17.24	17.93
θ	21.46	25.56
$\theta_{\max}$	26.71	36.43

TABLEAU 4.16: DURÉE DE VIE ESTIMÉE DU STATOR LG2-LG2A & BEAUHARNOIS

$\alpha = 10 \%$	Espérance _{min} de la durée de vie	Espérance de la durée de vie	Espérance _{max} de la durée de vie
LG2-LG2A	15 ans	19 ans	23 ans
Beauharnois	19 ans	29 ans	39 ans

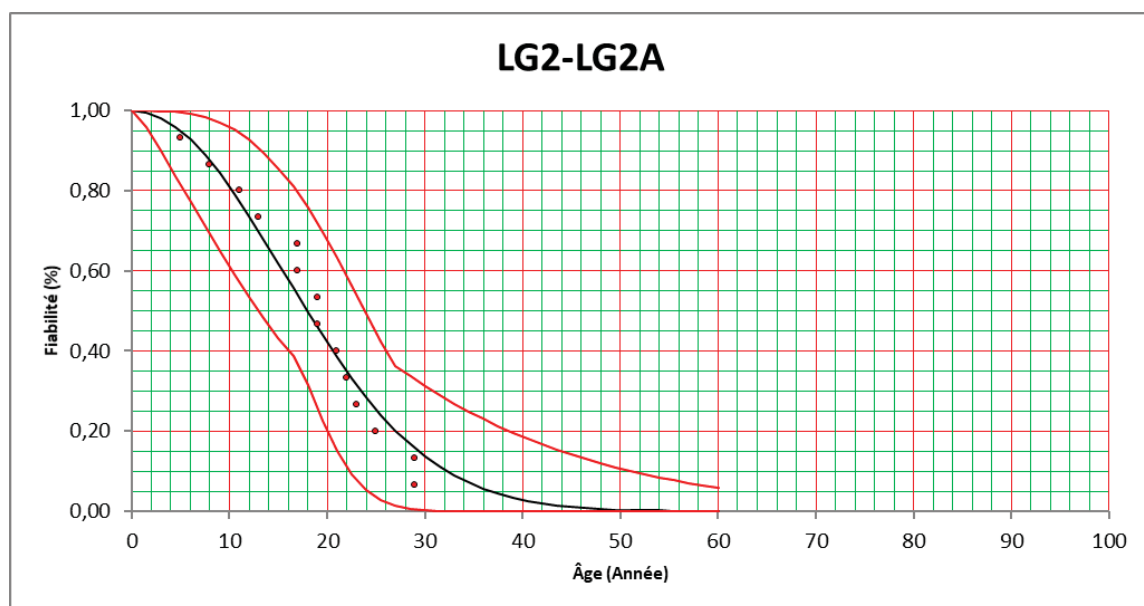


Figure 4.9: Courbe de vieillissement stator – centrale LG2-LG2A

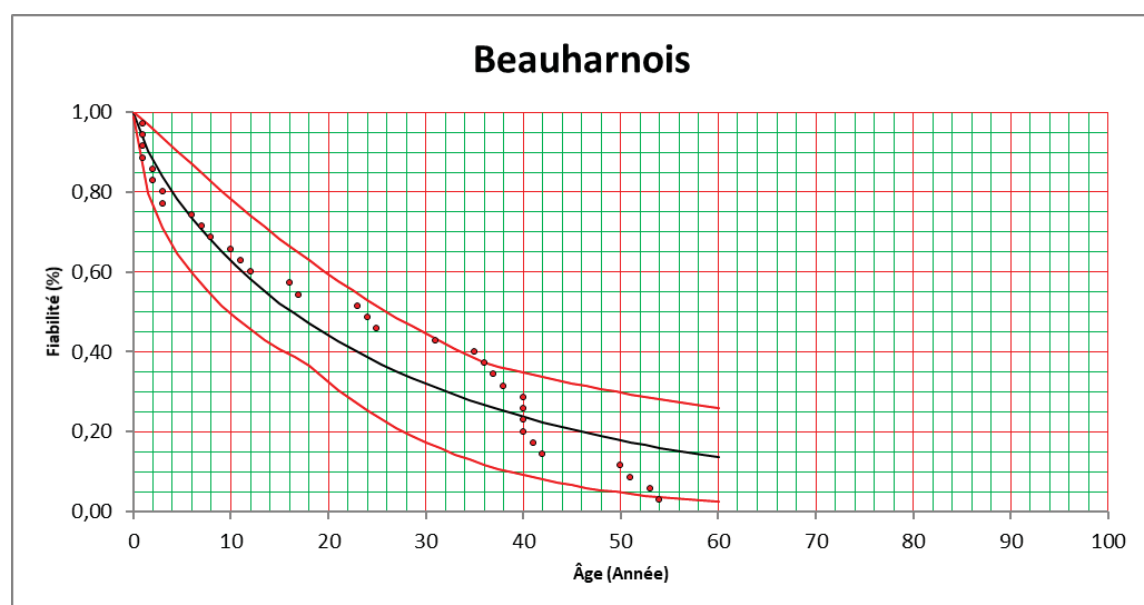


Figure 4.10: Courbe de vieillissement stator – centrale Beauharnois

4.2.3 Étape 15 : Test d'homogénéité des centrales

Le test-T statistique a été effectué pour apprécier le degré d'homogénéité entre les données des centrales LG2-LG2A et Beauharnois. La p-value résultante est égale à 0.169. Ainsi, au seuil $\alpha = 20 \%$, l'hypothèse que les données sont homogènes est acceptée.

La même analyse statistique effectuée précédemment a été répétée avec le mélange des données des deux centrales. Le tableau 4.17 présente les paramètres de la loi de survie Weibull. Le tableau 4.18 reporte les estimations d'un point de vue global de la durée de vie des stators. La figure 4.11 illustre la courbe de vieillissement du stator.

TABLEAU 4.17: PARAMÈTRES WEIBULL DES DONNÉES GLOBALES

Paramètres ($\alpha=20 \%$)	LG2-LG2A
$m_{\min}$	0.80
m	0.97
$m_{\max}$	1.18
$\theta_{\min}$	20.38
θ	24.72
$\theta_{\max}$	29.99

TABLEAU 4.18: DURÉE DE VIE ESTIMÉE GLOBALE DU STATOR

$\alpha = 20 \%$	Espérance _{min} de la durée de vie	Espérance de la durée de vie	Espérance _{max} de la durée de vie
Global	20 ans	25 ans	30 ans

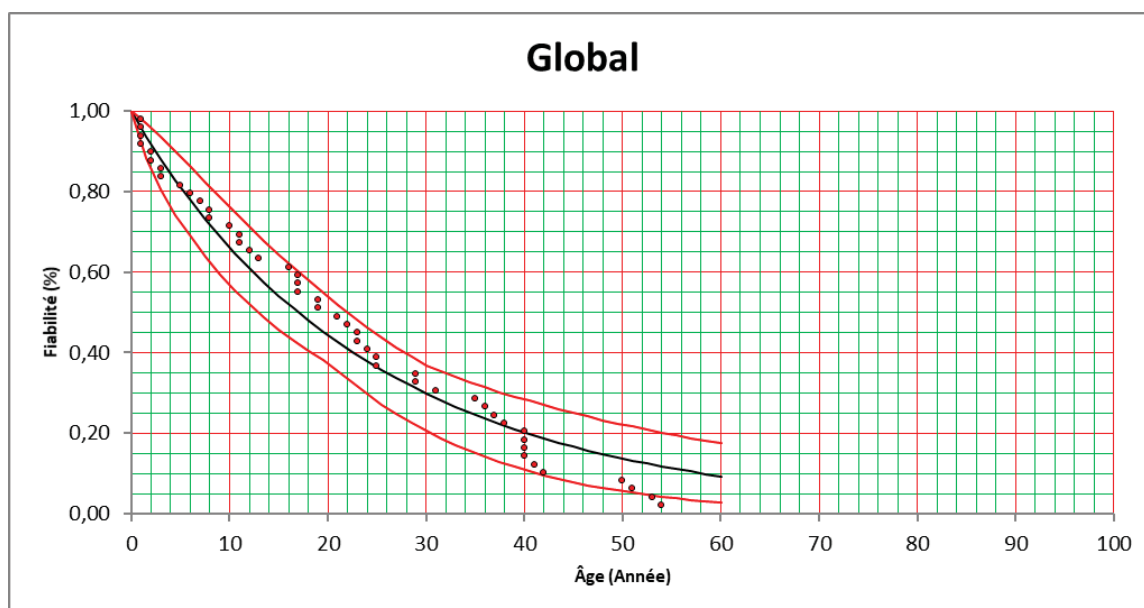


Figure 4.11: Courbe de vieillissement stator – Global

L'analyse statistique a permis d'évaluer la durée de vie du stator aux alentours de 25 ans avec un intervalle de confiance de plus ou moins 5 ans au seuil $\alpha = 20\%$. Les dates de mises en service des stators et l'hypothèse selon laquelle la réparation d'une défaillance critique est considérée comme un remplacement permettent d'estimer la DVR du stator. Le tableau 4.19 rapporte les DVR des stators des GTA de la centrale Beauharnois en considérant les résultats de l'analyse globale.

TABLEAU 4.19: DVR DES STATORS DE LA CENTRALE BEAUHARNOIS

Groupe	DVR (en années)
1	0
2	25
3	21
4	0
5	0
6	0
7	21
8	0
9	21
10	12
11	0
12	0
13	10
14	0
15	0
16	8
17	7
18	8
19	2
20	15
21	13
22	16
23	0
24	0
25	16
26	14
27	20
28	1
29	7
30	10
31	0
32	5
33	3
34	6
35	12
36	20

4.3 Extension de l'algorithme à d'autres systèmes et sous-systèmes

L'algorithme développé en tant que solution au problème d'intégrité des données de HQP a été appliquée au stator. Cependant, l'étude pourrait être faite sur n'importe quel autre composant générique de n'importe quel système. En effet, l'algorithme se montre capable de généraliser sur d'autres sous-systèmes en réglant certains paramètres :

- *Le choix du système dans les bases de données :*
Le système choisi peut être n'importe lequel des systèmes du GTA
- *Le modèle de classification des sous-systèmes :*
Un jeu de données doit être monté dans un premier temps constitué des descriptions des BT du système choisi et des véritables classes de sous-systèmes associées à ces BT. Ensuite, le modèle de classification doit être entraîné et optimisé sur ce jeu de données. Puis, importer le modèle dans l'algorithme de pairage pour remplacer le modèle de l'alternateur.
- *Le modèle de classification des modes de défaillance :*
Un jeu de données doit être monté dans un premier temps constitué des descriptions des BT du sous-système choisi et des véritables classes de modes de défaillance associées à ces BT. Ensuite le modèle de classification doit être entraîné et optimisé sur ce jeu de données. Puis, importer le modèle dans l'algorithme de pairage pour remplacer le modèle du stator.
- *Le choix du mode de défaillance:*
Choisir le mode de défaillance du sous-système considéré à retracer dans l'historique.

4.4 Conclusion

Ce quatrième chapitre a présenté l'application de la méthodologie énoncée au chapitre 3 et les résultats obtenus. L'idée de joindre SIPROD et CEGTA s'est montrée très prometteuse. En effet, SIPROD seule ne permet pas de distinguer correctement des interventions faites dans le cadre d'un entretien planifié ou d'une panne imprévue. Dès lors, les analyses faites avec SIPROD seule s'avèrent biaisées. CEGTA seule quant à elle ne donne pas suffisamment de détails sur les interventions effectuées durant les événements d'arrêt. L'algorithme développé a montré sa capacité à améliorer l'intégrité des données de HQP et d'en extraire automatiquement des caractéristiques importantes. Les défaillances critiques par CC stator des GTA ont ainsi pu être retracées. Entre autres, un constat a été fait selon lequel les défaillances critiques par CC stator des GTA sont relativement rares. La durée de vie des stators a ensuite été estimée à une valeur autour de 25 ans avec un intervalle de confiance de plus ou moins 5 ans. Cette estimation de la durée de vie est une base pour le calcul des DVR des stators de chaque groupe en fonction de leur âge. Finalement, quelques consignes ont été prodiguées pour l'extension de l'algorithme à d'autres sous-systèmes.

CONCLUSION

Ce mémoire a relaté toutes les étapes parcourues durant ce projet de recherche en passant du contexte de la recherche jusqu'aux résultats. La revue de la littérature a permis de recenser les écrits et de se faire une idée générale sur l'avancement de la communauté scientifique dans le domaine du pronostic, globalement influencé par les méthodes de prédiction, les décisions économiques, l'intégrité des données et la disponibilité des actifs physiques. Un algorithme a été développé pour le traitement des données de HQP permettant l'extraction de caractéristiques précises nécessaires pour l'estimation de la DVR des stators grâce à une analyse statistique.

Toute activité de recherche scientifique pertinente propose une valeur ajoutée à la communauté scientifique et technologique. HQP et les autres groupes de HQ en général font face à un sérieux problème d'intégrité des données, surtout au niveau de la jointure des bases de données comportant chacune des informations intéressantes. En effet, ces bases de données manquent pour la plupart de clé commune permettant une jointure directe à l'instar de SIPROD et CEGTA où la clé représente le numéro de BT inexistant dans CEGTA. Cette recherche propose une façon originale de joindre deux bases de données sans clé directe de jointure et d'en extraire automatiquement des caractéristiques spécifiques. Selon Biard et Nour (2021), parmi les publications scientifiques discutant de l'impact des outils de l'industrie 4.0 sur la gestion des actifs dans le secteur électrique, très peu s'adressent à la production, contrairement à la distribution et la transmission. Cette recherche exploite valablement l'IA qui est un outil de l'industrie 4.0 et s'applique au groupe de production d'une entreprise électrique.

N'ayant pas accès aux données de maintenance conditionnelle, il est difficile de distinguer une défaillance létale d'une défaillance non létale. En effet, l'étude se base uniquement sur des données textuelles et suppose qu'une défaillance est létale dépendamment de la durée de la remise en état.

De plus, si la description d'un BT est très vague ou inexistante, impossible de retracer l'existence de la défaillance. Ainsi sont les limites de la présente recherche.

Les travaux futurs consisteront à trouver une façon optimale d'associer les données historiques traitées extraites par l'algorithme aux données des systèmes de surveillance en continu des stators. Cela permettra non seulement d'intégrer le comportement historique des stators, mais également de prendre en considération des données conditionnelles permettant de mieux cerner les défaillances létales.

Par la suite, la prédiction de la fin de vie des autres composants du GTA serait intéressante pour une analyse holistique de chaque système. Une analyse économique justifiant les gains du remplacement de ces composants en fin de vie estimée se révèle un horizon de recherche intéressant.

RÉFÉRENCES

- Ahmadzadeh, F., & Lundberg, J. (2014). Remaining useful life estimation: review. *International Journal of System Assurance Engineering and Management*, 5(4), 461-474. doi: 10.1007/s13198-013-0195-0
- Albert, I., Donnet, S., Guihenneuc-Jouyaux, C., Low-Choy, S., Mengersen, K., & Rousseau, J. (2012). Combining Expert Opinions in Prior Elicitation. *Bayesian Anal.*, 7(3), 503-532. doi: 10.1214/12-BA717
- Altinel, B., & Ganiz, M. C. (2018). Semantic text classification: A survey of past and recent advances. *Information Processing and Management*, 54(6), 1129-1153. doi: 10.1016/j.ipm.2018.08.001
- Ayu, K., & Yunusa-Kaltungo, A. (2020). A Holistic Framework for Supporting Maintenance and Asset Management Life Cycle Decisions for Power Systems. *Energies*, 13(8). doi: 10.3390/en13081937
- Basri, E. I., Abdul Razak, I. H., Ab-Samat, H., & Kamaruddin, S. (2017). Preventive maintenance (PM) planning: a review. *Journal of Quality in Maintenance Engineering*, 23(2), 114-143. doi: 10.1108/JQME-04-2016-0014
- Behera, S., & Misra, R. (2021). Generative adversarial networks based remaining useful life estimation for IIoT. *Computers and Electrical Engineering*, 92. doi: 10.1016/j.compeleceng.2021.107195
- Biard, G., & Nour, G. A. (2021). Industry 4.0 Contribution to Asset Management in the Electrical Industry. *Sustainability*, 13(18). doi: 10.3390/su131810369
- Biard, G., Vaillancourt, R., Abdalnour, G., Langheit, C., Gaha, M., & Houle, G. (2021). Determining the lifecycle and future replacement cost of distribution network equipment.
- Chollet, F. o. (2021). *Deep Learning with Python, Second Edition*. New York: Manning Publications Co. LLC. Repéré à <https://learning.oreilly.com/library/view/~9781617296864/?ar?orpq&email=^u>

- Daigle, M. J., & Goebel, K. (2013). Model-Based Prognostics With Concurrent Damage Progression Processes. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 43(3). doi: 10.1109/TSMCA.2012.2207109
- Damodaran, D., Ravikumar, B., & Ramachandran, V. (2014). Bayesian software reliability model combining two priors and predicting total number of failures and failure time. *International Journal of Reliability, Quality and Safety Engineering*, 21(6). doi: 10.1142/S0218539314500314
- de Almeida Pais, J. E., Raposo, H. D. N., Farinha, J. T., Cardoso, A. J. M., & Marques, P. A. (2021). Optimizing the Life Cycle of Physical Assets through an Integrated Life Cycle Assessment Method. *Energies*, 14(19). doi: 10.3390/en14196128
- Deng, L., & Liu, Y. (2018). *Deep learning in natural language processing*. Repéré à <https://doi.org/10.1007/978-981-10-5209-5>
- Ding, S.-H., & Kamaruddin, S. (2015). Maintenance policy optimization—literature review and directions. *The International Journal of Advanced Manufacturing Technology*, 76(5-8), 1263-1283. doi: 10.1007/s00170-014-6341-2
- Fink, O., Wang, Q., Svensén, M., Dersin, P., Lee, W.-J., & Ducoffe, M. (2020). Potential, challenges and future directions for deep learning in prognostics and health management applications. *Engineering Applications of Artificial Intelligence*, 92. doi: 10.1016/j.engappai.2020.103678
- Ge, Y., Guo, L., & Dou, Y. (2019). Remaining useful life prediction of machinery based on K-S distance and LSTM neural network. *International Journal of Performability Engineering*, 15(3), 895-901. doi: 10.23940/ijpe.19.03.p18.895901
- Géron, A. (2019). *Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems* (2nd ed.). Sebastopol (Clif.) [etc.]: O'Reilly.
- Hydro-Québec. (2021a). Groupe turbine-alternateur. Repéré à <http://www.hydroquebec.com/comprendre/hydroelectricite/turbine-alternateur.html>
- Hydro-Québec. (2021b). Production d'électricité. Repéré à <https://www.hydroquebec.com/production/#:~:text=Hydro%2DQu%C3%A9bec%20Production%20produit%20de,sur%20les%20march%C3%A9s%20de%20gros>

- Hydro-Québec. (2022). Plan stratégique. Repéré à <https://www.hydroquebec.com/a-propos/plan-strategique.html>
- Jardine, A. K. S., Lin, D., & Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing*, 20(7), 1483-1510. doi: 10.1016/j.ymssp.2005.09.012
- Jiao, R., & Fang, F. (2019). *Research on fault diagnosis and prediction of power plant fans*. Communication présentée au Proceedings - 2019 International Conference on Sensing, Diagnostics, Prognostics, and Control, SDPC 2019. Repéré à <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85091502651&doi=10.1109%2fSDPC.2019.00119&partnerID=40&md5=09994175367259bbdd738dd7da42ce2a>
- Kang, Y., Cai, Z., Tan, C.-W., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. *Journal of Management Analytics*, 7(2), 139-172. doi: 10.1080/23270012.2020.1756939
- Khalifa, M., Khan, F., & Thorp, J. (2015). Risk-based maintenance and remaining life assessment for gas turbines. *Journal of Quality in Maintenance Engineering*, 21(1), 100-111. doi: 10.1108/JQME-12-2012-0047
- Khan, S., & Yairi, T. (2018). A review on the application of deep learning in system health management. *Mechanical Systems and Signal Processing*, 107, 241-265. doi: 10.1016/j.ymssp.2017.11.024
- Kim, S., Kim, N. H., & Choi, J.-H. (2020). Prediction of remaining useful life by data augmentation technique based on dynamic time warping. *Mechanical Systems and Signal Processing*, 136. doi: 10.1016/j.ymssp.2019.106486
- Koziel, S., Hilber, P., Westerlund, P., & Shayesteh, E. (2021). Investments in data quality: Evaluating impacts of faulty data on asset management in power systems. *Applied Energy*, 281. doi: 10.1016/j.apenergy.2020.116057
- Laayouj, N., & Jamouli, H. (2017). Prognosis of degradation based on a new dynamic method for remaining useful life prediction. *Journal of Quality in Maintenance Engineering*, 23(2), 239-255. doi: 10.1108/JQME-03-2016-0012

- Levin, M., Kalal, T. T., & Rodin, J. (2019). Reliability Concepts In. *Improving product reliability and software quality : strategies, tools, process and implementation* (Second edition, pp. 50-51). Repéré à <https://doi.org/10.1002/9781119179429>
- Lewis, E. E. (1996). *Introduction to reliability engineering* (2nd ed.). New York, N.Y.: J. Wiley and Sons.
- Li, C. J., & Lee, H. (2005). Gear fatigue crack prognosis using embedded model, gear dynamic model and fracture mechanics. *Mechanical Systems and Signal Processing*, 19(4), 836-846. doi: 10.1016/j.ymssp.2004.06.007
- Li, R., Arzaghi, E., Abbassi, R., Chen, D., Li, C., Li, H., & Xu, B. (2020). Dynamic maintenance planning of a hydro-turbine in operational life cycle. *Reliability Engineering & System Safety*, 204. doi: 10.1016/j.res.2020.107129
- Li, R., Verhagen, W. J. C., & Curran, R. (2020). A systematic methodology for Prognostic and Health Management system architecture definition. *Reliability Engineering & System Safety*, 193. doi: 10.1016/j.res.2019.106598
- Li, X., Jia, X., Wang, Y., Yang, S., Zhao, H., & Lee, J. (2020). Industrial Remaining Useful Life Prediction by Partial Observation Using Deep Learning With Supervised Attention. *IEEE/ASME Transactions on Mechatronics*, 25(5). doi: 10.1109/TMECH.2020.2992331
- Liu, J., & Geng, G. (2016). *Fault Prediction for Power Plant Equipment Based on Support Vector Regression*. Communication présentée au Proceedings - 2015 8th International Symposium on Computational Intelligence and Design, ISCID 2015. Repéré à <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84978144228&doi=10.1109%2fISCID.2015.130&partnerID=40&md5=660fdb87937c8cc375714caa15d9675>
- Mosallam, A., Medjaher, K., & Zerhouni, N. (2016). Data-driven prognostic method based on Bayesian approaches for direct remaining useful life prediction. *JOURNAL OF INTELLIGENT MANUFACTURING*, 27(5), 1037-1048.
- Navas, M. A., Sancho, C., & Carpio, J. (2017). Reliability analysis in railway repairable systems. *International Journal of Quality & Reliability Management*, 34(8), 1373-1398. doi: 10.1108/IJQRM-06-2016-0087

- Okoh, C., Roy, R., Mehnen, J., & Redding, L. (2014). Overview of Remaining Useful Life Prediction Techniques in Through-life Engineering Services. *Procedia CIRP*, 16, 158-163. doi: 10.1016/j.procir.2014.02.006
- Prajapati, A., Bechtel, J., & Ganesan, S. (2012). Condition based maintenance: a survey. *Journal of Quality in Maintenance Engineering*, 18(4), 384-400. doi: 10.1108/13552511211281552
- Rajpathak, D., & De, S. (2016). A data- and ontology-driven text mining-based construction of reliability model to analyze and predict component failures. *Knowledge and Information Systems : An International Journal*, 46(1), 87-113. doi: 10.1007/s10115-014-0806-3
- Robinson, E. I., Marzat, J., & Raïssi, T. (2017). Model-based prognosis using an explicit degradation model and Inverse FORM for uncertainty propagation. *IFAC PapersOnLine*, 50(1), 14242-14247. doi: 10.1016/j.ifacol.2017.08.1815
- Shahin, K. I., Simon, C., & Weber, P. (2020). Bootstrap Confidence Interval on IOHMM Parameters for System Health Diagnostic Under Multiple Operating Conditions. *IFAC PapersOnLine*, 53(2), 10297-10303. doi: 10.1016/j.ifacol.2020.12.2764
- Sikorska, J. Z., Hodkiewicz, M., & Ma, L. (2011). Prognostic modelling options for remaining useful life estimation by industry. *Mechanical Systems and Signal Processing*, 25(5), 1803-1836. doi: 10.1016/j.ymssp.2010.11.018
- Sun, B., Yang, Z., Balakrishnan, N., Chen, C., Tian, H., & Luo, W. (2020). An adaptive Bayesian melding method for reliability evaluation via limited failure data: An application to the servo turret. *Applied Sciences (Switzerland)*, 10(21), 1-16. doi: 10.3390/app10217591
- Teixeira, H. N., Lopes, I., & Braga, A. C. (2020). Condition-based maintenance implementation: a literature review. *Procedia Manufacturing*, 51, 228-235. doi: <https://doi.org/10.1016/j.promfg.2020.10.033>
- Thushara, G., Sean, H., Laura, K., Christopher, D., Ryan, J., Christopher, M., ... Andy, W. (2020). A Machine Learning Evaluation of Maintenance Records for Common Failure Modes in PV Inverters. 8, 211610-211620. doi: 10.1109/ACCESS.2020.3039182

- Waghmode, L. Y., & Patil, R. B. (2016). Reliability analysis and life cycle cost optimization: a case study from Indian industry. *International Journal of Quality & Reliability Management*, 33(3), 414-429. doi: 10.1108/IJQRM-11-2014-0184
- Wang, F.-K., & Mamo, T. (2019). Hybrid approach for remaining useful life prediction of ball bearings. *Quality and Reliability Engineering International*, 35(7), 2494-2505. doi: 10.1002/qre.2538
- Wang, Y., Zhao, Y., & Addepalli, S. (2020). Remaining Useful Life Prediction using Deep Learning Approaches: A Review. *Procedia Manufacturing*, 49, 81-88. doi: 10.1016/j.promfg.2020.06.015
- Wong, E. L., Jardine, A. K. S., & Banjevic, D. (2011, 24-27 Jan. 2011). *Incorporating repair information into maintenance optimization models for repairable systems*. Communication présentée au 2011 Proceedings - Annual Reliability and Maintainability Symposium
- Ye, C., Wang, H., Lu, W., & Li, J. (2020). Effective Bayesian-network-based missing value imputation enhanced by crowdsourcing. *Knowledge-Based Systems*, 190. doi: 10.1016/j.knosys.2019.105199
- Zhao, W., Egusquiza, M., Valero, C., Valentín, D., Presas, A., & Egusquiza, E. (2020). On the use of artificial neural networks for condition monitoring of pump-turbines with extended operation. *Measurement*, 163. doi: 10.1016/j.measurement.2020.107952
- Zhao, Z., Wu, J., Li, T., Sun, C., Yan, R., & Chen, X. (2021). Challenges and Opportunities of AI-Enabled Monitoring, Diagnosis & Prognosis: A Review. *Chinese Journal of Mechanical Engineering (English Edition)*, 34(1). doi: 10.1186/s10033-021-00570-7
- Zhou, M., Duan, N., Liu, S., & Shum, H.-Y. (2020). Progress in Neural NLP: Modeling, Learning, and Reasoning. 6(3), 275-290. doi: 10.1016/j.eng.2019.12.014

ANNEXE I

Graphiques comparatifs des données de SIPROD et CEGTA

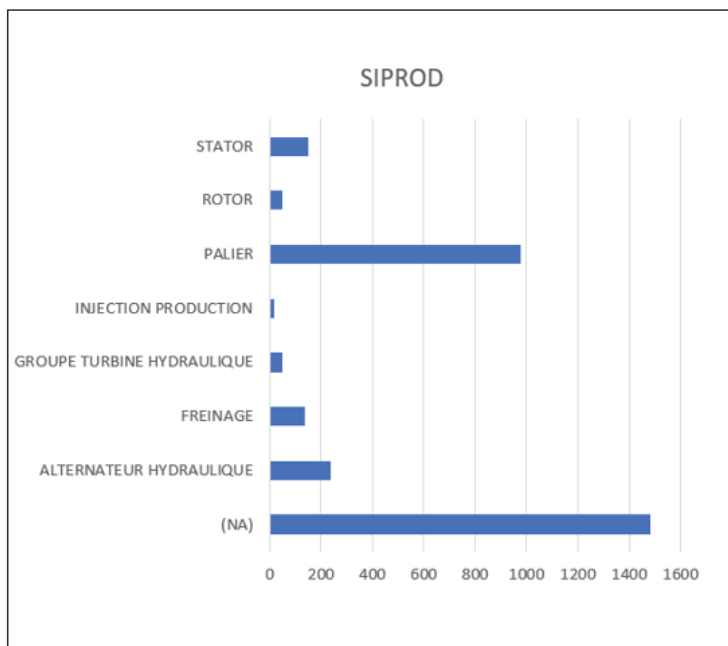


Figure A1.1 Nombre de données Alternateur - SIPROD - Beauharnois

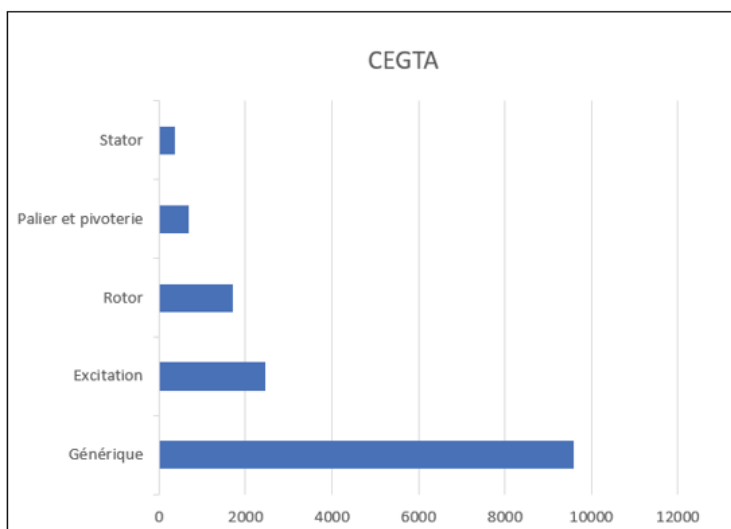


Figure A1.2: Nombre de données Alternateur - CEGTA - Beauharnois

Les figures A1.1 et A1.2 illustrent respectivement la représentativité des données par sous-systèmes du système alternateur de la centrale Beauharnois dans les bases de données SIPROD et CEGTA. On précise que dans SIPROD il s'agit des données de maintenance corrective et conditionnelle seulement, et dans CEGTA les données d'états d'arrêts forcés et d'arrêts pour entretien. L'analyse de SIPROD montre qu'environ 48% des données sont non attribuées (NA). L'analyse de CEGTA montre quant à elle qu'environ 65% des données sont classées dans la catégorie générique, c'est-à-dire une catégorie générale non assignée. L'intelligence artificielle a permis de réduire dans SIPROD le nombre d'instances NA, tout en corrigeant également les catégories mal attribuées. Le croisement du SIPROD nettoyé à CEGTA a permis de réduire dans CEGTA le nombre d'instances de la catégorie générique.

ANNEXE II

Code Python de l'algorithme de traitement des données

```

import numpy as np

import pandas as pd

##importation des BD##

CEGTA = pd.read_excel('CEGTA_ProfMaint_1980-01-01_2020-12-31.xlsx')

SIPROD = pd.read_excel('LG2A avec PBI connecté SQLSERVEUR.xlsx', sheet_name='Alternateur')

df1 = SIPROD.copy()

df2 = CEGTA.copy()

#Filtrage de CEGTA par codes ACE selectionnés: Alternateur

codes =
[42100,42110,42111,42112,42113,42114,42115,42120,42123,42124,42125,42126,42200,42210,42220,4
2230,42240,42260,42170,42171,42172,42174,42176,42177,42178, 41100, 41110]

df2 = df2[df2['Code Cause Indisp.'].isin(codes)]

df2 = df2[df2['État'].isin([21.1,21.2,21.4,21.7,22,23,24,25])]

#Affichage du nombre d'évènements Alternateur SIPROD et CEGTA

print('CEGTA:', df2.shape[0], 'lignes')

print('sigma-P:', df1.shape[0], 'lignes')

groupe = 10 #Choix du groupe

df1_s = df1[df1['SousRegroupement'] == groupe]

df2_s = df2[df2['Gr.'] == groupe]

print('sigma-p', df1_s.shape[0])

print('CEGTA', df2_s.shape[0])

##Juxtaposition temporelle des deux BD##

synthese = df1_s[['CreatedOn', 'HresReellesTotales']]

synthese['Durée (Hr)'] = np.nan

match = []

for i in df1_s.index:

```

```

for j in df2_s.index:
    if i==j:
        match.append(j)
print(match)
test = synthese.copy()
index_df2_s = df2_s.index.tolist()
ajust = [n for n in range(30000,30051)]
len_ = len(match)
ajust_2 = ajust[:len_]
index_df2_s_ajust = index_df2_s + ajust_2
if not match:
    test = test.reindex(test.index.union(index_df2_s))
else:
    test = test.reindex(test.index.union(index_df2_s_ajust))
    for index_, index_ajust in zip(match, ajust_2):
        df2_s.rename(index={index_: index_ajust}, inplace=True)
residual = [i for i in index_df2_s_ajust if i not in match]
for i in residual:
    test['CreatedOn'].loc[i] = df2_s['Date de début'].loc[i]
    test['Durée (Hr)'].loc[i] = df2_s['Durée (Hr)'].loc[i]
test = test.sort_values('CreatedOn')
index_SIPROD = [i for i in test.index if pd.isna(test['Durée (Hr)'].loc[i]) is True]
index_CEGTA = [i for i in test.index if pd.isna(test['Durée (Hr)'].loc[i]) is not True]
##Configuration des infos-bulle du graphique dynamique de visualisation##
dict_text_SIPROD = {}
for i in index_SIPROD:
    cout = 'Coût: ' + str(df1_s['CoutsReelsTotals'].loc[i])+'$' + '\n' + '<br>'
    description_panne = str('Descr.: ' + str(df1_s['BT Description'].loc[i]) + '\n') + '<br>'
    bt = 'BT: ' + str(df1_s['Numero'].loc[i]) + '\n' + '<br>'
    composante_prim = 'Composante prim: ' + str(df1_s['Composante primaire desc.'].loc[i]) + '\n' + '<br>'

```

```

composante_sec = 'Composante sec: ' + str(df1_s['Composante secondaire desc.'].loc[i]) + '\n' + '<br>'
equ_genre = 'Equ genre: ' + str(df1_s['Equ Genre desc'].loc[i]) + '\n' + '<br>'
cat_panne = 'Cat panne: ' + str(df1_s['Catégorie de panne desc.'].loc[i]) + '\n' + '<br>'

texte_SIPROD = cat_panne + equ_genre + composante_prim + composante_sec + cout + bt +
description_panne

#texte_SIPROD = texte_SIPROD
dict_text_SIPROD[i] = texte_SIPROD

dict_text_CEGTA = {}
for i in index_CEGTA:
    if len(str(df2_s['Faits Saillants Remarques'].loc[i])) < 110:
        date_fin = 'Date de fin: ' + str(df2_s['Date de fin'].loc[i]) + '\n' + '<br>'
        etat = 'État: ' + str(df2_s['État'].loc[i]) + '\n' + '<br>'
        desc_etat = 'Descr_état: ' + str(df2_s["Description de l'état"].loc[i]) + '\n' + '<br>'
        ace = 'ACE: ' + str(df2_s['Composant ACE'].loc[i]) + '\n' + '<br>'
        faits_saillants_3 = str('Descr 3: ' + str(df2_s['Faits Saillants Remarques'].loc[i]) + '\n') + '<br>'
        faits_saillants_2 = str('Descr 2: ' + str(df2_s['Faits Saillants Observations'].loc[i])[:100] + '\n') + '<br>'
        faits_saillants_1 = str('Descr 1: ' + str(df2_s['Faits Saillants Description'].loc[i])[:100] + '\n') + '<br>'
        texte_CEGTA = date_fin + etat + desc_etat + ace + faits_saillants_1 + faits_saillants_2 +
        faits_saillants_3
        dict_text_CEGTA[i] = texte_CEGTA
    else:
        date_fin = 'Date de fin: ' + str(df2_s['Date de fin'].loc[i]) + '\n' + '<br>'
        etat = 'État: ' + str(df2_s['État'].loc[i]) + '\n' + '<br>'
        desc_etat = 'Descr_état: ' + str(df2_s["Description de l'état"].loc[i]) + '\n' + '<br>'
        ace = 'ACE: ' + str(df2_s['Composant ACE'].loc[i]) + '\n' + '<br>'
        faits_saillants_3 = str('Descr 3: ' + str(df2_s['Faits Saillants Remarques'].loc[i]) + '\n') #+ '<br>'
        faits_saillants_2 = str('Descr 2: ' + str(df2_s['Faits Saillants Observations'].loc[i]) + '\n')[:100] + '<br>'
        faits_saillants_1 = str('Descr 1: ' + str(df2_s['Faits Saillants Description'].loc[i]) + '\n')[:100] + '<br>'
        b1 = faits_saillants_3.split(' ')
        len_b1 = len(b1)
        c1 = b1[:int(len_b1/2)]

```

```

d1 = b1[int(len_b1/2):]
e1 = (' '.join(c1))[:100] + '<br>'
f1 = (' '.join(d1))[:100] + '<br>'
texte_CEGTA = date_fin + etat + desc_etat + ace + faits_saillants_1 + faits_saillants_2 + e1 + f1
#texte_CEGTA = texte_CEGTA
dict_text_CEGTA[i] = texte_CEGTA

test['Détails'] = np.nan

for key in dict_text_SIPROD.keys():
    test['Détails'].loc[key] = dict_text_SIPROD[key]

for key in dict_text_CEGTA.keys():
    test['Détails'].loc[key] = dict_text_CEGTA[key]

import re
import nltk

from nltk.corpus import stopwords
from nltk.stem import SnowballStemmer

def acronym(texte): #Fonction de traitement des acronymes pour la fouille de texte
    liste = []
    for word in texte.split(' '):
        if word.count('.') > 2 and len(word) <= 20:
            word = word.replace('.', '')
        if len(word) == 3 and word[1] == '.':
            word = word.replace(word[1], '')
        if len(word) == 4 and word[1] == '.' and word[3] == '.':
            word = word.replace('.', '')
        if len(word) > 20 and word.count('.') > 2:
            word = word
        liste.append(word)
    output = ' '.join(liste)
    return output

def preprocess(text): # Fonction de traitement de texte

```

```

text = str(text).lower() # Tout en lettres minuscules pour faciliter le traitement
text = re.sub(r'\s+', ' ', text, flags=re.I)
text = acronym(text)
text = re.sub("[\d\W]+", " ", text) # Suppression de tous les caractères non alphabétiques
text = re.sub("[_]+", "", text)
text = re.sub(r'\s+', ' ', text, flags=re.I) #suppression des espaces en trop
text = [word for word in text.split() if len(word)>1] # Tokens en supprimant les mots dont len=1
stopword = stopwords.words('french') + ['rp', 'di', 'om', 'df', 'oe', 'ie', 'rp', 'mm', 'me', 'bt', 'vf', 'gr', 'vg',
'pt', 'dt', 'nor', 'er', 'mx', 'mw']
stopword = [word for word in stopword if word not in ['te']]
text = [word for word in text if word not in stopword] # Retrait des mots communs
return text

SIPROD_s = df1_s.copy()
CEGTA_s = df2_s.copy()

s = SIPROD_s[['CreatedOn', 'Numero', 'Equ Genre desc', 'Catégorie de panne desc.', 'BT Description',
'Composante primaire desc.', 'Composante secondaire desc.', 'HresReellesTotales', 'CoutsReelsTotals']]

c = CEGTA_s[['Date de début', 'Date de fin', 'Composant ACE', 'Faits Saillants Description', 'Faits Saillants
Observations', 'Faits Saillants Remarques', 'État']].reset_index()

c['Faits Saillants Description'].fillna('Empty', inplace=True)
c['Faits Saillants Observations'].fillna('Empty', inplace=True)
c['Faits Saillants Remarques'].fillna('Empty', inplace=True)

from datetime import timedelta

##Ajustement temporel##

timedelta1 = timedelta(hours=48) #SIPROD ouvert 48h avant le début de CEGTA
timedelta2 = timedelta(hours=48) #SIPROD ouvert 48h après la fermeture de CEGTA

##Pairage##

for s_index, s_line in s.iterrows():
    for c_index, c_line in c.iterrows():
        if c_line['État'] != 25:
            if (c_line['Date de début'] - timedelta1 <= s_line['CreatedOn']) and (s_line['CreatedOn'] <=
c_line['Date de fin'] + timedelta2) :
                s.at[s_index, 'Date_debut_CEGTA'] = c_line['Date de début']

```

```

s.at[s_index,'Date_fin_CEGTA'] = c_line['Date de fin']
s.at[s_index,'Composant ACE CEGTA'] = c_line['Composant ACE']
s.at[s_index,'index CEGTA'] = c_line['index']
s.at[s_index,'État'] = c_line['État']
s.at[s_index,'Faits Saillants Description'] = c_line['Faits Saillants Description']
s.at[s_index,'Faits Saillants Observations'] = c_line['Faits Saillants Observations']
s.at[s_index,'Faits Saillants Remarques'] = c_line['Faits Saillants Remarques']

for s_index,s_line in s.iterrows():
    for c_index, c_line in c.iterrows():
        if c_line['État'] == 25:
            if s_line['État'] not in [21.1, 21.2, 21.4, 21.7, 22, 23, 24]:
                if (c_line['Date de début'] - timedelta1 <= s_line['CreatedOn']) and (s_line['CreatedOn'] <=
c_line['Date de fin'] + timedelta2) :
                    s.at[s_index,'Date_debut_CEGTA'] = c_line['Date de début']
                    s.at[s_index,'Date_fin_CEGTA'] = c_line['Date de fin']
                    s.at[s_index,'Composant ACE CEGTA'] = c_line['Composant ACE']
                    s.at[s_index,'index CEGTA'] = c_line['index']
                    s.at[s_index,'État'] = c_line['État']
                    s.at[s_index,'Faits Saillants Description'] = c_line['Faits Saillants Description']
                    s.at[s_index,'Faits Saillants Observations'] = c_line['Faits Saillants Observations']
                    s.at[s_index,'Faits Saillants Remarques'] = c_line['Faits Saillants Remarques']

def nearest(items, pivot):
    return min(items, key=lambda x: abs(x - pivot))

s['nearest CEGTA date'] = np.nan

for s_index,s_line in s.iterrows():
    if s_line['État'] not in [21.1, 21.2, 21.4, 21.7, 22, 23, 24, 25]:
        nearest_CEGTA_date = nearest(c['Date de début'], s.at[s_index,'CreatedOn'])
        if abs(nearest_CEGTA_date - s.at[s_index,'CreatedOn']) <= timedelta(weeks=5):
            s.at[s_index,'nearest CEGTA date'] = nearest_CEGTA_date

dico = {}

```

```

for date, index in zip(c['Date de début'], c['index']):
    dico[date] = index

ss=s[s['nearest CEGTA date'].notnull()]

for s_index in ss.index:
    ss['index CEGTA'].loc[s_index] = dico[ss['nearest CEGTA date'].loc[s_index]]

for index in ss.index:
    for c_index, c_line in c.iterrows():
        if c_line['index'] == ss['index CEGTA'].loc[index]:
            ss['index CEGTA'].loc[index] = ss['index CEGTA'].loc[index]
            ss['Date_debut_CEGTA'].loc[index] = c_line['Date de début']
            ss['Date_fin_CEGTA'].loc[index] = c_line['Date de fin']
            ss['Composant ACE CEGTA'].loc[index] = c_line['Composant ACE']
            ss['État'].loc[index] = c_line['État']
            ss['Faits Saillants Description'].loc[index] = c_line['Faits Saillants Description']
            ss['Faits Saillants Observations'].loc[index] = c_line['Faits Saillants Observations']
            ss['Faits Saillants Remarques'].loc[index] = c_line['Faits Saillants Remarques']

s2 = s[s['État'].isin([21.1, 21.2, 21.4, 21.7, 22, 23, 24, 25])]#Filtre Pour garder uniquement les potentiels
matches

s_i = s2.copy()

for s_index,s_line in s2.iterrows():
    level_1 = s_line['État'] == 25

    level_2 = any(item in preprocess(s_line['Composant ACE CEGTA']) for item in
preprocess(s_line['Catégorie de panne desc.']))

    level_3 = any(item in preprocess(s_line['BT Description']) for item in preprocess(s_line['Faits Saillants
Remarques']))

    level_31 = any(item in preprocess(s_line['BT Description']) for item in preprocess(s_line['Faits Saillants
Observations']))

    level_32 = any(item in preprocess(s_line['BT Description']) for item in preprocess(s_line['Faits Saillants
Description']))

    level_4 = any(item in preprocess(s_line['Composant ACE CEGTA']) for item in preprocess(s_line['Equ
Genre desc.']))

    level_5 = s_line['Composant ACE CEGTA'] == 'ALTERNATEUR-GENERAL'

```

```

if s_line['CreatedOn'].strftime("%Y-%m-%d") == s_line['Date_debut_CEGTA'].strftime("%Y-%m-%d"):
    s2.at[s_index, 'Match'] = True

if s_line['CreatedOn'] >= s_line['Date_debut_CEGTA'] and s_line['CreatedOn'] <=
s_line['Date_fin_CEGTA']:
    s2.at[s_index, 'Match'] = True

if level_1 :
    s2.at[s_index, 'Match'] = True

if level_5:
    s2.at[s_index, 'Match'] = True

if level_2 or level_3 or level_31 or level_32 or level_4:
    s2.at[s_index, 'Match'] = True

def second_nearest(items, pivot):
    liste = [(abs(x - pivot), x) for x in items]
    liste.sort(key=lambda tup: tup[0])
    second_tuple = liste[1]
    return second_tuple[1]

ss2 = s2[s2['nearest CEGTA date'].notnull()]
ss2['second nearest CEGTA date'] = np.nan

for s_index, s_line in ss2.iterrows():
    if s_line['Match'] != True:
        second_nearest_CEGTA_date = second_nearest(c['Date de début'], s_line['CreatedOn'])
        if abs(second_nearest_CEGTA_date - s.at[s_index, 'CreatedOn']) <= timedelta(weeks=5):
            ss2.at[s_index, 'second nearest CEGTA date'] = second_nearest_CEGTA_date

for s_index in ss2[ss2['second nearest CEGTA date'].notnull()].index:
    ss2['index CEGTA'].loc[s_index] = dico[ss2['second nearest CEGTA date'].loc[s_index]]

for index in ss2.index:
    for c_index, c_line in c.iterrows():
        if c_line['index'] == ss2['index CEGTA'].loc[index]:
            s2['index CEGTA'].loc[index] = ss2['index CEGTA'].loc[index]
            s2['Date_debut_CEGTA'].loc[index] = c_line['Date de début']

```

```

s2['Date_fin_CEGTA'].loc[index] = c_line['Date de fin']
s2['Composant ACE CEGTA'].loc[index] = c_line['Composant ACE']
s2['État'].loc[index] = c_line['État']
s2['Faits Saillants Description'].loc[index] = c_line['Faits Saillants Description']
s2['Faits Saillants Observations'].loc[index] = c_line['Faits Saillants Observations']
s2['Faits Saillants Remarques'].loc[index] = c_line['Faits Saillants Remarques']

for s_index,s_line in s2.iterrows():
    a = preprocess(s_line['Composant ACE CEGTA'])
    b = preprocess(s_line['Catégorie de panne desc.'])
    c = preprocess(s_line['BT Description'])
    d = preprocess(s_line['Faits Saillants Remarques'])
    e = preprocess(s_line['Equ Genre desc'])
    f = ['altern']
    a = [i for i in a if i not in f]
    b = [i for i in b if i not in f]
    c = [i for i in c if i not in f]
    d = [i for i in d if i not in f]
    e = [i for i in e if i not in f]
    level_1 = s_line['État'] == 25
    level_2 = any(item in a for item in b)
    level_3 = any(item in c for item in d)
    level_4 = any(item in a for item in e)
    level_5 = s_line['Composant ACE CEGTA'] == 'ALTERNATEUR-GENERAL'
    if s_line['CreatedOn'].strftime("%Y-%m-%d") == s_line['Date_debut_CEGTA'].strftime("%Y-%m-%d"):
        s2.at[s_index,'Match'] = True
    if s_line['CreatedOn'] >= s_line['Date_debut_CEGTA'] and s_line['CreatedOn'] <=
s_line['Date_fin_CEGTA']:
        s2.at[s_index,'Match'] = True
    if level_1 :
        s2.at[s_index,'Match'] = True

```

```

if level_5:
    s2.at[s_index,'Match'] = True

if level_2 or level_3 or level_31 or level_32 or level_4:
    s2.at[s_index,'Match'] = True

for s_index,s_line in s2[s2['nearest CEGTA date'].notnull()].iterrows():
    if s_line['CoutsReelsTotals'] >= 10000 and s_line['Match'] != True:
        s2.at[s_index,'index CEGTA'] = dico[s2['nearest CEGTA date'].loc[s_index]]
        s2['Date_debut_CEGTA'].loc[s_index] = s_i['Date_debut_CEGTA'].loc[s_index]
        s2['Date_fin_CEGTA'].loc[s_index] = s_i['Date_fin_CEGTA'].loc[s_index]
        s2['Composant ACE CEGTA'].loc[s_index] = s_i['Composant ACE CEGTA'].loc[s_index]
        s2['État'].loc[s_index] = s_i['État'].loc[s_index]
        s2['Faits Saillants Remarques'].loc[s_index] = s_i['Faits Saillants Remarques'].loc[s_index]
        s2['Faits Saillants Description'].loc[s_index] = s_i['Faits Saillants Description'].loc[s_index]
        s2['Faits Saillants Observations'].loc[s_index] = s_i['Faits Saillants Observations'].loc[s_index]
        s2.at[s_index,'Match'] = True

for s_index,s_line in s2[s2['nearest CEGTA date'].isnull()].iterrows():
    if s_line['CoutsReelsTotals'] >= 10000 and s_line['Match'] != True:
        s2.at[s_index,'Match'] = True

print(len(s2[s2['Match']==True]), 'sur', df1_s.shape[0], 'événements SIPROD pairés à CEGTA')# Nombre de
points SIPROD pairés

print('Correspond à {:.1f}% de SIPROD du
groupe'.format(((len(s2[s2['Match']==True]))/df1_s.shape[0])*100))

s_match = s2[s2['Match']== True]

CEGTA_match_idx = s_match['index CEGTA'].tolist()

CEGTA_match_idx = list(dict.fromkeys(CEGTA_match_idx))

print(len(CEGTA_match_idx), 'sur', df2_s.shape[0], 'événements CEGTA pairés à SIPROD')# Nombre de
points CEGTA pairés

print('Correspond à {:.1f}% de CEGTA du groupe'.format(((len(CEGTA_match_idx)/df2_s.shape[0])*100)))

SIPROD_match_idx = s_match.index.tolist()

test2 = test.copy()

for index, line in test2.iterrows():

```

```

if index in (SIPROD_match_idx + CEGTA_match_idx):
    test2.at[index, 'HresReellesTotales'] = np.nan
    test2.at[index, 'Durée (Hr)'] = np.nan
    test2.at[index, 'SIPROD_match'] = test.at[index, 'HresReellesTotales']
    test2.at[index, 'CEGTA_match'] = test.at[index, 'Durée (Hr)']
CEGTA_25 = s_match[s_match['État']==25]['index CEGTA'].tolist()
CEGTA_25 = list(dict.fromkeys(CEGTA_25))
SIPROD_25 = s_match[s_match['État']==25].index.tolist()
test3 = test2.copy()
for index, line in test3.iterrows():
    if index in (SIPROD_25 + CEGTA_25):
        test3.at[index, 'SIPROD_match'] = np.nan
        test3.at[index, 'CEGTA_match'] = np.nan
        test3.at[index, 'SIPROD_match_25'] = test.at[index, 'HresReellesTotales']
        test3.at[index, 'CEGTA_match_25'] = test.at[index, 'Durée (Hr)']

##Graphique dynamique (Outil de visualisation)##

test3.iplot(y=['HresReellesTotales', 'Durée (Hr)', 'SIPROD_match', 'CEGTA_match', 'SIPROD_match_25',
'CEGTA_match_25'], x='CreatedOn', mode='markers', xTitle='Date', yTitle='durée (h)',text = 'Détails',
title='Pairage SIPROD - CEGTA GR'+str(groupe))#, color=['orange', 'blue', 'green', 'red',]) #Outil de
visualisation de la juxtaposition temporelle des deux BD et du pairage

def token_to_text(tokens):
    text = ' '.join(tokens)
    return text

import joblib

##Combinaison pairage et classification automatique##

model = joblib.load("model_2.pkl") #Importation du modèle de classification automatique de l'alternateur
encodage = {0:'Barres blindées', 1:'Excitation', 2:'Freinage', 3:'Général',
4:'Palier & Injection', 5:'Protection incendie', 6:'Rotor',
7:'Sectionneurs - Disjoncteurs', 8:'Stator',
9:'refroidissement général'}

s_match_2 = s_match.copy()

```

```

X_SIPROD = s_match_2['BT Description']
X_CEGTA = s_match_2['Faits Saillants Remarques']
X_SIPROD = X_SIPROD.apply(lambda x: preprocess(x))
X_CEGTA = X_CEGTA.apply(lambda x: preprocess(x))
X_SIPROD = X_SIPROD.apply(lambda x: token_to_text(x))
X_CEGTA = X_CEGTA.apply(lambda x: token_to_text(x))
predictions_SIPROD = model.predict(X_SIPROD)
predictions_SIPROD_mdf = [encodage[n] for n in predictions_SIPROD]
s_match_2['prediction SIPROD'] = predictions_SIPROD_mdf

##Tri stator##

stator = s_match_2[s_match_2['Composant ACE CEGTA'].isin(['BOBINAGE DU STATOR', "CALES
D'ENCOCHE DES ENROULEMENTS", 'NOYAU MAGNETIQUE (LAMINATIONS) DU STATOR', "STATOR DE
L'ALTERNATEUR", "SYST. DE REFROIDISSEMENT DU STATOR"])]

| s_match_2['prediction SIPROD'].isin(['Stator'])]

model_stator = joblib.load("model_stator.pkl") #importation du modèle de classification automatique du
stator

X_SIPROD1 = stator['BT Description']
X_CEGTA3 = stator['Faits Saillants Remarques']
X_CEGTA2 = stator['Faits Saillants Observations']
X_CEGTA1 = stator['Faits Saillants Description']
X_SIPROD1 = X_SIPROD1.apply(lambda x: preprocess(x))
X_CEGTA1 = X_CEGTA1.apply(lambda x: preprocess(x))
X_CEGTA2 = X_CEGTA2.apply(lambda x: preprocess(x))
X_CEGTA3 = X_CEGTA3.apply(lambda x: preprocess(x))
X_SIPROD1 = X_SIPROD1.apply(lambda x: token_to_text(x))
X_CEGTA1 = X_CEGTA1.apply(lambda x: token_to_text(x))
X_CEGTA2 = X_CEGTA2.apply(lambda x: token_to_text(x))
X_CEGTA3 = X_CEGTA3.apply(lambda x: token_to_text(x))
predictions_SIPROD1 = model_stator.predict(X_SIPROD1)
predictions_CEGTA1 = model_stator.predict(X_CEGTA1)
predictions_CEGTA2 = model_stator.predict(X_CEGTA2)

```

```

predictions_CEGTA3 = model_stator.predict(X_CEGTA3)
encodage_stator = {0:'Court-circuit', 1:'Général', 2:'Refroidissement'}
predictions_SIPROD_stator = [encodage_stator[n] for n in predictions_SIPROD1]
predictions_CEGTA1_stator = [encodage_stator[n] for n in predictions_CEGTA1]
predictions_CEGTA2_stator = [encodage_stator[n] for n in predictions_CEGTA2]
predictions_CEGTA3_stator = [encodage_stator[n] for n in predictions_CEGTA3]
stator['mdf stator SIPROD'] = predictions_SIPROD_stator
stator['mdf stator CEGTA1'] = predictions_CEGTA1_stator
stator['mdf stator CEGTA2'] = predictions_CEGTA2_stator
stator['mdf stator CEGTA3'] = predictions_CEGTA3_stator
for index, line in stator.iterrows():
    if line['prediction SIPROD'] not in ['Stator', 'Général']:
        stator.at[index, 'mdf stator SIPROD'] = 's/o'

    if line['Composant ACE CEGTA'] not in ['BOBINAGE DU STATOR', 'CALES D'ENCOCHE DES
    ENROULEMENTS', 'NOYAU MAGNETIQUE (LAMINATIONS) DU STATOR', 'STATOR DE L'ALTERNATEUR',
    'SYST. DE REFROIDISSEMENT DU STATOR', 'ALTERNATEUR-GENERAL']:
        stator.at[index, 'mdf stator CEGTA1'] = 's/o'
        stator.at[index, 'mdf stator CEGTA2'] = 's/o'
        stator.at[index, 'mdf stator CEGTA3'] = 's/o'
index_CEGTA_stator = stator['index CEGTA'].unique().tolist()
pairage_stator = s_match_2[s_match_2['index CEGTA'].isin(index_CEGTA_stator)]
pairage_stator['mdf stator SIPROD'] = np.nan
pairage_stator['mdf stator CEGTA1'] = np.nan
pairage_stator['mdf stator CEGTA2'] = np.nan
pairage_stator['mdf stator CEGTA3'] = np.nan
for index in pairage_stator.index:
    if pairage_stator['prediction SIPROD'].loc[index] == 'Stator' or pairage_stator['Composant ACE
    CEGTA'].loc[index] in ['BOBINAGE DU STATOR', 'CALES D'ENCOCHE DES ENROULEMENTS', 'NOYAU
    MAGNETIQUE (LAMINATIONS) DU STATOR', 'STATOR DE L'ALTERNATEUR', 'SYST. DE REFROIDISSEMENT
    DU STATOR']:
        pairage_stator['mdf stator SIPROD'].loc[index] = stator['mdf stator SIPROD'].loc[index]
        pairage_stator['mdf stator CEGTA1'].loc[index] = stator['mdf stator CEGTA1'].loc[index]

```

```

    paireage_stator['mdf stator CEGTA2'].loc[index] = stator['mdf stator CEGTA2'].loc[index]
    paireage_stator['mdf stator CEGTA3'].loc[index] = stator['mdf stator CEGTA3'].loc[index]
else:
    paireage_stator['mdf stator SIPROD'].loc[index] = 's/o'
    paireage_stator['mdf stator CEGTA1'].loc[index] = 's/o'
    paireage_stator['mdf stator CEGTA2'].loc[index] = 's/o'
    paireage_stator['mdf stator CEGTA3'].loc[index] = 's/o'
pairages = []
for index in index_CEGTA_stator:
    paireage = paireage_stator[paireage_stator['index CEGTA'] == index]
    pairages.append(paireage)
def decision (n):
    proportions_effectif = pairages[n]['prediction SIPROD'].value_counts()/len(pairages[n]['prediction SIPROD'])
    proportions_effectif = proportions_effectif.to_frame().reset_index()
    effectif = pairages[n]['prediction SIPROD'].value_counts()
    proportions_effectif['Effectif'] = [n for n in effectif]
    proportions_effectif.sort_values(by='index', inplace=True)
    proportions_couts = pairages[n].groupby('prediction SIPROD').sum()['CoutsReelsTotals']/pairages[n]['CoutsReelsTotals'].sum()
    couts = pairages[n].groupby('prediction SIPROD').sum()['CoutsReelsTotals']
    mdf_stator = pairages[n][(pairages[n]['prediction SIPROD']=='Stator')|(pairages[n]['mdf stator SIPROD']=='Court-circuit')].groupby('mdf stator SIPROD').sum()['CoutsReelsTotals']
    mdf_stator = mdf_stator.to_frame().reset_index()
    decision = pd.DataFrame({'Entité': [n for n in proportions_effectif['index']], 'Effectif':[n for n in proportions_effectif['Effectif']], 'Proportion effectif':[round(n,2) for n in proportions_effectif['prediction SIPROD']], 'Coûts': [n for n in couts], 'Proportion coûts':[round(n,2) for n in proportions_couts]})
    decision.sort_values(by='Coûts', inplace=True, ascending=False)
    couts_stator = decision['Proportion coûts'].where(decision['Entité']=='Stator').dropna()
    couts_general = decision['Proportion coûts'].where(decision['Entité']=='Général').dropna()
    couts_max = max([n for n in decision['Proportion coûts']])
    index_max = decision.where(decision['Proportion coûts']==couts_max).dropna().index.tolist()

```

```

decision["Déclencheur CEGTA"] = False
if len(mdf_stator['CoutsReelsTotals']) > 0:
    index_mdf_stator_max = mdf_stator[mdf_stator['CoutsReelsTotals'] ==
max(mdf_stator['CoutsReelsTotals'])].index.tolist()
    decision['mdf stator'] = np.nan
    if "Stator" in decision['Entité'].tolist():
        decision['mdf stator'].loc[couts_stator.index] = mdf_stator['mdf stator
SIPROD'].loc[max(index_mdf_stator_max)]
    if "Stator" not in decision['Entité'].tolist() and "Général" in decision['Entité'].tolist() :
        decision['mdf stator'].loc[couts_general.index] = mdf_stator['mdf stator
SIPROD'].loc[max(index_mdf_stator_max)]
    if "Stator" in decision['Entité'].tolist():
        if 'Court-circuit' in decision['mdf stator'].tolist(): #decision['Coûts'].loc[(couts_stator.index)]>0 and
decision['mdf stator'].loc[(couts_stator.index)] == 'Court-circuit':
            idx = decision[decision['mdf stator']=='Court-circuit'].index.tolist()
            decision["Déclencheur CEGTA"].loc[idx] = True
            decision.sort_values(by='Déclencheur CEGTA', inplace=True, ascending=False)
        else:
            decision["Déclencheur CEGTA"].loc[(index_max)] = True
    else:
        decision["Déclencheur CEGTA"].loc[(index_max)] = True
return decision
true_stator = []
for n in range(0,len(pairages)):
    if decision(n)['Entité'].iloc[0]=='Stator' or decision(n)[decision(n).columns[-1]].iloc[0]=='Court-circuit':
        true_stator.append(n)
        print(f'{n} remplit la condition')
print(true_stator)
##Tri des courts-circuits stator##
cc = []
for n in true_stator:
    if decision(n)[decision(n).columns[-1]].iloc[0] == 'Court-circuit':

```

```
        print(f'{n} remplit la condition')
        cc.append(n)

print(cc)

plus=[]

for i in range(0, len(pairages)):

    if 'Court-circuit' in pairages[i][['mdf stator CEGTA1', 'mdf stator CEGTA2', 'mdf stator
CEGTA3']].iloc[0].tolist():

        plus.append(i)

res = [i for i in plus if i not in cc]

cc.extend(res)

print(cc)
```

ANNEXE III

Code python de la classification automatique des sous-systèmes de l’alternateur

```

import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import re

data = pd.read_excel('df alternateur.xlsx')
df = data.copy()

df.shape

plt.figure(figsize=(12,8))
df['Classe'].value_counts().plot.pie(autopct="%.1f%%")

import nltk

from nltk.corpus import stopwords
from nltk.stem import SnowballStemmer

df['tokens'] = df['Texte'].apply(lambda x: preprocess(x))

from sklearn.preprocessing import LabelEncoder

encoder = LabelEncoder()

df['target'] = encoder.fit_transform(df['Classe'])

from wordcloud import WordCloud

##Visualisation des nuages de mots##

classes = [cat_0, cat_1, cat_2, cat_3, cat_4, cat_5, cat_6, cat_7, cat_8, cat_9]

titles = ['Barres blindées', 'Excitation',

          'Freinage', 'Général', 'Palier & Injection', 'Protection incendie', 'Rotor',

          'Sectionneurs - Disjoncteurs', 'Stator', 'Refroidissement-général']

for i,j in zip(classes, titles):

    plt.figure(figsize=(10,7))

    separator = ','

```

```

wordcloud = WordCloud(background_color='white', width=800, height=400).generate(separator.join(i))
plt.imshow(wordcloud, interpolation='bilinear')

plt.title(j)

plt.axis("off")

plt.show()

from sklearn.model_selection import train_test_split

##Séparation du jeu de données (80% entraînement - 20% test)
X_train, X_test, y_train, y_test = train_test_split(df['tokens'], df['target'], test_size=0.2, random_state=42 )

from sklearn.feature_extraction.text import CountVectorizer

count_vec = CountVectorizer()

X_train = X_train.apply(lambda x: token_to_text(x))
X_test = X_test.apply(lambda x: token_to_text(x))

vector = count_vec.fit_transform(X_train)

##importation des algorithmes de ML##

from sklearn.naive_bayes import MultinomialNB

from sklearn.linear_model import SGDClassifier

from sklearn.tree import DecisionTreeClassifier

from sklearn.ensemble import RandomForestClassifier

from sklearn.linear_model import LogisticRegression

clf_nb = MultinomialNB()

clf_sgd = SGDClassifier(random_state=42)

clf_tree = DecisionTreeClassifier(random_state=42)

clf_rf = RandomForestClassifier(random_state=42)

clf_lr = LogisticRegression(random_state=42)

from sklearn.model_selection import cross_val_score

score1 = cross_val_score(clf_nb, vector, y_train, cv=5)

print(score1.mean())

score2 = cross_val_score(clf_sgd, vector, y_train, cv=5)

print(score2.mean())

score3 = cross_val_score(clf_tree, vector, y_train, cv=5)

```

```

print(score3.mean())

score4 = cross_val_score(clf_rf, vector, y_train, cv=5)

print(score4.mean())

score5 = cross_val_score(clf_lr, vector, y_train, cv=5)

print(score5.mean())

# Les trois meilleurs algos sont SGDClassifier, RandomForest et LogisticRegression

##Optimisation des hyperparamètres##

from sklearn.pipeline import Pipeline

text_clf_sgd = Pipeline([

    ('vect', CountVectorizer()),

    ('tfidf', TfidfTransformer()),

    ('clf_sgd', SGDClassifier(random_state=42))

])

text_clf_rf = Pipeline([

    ('vect', CountVectorizer()),

    ('tfidf', TfidfTransformer()),

    ('clf_rf', RandomForestClassifier(random_state=42))])

text_clf_lr = Pipeline([

    ('vect', CountVectorizer()),

    ('tfidf', TfidfTransformer()),

    ('clf_lr', LogisticRegression(random_state=42))])

from sklearn.model_selection import GridSearchCV

parameters_sgd = {

    'vect__ngram_range': [(1,1),(1,2)],

    'vect__min_df': [n for n in range(1,6)],

    'tfidf__use_idf': (True, False),

    'clf_sgd__alpha': (1e-2, 1e-3, 1e-4, 1e-5, 1e-6),

    'clf_sgd__penalty': ['l1','l2']

}

```

```

gs_clf_sgd = GridSearchCV(text_clf_sgd, parameters_sgd, cv=5, n_jobs=-1)

gs_clf_sgd = gs_clf_sgd.fit(X_train, y_train)

parameters_rf = {
    'vect__ngram_range': [(1,1)],
    'vect__min_df': [n for n in range(1,5)],
    'tfidf__use_idf': (True, False),
    'clf_rf__bootstrap': (True, False),
    'clf_rf__criterion': ['gini','entropy'],
}

gs_clf_rf = GridSearchCV(text_clf_rf, parameters_rf, cv=5, n_jobs=-1)
gs_clf_rf = gs_clf_rf.fit(X_train, y_train)

parameters_lr = {
    'vect__ngram_range': [(1,1)],
    'vect__min_df': [1,2,3],
    'tfidf__use_idf': (True, False),
    'clf_lr__C': [1, 3, 5, 7, 10],
    'clf_lr__penalty': ['l1','l2'],
    'clf_lr__max_iter': [1000]
}

gs_clf_lr = GridSearchCV(text_clf_lr, parameters_lr, cv=5, n_jobs=-1)

gs_clf_lr = gs_clf_lr.fit(X_train, y_train)

##Courbes d'apprentissage##

from sklearn.model_selection import learning_curve

N, train_score, val_score = learning_curve(gs_clf_sgd.best_estimator_, X_train, y_train,
                                          train_sizes=np.linspace(0.1, 1, 10), cv=5)

print(N)

```

```

plt.plot(N, train_score.mean(axis=1), label='train')
plt.plot(N, val_score.mean(axis=1), label='validation')
plt.xlabel('train_sizes')
plt.title('learning curve SGDClassifier')
plt.legend()

N, train_score, val_score = learning_curve(gs_clf_rf.best_estimator_, X_train, y_train,
                                          train_sizes=np.linspace(0.1, 1, 10), cv=5)

print(N)

plt.plot(N, train_score.mean(axis=1), label='train')
plt.plot(N, val_score.mean(axis=1), label='validation')
plt.xlabel('train_sizes')
plt.title('learning curve Random Forest')
plt.legend()

N, train_score, val_score = learning_curve(gs_clf_lr.best_estimator_, X_train, y_train,
                                          train_sizes=np.linspace(0.1, 1, 10), cv=5)

print(N)

plt.plot(N, train_score.mean(axis=1), label='train')
plt.plot(N, val_score.mean(axis=1), label='validation')
plt.xlabel('train_sizes')
plt.title('learning curve logistic regression')
plt.legend()

##Rapport de classification et matrice de confusion##

from sklearn.metrics import classification_report, confusion_matrix
predicted_sgd = gs_clf_sgd.best_estimator_.predict(X_test)
print(classification_report(y_test, predicted_sgd, target_names = ['Barres blindées', 'Excitation',
    'Freinage', 'Général', 'Palier & Injection', 'Protection incendie', 'Rotor',
    'Sectionneurs - Disjoncteurs', 'Stator', 'refroidissement général']))

from sklearn.metrics import plot_confusion_matrix
plot_confusion_matrix(gs_clf_sgd.best_estimator_, X_test, y_test, colorbar=False)

```

```

predicted_lr = gs_clf_lr.predict(X_test)
print(classification_report(y_test, predicted_lr, target_names = ['Barres blindées', 'Excitation',
    'Freinage', 'Général', 'Palier & Injection', 'Protection incendie', 'Rotor',
    'Sectionneurs - Disjoncteurs', 'Stator', 'refroidissement général']))
plot_confusion_matrix(gs_clf_lr.best_estimator_, X_test, y_test, colorbar=False)
predicted_rf = gs_clf_rf.predict(X_test)
print(classification_report(y_test, predicted_rf, target_names = ['Barres blindées', 'Excitation',
    'Freinage', 'Général', 'Palier & Injection', 'Protection incendie', 'Rotor',
    'Sectionneurs - Disjoncteurs', 'Stator', 'refroidissement général']))
plot_confusion_matrix(gs_clf_rf.best_estimator_, X_test, y_test, colorbar=False)
#Régression logistique, meilleur score sur le jeu de test est retenu#
model_2 = gs_clf_lr.best_estimator_
##Sauvegarde du modèle##
import joblib
joblib.dump(model_2, "model_2.pkl")

```

ANNEXE IV

Code Python de la classification automatique des modes de défaillance du stator

```

import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import re

data = pd.read_excel('df stator.xlsx')
df = data.copy()

df.shape

plt.figure(figsize=(12,8))

df['Classe'].value_counts().plot.pie(autopct="%.1f%%")

import nltk

from nltk.corpus import stopwords
from nltk.stem import SnowballStemmer

df['tokens'] = df['Texte'].apply(lambda x: preprocess(x))

from sklearn.preprocessing import LabelEncoder

encoder = LabelEncoder()

df['target'] = encoder.fit_transform(df['Classe'])

from wordcloud import WordCloud

classes = [cat_0, cat_1, cat_2]

titles = ['Court-circuit', 'Général', 'Refroidissement']

for i,j in zip(classes, titles):

    plt.figure(figsize=(10,7))

    separator = ','

    wordcloud = WordCloud(background_color='white', width=800, height=400).generate(separator.join(i))

    plt.imshow(wordcloud, interpolation='bilinear')

    plt.title(j)

    plt.axis("off")

    plt.show()

```

```

def token_to_text(tokens):
    text = ' '.join(tokens)
    return text

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(df['tokens'], df['target'], test_size=0.2, random_state=42 )
from sklearn.feature_extraction.text import CountVectorizer
count_vec = CountVectorizer()
X_train = X_train.apply(lambda x: token_to_text(x))
X_test = X_test.apply(lambda x: token_to_text(x))
vector = count_vec.fit_transform(X_train)

from sklearn.naive_bayes import MultinomialNB
from sklearn.linear_model import SGDClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression

clf_nb = MultinomialNB()
clf_sgd = SGDClassifier(random_state=42)
clf_tree = DecisionTreeClassifier(random_state=42)
clf_rf = RandomForestClassifier(random_state=42)
clf_lr = LogisticRegression(random_state=42)

from sklearn.model_selection import cross_val_score
score1 = cross_val_score(clf_nb, vector, y_train, cv=5)
print(score1.mean())

score2 = cross_val_score(clf_sgd, vector, y_train, cv=5)
print(score2.mean())

score3 = cross_val_score(clf_tree, vector, y_train, cv=5)
print(score3.mean())

score4 = cross_val_score(clf_rf, vector, y_train, cv=5)
print(score4.mean())

score5 = cross_val_score(clf_lr, vector, y_train, cv=5)

```

```

print(score5.mean())

from sklearn.pipeline import Pipeline

text_clf_sgd = Pipeline([
    ('vect', CountVectorizer()),
    ('tfidf', TfidfTransformer()),
    ('clf_sgd', SGDClassifier(random_state=42))
])

text_clf_rf = Pipeline([
    ('vect', CountVectorizer()),
    ('tfidf', TfidfTransformer()),
    ('clf_rf', RandomForestClassifier(random_state=42))])

text_clf_lr = Pipeline([
    ('vect', CountVectorizer()),
    ('tfidf', TfidfTransformer()),
    ('clf_lr', LogisticRegression(random_state=42))])

text_clf_tree = Pipeline([
    ('vect', CountVectorizer()),
    ('tfidf', TfidfTransformer()),
    ('clf_tree', DecisionTreeClassifier(random_state=42))])

from sklearn.model_selection import GridSearchCV

parameters_sgd = {
    'vect__ngram_range': [(1,1)],
    'vect__min_df': [n for n in range(1,6)],
    'tfidf__use_idf': (True, False),
    'clf_sgd__alpha': (1e-2, 1e-3, 1e-4, 1e-5, 1e-6),
    'clf_sgd__penalty': ['l1','l2']
}

gs_clf_sgd = GridSearchCV(text_clf_sgd, parameters_sgd, cv=5, n_jobs=-1)

```

```

gs_clf_sgd = gs_clf_sgd.fit(X_train, y_train)

parameters_rf = {
    'vect__ngram_range': [(1,1)],
    'vect__min_df': [n for n in range(1,6)],
    'tfidf__use_idf': (True, False),
    'clf_rf__bootstrap': (True, False),
    'clf_rf__criterion': ['gini','entropy'],
}

gs_clf_rf = GridSearchCV(text_clf_rf, parameters_rf, cv=5, n_jobs=-1)
gs_clf_rf = gs_clf_rf.fit(X_train, y_train)

parameters_lr = {
    'vect__ngram_range': [(1,1)],
    'vect__min_df': [1,2,3],
    'tfidf__use_idf': (True, False),
    'clf_lr__C': [1, 3, 5, 7, 10],
    'clf_lr__penalty': ['l1','l2'],
    'clf_lr__max_iter': [1000]
}

gs_clf_lr = GridSearchCV(text_clf_lr, parameters_lr, cv=5, n_jobs=-1)

gs_clf_lr = gs_clf_lr.fit(X_train, y_train)

parameters_tree = {
    'vect__ngram_range': [(1,1)],
    'vect__min_df': [n for n in range(1,6)],
    'tfidf__use_idf': (True, False),
    'clf_tree__criterion': ['gini','entropy'],
}

gs_clf_tree = GridSearchCV(text_clf_tree, parameters_tree, cv=5, n_jobs=-1)
gs_clf_tree = gs_clf_tree.fit(X_train, y_train)

```

```

from sklearn.metrics import classification_report, confusion_matrix

predicted_sgd = gs_clf_sgd.best_estimator_.predict(X_test)

print(classification_report(y_test, predicted_sgd, target_names = ['Court-circuit', 'Général',
'Refroidissement']))

from sklearn.metrics import plot_confusion_matrix

plot_confusion_matrix(gs_clf_sgd.best_estimator_, X_test, y_test, colorbar=False)

from sklearn.metrics import classification_report

predicted_rf = gs_clf_rf.best_estimator_.predict(X_test)

print(classification_report(y_test, predicted_rf, target_names = ['Court-circuit', 'Général',
'Refroidissement']))

plot_confusion_matrix(gs_clf_rf.best_estimator_, X_test, y_test, colorbar=False)

predicted_lr = gs_clf_lr.best_estimator_.predict(X_test)

print(classification_report(y_test, predicted_lr, target_names = ['Court-circuit', 'Général',
'Refroidissement']))

plot_confusion_matrix(gs_clf_lr.best_estimator_, X_test, y_test, colorbar=False)

predicted_tree = gs_clf_tree.best_estimator_.predict(X_test)

print(classification_report(y_test, predicted_tree, target_names = ['Court-circuit', 'Général',
'Refroidissement']))

plot_confusion_matrix(gs_clf_tree.best_estimator_, X_test, y_test, colorbar=False)

model_stator = gs_clf_rf.best_estimator_

import joblib

joblib.dump(model_stator, "model_stator.pkl")

```