République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université A/Mira de Béjaia
Faculté des Sciences Exactes
Département d'Informatique

MÉMOIRE DE MASTER

En Informatique

Option : ***Système d'Information Avancé***

Thème

Développement d'un système résilient de prédiction de défaillances pour les trains autonomes en utilisant l'apprentissage automatique.

Présenté par : M. **AGGOUNE Anouar**
M. **AMRANI Khalil**

Soutenu le : 02 juillet 2024

Devant le jury composé de :

Président	**M. L. KHENOUS**	**MCB**	**U. A/Mira Béjaia.**
Examinatrice	**Mme. D. KESSIRA**	**MAA**	**U. A/Mira Béjaia.**
Encadrant	**M. M. MOHAMMEDI**	**MCA**	**U. A/Mira Béjaia.**
Co-encadrante	**Mme. L.YAHIAOUI**	**Doctorante**	**U. A/Mira Béjaia.**

Béjaia, Juillet 2024.

⁂ *Remerciements* ⁂

Nous remercions Dieu le tout Puissant qui nous a donné la force et la volonté d'accomplir ce travail.

Nous souhaitons exprimer notre gratitude profonde à tous ceux qui ont contribué, de près ou de loin, à l'aboutissement de ce travail.

Nous remercions sincèrement notre encadrant, docteur Mohamed MOHAMMEDI, pour sa disponibilité, ses conseils avisés et son soutien constant tout au long de ce projet. Ses orientations précieuses ont été déterminantes pour la réalisation de ce mémoire.

Nous exprimons également nos vifs remerciements à notre co-encadrante, Mme. Lydia YAHIAOUI, pour son soutien, ses suggestions pertinentes et son encouragement continu. Sa collaboration et son expertise ont été d'une grande aide dans la concrétisation de ce travail.

Nos remerciements vont aussi au président et aux membres du jury pour l'honneur qu'ils nous font en acceptant d'évaluer notre travail. Nous apprécions grandement leur temps et leur expertise.

Nous tenons à remercier chaleureusement tous les enseignants et enseignantes qui ont enrichi notre parcours académique par leurs connaissances et leur dévouement. Leur contribution à notre formation est inestimable.

Enfin, nous adressons nos plus sincères remerciements à nos parents et familles, ainsi que nos amis pour leur soutien inébranlable et leur encouragement tout au long de notre cursus. Leur présence et leurs encouragements ont été une source de motivation essentielle.

Merci à toutes et à tous.

⁕ *Dédicaces* ⁕

Je dédie ce modeste travail :
À mes chers parents, pour leur soutien inconditionnel et
leur tolérance durant toutes mes années d'études.
À mes frères, Riyadh et Nassim
pour leur soutien, leurs encouragements et leur présence tout au long de mon parcours.
À mes sœurs, Amel et Nouria
pour leur amour et leur soutien.
À mes neveux,
qui apportent joie et inspiration.
À notre encadrant, Docteur Mohamed MOHAMMEDI, pour sa motivation
et son guidage constant tout au long de ce travail.
À notre co-encadrante, Mme. Lydia YAHIAOUI,
pour ses précieux conseils et son soutien continu.
À tous mes amis, pour les nombreux bons moments partagés, leurs encouragements constants, et leur soutien indéfectible.
Leur amitié et leur présence ont été une source de motivation et de joie tout au long de ce parcours.
À tous ceux qui m'ont aidé de près ou de loin
dans la réalisation de ce travail.
À tous mes camarades et toute la promotion Système d'Information Avancé,
pour leur camaraderie et leur soutien.
À tous ceux qui œuvrent pour que notre nation soit meilleure.
Je vous remercie tous pour votre soutien inconditionnel, votre amour et votre inspiration qui ont rendu possible l'aboutissement de ce travail.

AGGOUNE Anouar

⁕ *Dédicaces* ⁕

Je dédie ce modeste travail :
À mes très chers parents, pour leur amour inconditionnel, leur soutien indéfectible
et leurs sacrifices qui m'ont permis d'arriver où j'en suis aujourd'hui.
Vous êtes ma plus grande source d'inspiration et de motivation.
À mes frères bien-aimés, Younes, Tarek et Zakaria
pour leur présence, leur complicité et leurs encouragements constants.
Votre fraternité est un trésor inestimable.
À mes chères sœurs, Sara et Mouna pour leur affection, leur soutien moral et leur foi en moi. Votre douceur et votre force m'ont toujours guidé.
À notre encadrant, Docteur Mohamed MOHAMMEDI,
pour sa patience, sa disponibilité et ses précieux conseils qui ont été déterminants dans la réalisation de ce travail.
À notre co-encadrante, Mme. Lydia YAHIAOUI,
pour ses précieux conseils et son soutien continu.
À tous mes amis, pour les moments inoubliables partagés et leur soutien sans faille.
Votre amitié a illuminé ce parcours.
À mes camarades de la promotion Système d'Information Avancé,
pour leur esprit d'entraide et la solidarité qui nous a unis.
À tous ceux qui, de près ou de loin,
ont contribué à l'aboutissement de ce travail.
Ce mémoire est le fruit de vos encouragements et de votre confiance.
Merci du fond du cœur.

AMRANI Khalil

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ACRONYMES

APU	Air Production Unit (Unité de Production d'Air)
AUC	Area Under Curve (Surface Sous la Courbe)
ATO	Automatic Train Operation (Opération Automatique des Trains)
ATP	Automatic Train Protection (Protection Automatique des Trains)
CBR	Case-Based Reasoning (Raisonnement Basé sur les Cas)
CNN	Réseau de Neurones Convolutionnels
DNN	Deep Neural Network (Réseau de Neurones Profond)
FAR	False Alarm Rate (Taux de Fausses Alertes)
FDR	False Discovery Rate (Taux de Fausses Découvertes)
FN	False Negative (Faux Négatif)
FNN	Fuzzy Neural Network (Réseau de neurones flousflous)
GAN	Generative Adversarial Network (Réseau Génératif Antagoniste)
GoA	Grade of Automation (Niveau d'Automatisation)
IA	Intelligence Artificielle
IoT	Internet des Objets
KBS	Knowledge-Based System (Système à Base de Connaissances)
LSTM	Long Short-Term Memory (Mémoire à Long et Court Terme)
MAE	Mean Absolute Error (Erreur Absolue Moyenne)
MDP	Markov Decision Process (Processus de Décision Markovien)
MSE	Mean Squared Error (Erreur Quadratique Moyenne)
MTTF	Mean Time To Failure (Temps Moyen Avant Panne)
PSO	Particle Swarm Optimization (Optimisation par Essaim de Particules)
RBML	Rule-Based Machine Learning (Apprentissage Automatique Basé sur des Règles)
RF	Random Forest (Forêt Aléatoire)
RNN	Réseau Neuronal Récurrent

ROC Receiver Operating Characteristic (Caractéristique de Fonctionnement du Récepteur)

SVM Support Vector Machine (Machine à Vecteurs de Support)

SVR Support Vector Regression (Régression par Vecteurs de Support)

T-S FNN Takagi-Sugeno Fuzzy Neural Network (Réseau de Neurones Flou Takagi-Sugeno)

UITP Union Internationale des Transports Publics

VP True Positive (Vrai Positif)

VN True Negative (Vrai Négatif)

XGBoost eXtreme Gradient Boosting

INTRODUCTION GÉNÉRALE

Les avancées technologiques récentes dans les domaines de l'Intelligence Artificielle (IA), l'Internet des Objets (IoT) et les systèmes embarqués ont ouvert la voie au développement de systèmes de prédiction de défaillance résilients et sophistiqués [1]. L'IA, en exploitant des techniques d'apprentissage automatique avancées telles que les réseaux de neurones convolutifs et les autoencodeurs, permet d'analyser d'immenses quantités de données et de détecter des modèles complexes indicateurs de défaillances potentielles [2]. Parallèlement, l'IoT et les systèmes embarqués facilitent la collecte en temps réel de données provenant de multiples capteurs et composants, offrant ainsi une vue d'ensemble complète de l'état du système [3]. L'intégration de ces technologies a rendu possible le développement de systèmes de prédiction de défaillance résilients, capables de s'adapter aux conditions changeantes, de tolérer les pannes partielles et de maintenir leur fonctionnalité malgré les perturbations. Ces systèmes visent à améliorer considérablement la fiabilité et la sécurité des opérations, en particulier dans des domaines critiques comme le transport ferroviaire.

Les systèmes de transport ferroviaire conventionnels reposent largement sur la surveillance et le contrôle humains, pour assurer la sécurité et l'efficacité des opérations. Cependant, avec l'émergence des trains autonomes, la détection et la prévention des défaillances doivent être entièrement automatisées et intégrées dans les systèmes de contrôle et de prise de décision. Les défaillances peuvent survenir à différents niveaux, allant des composants matériels aux logiciels de commande en passant par les systèmes de communication et de navigation [4]. Un système de prédiction de défaillance résilient doit être capable de détecter les anomalies, d'identifier leurs causes sous-jacentes et de prendre des mesures correctives appropriées en temps réel. De plus, ce système doit être conçu pour être robuste face aux conditions environnementales variables, aux perturbations externes et aux imprévus qui peuvent survenir lors de l'exploitation des trains.

Les trains autonomes sont équipés de nombreux capteurs intégrés qui collectent en continu des données sur l'état des différents composants, tels que les moteurs, les freins, les systèmes de signalisation et de contrôle. Ces données sont transmises en temps réel à une unité de traitement centrale, où elles sont analysées par le système de prédiction de défaillance. Cependant, la transmission et le stockage de ces données soulèvent des préoccupations majeures en matière de sécurité. Toute altération ou corruption des données pourrait compromettre gravement la fiabilité du système de prédiction de

défaillance et entraîner des décisions erronées, avec des conséquences potentiellement catastrophiques. Par conséquent, des mesures de sécurité robustes doivent être mises en place pour garantir l'intégrité et la confidentialité des données tout au long de leur cycle de vie, depuis leur collecte jusqu'à leur analyse.

Dans le cadre de notre mémoire, l'objectif est de proposer une approche novatrice qui consiste à prédire les défaillances potentielles dans les métros jusqu'à trois jours avant leur occurrence effective. Notre système a été conçu en mettant l'accent sur la résilience, garantissant sa fiabilité et son efficacité dans diverses conditions opérationnelles. Cette résilience se manifeste par sa robustesse face aux données bruitées ou incomplètes, son adaptabilité aux changements des conditions d'exploitation, et sa capacité à gérer efficacement les cas de défaillance. Cette approche repose sur l'utilisation de l'algorithme XGBoost [5], un modèle d'apprentissage automatique basé sur les arbres de décision boostés. En exploitant des données massives collectées par de multiples capteurs installés sur l'unité de production d'air d'un train de métro, notre modèle entraîné vise à détecter et caractériser les schémas précurseurs plusieurs jours avant qu'ils ne se produisent. Cette capacité de prédiction anticipée offre un délai précieux pour mettre en œuvre des mesures préventives et des interventions de maintenance.

Ce mémoire est organisée comme suit : le **Chapitre 1** présente les généralités sur les trains autonomes et la prédiction de défaillance, fournissant les bases nécessaires à la compréhension du sujet. Le **Chapitre 2** effectue une revue approfondie de la littérature existante sur les systèmes de prédiction de défaillance résilients pour les trains autonomes, mettant en évidence les avancées récentes et les défis à relever. Le **Chapitre 3** décrit en détail la conception et le développement de notre système de prédiction de défaillance basé sur XGBoost, en couvrant les aspects liés à la collecte et au prétraitement des données, à l'entraînement du modèle, ainsi qu'à l'intégration du système dans l'environnement des trains autonomes. Enfin, le **Chapitre 4** présente les expérimentations menées et les résultats obtenus, évaluant les performances de notre système et discutant de ses implications pour les opérations ferroviaires autonomes.

Enfin, ce document s'achève par une conclusion générale et des perspectives de recherche. Cette partie finale récapitule les points clés de notre étude sur la prédiction de défaillances dans les systèmes de métro. Nous y réfléchissons sur les implications potentielles de cette approche pour l'amélioration de la maintenance prédictive et de la sécurité dans les transports ferroviaires. Les perspectives de recherche présentent des pistes prometteuses pour poursuivre et élargir cette étude.

CHAPITRE 1

PRÉ-REQUIS THÉORIQUES SUR LES TRAINS AUTONOMES ET LA PRÉDICTION DE DÉFAILLANCES

1.1 Introduction

Avec les avancées technologiques récentes, y compris l'IA, l'IOT et les systèmes intégrés, les trains autonomes deviennent de plus en plus populaires dans le secteur ferroviaire. Ces trains automatisés fonctionnent sans intervention humaine directe, ils sont pilotés par un système sophistiqué qui gère la navigation, la sécurité et l'entretien. L'automatisation ferroviaire représente une avancée révolutionnaire susceptible d'améliorer considérablement l'efficacité opérationnelle. Cependant, elle soulève également de nouveaux défis tels que la maintenance et la gestion des pannes.

Une autre fonction importante des trains autonomes est la capacité de prédire les pannes. Cela nous offrira la possibilité de prévoir les pannes à l'avance, par conséquent, la possibilité de prendre des mesures de sécurité avant qu'une panne coûteuse et potentiellement catastrophique ne se produise. La prévision des pannes repose sur le traitement des données collectées par les multiples capteurs installés sur les trains. Il serait possible d'inclure dans ces données les performances du moteur, les performances du système de freinage, les performances du système de signalisation et l'état de la route. L'analyse de ces données collectées, en passant par plusieurs étapes de traitements, nous permettra de réaliser un système de prédiction d'anomalies, qui va servir à plannifer d'avance la maintenance du composant en question, ce qui faciliterait un service sûr et sans accidents des trains.

Dans ce chapitre, nous commençerons par définir le concept des trains autonomes, et ce qui les rend différents d'autres trains, puis, nous allons définir l'unité de production d'air, l'un des composants essentiels des trains. Ensuite, nous allons parler de la maintenance prédictive en générale, ainsi que dans le domaine ferroviaire. Enfin, nous discuterons de l'importance de la prédiction de défaillances pour la sécurité, ainsi que certains obstacles que nous puissions rencontrer lors de la réalisation de notre système.

1.2 Progrés technologiques industriels

L'évolution technologique dans le secteur ferroviaire a permis des avancées significatives vers l'automatisation et l'efficacité des systèmes de transport. La détection d'anomalies et la maintenance prédictive jouent un rôle crucial dans l'amélioration de la sécurité et de la fiabilité des services ferroviaires. Cette section explore les bases de ces technologies avant de se concentrer spécifiquement sur les trains autonomes, la détection d'anomalies et les stratégies de maintenance.

1.2.1 Trains autonomes

Un train autonome est un train qui est tracté et freiné par un autopilote, en suivant les instructions de signalisation et le calendrier. L'auto-pilotage peut être réalisé avec un conducteur ou sans l'accompagner [6]. Ces trains reposent sur des systèmes automatisés avancés pour la navigation, la detection des obstacles et la gestion de la viteses.

L'ensemble de systèmes visant à automatiser les opérations des trains est appellé ATO (Automatic Train Operation), est couplé avec la Protection Automatique du Train (ATP), qui supervise le train en détectant les excès de vitesse et les franchissements de points dangereux, et active le frein d'urgence en cas d'incident.

L'Association Internationale du Transport Public (UITP) a défini quatre niveaux d'automatisation, appelés Grades d'Automatisation (GoA), allant de GoA1 à GoA4 [7], comme illustré dans la Figure 1.1.

Niveau d'automatisation	Type d'opération du train	Mise en marche du train	Arrêt du train	Fermeture des portes	Opération en cas de perturbation
GoA1 	ATP* avec conducteur	Conducteur	Conducteur	Conducteur	Conducteur
GoA2 	ATP et ATO* avec conducteur	Automatique	Automatique	Conducteur	Conducteur
GoA3 	Sans conducteur	Automatique	Automatique	Agent de bord	Agent de bord
GoA4 	**Exploitation Automatique des Trains (Sans surveillance)**	Automatique	Automatique	Automatique	Automatique

FIGURE 1.1 – Niveaux d'automatisation [8].

1.2.2 Unité de production d'air (APU)

L'unité de production d'air assure la production d'air comprimé, qui est utilisée dans les sacs à air de la suspension secondaire, afin de créer un coussin à air qui réduit la vibration des cabines du véhicule, et le nivele de manière à maintenir sa hauteur constante par rapport à la route [9].

Les trois principales zones de l'APU sont illustrées dans le diagramme de la Figure 1.2.

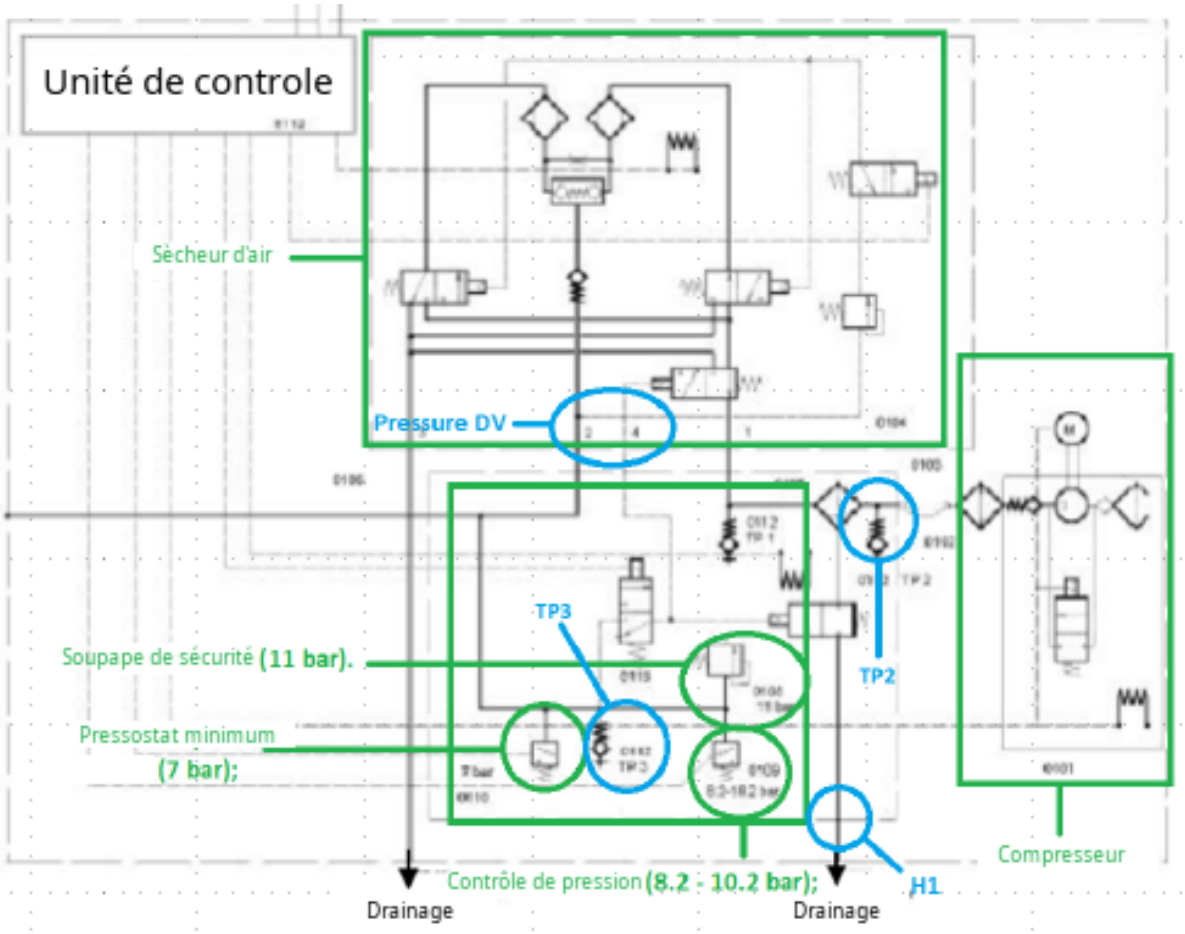

FIGURE 1.2 – Illustration d'une Unité de Production d'Air (APU) [9].

La figure met en évidence les trois principales zones de l'APU : le compresseur d'air, le séchage de l'air et le contrôle de la pression. Chaque zone joue un rôle crucial dans le fonctionnement global de l'APU, assurant ainsi une performance optimale et une stabilité accrue du véhicule ferroviaire.

1.2.3 Détection d'anomalies

En dépit des avancées technologiques, l'industrie n'est pas exempt de défis, des anomalies peuvent survenir, perturbant ainsi le fonctionnement optimal des systèmes. La détection et la gestion de ces anomalies sont cruciales pour garantir la fiabilité et la sécurité continues des trains autonomes.

Anomalies

Dans le contexte industriel, une anomalie peut être définie comme une déviation du comportement standard ou de la norme attendue. Les anomalies peuvent indiquer des problèmes critiques tels que des défauts, des dysfonctionnements ou des inefficacités opérationnelles nécessitant une attention particulière [10].

Les anomalies dans les séries temporelles peuvent être divisées en 3 catégories :

— **Anomalies globales** : Aussi connues sous le nom d'anomalies ponctuelles, les anomalies globales sont celles qui se manifestent en dehors des données. Ce sont les points de données qui diffèrent le plus par rapport aux autres données dans un ensemble donné [10].

La Figure 1.3 illustre une anomalie globale.

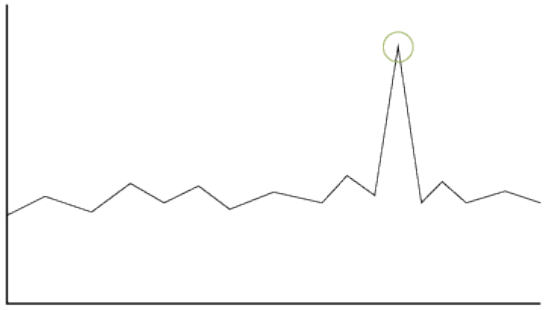

FIGURE 1.3 – Illustration d'une anomalie globale [10].

— **Anomalies contextuelles** : Aussi connues sous le nom d'anomalies conditionnelles, ces anomalies diffèrent grandement des autres données dans le même jeu de données, en fonction d'un contexte spécifique ou d'une condition unique [10].

La Figure 1.4 est une illustration d'une anomalie contextuelle.

FIGURE 1.4 – Illustration d'une anomalie contextuelle [10].

— **Anomalies collectives** : Points de données qui diffèrent significativement du reste de l'ensemble de données. Pris individuellement, ils ne sont pas nécessairement des valeurs aberrantes, mais combinés avec un autre ensemble de données temporelles, ils agissent collectivement comme des anomalies [10].

La Figure 1.5 illustre une anomalie collective.

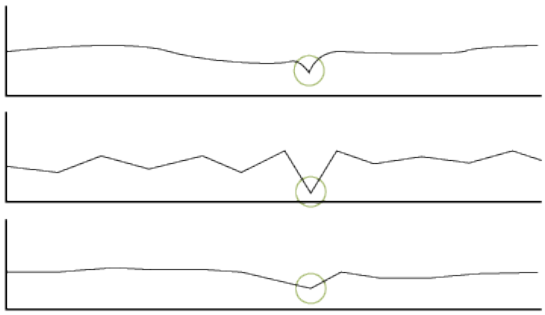

FIGURE 1.5 – Illustration d'une anomalie collective [10].

Les techniques de détection des anomalies suivent des approches différentes, et cela en se basant sur trois scénarios principaux [11] :

— L'ensemble de données n'a pas de caractéristiques temporaires et les instances sont indépendantes.
— Le jeu de données suit un ordre séquentiel et il est supposé que les données sont générées suivant des flux.
— Le jeu de données est produit sous forme de séries temporelles.

1.2.4 Maintenance

La maintenance est le processus par lequel les composants, l'équipement, ou les systèmes sont entretenus pour un fonctionnement approprié en tout temps [12]. Avec la révolution technologique, plusieurs approches de maintenance ont été proposées. En général, les trois principales techniques de maintenance peuvent être divisées comme suit :

a) Maintenance Corrective

Il s'agit de la méthode la plus simple et la plus ancienne, appelée "run-to-failure". L'idée est d'intervenir uniquement après une défaillance de la machine ou de l'équipement. Cela implique presque toujours des temps d'arrêt élevés et imprévus, ainsi que des coûts élevés pour le personnel de maintenance. Cette méthode génère souvent des situations critiques et coûteuses pour les entreprises [12].

b) Maintenance Preventive

Elle consiste à planifier le remplacement régulier des composants et/ou équipements. Se basant sur les données historiques de défaillance et/ou les informations fournies par le fabricant, le MTTF (Mean Time To Failure) est calculé et utilisé par l'équipe de maintenance pour proposer un plan d'action préventive. Bien que cette approche prévienne les arrêts imprévus, elle nécessite généralement des coûts supplémentaires et entraîne une durée de vie inutilisée augmentée [12].

c) Maintenance Prédictive

La maintenance prédictive est une méthode qui analyse les conditions réelles et le fonctionnement des équipements et des systèmes, afin de déterminer le temps moyen jusqu'à la défaillance ou la perte d'efficacité réelle. Elle aide à réduire les pannes imprévues, à repérer les problèmes avant qu'ils ne se développent, et à prévenir les réparations importantes en repérant les problèmes à temps [13].

Les étapes d'un model de maintenance prédictive est illustré dans la Figure 1.6.

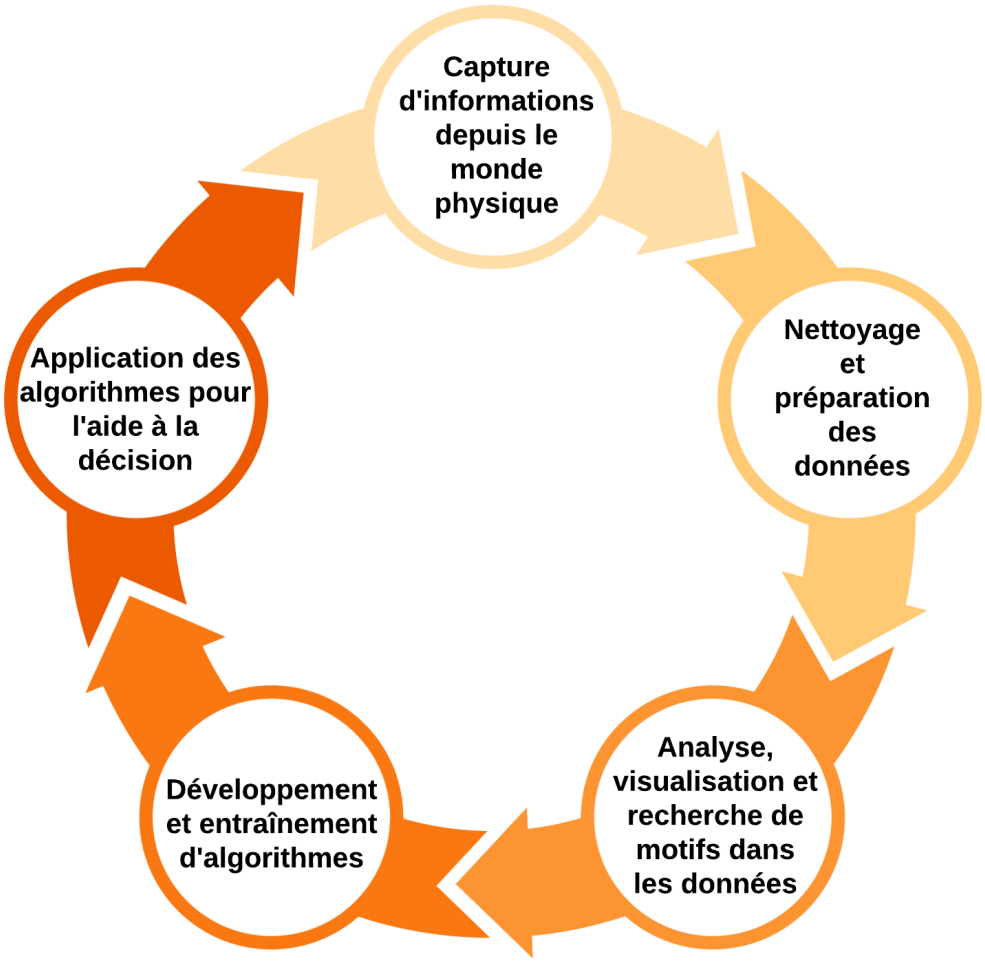

FIGURE 1.6 – Cycle de la maintenance prédictive.

Les trois (3) techniques de maintenance sont illustrés dans la Figure 1.7.

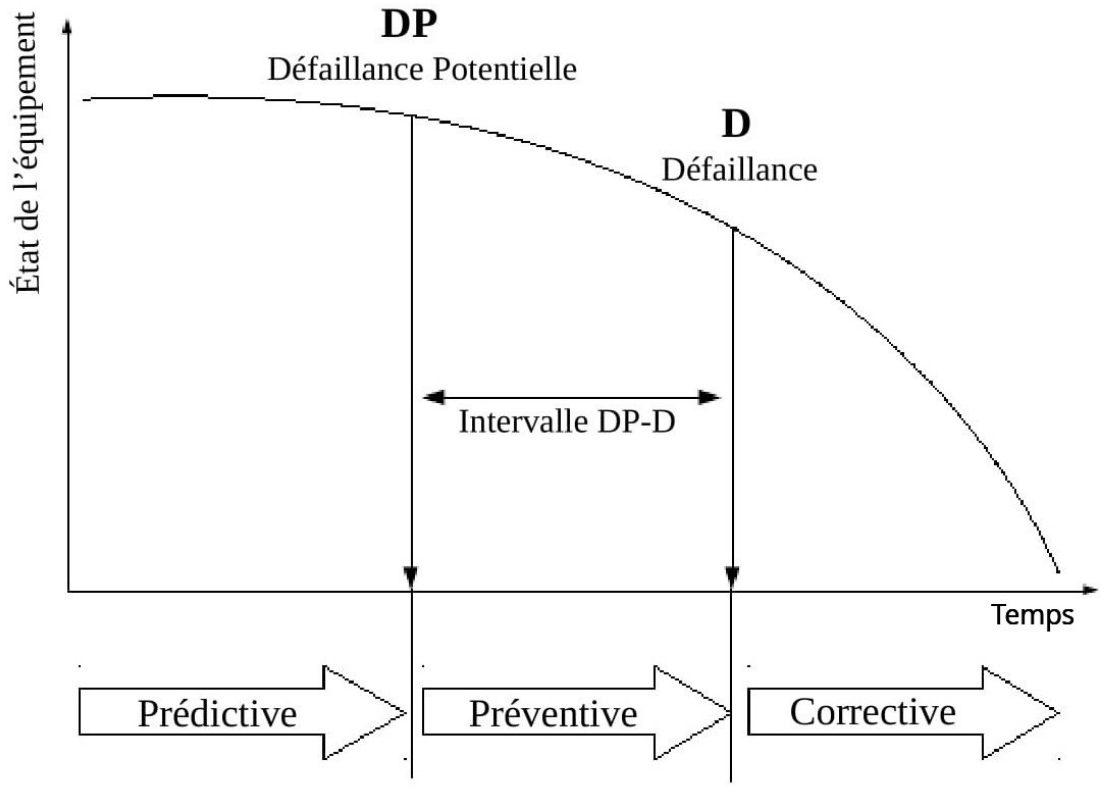

FIGURE 1.7 – Illustration des approches de maintenance [14].

Cette Figure présente une visualisation des trois principales approches de maintenance utilisées dans les systèmes industriels : corrective, préventive et prédictive. Chacune de ces approches se distingue par sa méthode, son temps d'intervention et son application spécifique ; influençant ainsi la gestion des équipements et des systèmes.

1.3 Les approches utilisées pour la maintenance prédictive

La maintenance prédictive repose sur diverses approches et technologies pour surveiller et analyser les conditions de fonctionnement des équipements et systèmes. Les approches couramment utilisées incluent des techniques d'apprentissage automatique classique, apprentissage profond et des approches basées sur données. Chacune de ces méthodes offre des avantages spécifiques et contribue à une gestion de la maintenance plus efficace.

1.3.1 Méthodes d'apprentissage automatiques classiques

On peut mentionner parmi les techniques d'apprentissage machine classiques les réseaux de neurones à rétropropagation (DNN), les forêts d'arbres décisionnels (RF : Random Forests), les machines à vecteurs de support (SVM) et la régression logistique. Ces algorithmes ont été utilisés souvent afin de classer les défauts, prédire les défaillances et estimer la durée de vie résiduelle des composants [12].

Parmi les techniques avancées d'apprentissage automatique, le boosting s'est distingué comme une approche particulièrement efficace.

Le Boosting

Le boosting est une méthode d'apprentissage automatique, qui consiste à combiner de manière itérative plusieurs modèles faibles pour en créer un modèle final plus puissant. L'idée principale est d'entraîner une séquence de modèles simples, où chaque modèle successeur cherche à corriger les erreurs du modèle predecesseur, en accordant plus d'importance aux exemples mal classés [15].

Le boosting consiste à entraîner chaque modèle faible sur une version ajustée du jeu de données d'entraînement. Les exemples mal classés par le modèle predecesseur se voient attribuer un poids plus élevé, ce qui fait que le modèle successeur se concentre davantage sur ces exemples difficiles. Ce processus est répété de manière itérative jusqu'à ce qu'on obtienne un modèle final, qui regroupe tous les modèles faibles [16].

Cette approche permet de créer un modèle final très précis et robuste, en tirant parti de la force de chaque modèle faible, tout en corrigeant leurs faiblesses respectives. Le boosting a montré d'excellentes performances dans de nombreuses tâches d'apprentissage automatique, notamment la classification et la régression.

La Figure 1.8 montre le schéma du fonctionnement du boosting.

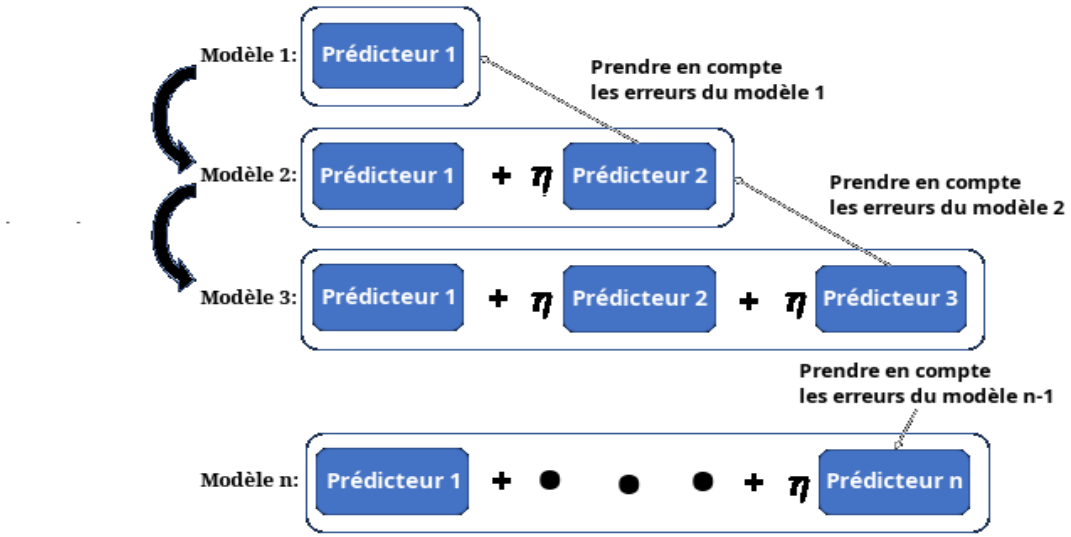

FIGURE 1.8 – Schéma du fonctionnement du boosting [17].

Dans ce contexte, eXtreme Gradient Boosting (XGBoost) a émergé comme l'un des algorithmes les plus performants et les plus populaires [5].

XGBoost

XGBoost, abréviation de "eXtreme Gradient Boosting", est une implémentation optimisée et distribuée de l'algorithme de boosting de gradient. Il a été développé par Tianqi Chen et Carlos Guestrin de l'Université de Washington [5].

D'après Chen et Guestrin [5], XGBoost est "un outil d'optimisation de pointe conçu pour être hautement efficace, flexible et portable. Il implémente les machines à vecteurs de support, les forêts d'arbres décisionnels et la linéarisation de gradient par arbre". Le succès de XGBoost réside dans ses performances exceptionnelles, et sa capacité à gérer de manière efficace les jeux de données de grande taille et de grande dimensionnalité. Il a été largement utilisé dans diverses applications, telles que la reconnaissance d'images, la prédiction de clics publicitaires et la détection de fraudes, entre autres.

1.3.2 Méthodes d'apprentissage profond

Les méthodes d'apprentissage profond ont récemment connu une croissance de leur popularité en raison de leur capacité à extraire automatiquement des caractéristiques pertinentes à partir de données brutes. Les réseaux de neurones convolutifs (CNN) ont été employés afin de repérer les imperfections dans les engrenages planétaires et les roulements en se basant sur des signaux vibratoires [12]. En outre, les réseaux de mémoire longue et courte durée sont encore utilisés pour les données basées sur les séries temporelles.

En plus des méthodes mentionnées ci-dessus, d'autres approches avancées peuvent être utilisées pour réaliser des systèmes de maintenance prédictive, comme les réseaux antagonistes génératifs (GAN) et les autoencodeurs variationnels pour traiter les problèmes de données insuffisamment étiquetées, en plus de la modélisation des défaillances à partir de données de capteurs.

1.4 Contexte et enjeux des trains autonomes

Les trains autonomes sont une évolution majeure pour l'industrie ferroviaire, ouvrant la voie vers une mobilité plus sûre, plus efficace et plus durable. Cependant, leur déploiement à grande échelle soulève des défis de taille. D'abord, suite à l'absense de l'intervention humaine, il est essentiel d'avoir des systèmes de contrôle et de surveillance extrêmement fiables, afin d'assurer une exploitation sans faille. Il est essentiel de détecter rapidement toute anomalie ou défaillance éventuelle, afin d'éviter les accidents aux conséquences dramatiques. Par la suite, l'utilisation de réseaux entièrement automatisés requiert une coordination, en plus d'une gestion des flux de trafic d'une précision extrême, dans le but de garantir une ponctualité excellente.

Par ailleurs, les défis liés à la cybersécurité deviennent essentiels pour ces systèmes étroitement liés et automatisés, qui sont exposés aux cybermenaces. Enfin, la question de l'acceptabilité sociale de cette nouvelle technologie pose un défi, demandant des efforts de sensibilisation pour rassurer le grand public de la fiabilité et de la sécurité de ces trains "sans conducteur". En dépit de ces défis complexes, les trains

autonomes offrent la possibilité d'une mobilité plus écologique, plus sécurisée et plus efficace, répondant ainsi aux exigences de développement durable et de sécurité qui influencent l'avenir des transports.

1.5 Importance de la prédiction de défaillance pour la sécurité

La sécurité représente un enjeu primordial pour l'industrie ferroviaire, où une simple défaillance peut causer des conséquences catastrophiques. La prédiction de défaillance, rendue possible grâce aux avancées de l'apprentissage automatique, s'est montré crutiale pour la protection contre les accidents potentiellement meurtriers. Les composants ferroviaires critiques tels que les rails, les aiguillages, les freins ou les moteurs sont soumis à des conditions d'utilisation extrêmes, ce qui les exposent à une usure et une dégradation constantes. Une défaillance soudaine de l'un de ces éléments pourrait provoquer un déraillement, une collision ou toute autre catastrophe mettant en péril la vie des passagers et des employés.

C'est ici que la prédiction de défaillance entre en jeu, en tirant parti des données massives collectées par les multiples capteurs déployés sur les trains et l'infrastructure ferroviaire. Ce qui permet d'anticiper les défaillances avant qu'elles ne surviennent, offrant une fenêtre d'intervention cruciale pour effectuer une maintenance préventive ciblée. En évitant les pannes inattendues, ces systèmes de prédiction contribuent à prévenir les scénarios catastrophiques, et à assurer un niveau de sécurité optimal pour les usagers du rail.

Au-delà de sauver des vies, la prédiction de défaillance apporte également des bénéfices opérationnels et économiques substantiels, en réduisant les temps d'immobilisation non planifiés et les coûts de réparation d'urgence. Cependant, son rôle primordial reste la préservation de l'intégrité du système ferroviaire, ainsi que la protection des vies humaines contre les conséquences dévastatrices d'une défaillance majeure.

1.6 Conclusion

En conclusion, les trains autonomes représentent une avancée technologique significative dans le domaine du transport ferroviaire, offrant des avantages en termes de sécurité, d'efficacité et de fiabilité. Les données des capteurs sont collectées et analysées en temps réel, ce qui permet de surveiller les performances des trains et de prédire les défaillances avant qu'elles ne se produisent. La capacité à anticiper les problèmes et à intervenir de manière proactive est essentielle pour garantir la sécurité des opérations ferroviaires autonomes. Cependant, plusieurs obstacles et défis nous opposent, ce qui rend la mise en place de ces systèmes difficile, en particulier en ce qui concerne la sécurité des données et la précision des modèles de prédiction. Il serait judicieux de se focaliser dans les recherches et développements à venir dans ce domaine, sur l'amélioration de la résilience des systèmes de prédiction de défaillance, encore, sur l'intégration de technologies avancées, afin de répondre aux exigences croissantes du secteur.

Dans le chapitre suivant, nous examinerons l'état actuel des systèmes de prédiction de défaillance propsés, en examinant les méthodes actuelles et les perspectives à venir pour améliorer la sécurité et la fiabilité des trains intelligents.

CHAPITRE 2

ÉTAT DE L'ART SUR LES SYSTÈMES RÉSILIENTS DE PRÉDICTION DE DÉFAILLANCE POUR LES TRAINS AUTONOMES

2.1 Introduction

Les systèmes de prédiction de défaillance pour les trains autonomes font face à de nombreux défis et les solutions proposées sont également nombreuses. Le défi majeur réside dans la prédiction précise des défaillances, car les erreurs de prédiction peuvent non seulement engendrer des coûts opérationnels élevés, mais aussi compromettre la sécurité des passagers. Ainsi, il est essentiel de développer de nouvelles techniques de prédiction en utilisant l'apprentissage automatique.

Dans ce chapitre nous commençerons par une revue des techniques appliquées à la prédiction de défaillances, en mettant l'accent sur les approches les plus prometteuses et les plus largement utilisées dans ce domaine. Nous discuterons également des avantages et des inconvénients de ces techniques, ainsi que des défis associés à leur mise en œuvre dans un contexte réel. Enfin, nous identifierons les lacunes dans la recherche existante et les opportunités pour de futures recherches. Nous discuterons des directions potentielles pour de nouvelles recherches qui pourraient contribuer à améliorer encore la prédiction de défaillance dans les systèmes des trains autonomes.

2.2 Critères d'évaluation des systèmes existants

Dans cette section, nous examinons les critères essentiels pour évaluer les performances des systèmes de prédiction de défaillance destinés aux trains autonomes. L'évaluation de ces systèmes repose sur une série de critères variés, chacun apporte un éclairage spécifique sur leur efficacité et leur robustesse. Nous analyserons en détail ces critères afin de comprendre comment les systèmes existants se positionnent en termes de qualité, de fiabilité et d'adéquation aux besoins opérationnels des trains autonomes.

2.2.1 Exactitude

L'exactitude est une mesure de performances qui évalue la proportion des prédictions correctes par rapport au nombre total d'instances évaluées. Il s'agit d'un des critères essentiels employés afin d'évaluer les résultats d'un modèle de classification. Une exactitude élevée suggère que le modèle a classé de manière précise la plupart des instances, ce qui laisse entendre qu'il est capable de distinguer de manière fiable entre les différentes classes [18]. Cela peut être représenté mathématiquement comme suit :

$$\text{Exactitude} = \frac{VP + VN}{VP + FP + VN + FN} \tag{2.1}$$

2.2.2 Précision

Dans le domaine de l'apprentissage automatique, la précision est évaluée en fonction de la proportion des résultats réels positifs par rapport au nombre total de prédictions positives du modèle. Donc, nous pouvons définir la précision de la manière suivante :

$$\text{Précision} = \frac{VP}{VP + FP} \tag{2.2}$$

Un modèle qui ne génère aucune faux positif reçoit une précision de 1.0 [19].

2.2.3 Rappel

La mesure de performance appelée Rappel (Recall) évalue la capacité d'un modèle à repérer de manière précise toutes les instances positives parmi le nombre total d'instances positives dans l'ensemble de données [20], on peut définir le rappel de la manière suivante :

$$\text{Rappel} = \frac{VP}{VP + FN} \tag{2.3}$$

2.2.4 Temps de réponse

Le temps de réponse d'un système de prédiction d'anomalies désigne la durée requise pour que le système réponde à une demande ou détecte et informe les chercheurs d'une anomalie. Elle mesure l'efficacité du système dans la détection rapide des anomalies et dans la mise en place des mesures appropriées [21].

2.2.5 Robustesse

La robustesse fait référence à la capacité d'un algorithme d'apprentissage automatique à faire face à des circonstances difficiles. Dans le domaine de l'intelligence artificielle, la robustesse est une mesure qui évalue la capacité d'un modèle ou d'un algorithme à maintenir des performances élevées même en cas de données bruitées, d'erreurs de mesure, de valeurs aberrantes ou d'autres perturbations [22].

2.2.6 Scalabilité

La scalabilité désigne la faculté d'un système, d'un réseau ou d'un projet à gérer la croissance sans modifier ses principes directeurs, c'est-à-dire à aménager davantage de matériel avec un coût minimal [23]. Dans le domaine de l'intelligence artificielle, la scalabilité mesure la capacité des systèmes à gérer des ensembles de données de plus en plus volumineux, ce qui est crucial pour des applications comme

l'analyse en temps réel de données, l'apprentissage sur des ensembles de données massifs, ainsi que la mise en place de modèles sur des plateformes distribuées.

2.2.7 F1-Score

F1-score est une métrique d'évaluation des performances des modèles de classification. On le retrouve principalement dans les situations où des données déséquilibrées sont utilisées, telles que la détection de fraudes ou la prédiction d'incidents graves.

Le F1-score permet d'exprimer en une seule mesure les valeurs de précision et de rappel. En mathématiques, on définit le F1-score comme la moyenne harmonique de la précision et du rappel [24], ce qui se traduit par l'équation suivante :

$$\mathsf{F1-score} = \frac{\mathsf{VP}}{\mathsf{VP} + \frac{1}{2}(\mathsf{FN} + \mathsf{FP})} \tag{2.4}$$

2.2.8 Spécificité

La spécificité représente la proportion de vrais négatifs correctement identifiés parmi tous les cas réellement négatifs [25]. La formule de la spécificité est la suivante :

$$\text{Spécificité} = \frac{\mathsf{VN}}{\mathsf{VN} + \mathsf{FP}} \tag{2.5}$$

Cette métrique mesure la capacité du test à identifier correctement les cas négatifs.

2.2.9 Area Under the Curve - Receiver Operating Characteristic (AUC ROC)

L'AUC ROC est une mesure de la performance d'un modèle de classification qui représente la capacité du modèle à distinguer entre les classes. Elle correspond à l'aire sous la courbe ROC, qui trace le taux de vrais positifs en fonction du taux de faux positifs à différents seuils de classification [26].

2.3 Taxonomie des travaux examinés

Afin de rendre l'analyse des divers travaux étudiés plus facile, nous les avons répartis en trois grandes catégories en fonction de leur approche principale. La catégorie initiale comprend les « Systèmes de prédiction et d'analyse des défaillances », qui se concentre sur la réalisation de systèmes permettant de prédire et d'analyser les éventuelles défaillances des composants ou sous-systèmes. Dans la seconde catégorie, on retrouve les « Systèmes de surveillance de l'état du système et de détection des anomalies », qui ont pour but de surveiller en temps réel l'état des divers éléments du système, pour repérer les anomalies ou les dégradations éventuelles. Finalement, la troisième catégorie comprend les "Systèmes axés sur l'évaluation des risques et l'amélioration de la maintenance", qui associent fréquemment des méthodes de prédiction de pannes à des modèles d'optimisation, afin de planifier de manière efficace les activités de maintenance, de surcôit, réduire les risques liés aux pannes. La classification des travaux examinés est illustrée par la taxonomie présentée dans la Figure 2.1.

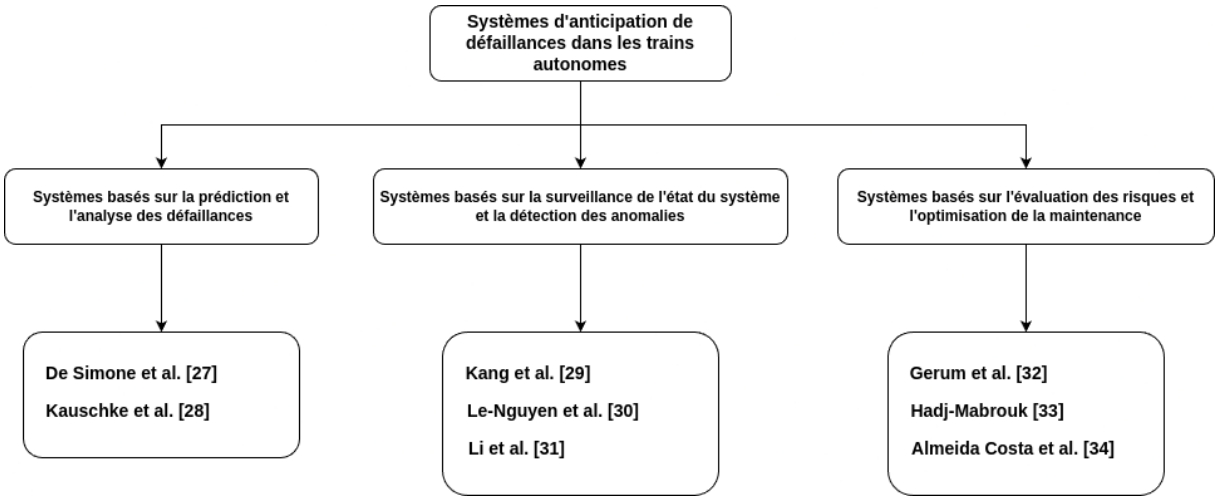

FIGURE 2.1 – Taxonomie des travaux mis en revue.

2.4 Travaux antérieurs

2.4.1 Systèmes basés sur la prédiction et l'analyse des défaillances

LSTM-based failure prediction for railway rolling stock equipment

De Simone et al. [27], ont proposé une méthodologie basée sur les réseaux de neurones LSTM (Long Short-Term Memory) pour la maintenance prédictive des équipements de matériel roulant ferroviaire. L'objectif principal est de développer une approche permettant d'apprendre correctement les dépendances à long terme dans les données évoluant graduellement, caractéristiques des systèmes ferroviaires, en prédisant les défaillances des composants critiques spécifiquement le sous-système de refroidissement du convertisseur de traction. La méthodologie se décompose en trois étapes clés. Premièrement, une exploration des données est réalisée à l'aide de tests statistiques comme le test de Mann-Kendall et l'analyse de covariance afin d'identifier les tendances et les corrélations entre les caractéristiques et les défaillances. Deuxièmement, un prétraitement rigoureux des données brutes est effectué, incluant le nettoyage, le filtrage des pics considérés comme du bruit, l'extraction des itinéraires des trains et des séquences temporelles pertinentes. Enfin, l'analyse des défaillances proprement dite est mise en œuvre via des réseaux LSTM entraînés pour deux tâches distinctes mais complémentaires : une tâche de classification multiclasses attribuant un niveau de sévérité (Bon, Mineur, Majeur) à chaque séquence, et une tâche de prévision visant à prédire les tendances futures du signal de défaillance sur une période de temps spécifique.

Predicting Cargo Train Failures : A Machine Learning Approach for a Lightweight Prototype

Kauschke et al. [28], ont proposé une approche en deux niveaux pour prédire les défaillances de convertisseurs de puissance sur les locomotives de fret. Leur objectif est de développer un système léger de maintenance prédictive qui peut être implémenté sur les systèmes existants sans nécessiter de modifications coûteuses. Au premier niveau, ils entraînent un classificateur d'instances sur les messages de diagnostic générés par les engins, en étiquetant les instances comme positives (indiquant une défaillance

potentielle) ou négatives en fonction d'une fenêtre glissante avant la défaillance réelle. Au deuxième niveau, un méta-classificateur prend les sorties du classificateur d'instances et classe l'ensemble du trajet comme un incident ou non, en fonction du pourcentage d'instances positives. Ils utilisent des techniques d'apprentissage automatique simples, notamment des classificateurs basés sur des règles et des arbres de décision, qui peuvent être facilement implémentés dans les systèmes existants. L'approche de Kauschke et al. est évaluée par une méthode de validation croisée leave-one-tour-out sur des données comprenant 40 trajets avec défaillance et 140 sans défaillance. Cette méthode légère et flexible vise à fournir des prédictions fiables de défaillance avec suffisamment d'avance, tout en minimisant les faux positifs coûteux. Elle permet de gagner la confiance des ingénieurs avant une intégration plus poussée dans les systèmes existants. Les auteurs envisagent également d'autres améliorations futures, comme l'optimisation des paramètres, l'utilisation d'ensembles de classificateurs et l'exploration d'algorithmes plus complexes une fois la confiance établie.

2.4.2 Systèmes basés sur la surveillance de l'état du système et la détection des anomalies

A method of online anomaly perception and failure prediction for high-speed automatic train protection system

Kang et al. [29], ont proposé une approche novatrice pour l'exploitation et la maintenance intelligente des systèmes de protection automatique des trains (ATP) utilisés sur les lignes ferroviaires à grande vitesse. Leur objectif principal est d'améliorer les capacités de protection et l'efficacité opérationnelle de ces systèmes ATP cruciaux pour la sécurité. Leur contribution se décline en trois volets. Premièrement, ils développent un modèle de perception d'anomalies en ligne basé sur les réseaux de neurones récurrents LSTM (long short-term memory). Ce modèle traite les données de journalisation des ATPs, en analysant les séquences de clés de journaux et les vecteurs de valeurs de paramètres pour détecter en temps réel les anomalies par rapport au comportement normal appris.

Deuxièmement, une méthode de prédiction des défaillances à venir est proposée, s'appuyant sur l'analyse des séries chronologiques des taux de défaillance historiques. La série temporelle unidimensionnelle des taux de défaillance est reconstruite dans un espace de phase multidimensionnel. Puis, la régression à vecteurs de support (SVR) est utilisée pour modéliser le mappage complexe entre les points de phase et le taux de défaillance futur à prédire. Des algorithmes d'optimisation comme la validation croisée, le PSO (Particle Swarm Optimization) et les algorithmes génétiques permettent de trouver les paramètres optimaux du modèle SVR.

Enfin, en intégrant les modules de perception d'anomalies et de prédiction de défaillances, une architecture globale de maintenance intelligente des ATPs est conçue. Cette architecture combine des sous-systèmes embarqués et sol, ainsi qu'une transmission des données véhicule-sol, afin de réaliser une maintenance affinée sur tout le cycle de vie.

Real-time learning for real-time data : online machine learning for predictive maintenance of railway systems

Le-Nguyen et al. [30] ont proposé une approche d'apprentissage en ligne innovante pour la maintenance prédictive des systèmes ferroviaires. Leur objectif principal était d'exploiter les flux de données en

temps réel provenant des capteurs embarqués pour détecter les dégradations et prédire les défaillances des équipements. Pour atteindre cet objectif, les auteurs ont conçu un pipeline composé de cinq modules interconnectés. Le premier module, appelé InterCE, vise à identifier de manière semi-autonome les segments de données cycliques représentant les comportements opérationnels des systèmes à partir des signaux bruts des capteurs. Cette étape utilise un ensemble d'extracteurs et l'apprentissage actif en interaction avec des experts. Une fois les cycles identifiés, le deuxième module, basé sur un auto-encodeur à mémoire longue à court terme (LSTM-AE), extrait les caractéristiques pertinentes sous forme de vecteurs de caractéristiques. Le LSTM-AE a démontré sa capacité à préserver efficacement les informations contenues dans les cycles d'origine. Le module central du pipeline est le module de détection de l'état de santé (CHMOC). Ce module utilise l'algorithme de clustering en ligne DenStream pour partitionner continuellement le flux de vecteurs de caractéristiques en différents clusters représentant les profils de santé des systèmes. Un score de dégradation est calculé pour chaque système en fonction de son appartenance aux clusters d'anomalies. Si une dégradation est détectée par CHMOC, le quatrième module est chargé d'estimer la durée de vie restante du système concerné. Enfin, le cinquième module filtre les fausses alertes afin de réduire la fréquence des inspections inutiles.

A Novel Method for Aging Prediction of Railway Catenary Based on Improved Kalman Filter

Li et al. [31], ont proposé une nouvelle méthode pour la prédiction du vieillissement des caténaires ferroviaires basée sur un filtre de Kalman amélioré. Les caténaires ferroviaires désignent les systèmes de câbles aériens utilisés pour fournir l'alimentation électrique aux trains électrifiés. Leur objectif est d'assurer un fonctionnement régulier des trains électrifiés en prédisant avec précision le vieillissement des caténaires malgré les interférences environnementales telles que le passage du pantographe ou le vent.

Li et al. remplacent l'estimation a priori de l'état dans le filtre de Kalman traditionnel par un réseau neuro-flou Takagi-Sugeno (T-S FNN). Ce T-S FNN utilise le temps, la température et le déplacement historique comme entrées pour prédire le déplacement de la caténaire. Un algorithme d'apprentissage adaptatif non supervisé est proposé pour former efficacement le T-S FNN. Cet algorithme sélectionne dynamiquement les règles floues pertinentes, permettant au réseau d'utiliser moins de règles tout en maintenant une précision adéquate, réduisant ainsi la complexité de calcul. De plus, la covariance d'erreur de mesure du filtre de Kalman est remplacée par les données de vibration de la caténaire au lieu d'utiliser uniquement le bruit interne du capteur. Cette modification permet de mieux refléter les interférences impactant les mesures de déplacement. Le déplacement prédit par le T-S FNN est ensuite comparé au déplacement théorique calculé à partir de l'expansion thermique des métaux. Si l'écart entre les deux persiste au-delà d'un certain seuil pendant une durée spécifique, la caténaire est considérée comme vieillie.

2.4.3 Systèmes basés sur l'évaluation des risques et l'optimisation de la maintenance

Data-driven predictive maintenance scheduling policies for railways

Gerum et al. [32], ont proposé une approche intégrée pour prédire les défauts ferroviaires et planifier de manière optimale les inspections et activités de maintenance en se basant sur ces prédictions. L'objectif est d'optimiser les coûts à long terme tout en assurant la sécurité et l'efficacité des infrastructures ferro-

viaires. Leur approche commence par regrouper les segments de voies en clusters similaires basés sur leurs caractéristiques de défauts afin de disposer de suffisamment de données pour la prédiction. Ensuite, ils utilisent des modèles d'apprentissage machine comme les forêts aléatoires (RF) et les réseaux de neurones récurrents (RNN) pour prédire les défauts de voie et de géométrie en utilisant des données couramment disponibles. Les prédictions alimentent un modèle de processus de décision markovien pour déterminer les politiques d'inspection et de maintenance optimales minimisant les coûts actualisés. En cas de contraintes sur les équipes d'inspection, ils formulent le problème comme un bandit manchot étendu (Restless bandit) et utilisent la politique d'indice de Whittle pour prioriser les segments à inspecter, De plus, ils mettent en place un mécanisme de mise à jour continue des probabilités de transition pour s'adapter aux évolutions de l'environnement. Cette méthodologie offre une approche systématique et adaptable pour la gestion proactive des infrastructures ferroviaires, ouvrant la voie à des économies significatives et à une meilleure sécurité.

Analysis and prediction of railway accident risks using machine learning

Hadj-Mabrouk [33], a proposé une implémentation d'une méthode hybride pour la prévention des accidents ferroviaires. Cette méthode est basée sur plusieurs algorithmes et utilise divers modes de raisonnement. L'objectif est de développer deux outils complémentaires, "Acasya" et "Sautrel", pour aider à l'analyse et à l'évaluation de la sécurité. "Acasya" est destiné à l'Analyse de la Sécurité Fonctionnelle (ASF) et "Sautrel" est destiné à l'analyse des logiciels critiques pour la sécurité, spécifiquement l'Analyse des Erreurs et des Effets du Logiciel (AEEL). Ces outils visent à identifier automatiquement les règles de sécurité pertinentes et à proposer des solutions et des recommandations pour traiter les problèmes de sécurité. La méthode commence par l'acquisition de connaissances pour recueillir des informations sur la sécurité ferroviaire et les scénarios d'accidents potentiels. Ensuite, elle utilise l'apprentissage par classification de concepts pour regrouper les scénarios d'accidents en classes homogènes. Le RBML (Rule-Based Machine Learning) est utilisé pour identifier automatiquement les règles de sécurité pertinentes à partir d'une base de scénarios historiques. Les règles de production induites par l'apprentissage machine sont transférées au KBS (Knowlege-Based System) pour former la base de connaissances de l'outil de support à l'évaluation de la sécurité. Enfin, le CBR (Case-Based Reasoning) est utilisé pour trouver le cas le plus similaire à un nouveau problème de sécurité et proposer des solutions appropriées. La méthode a été évaluée à l'aide de trois systèmes d'apprentissage : "Clasca", "Charade" et "ReCall". Malgré leur utilité, plusieurs lacunes ont été notées. Par exemple, le système "Clasca" a besoin d'une base d'apprentissage plus représentative et d'une nouvelle approche pour gérer la sensibilité du système à l'ordre d'arrivée des scénarios d'exemple. Le système "ReCall" a eu des problèmes avec le calcul de la similarité, les stratégies d'adaptation et la gestion des valeurs manquantes. Certaines règles générées par le système "Charade" n'étaient pas directement utiles pour l'évaluation de la sécurité.

A data-driven maintenance policy for railway wheelset based on survival analysis and Markov decision process

Almeida Costa et al. [34], ont proposé une approche par processus de décision markoviens (MDP) pour déterminer une politique de maintenance optimale des roues de train d'une compagnie ferroviaire portugaise. L'objectif était de minimiser les coûts de maintenance à long terme, en tenant compte de la dégradation par usure du diamètre des roues et de l'occurrence de dommages nécessitant un reprofilage ou

un remplacement. Leur modèle MDP bidimensionnel définit un espace d'états discrets basé sur le diamètre de la roue et le kilométrage depuis le dernier reprofilage, avec des états supplémentaires pour représenter les roues endommagées. Trois actions de maintenance sont considérées : ne rien faire, renouvellement ou reprofilage. Pour estimer les matrices de transitions probabilistes entre états, les auteurs ont utilisé différentes approches selon l'action. Pour "ne rien faire", une approche markovienne simple modélise la dégradation du diamètre, tandis qu'un modèle de survie de Cox estime la probabilité d'occurrence de dommages. Pour le renouvellement, la transition est certaine vers l'état initial. Pour le reprofilage, des distributions empiriques sont utilisées selon que le reprofilage corrige un dommage ou une simple usure. Parmi les limites, les auteurs citent le fait de ne pas considérer d'autres paramètres géométriques des roues, les incertitudes de mesure, ainsi que l'aspect de la planification des opérations de maintenance.

2.5 Étude comparative

Dans cette section, nous allons comparer différentes approches et méthodes utilisées dans la prédiction de défaillances. Cette étude comparative permet d'identifier les points forts et les limitations de chaque technique. Les critères d'évaluation suivants seront utilisés pour mesurer l'efficacité et la pertinence de chaque méthode.

2.5.1 Critères d'évaluation utilisés

Certains des travaux de recherche examinés n'ont pas fourni d'évaluations quantitatives classiques, telles que la précision, le rappel ou le F1-score. Dans ces cas, d'autres critères d'évaluation ont été utilisés pour évaluer les performances des approches proposées.

Dans l'étude de Le-Nguyen et al. [30], le module d'extraction de cycles (InterCE) a été évalué en calculant le score Kappa entre les cycles identifiés par InterCE et ceux identifiés par un système expert. Un score Kappa supérieur à 0,8 a été considéré comme satisfaisant. Par ailleur plus, le nombre de requêtes nécessaires à InterCE pour traiter un volume de données de 100 000 fichiers a été utilisé comme métrique d'évaluation.

Dans la même étude, les performances de l'auto-encodeur LSTM ont été évaluées en utilisant la mesure nDCG@k (Normalized Discounted Cumulative Gain), qui désigne la capacité du modèle à classer correctement les instances pertinentes. Les scores nDCG@k supérieurs à 0,9 ont été considérés comme des performances satisfaisantes pour des valeurs de k allant de 1000 à 10000.

Li et al. [31], ont utilisé l'écart type de la différence entre les valeurs filtrées et théoriques comme critère d'évaluation pour leur modèle de prédiction du vieillissement de la caténaire ferroviaire. Une faible valeur de l'écart type indique une meilleure précision du modèle par rapport aux méthodes de filtrage traditionnelles.

2.5.2 Résultats et performances des systèmes examinés

Le tableau 2.1 présente une comparaison détaillée des résultats et des performances des différents systèmes examinés dans cette étude. Il montre les forces et les faiblesses de chaque approche, en utilisant diverses métriques d'évaluation pertinentes pour ce domaine. Cette analyse comparative permet d'identifier les techniques les plus prometteuses, ainsi que de mettre en évidence les aspects à améliorer, afin de développer des systèmes de prédiction de défaillance plus robustes et fiables pour les trains autonomes.

Auteurs	Solution proposée	Méthode utilisée	Évaluation
De Simone et al. [27]	Une méthodologie basée sur des algorithmes d'apprentissage profond LSTM.	Utilisation de réseaux LSTM en deux tâches : classification multi-classes (Bon, Mineur, Majeur) pour la prédiction de défaillances, et prévision de la tendance future du signal de défaillance.	• Tâche de classification : Précision : 99,45%, FDR : 99,42%, FAR : 0,35% • Tâche de prévision : Précision : 99,71%, MAE : 0,0184%
S. Kauschke et al. [28]	Une approche en deux niveaux en utilisant un classificateur d'instances suivi d'un méta-classificateur pour évaluer l'ensemble du trajet.	•Classificateurs d'instances : JRIP (règles) et J48 (arbres de décision) avec différentes techniques d'échantillonnage pour équilibrer les classes. •Meta-classificateur : seuil sur le pourcentage d'instances positives pour prédire une défaillance du trajet.	•Exactitude : 80,56%, •Rappel : 12.5%
Kang et al. [29]	•Modèle LSTM pour la perception en ligne des anomalies dans les journaux et les paramètres du système ATP. •Modèle SVR pour la prédiction des tendances futures du taux de défaillance à partir des données historiques.	•Parsing des journaux pour extraire les clés et les vecteurs de paramètres. •Entraînement du LSTM sur ces données pour détecter les anomalies. •Reconstruction d'espace de phase des taux de défaillance historiques. •Application de la SVR pour modéliser la relation entre les points de phase et les points futurs.	•Perception en ligne des anomalies : Précision : 0,966%, Rappel : 0,996%, F-mesure : 0,981% •Prédiction des tendances futures : MSE : 0,081%, R2 : 0,987%
Le-Nguyen et al. [30]	Un pipeline composé de cinq modules (Cycle extraction, Feature learning, Health detection, Prognostics, et Filtering) utilisant l'apprentissage en ligne.	•Utilisation d'une methode InterCE pour extraire et étiqueter de manière semi-autonome les cycles à partir des signaux de capteurs bruts. •Un auto-encodeur LSTM pour apprendre les caractéristiques des cycles. •CHMOC, une approche utilisée pour surveiller l'état de santé des systèmes.	InterCE (Cycle extraction) : •Le score Kappa entre InterCE et un système expert était supérieur à 0,8. •Le nombre de requêtes nécessaires à InterCE était de 4000 par rapport à un volume de données de 100 000 fichiers. Auto-encodeur LSTM : •Les scores nDCG@k dépassent 0,9 pour k allant de 1000 à 10000.

Approche	Solution proposée	Méthode utilisée	Évaluation
Li et al. [31]	Un modèle de prédiction du vieillissement de la caténaire ferroviaire basé sur une combinaison d'un filtre de Kalman et du T-S FNN.	Utilisation d'un modèle T-S FNN basé sur des données de vibration pour déterminer la covariance d'erreur de mesure dans le filtre de Kalman.	Écart type de la différence entre les valeurs filtrées et théoriques : Filtre de Kalman standard : 187,9107, Filtre de Kalman non scindé (UKF) : 111,4947, KalmanNet : 52,9625, Méthode proposée : 6,1236
Gerum et al. [32]	Une méthode qui combine deux techniques : forêts aléatoires et RNN avec une optimisation des activités de maintenance.	• Prédiction des défauts : Forêts aléatoires, réseaux de neurones récurrents et régression ordinale multinomiale. • Optimisation et planification : Processus de décision markovien, bandits fainéants avec indices de Whittle.	•Précision globale de 82% pour la prédiction des défauts rouges. •Précision globale de 62% pour la prédiction des défauts jaunes.
Hadj-Mabrouk [33]	Une méthode hybride qui combine plusieurs algorithmes : l'apprentissage par classification de concepts, l'apprentissage par règles et le raisonnement à partir de cas.	• "Clasca" : un algorithme utilisé pour regrouper et classifier les incidents historiques. • "Charade" : un système conçu pour générer des règles d'évaluation des scénarios. • Raisonnement à partir de cas (CBR) avec l'outil "ReCall" pour proposer des mesures de protection.	N/A
Almeida Costa et al. [34]	La mise en place d'une politique de maintenance pour les roues de train, basée sur MDP.	Une modélisation d'un MDP , avec une implémentation d'une analyse de survie par le modèle de Cox afin de modéliser la probabilité de dommages.	N/A

Tableau 2.1 – Comparaison des résultats et performances des systèmes examinés.

2.6 Synthèse

Les méthodes de maintenance prédictive analysées présentent diverses forces et faiblesses. L'approche en ligne CHMOC (Continuous Health Monitoring using Online Clustering) [30], permet de capturer efficacement les regroupements de données tout en fonctionnant en temps réel, Le modèle peut s'adapter aux changements de distribution des données au fil du temps, mais manque d'interprétabilité, en plus, nécessite une validation par des experts du domaine. L'utilisation conjointe de l'analyse de survie et des processus de décision markoviens [34], offre une politique de maintenance optimale, de même, l'approche est considérée data-driven, et utilise des données réelles d'inspections et de maintenance, cependant, elle manque d'évaluation quantitative. En outre, la fonction de coût considérée ne prend pas en compte les coûts opérationnels. Les méthodes d'apprentissage automatique comme Clasca, Charade et ReCall [33], permettent d'analyser les risques d'accidents ferroviaires, en plus de ça, ces dernières proposent des mesures de prévention et de protection appropriées face à de nouveaux risques détectés. Néanmoins, ces méthodes nécessitent une validation par des experts, ainsi qu'une base de connaissances historiques importante pour être efficace. Les forêts aléatoires et réseaux neuronaux récurrents [32], prédisent efficacement les défauts, malgré cela, ces approches ont tendance à les sous-estimer, une approche hybride avec une fonction de perte asymétrique permet de réduire ce problème. L'algorithme de filtrage de Kalman amélioré [31], réduit considérablement les faux positifs pour la prédiction de vieillissement du caténaire ferroviaire, par rapport aux méthodes classiques, par contre, un compromis entre le temps de prévision avant défaillance et le taux de vrais/faux positifs est nécéssaire. La méthode LSTM [27], se distingue par sa capacité à traiter des séquences temporelles longues, ainsi que sa robustesse grâce à une mémoire interne efficace, offrant une perception des anomalies très précise, surpassant ainsi les méthodes traditionnelles. En revanche, elle nécessite des réglages fins et une grande quantité de données étiquetées.

2.7 Conclusion

Dans ce chapitre, nous avons discuté les recherches actuelles, abordant les systèmes de prédiction de défaillance résistants pour les trains autonomes. Différentes méthodes utilisent des approches distinctes, c'est pourquoi nous avons réalisé une classification de ces différentes méthodes. Nous avons parlé des méthodes d'apprentissage automatique classiques, des méthodes d'apprentissage profond et encore d'autre méthodes avancées, puis nous avons réalisé une taxonomie des travaux de recherche examinés. Nous avons classé ces travaux en trois classes distinctes : les systèmes axés sur la prédiction et l'analyse des défaillances, les systèmes axés sur la surveillance de l'état du système et la détection des anomalies, puis les systèmes axés sur l'évaluation des risques et l'optimisation de la maintenance. Les avantages et les limites de chaque méthode ont également été discutés, en mettant en évidence leurs résultats en ce qui concerne la précision, le temps de réponse, la solidité et d'autres critères d'évaluation pertinents.

Malgré les progrès notables réalisés, il reste encore des obstacles à surmonter afin de concevoir des systèmes de prédiction robustes, flexibles et adaptés aux exigences spécifiques des trains autonomes.

Le prochain chapitre présentera notre propre proposition pour un système de détection de défaillance, se basant sur l'algorithme de XGBoost, pour prédire les périodes précédant directement les anomalies.

CHAPITRE 3

SYSTÈME D'ANTICIPATION DE DÉFAILLANCES DANS LES TRAINS À L'AIDE DE L'APPROCHE XGBOOST

3.1 Introduction

La prédiction de défaillance est un élément crucial pour assurer la sécurité et la fiabilité des trains autonomes. Les trains autonomes représentent une avancée technologique majeure, avec des avantages considérables en termes de sécurité, d'efficacité énergétique et de réduction des coûts opérationnels. Cependant, leur complexité croissante et l'interdépendance de leurs systèmes nécessitent des mécanismes sophistiqués de surveillance et de maintenance. Dans ce chapitre, nous proposons un système de prédiction innovant qui se concentre sur la détection d'anomalies quelques jours avant leur occurrence, basé sur une méthode d'apprentissage automatique. L'approche que nous utiliserons, le XGBoost (eXtreme Gradient Boosting), nous permettra de prédire les pré-failures, c'est-à-dire les moments critiques précédant les anomalies, afin de prendre des mesures préventives efficaces. Cette méthode s'appuie sur l'analyse de données réelles collectées à partir des capteurs des trains, offrant ainsi une solution robuste et fiable pour anticiper les défaillances potentielles et optimiser les opérations de maintenance.

Dans ce chapitre, nous commençerons par présenter la motivation derrière notre choix d'utiliser le XGBoost, encore, les avantages qu'elle offre par rapport aux autres méthodes. Ensuite, nous décrirons l'architecture globale de notre système de prédiction de défaillance, en détaillant les différentes étapes de cette dernière, de la collecte et du prétraitement des données, à l'intégration des mécanismes de résilience.

3.2 Motivation

Pour développer un système de prédiction de défaillance pour les trains autonomes, plusieurs facteurs ont été des sources d'inspiration et de motivation. Les trains autonomes représentent une avancée technologique majeure dans le domaine du transport ferroviaire, promettant une efficacité accrue et une sureté améliorée. Cependant, ces systèmes complexes et interdépendants présentent également des défis importants en termes de maintenance et de gestion des défaillances.

L'une des principales motivations a été de répondre à la nécessité de systèmes de surveillance et de maintenance sophistiqués pour garantir la sureté des trains autonomes et des passagers. La détection précoce des anomalies et des défaillances est cruciale pour éviter des interruptions de service coûteuses et des incidents potentiellement dangereux. En intégrant des techniques avancées d'apprentissage automatique, nous visons à créer un système capable de prédire les anomalies plusieurs jours avant leur occurrence, permettant ainsi des interventions préventives.

Un autre facteur clé est l'optimisation des opérations de maintenance. La maintenance prédictive permet de planifier les interventions de manière plus efficace, réduisant les temps d'arrêt et les coûts associés à des réparations d'urgence. Cela conduit non seulement à une meilleure utilisation des ressources, mais aussi à une augmentation de la durée de vie des équipements.

Enfin, le développement de ce système de prédiction de défaillance s'inscrit dans une vision plus large d'innovation et d'amélioration continue dans le domaine ferroviaire. En adoptant des approches basées sur l'apprentissage automatique, nous contribuons à l'évolution des technologies ferroviaires et à la création de solutions de pointe pour les défis actuels et futurs.

3.3 Organigramme de notre proposition

La Figure 3.1 illustre l'organigramme des diverses étapes suivies dans notre démarche.

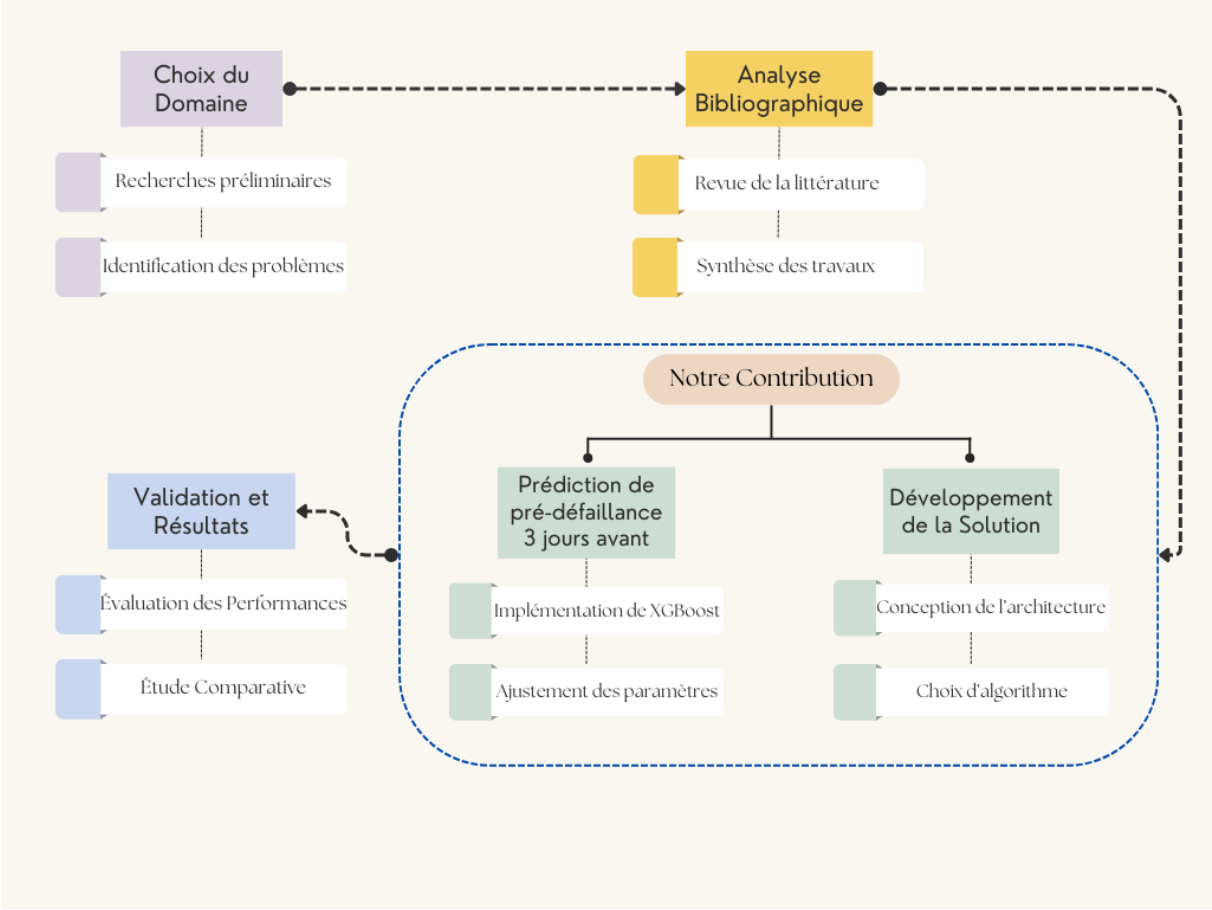

FIGURE 3.1 – Organigramme de notre contribution.

L'organigramme présente les différentes étapes de notre travail. Tout d'abord, le "Choix du Domaine" implique des recherches préliminaires et l'identification des problèmes spécifiques à aborder. Nous avons décidé de travailler sur le développement d'un système de prédiction de défaillances pour les trains autonomes. Ensuite, une "Analyse Bibliographique" est effectuée, comprenant une revue de la littérature existante et une synthèse des travaux antérieurs.

Notre contribution se divise en deux parties principales. La première est la "Prédiction de pré-défaillance 3 jours avant", qui comprend l'implémentation de l'algorithme XGBoost et l'ajustement de ses paramètres pour optimiser les prédictions, afin de prédire les instances de pré-défaillances quelques jours avant leur survenue. La seconde partie est le "Développement de la Solution", où l'architecture du système est conçue et les algorithmes appropriés sont choisis pour la mise en œuvre.

Enfin, les résultats sont validés à travers une évaluation des performances du système et une étude comparative avec une autre méthode. Cette approche méthodique assure que la solution proposée est à la fois efficace et fiable pour la prédiction des pré-défaillances.

3.4 Description de données

Dans cette section, nous allons discuter le jeu de données que nous avons utilisé dans notre étude, nous parlerons de la source des données utilisées, la méthode utilisée pour leur collection, ensuite, nous allons donner un petit aperçu des données fournies.

3.4.1 Collecte de données

Le jeu de données utilisé pour notre étude est le MetroPT-3 Dataset, fourni par l'INESC TEC - Laboratory of Artificial Intelligence and Decision Support au Portugal [35]. Ce jeu de données contient des lectures de pression, température, courant du moteur et signaux des vannes d'admission d'air provenant de l'unité de production d'air (APU) d'un compresseur d'un train en contexte opérationnel. Elles consistent en des données de séries temporelles multivariées obtenues à partir de plusieurs capteurs analogiques et numériques, installés sur le compresseur d'un train. Les données couvrent la période de février à août 2020.

La surveillance et la collecte de données, tels que le comportement temporel et les événements de pannes, ont été obtenus à partir des enregistrements générés par les capteurs. Les données ont été enregistrées à une fréquence de 1 Hz par un dispositif mis en place.

3.4.2 Aperçu du jeu de données

Ce jeu de données peut être utilisé pour les prédictions d'anomalies, l'estimation du temps de vie résiduelle et d'autres tâches.

Caractéristiques du jeu de données :
- Séries temporelles multivariées
- Nombre d'instances : 15 169 480
- Caractéristiques des attributs : réelles
- Nombre d'attributs : 15

— Pistes associées : classification, régression

— Valeurs manquantes : N/A

Le jeu de données comprend sept (7) capteurs analogiques et huit (8) capteurs numériques, chacun joue un rôle dans la surveillance du compresseur. Les attributs sont montrés dans le Tableau 3.1.

Caractéristique	**Description**	**Type**
TP2 (bar)	Pression du compresseur	Analogique
TP3 (bar)	Pression générée au panneau pneumatique	Analogique
H1 (bar)	Pression due à la chute lors de la décharge du filtre du séparateur cyclonique	Analogique
DV pressure (bar)	Pression de décharge des déshydrateurs d'air	Analogique
Reservoirs (bar)	Pression en aval des réservoirs	Analogique
Motor Current (A)	Courant du moteur triphasé	Analogique
Oil Temperature (^{o}C)	Température de l'huile du compresseur	Analogique
COMP - Signal électrique	Signal électrique de la vanne d'admission d'air du compresseur	Numérique
DV electric	Signal électrique de la vanne de sortie du compresseur	Numérique
TOWERS	Signal électrique définissant la tour de séchage de l'air	Numérique
MPG	Signal électrique de démarrage du compresseur	Numérique
LPS	Signal électrique de détection de pression basse	Numérique
Pressure Switch	Signal électrique de détection de décharge dans les tours de séchage de l'air	Numérique
Oil Level	Signal électrique de détection du niveau d'huile du compresseur	Numérique
Caudal Impulse	Signal électrique comptant les impulsions de flux d'air de l'APU aux réservoirs	Numérique

Tableau 3.1 – Description des variables.

La Figure 3.2 est une illustration de quelques lignes de jeu de données utilisé.

TP2	TP3	H1	DV_pressure	Reservoirs	Oil_temperature	Motor_current	COMP	DV_eletric	Towers	MPG	LPS	Pressure_switch	Oil_level	Caudal_impulses	pre_failure
0.0	0.9	0.91	0.0	0.9	0.52	0.0	1.0	0.0	1.0	1.0	0.0	1.0	1.0	1.0	0.0
0.0	0.9	0.91	0.0	0.9	0.52	0.0	1.0	0.0	1.0	1.0	0.0	1.0	1.0	1.0	0.0
0.0	0.9	0.91	0.0	0.9	0.52	0.0	1.0	0.0	1.0	1.0	0.0	1.0	1.0	1.0	0.0
0.0	0.9	0.91	0.0	0.9	0.52	0.0	1.0	0.0	1.0	1.0	0.0	1.0	1.0	1.0	0.0
0.0	0.9	0.9	0.0	0.9	0.52	0.0	1.0	0.0	1.0	1.0	0.0	1.0	1.0	1.0	0.0

Figure 3.2 – Illustration de jeu de données.

3.4.3 Rapport d'anomalies

Le jeu de données n'est pas initialement étiqueté, mais il est fourni avec des rapports de défaillance répertoriés, comme illustré dans le Tableau 3.2.

Nr.	Heure de Début	Heure de Fin	Anomalie	Rapport
1	18/04/2020 00 :00	18/04/2020 23 :59	Fuite d'air	Maintenance effectuée le 30/04 à 12 :00
2	29/05/2020 23 :30	30/05/2020 06 :00	Fuite d'air	
3	05/06/2020 10 :00	07/06/2020 14 :30	Fuite d'air	Maintenance effectuée le 08/06 à 16 :00
4	15/07/2020 14 :30	15/07/2020 19 :00	Fuite d'air	Maintenance effectuée le 16/07 à 00 :00

TABLEAU 3.2 – Rapports d'anomalies pour l'APU.

3.5 Notre proposition

Pour remédier à la problématique de prédiction de défaillances dans un train, nous avons décidé d'entraîner un modèle sur le jeu de données utilisé. Ce modèle doit être capable de distinguer entre les schémas de données défaillantes et ceux des données saines. Afin de ne pas traiter notre travail comme une simple classification, car notre objectif est de prédire les pannes à l'avance, nous avons choisi de prédire les instances précédant les pannes, c'est-à-dire les préfailures, et non les défaillances directement.

3.5.1 Description du systéme proposé

Le système que nous avons proposé est conçu pour analyser les données provenant des capteurs installés dans les trains autonomes en utilisant des algorithmes de machine learning avancés.

Notre approche vise à détecter des schémas anormaux indiquant des pré-défaillances. En cas de détection d'une pré-défaillance, le système déclenche immédiatement une alerte destinée à l'équipe de maintenance, permettant ainsi une intervention proactive avant que des problèmes graves ne surviennent.

La Figure 3.3 illustre le fonctionnement de notre système.

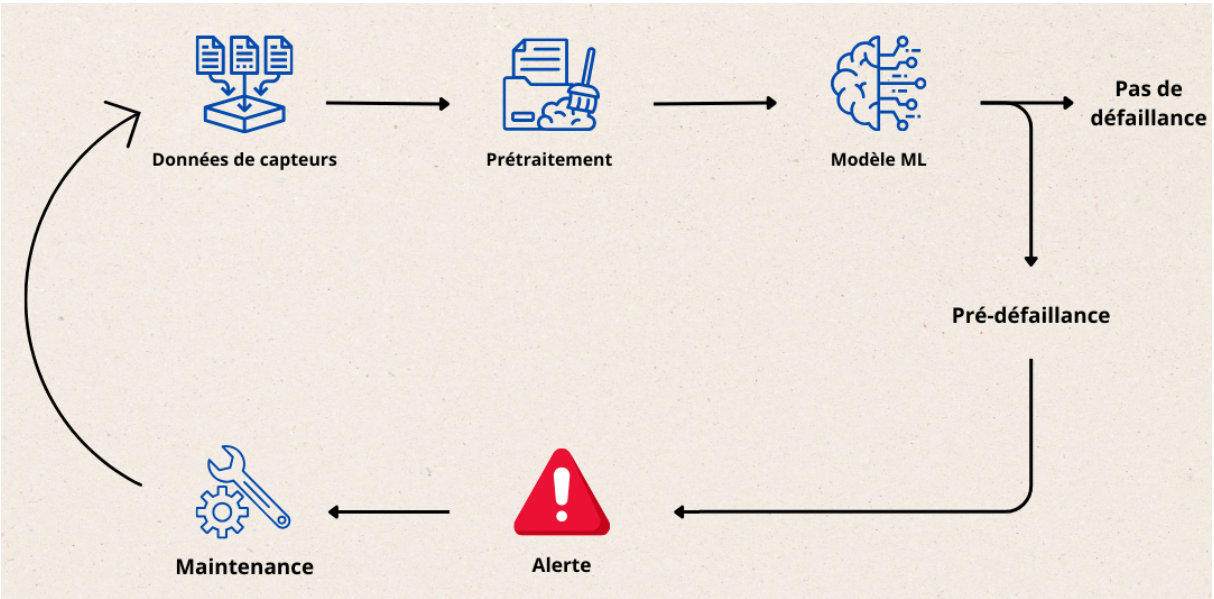

FIGURE 3.3 – Description du système proposé.

3.5.2 Prétraitement de données

Le jeu de données que nous utilisons n'a aucune valeur manquante, ce qui fait qu'aucun nettoyage supplémentaire n'est nécessaire. Cependant, plusieurs étapes de prétraitement ont été appliquées pour préparer les données à l'apprentissage automatique.

a. Normalisation de données

La normalisation est une étape cruciale dans le prétraitement des données, particulièrement pour les algorithmes d'apprentissage automatique sensibles à l'échelle des caractéristiques. Nous avons choisi d'utiliser le MinMaxScaler pour normaliser nos données.

MinMaxScaler

Le MinMaxScaler est une technique de prétraitement utilisée en apprentissage automatique pour mettre à l'échelle et normaliser les données. Cette méthode transforme les caractéristiques en les mettant à l'échelle dans une plage donnée, généralement entre 0 et 1. La transformation est donnée par la formule suivante :

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \tag{3.1}$$

Où X représente les valeurs originales de la caractéristique, X_{min} et X_{max} sont respectivement les valeurs minimales et maximales de cette caractéristique. Cette méthode de mise à l'échelle garantit que toutes les caractéristiques ont la même échelle, ce qui peut améliorer les performances et la vitesse de convergence de nombreux algorithmes d'apprentissage automatique [36].

Justification de l'utilisation du MinMaxScaler

L'utilisation du MinMaxScaler sur nos données est justifiée par les raisons suivantes :

— **Uniformisation de l'échelle des caractéristiques** : Les différentes caractéristiques de notre jeu de données, telles que la pression, la température et le courant moteur, sont enregistrées sur des échelles différentes. En utilisant le MinMaxScaler, nous transformons ces caractéristiques pour qu'elles partagent une même échelle, ce qui simplifie l'apprentissage pour le modèle [36].
— **Préservation des relations proportionnelles** : Contrairement à d'autres méthodes de normalisation, le MinMaxScaler préserve les relations proportionnelles entre les valeurs, ce qui est crucial pour certaines caractéristiques où les écarts relatifs sont significatifs [36].
— **Préparation optimale pour les algorithmes de ML** : De nombreux algorithmes d'apprentissage automatique, y compris XGBoost que nous utilisons, fonctionnent mieux lorsque les données sont normalisées. Le MinMaxScaler améliore la convergence et la performance du modèle en réduisant les biais introduits par des échelles de caractéristiques différentes.

b. Sélection des caractéristiques

Afin d'améliorer les performances du modèle, nous avons procédé à la sélection des caractéristiques pertinentes. Afin d'accomplir cela, nous avons analysé le pattern de données de chaque capteur, ce qui nous a permis de determiner les changements dans ces différentes données.

Le rapport d'anomalies fourni, nous a permis de visualiser quelques modèles de données différents. Au cours de notre analyse, nous avons constaté que les données de certains capteurs n'ont pas montré une déviation du comportement ordinaire dans les périodes d'anomalie, comme illustré dans les Figures 3.4 et 3.5 qui représentent respectivement les changements des modèles de données des capteurs du niveau d'huile (Oil_level) et des impulsions de débit (Caudal_impulses).

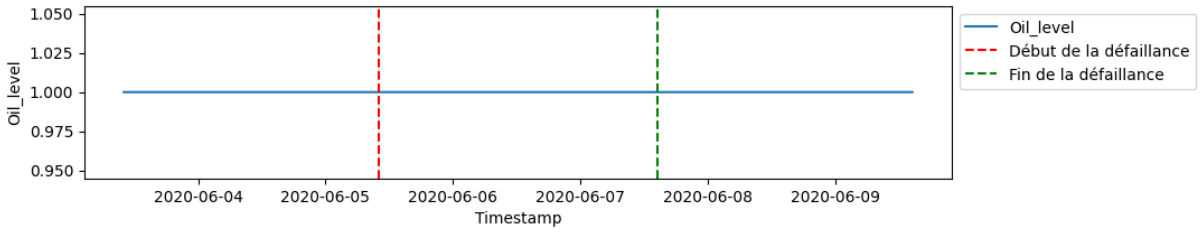

FIGURE 3.4 – Modèle de données du capteur pour la caractéristique Oil_level.

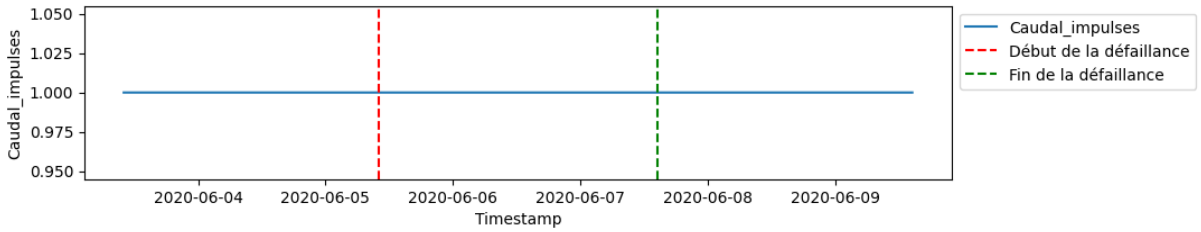

FIGURE 3.5 – Modèle de données du capteur pour la caractéristique Caudal_impulses.

Par contre, comme on peut le constater dans les Figures 3.6 et 3.7, le changement est apparent dans les modèles de données des capteurs de pression de décharge des déshydrateurs d'air (DV_pressure) et de pression générée au panneau pneumatique (TP3). Ces figures révèlent des variations significatives par rapport au comportement normal des capteurs durant les périodes d'anomalie. Ces déviations mettent en évidence l'impact des anomalies sur ces paramètres spécifiques et soulignent l'importance de surveiller attentivement ces indicateurs pour détecter précocement les défaillances potentielles du système.

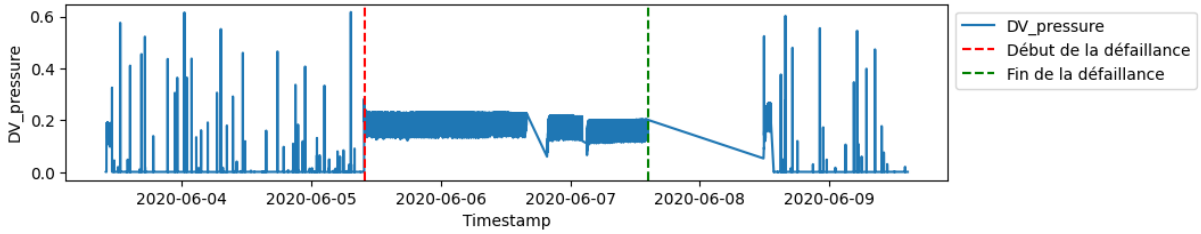

FIGURE 3.6 – Modèle de données du capteur pour la caractéristique DV_pressure.

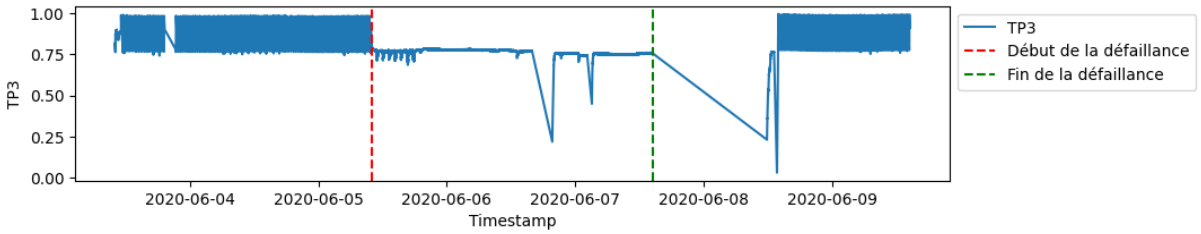

FIGURE 3.7 – Modèle de données du capteur pour la caractéristique TP3.

Analyse des corrélations

Pour approfondir notre compréhension de l'importance de chaque capteur dans notre prédiction, nous avons étudié les corrélations entre les caractéristiques, puis, nous avons selectionné ceux qui présentent une corrélation supérieure à 0.75, le Tableau 3.3 montre les caractéristiques choisies et leurs corrélations respectives, par rapport aux autres caractéristiques.

Caractéristique	**Corrélation**
TP2	0.945972
TP3	0.999993
H1	0.970599
Reservoirs	0.999993
COMP	0.984101
DV_eletric	0.945972
MPG	0.984101

TABLEAU 3.3 – Corrélation des caractéristiques.

Aprés la selection des caractéristiques qui montrent une forte corrélation, nous avons établi une matrice de corrélation entre ces dernières, ce qui est illustré dans la Figure 3.8.

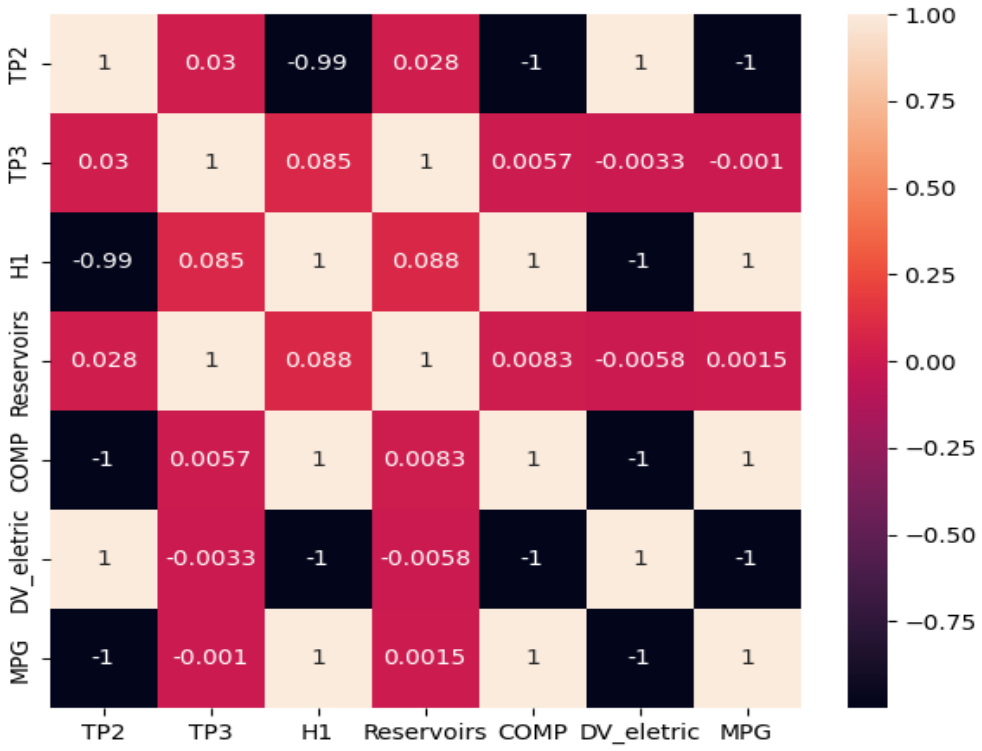

FIGURE 3.8 – Matrice de corrélation.

La Figure illustre la matrice de corrélation des différentes caractéristiques sélectionnées pour notre modèle de prédiction des défaillances. Cette matrice montre les coefficients de corrélation entre chaque paire de caractéristiques, indiquant la force et la direction de la relation linéaire entre elles.

Les valeurs de corrélation proches de 1 ou de -1 signifient une forte relation positive ou négative, respectivement, tandis que des valeurs proches de 0 indiquent une absence de relation linéaire significative.

Selon les résultats de la matrice de corrélation, il a été constaté que les caractéristiques [TP2, H1, Comp, MPG et DV_eletric], sont fortement corrélées entre elles. Cette corrélation significative suggère que ces caractéristiques fournissent des informations redondantes, ce qui peut causer des problèmes de multicolinéarité dans notre modèle de prédiction.

Multicolinéarité

La multicolinéarité désigne une situation dans laquelle deux ou plusieurs variables indépendantes dans un modèle de régression, sont fortement corrélées. Cette corrélation élevée entre les variables peut rendre difficile l'estimation des coefficients individuels de la régression [37].

Synthèse

Lors de la visualisation des données des capteurs, nous avons observé que certains capteurs, tels que le niveau d'huile (Oil_level) et les impulsions de débit (Caudal_impulses), ne montraient pas de chan-

gements apparents. Ces capteurs semblaient constants et ne fournissaient pas de variations significatives qui pourraient être utilisées pour prédire des défaillances.

Ensuite, en analysant les corrélations entre les différentes caractéristiques, nous avons constaté que seules certaines caractéristiques présentaient des corrélations élevées. Particulièrement, les caractéristiques TP2, H1, Comp, MPG et DV_eletric. Ce qui signifie que l'inclusion de toutes ces caractéristiques dans le modèle, pourrait ne pas ajouter de valeur supplémentaire, et pourrait même compliquer le modèle inutilement.

En conclusion, bien que certaines caractéristiques pourraient ne pas avoir une importance significative, TP2 se démarque comme une caractéristique clé qui mérite d'être incluse en raison de son importance potentielle dans la prédiction des défaillances, car cette dernière represente la mesure réelle de la pression sur le compresseur APU.

c. Étiquetage des données

Dans cette étape, nous utilisons le rapport d'anomalies fourni, pour étiqueter les occurences d'anomalies dans nos données, pour cela, nous avons crée une nouvelle colonne dans notre jeu de données qu'on a appelé 'Anomalie', cet nouvelle colonne prend une valeur de 1 tout au long de la période de chaqu'une des pannes reportées, et 0 pour le reste des données.

En utilisant le résultat de l'étiquetage fait, nous avons créé une autre colonne appelée 'Préfailure'. Cette colonne prend une valeur de 1 pour tous les timestamps couvrant exactement les trois jours avant chaque défaillance. Le reste des données, en dehors de cette période de pré-failure, est marqué avec une valeur de 0.

Les débuts et fins des périodes couvrant les préfailures sont illustrés dans le Tableau 3.4.

Anomalie	**Début de préfailure**	**Fin de préfailure**	**Début d'anomalie**	**Fin d'anomalie**
1	15/04/2020 00 :00	18/04/2020 00 :00	18/04/2020 00 :00	18/04/2020 23 :59
2	26/05/2020 23 :30	29/05/2020 23 :30	29/05/2020 23 :30	30/05/2020 06 :00
3	02/06/2020 10 :00	05/06/2020 10 :00	05/06/2020 10 :00	07/06/2020 14 :30
4	12/07/2020 14 :30	15/07/2020 14 :30	15/07/2020 14 :30	15/07/2020 19 :00

Tableau 3.4 – Début et fin de préfailures.

Pour que le modèle puisse atteindre son objectif de prédiction, il était crucial qu'il s'entraîne sur les instances pré-failure. C'est ainsi que nous avons éliminé toutes les instances situées dans les périodes d'anomalies, ce qui a engendré l'élimination de 31,500 instances parmi les 15,169,480 existantes.

d. Extraction de caractéristiques

Étant donné que notre jeu de données est une série temporelle, l'une des caractéristiques est le timestamp de chaque instance. Ainsi, à partir des timestamps, nous avons procédé à une extraction approfondie des caractéristiques pertinentes. L'importance de cette étape réside dans la capacité à transformer les données brutes en informations exploitables pour notre modèle.

Les caractéristiques extraites incluent, le jour, le mois, l'année, l'heure, la minute et la seconde de chaque instance de notre jeu de données. Ce qui fait que nous avons ajouté six nouvelles colonnes.

Pour améliorer les performances du modèle, nous avons ajouté deux attributs supplémentaires, basés sur les deux caractéristiques qui ont montré les plus fortes corrélations avec la cible (*pre_faillure*) :

— Moyenne mobile de la température de l'huile sur 3 jours : **Oil_temp_avg_3day**.

— Moyenne mobile de la pression DV sur 3 jours : **dv_pressure_avg_3day**.

e. Division du jeu de données

Après avoir extrait les caractéristiques pertinentes de notre jeu de données, nous avons procédé à la division de celui-ci en ensembles d'entraînement et de test. Cette étape est cruciale pour évaluer objectivement les performances de notre modèle. Voici les étapes détaillées suivies pour cette division :

Création des données X et y

Nous avons sélectionné les colonnes pertinentes de notre dataframe pour créer les données d'entraînement X. La colonne cible y a été définie comme étant la colonne 'préfailure', qui indique les instances de préfailures.

La Figure 3.9 représente le contenu de X et y.

```
X = new_df[['TP2','TP3','Reservoirs','Oil_temperature','Motor_current',
            'Pressure_switch','Towers','DV_pressure','second','minute','hour',
            'day','month','year','dv_pressure_avg_3day','Oil_temp_avg_3day']]

y = new_df['pre_failure']
```

FIGURE 3.9 – Création des données de X et y.

Division des données

Pour diviser les données en ensembles d'entraînement et de test, nous avons utilisé la fonction train_test_split de la bibliothèque scikit-learn.

train_test_split : Cette fonction permet de diviser les données de manière aléatoire, tout en respectant la proportion spécifiée pour l'ensemble de test [36]. Les paramètres incluent :

— X : Les caractéristiques d'entrée.

— y : La cible ou variable à prédire.

— test_size : Le pourcentage des données à inclure dans l'ensemble de test (30% dans notre cas, ce qui implique 70% de données pour l'ensemble d'entrainement).

— stratify : Permet de s'assurer que les proportions de chaque classe sont préservées dans les ensembles d'entraînement et de test.

Stratification

La stratification est une technique qui permet de s'assurer que la proportion des classes dans les ensembles d'entraînement et de test est similaire à celle du jeu de données original [36]. Étant donné que notre variable cible 'préfailure' est disproportionné (avec beaucoup plus de 0 que de 1), la stratification nous permet de maintenir ces proportions dans les deux ensembles, garantissant ainsi, une évaluation plus réaliste et robuste du modèle.

3.5.3 Modèles d'apprentissage automatique sélectionnés

Dans cette section, nous allons parler en détail, de l'algorithme d'apprentissage automatique que nous avons choisi pour notre modèle de prédiction de défaillances : XGBoost. Nous commencerons par une introduction générale au boosting, l'idée fondamentale derrière XGBoost, puis nous approfondirons la théorie et le fonctionnement de cet algorithme. Enfin, nous aborderons les aspects techniques de son implémentation et son utilisation pour entraîner notre modèle.

Nous avons choisi d'utiliser XGBoost (eXtreme Gradient Boosting) pour plusieurs raisons principales :

— **Performances supérieures** : XGBoost est célebre pour ses performances remarquable dans de nombreux défis en science des données. Sa capacité s'avére supérieure à celle d'autres algorithmes d'apprentissage automatique en termes de précision et de rapidité [5].

— **Gestion efficace des données manquantes et des valeurs aberrantes** : Il intègre des mécanismes de gestion des données manquantes et des valeurs aberrantes, ce qui le rend particulièrement adapté aux jeux de données complexes, tels que ce que nous rencontrons souvent dans le domaine des trains autonomes.

— **Régularisation intégrée** : Il utilise des techniques de régularisation telles que la pénalité de complexité du modèle et la réduction de dimensionnalité, ce qui aide à contrôler le surapprentissage et d'améliorer la généralisation du modèle.

— **Interface conviviale** :Il propose une interface conviviale pour plusieurs langages de programmation, tels que Python, R et Java, ce qui facilite son intégration dans nos pipelines de traitement de données.

Fonctionnement de XGBoost

XGBoost est connu par le principe du boosting de gradient, qui cherche à minimiser une fonction de perte en créant de manière additive un modèle composé d'une séquence d'arbres de décision faibles [5]. Mathématiquement, le modèle XGBoost peut être représenté comme suit :

$$\hat{y}i = \phi(x_i) = \sum k = 1^K f_k(x_i), f_k \in \mathcal{F} \tag{3.2}$$

Où $\hat{y}_i$ est la prédiction pour l'instance x_i, K est le nombre d'arbres de décision faibles dans le modèle, f_k est un arbre de décision faible, et $\mathcal{F}$ est l'espace des fonctions de régression des arbres de décision faibles. L'objectif est de minimiser une fonction de perte régularisée $\mathcal{L}$ définie comme suit :

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}i) + \sum k = 1^{K} \Omega(f_k) \tag{3.3}$$

Où l est une fonction de perte convexe qui mesure la différence entre la prédiction $\hat{y}_i$ et la valeur cible y_i, tandis que la fonction de régularisation Ω permet de contrôler la complexité du modèle pour éviter le surapprentissage. Le processus d'entraînement de XGBoost consiste à ajouter de manière itérative des arbres de décision faibles au modèle, de sorte que chaque nouvel arbre corrige les erreurs des arbres précédents. Pour apprendre un nouveau arbre f_t, XGBoost minimise la fonction de perte suivante :

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{3.4}$$

Où $\hat{y}_i^{(t-1)}$ est la prédiction du modèle avant l'ajout du nouvel arbre f_t. Cette minimisation est effectuée par une approximation de second ordre de la fonction de perte, ce qui permet d'obtenir une estimation rapide et précise des gains potentiels de chaque nœud de l'arbre. Cette approche, combinée à des techniques d'optimisation avancées et à une implémentation efficace, permet à XGBoost d'atteindre des performances élevées en termes de précision et de rapidité.

La Figure 3.10 montre le schéma du fonctionnement de l'algorithme XGBoost.

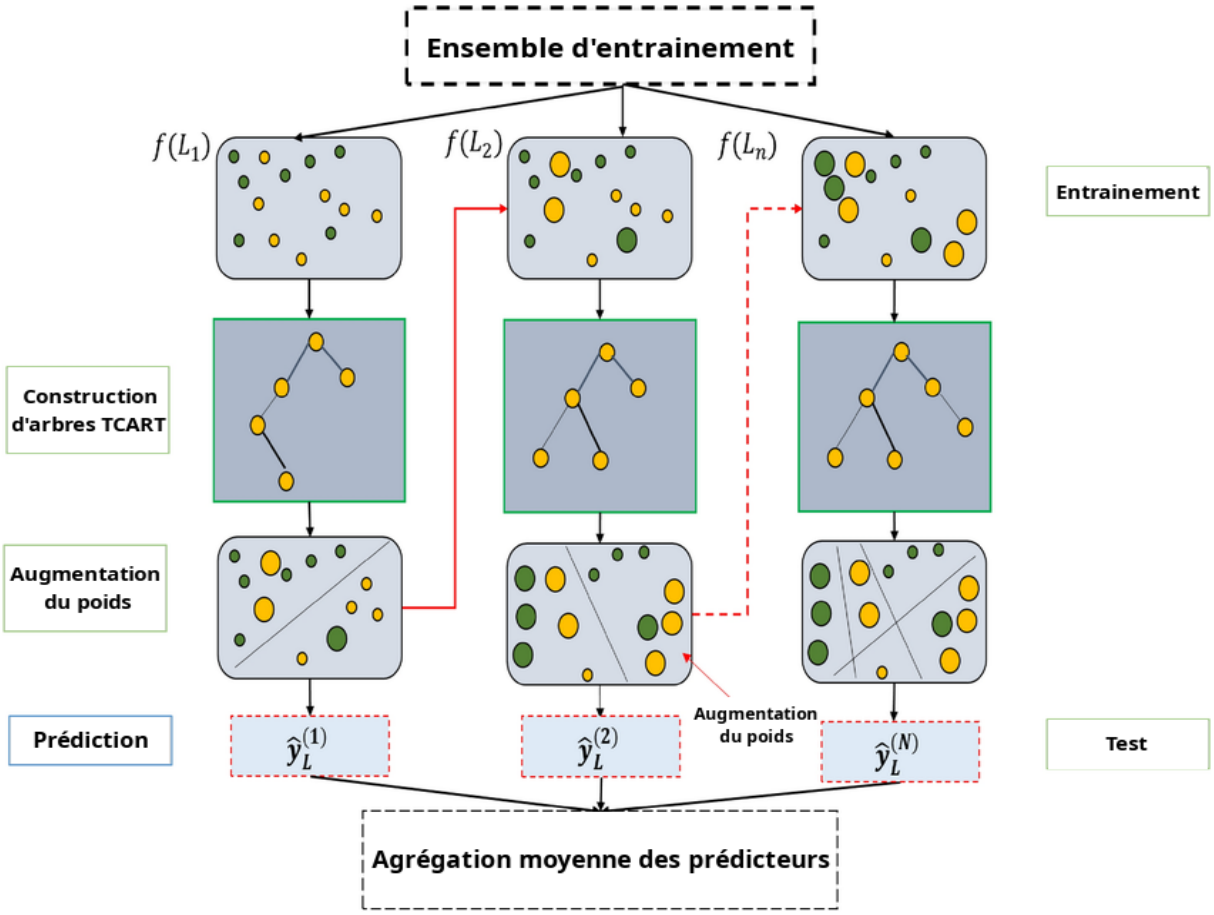

FIGURE 3.10 – Schéma du fonctionnement du XGBoost [38].

Aspects techniques de XGBoost

XGBoost présente plusieurs caractéristiques techniques qui contribuent à ses performances élevées :

— **Parallélisation et mise à l'échelle** : XGBoost utilise le parallélisme de différentes manières. Premièrement, il est capable de créer des arbres de décision simultanément sur plusieurs cœurs de processeur. Par la suite, il prend en charge le calcul distribué, permettant d'entraîner des modèles sur différentes machines, à l'aide de systèmes de traitement de données distribués, tels que Hadoop ou Spark. Afin de gérer efficacement de grands ensembles de données, la capacité d'évolutivité est nécessaire.

— **Optimisation du gain d'information** : XGBoost utilise une méthode approximative pour calculer rapidement le gain d'informations de chaque nœud de l'arbre, ce qui accélère considérablement la construction des arbres de décision. La méthode est basée sur une approximation de Taylor du second ordre de la fonction de perte.

— **Système de cache intelligent** : XGBoost met en cache les données de formation compressées, ainsi que les statistiques nécessaires à la création d'arbres de décision. Cela réduit considérablement les coûts de calcul et d'accès à la mémoire pendant la formation, en particulier pour les grands ensembles de données.

— **Prise en charge de différentes fonctions de perte** : XGBoost prend en charge différentes fonctions de perte, pour des types de tâches d'apprentissage automatiques différentes, notamment la régression logistique pour la classification, la régression linéaire est utilisée pour les modèles de régression, tandis que la régression logistique multiclasse est adoptée pour résoudre les problèmes impliquant plusieurs classes.

— **Élagage des arbres** : XGBoost intègre une méthode d'élagage d'arbre de décision, qui permet de réduire la complexité de l'arbre et d'améliorer la capacité de généralisation du modèle. Cela permet d'éviter le surajustement et d'améliorer les performances des problèmes multi-classes sur de nouveaux ensembles de données.

— **Interprétabilité des modèles** : Même si les arbres de décision de XGBoost sont difficiles, le modèle final peut être compris. Ce qu'on peut faire en regardant les scores de caractéristiques, et les structures d'arbres. Ce qui nous permet de voir facilement les facteurs importants qui affectent les prévisions du modèle [39].

— **Facilité d'utilisation** : XGBoost est disponible sous forme de bibliothèque open-source, avec des interfaces accueillantes pour plusieurs langages de programmation populaires, tels que Python, R et Java. De plus, il offre de nombreux paramètres configurables pour ajuster les performances du modèle, et cela en fonction des besoins spécifiques de l'application [5].

Ces aspects techniques, combinés à sa flexibilité et à son efficacité, font de XGBoost un choix judicieux pour une large gamme de tâches d'apprentissage automatique, y compris notre problème de prédiction de pré-défaillances dans les trains autonomes.

La Figure 3.11 illustre les avantages de l'algorithme XGBoost.

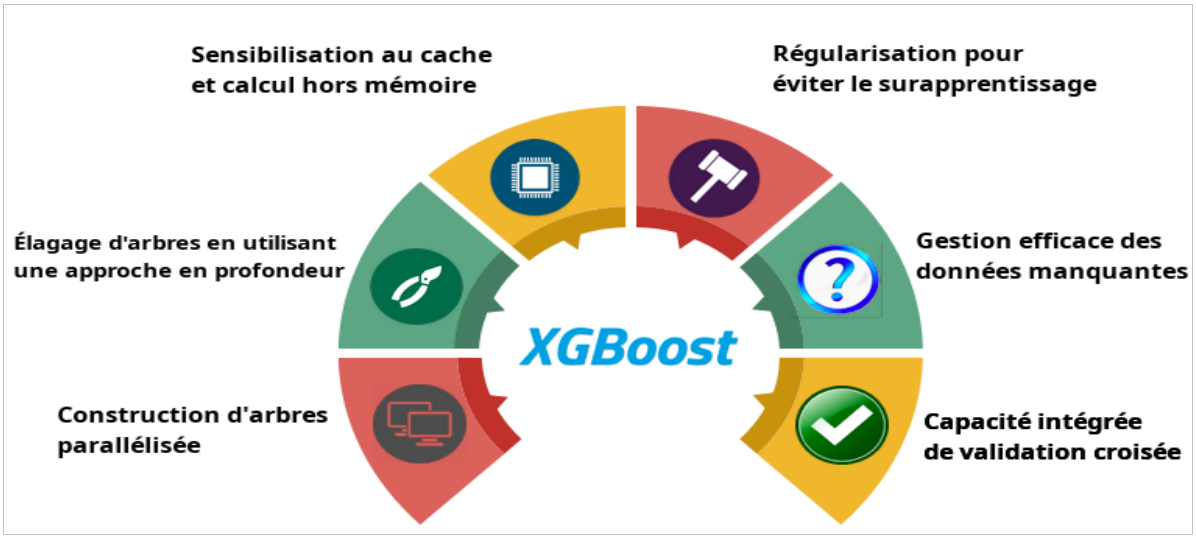

FIGURE 3.11 – Les avantages du XGBoost [17].

Utilisation de XGBoost pour entraîner notre modèle

Une fois que nous avons préparé nos données comme décrit dans les sections précédentes, nous pouvons maintenant commencer à entraîner notre modèle pour la prédiction des pré-défaillances. Voici les étapes principales :

- **Importation des bibliothèques nécessaires** : Nous importons les bibliothèques requises, notamment XGBoost, scikit-learn pour la division des données et l'évaluation des modèles, et d'autres bibliothèques utilitaires.
- **Définition des hyperparamètres** : Nous définissons les hyperparamètres de XGBoost, tels que le nombre d'arbres de décision, la profondeur maximale des arbres, le taux d'apprentissage, et les paramètres de régularisation. Ces hyperparamètres peuvent être ajustés pour optimiser les performances du modèle.
- **Création du modèle XGBoost** : Nous créons une instance du modèle XGBoost en spécifiant les hyperparamètres définis à l'étape précédente.
- **Entraînement du modèle** : Nous entraînons le modèle sur les données d'entraînement en utilisant la méthode Fit [5], et en fournissant les caractéristiques d'entrée X_train et les étiquettes cibles y_train.
- **Évaluation du modèle** : Après l'entraînement, nous évaluons les performances du modèle sur les données de test X_test et y_test en utilisant des métriques appropriées, telles que la précision, le rappel, le score F1 ou la courbe ROC-AUC [26].
- **Ajustement des hyperparamètres** : En cas de performances insataisfaisantes, nous pouvons ajuster les hyperparamètres de XGBoost et répéter les étapes de la création du modèle à son évaluation jusqu'à obtenir des résultats acceptables.
- **Sauvegarde du modèle final** : Après avoir évalué les performances du modèle et confirmé qu'il répond aux critères requis, nous procédons à sa sauvegarde pour une utilisation future.

3.6 Conclusion

L'adoption de XGBoost pour la prédiction des pré-failures, offre une amélioration substantielle en fournissant des prédictions plus précises et en temps réel des anomalies potentielles. Cette approche permet également d'optimiser les plans de maintenance en identifiant les défaillances avant qu'elles n'affectent le système.

Notre étude démontre que l'utilisation du XGBoost pour la détection précoce des anomalies dans les trains autonomes peut grandement améliorer la fiabilité et la sécurité de ces systèmes. En analysant des données réelles collectées à partir des capteurs des trains, nous avons pu développer un modèle capable d'identifier des schémas subtils et des signes précurseurs de défaillances. Cela permet aux opérateurs de prendre des mesures préventives en amont, évitant ainsi des interruptions de service coûteuses et des risques de sécurité.

Dans le chapitre suivant, nous présenterons une analyse détaillée des résultats obtenus à partir de notre modèle, et discuterons de son intégration dans un cadre de maintenance prédictive opérationnelle pour les trains autonomes. Nous explorerons également les possibilités d'amélioration, ainsi que les défis potentiels à surmonter pour une mise en œuvre efficace de cette approche dans des environnements réels.

CHAPITRE 4

SIMULATION ET ÉVALUATION DE PERFORMANCES

4.1 Introduction

Dans ce chapitre, nous présenterons les résultats de simulation et d'évaluation de performances de notre système de prédiction de défaillances basé sur XGBoost. Nous décrirons l'environnement de simulation utilisé, les paramètres de configuration et les métriques d'évaluation employées. Les résultats obtenus seront analysés en détail. Nous discuterons également de la résilience et de la robustesse de notre système face à différents scénarios de défaillances. Enfin, nous aborderons les avantages et les limites de notre solution, ainsi que les perspectives d'amélioration future.

4.2 Environement d'implémentation

Nous avons utilisé la pltafome kaggle pour implémenter notre approche.
Kaggle est une plateforme en ligne qui permet aux scientifiques des données de se lancer des défis sur des problèmes de machine learning. Elle propose des concours de science de données, des ensembles de données gratuits et une communauté pour apprendre et collaborer [40].

4.2.1 Outils et bibliothèques utilisés

Python :

Python est un langage de programmation interprété, interactif, orienté objet, et de haut niveau. Il est couramment utilisé pour l'analyse des données, le développement web, l'automatisation, et plus encore [41].

Matplotlib :

Matplotlib est une bibliothèque de traçage pour Python qui pérmis de créer des graphiques et des visualisations [44].

Scikit-learn (sklearn) :

Scikit-learn est une bibliothèque open-source de machine learning pour le langage de programmation Python. Elle comporte divers algorithmes de classification, régression et clustering, y compris le support vector machines, random forests, gradient boosting, k-means, et DBSCAN [43].

NumPy :

Numpy est une bibliothèque Python qui fournit des structures de données pour la manipulation de tableaux multidimensionnels et des fonctions mathématiques pour travailler avec ces tableaux [45].

4.3 Création du modèle

Dans cette section, nous allons explorer la configuration et l'optimisation de notre modèle de prédiction de pannes. Un aspect crucial de ce processus est la sélection et l'ajustement des hyperparamètres, qui jouent un rôle déterminant dans les performances et la robustesse du modèle.

4.3.1 Introduction aux Hyperparamètres

Les hyperparamètres sont des paramètres définis avant le processus d'entraînement du modèle et influencent directement son comportement et ses performances. Contrairement aux paramètres du modèle qui sont appris à partir des données d'entraînement, les hyperparamètres doivent être définis par l'utilisateur.

Pour notre modèle XGBoost, nous avons configuré plusieurs hyperparamètres clés pour optimiser les performances du modèle :

objective :

Définit la fonction de perte que le modèle essaie de minimiser. Pour notre problème de classification binaire, nous avons utilisé 'binary :logistic', qui correspond à la régression logistique pour une sortie binaire.

N_estimators :

Ce paramètre détermine le nombre de gradient boosted trees à construire. Nous avons initialement fixé ce paramètre à 10. Un nombre plus élevé d'estimateurs peut améliorer les performances mais augmente également le risque de surapprentissage et le temps de calcul.

Seed :

Permet de fixer la graine pour le générateur de nombres aléatoires, assurant ainsi la reproductibilité des résultats. Nous avons utilisé une graine de 17.

La Figure 4.1 représente les paramètres utilisés pour la création de notre modèle.

```
xgb_model = xgb.XGBClassifier(
    objective='binary:logistic',
    n_estimators=10,
    seed=17
)
```

FIGURE 4.1 – Paramètres de création du modèle.

4.4 Résultats des tests de prédiction de défaillances

Dans cette section, nous allons évaluer les performances de modèle de prédiction de défaillances proposé.

4.4.1 Résultats de modèle

Les résultats obtenus lors de l'évaluation de notre modèle de prédiction de défaillances sur le jeu de données de test, sont montrés dans le Tableau 4.1.

Classe	Précision	Rappel	F1-Score	Nombre d'instances
Sans défaillance	0.92	1.00	0.99	402030
Pré-défaillance	0.91	0.31	0.47	50456
Total	0.92	0.92	0.90	452486
Moyenne Macro	0.91	0.66	0.71	452486
Moyenne Pondérée	0.92	0.92	0.90	452486

TABLEAU 4.1 – Résultats des tests initiaux de prédiction de défaillances.

Nous pouvons observer que le modèle atteint une précision élevée de 0,92 pour la classe 0 (instances sans défaillance), avec un rappel parfait de 1,00 et un F1-Score qui atteint 0.99. Cependant, pour la classe 1 (instances de pré-défaillance), le rappel est plus faible à 0,31 et le F1-Score à 0.47, bien que la précision reste élevée à 0,91. Cela signifie que notre modèle a tendance à mieux détecter les instances sans défaillance, mais peut manquer certaines instances de pré-défaillance.

4.4.2 Résultats aprés création des caractéristiques

Après la création des nouvelles caractéristiques (**Oil_temp_avg_3day**, **dv_pressure_avg_3day**), les résultats obtenus sont illustrés dans le Tableau 4.2.

Classe	Précision	Rappel	F1-Score	Nombre d'instances
Sans défaillance	0.98	1.00	0.99	402030
Pré-défaillance	0.97	0.85	0.91	50456
Total	0.98	0.98	0.98	452486
Moyenne Macro	0.98	0.93	0.95	452486
Moyenne Pondérée	0.98	0.98	0.98	452486

TABLEAU 4.2 – Résultats des tests de prédiction de défaillance après ajout de nouvelles fonctionnalités.

Nous pouvons constater une amélioration significative des performances du modèle. La précision pour la classe 1 (instances de pré-défaillance) a augmenté à 0,97, et le rappel est passé à 0,85, soit une amélioration substantielle par rapport aux résultats précédents. La précision globale a également augmenté à 0,98, et le score F1 pondéré et global sont maintenant de 0,98, ce qui indique une performance excellente dans la classification des instances de pré-défaillance.

4.4.3 Analyse détaillée des performances du modèle

Dans cette section, nous allons examiner en profondeur les performances de notre modèle amélioré à travers deux outils d'évaluation essentiels : la matrice de confusion et la courbe Receiver Operating Characteristic (ROC).

Analyse de la matrice de confusion

La Figure 4.2 présente la matrice de confusion de notre modèle :

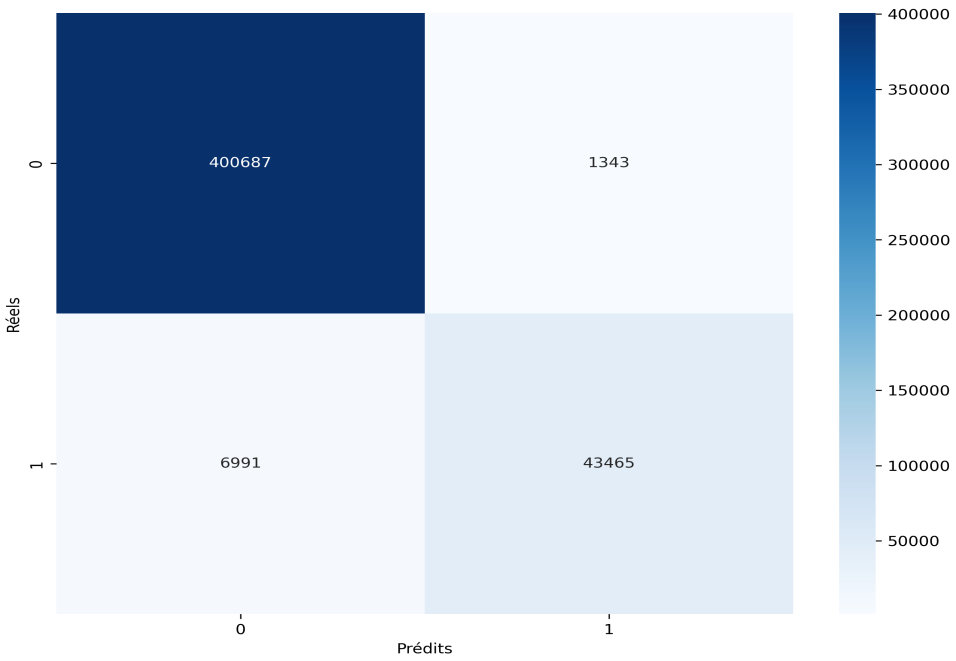

FIGURE 4.2 – Matrice de confusion.

Cette matrice nous offre un aperçu détaillé de la performance de notre modèle en termes de classification des instances normales et des pré-défaillances. Nous pouvons observer que :

- **Vrai Négatif (VN)** : 400687 instances ont été correctement classées comme n'étant pas des pré-défaillances.
- **Faux Positif (FP)** : 1343 instances ont été incorrectement classifiées comme des pré-défaillances alors qu'elles ne l'étaient pas.
- **Faux Négatif (FN)** : 6991 instances de pré-défaillances n'ont pas été détectées par le modèle.
- **Vrai Positif (VP)** : 43465 instances ont été correctement identifiées comme des pré-défaillances.

Cette matrice montre que le modèle a une excellente performance pour identifier les instances normales (VN élevé), et une bonne capacité à détecter les pré-défaillances (VP élevé). Cependant, il y a encore un certain nombre de faux négatifs, ce qui indique que certaines pré-défaillances ne sont pas détectées.

Analyse de la Courbe ROC

La Figure 4.3 est une illustration de la courbe ROC.

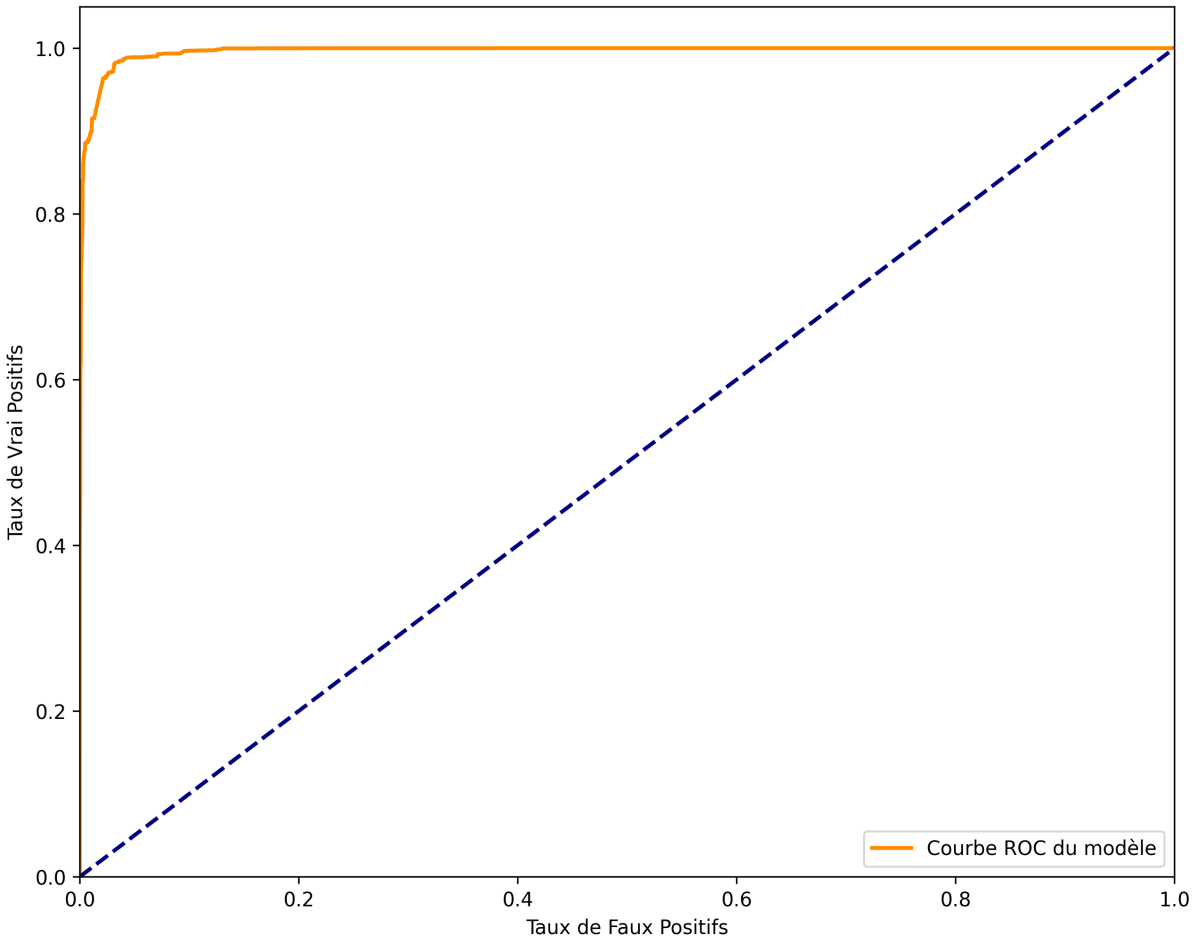

FIGURE 4.3 – Courbe ROC.

Voici les points clés à noter :

— L'axe X représente le taux de faux positifs (1 - Spécificité), tandis que l'axe Y représente le taux de vrais positifs (Rappel).
— La courbe orange représente la performance du modèle à différents seuils de classification.
— La ligne pointillée diagonale représente la performance d'un classificateur aléatoire (AUC = 0.5).
— L'aire sous la courbe (AUC) est de 0.9964, ce qui est extrêmement proche de 1. Cela indique une excellente performance du modèle.
— La courbe est très proche du coin supérieur gauche du graphique, ce qui est idéal car cela signifie que le modèle a un taux élevé de vrais positifs et un taux faible de faux positifs.

La courbe ROC confirme les excellentes performances du modèle observées dans la matrice de confusion. Avec une AUC de 0.9964, le modèle démontre une capacité remarquable à distinguer les instances de pré-défaillance des conditions normales de fonctionnement, sur l'ensemble des seuils de classification possibles.

4.5 Étude Comparative

Dans cette section, nous effectuons une étude comparative entre les résultats de notre système de prédiction de défaillances et les résultats rapportés dans une étude similaire. Cette comparaison est essentielle pour mettre en évidence l'efficacité et les avancées de notre approche par rapport aux méthodologies existantes.

4.5.1 Analyse Comparative

Selon la taxonomie des systèmes examinés dans le **Chapitre 2**, notre système peut être classé sous la catégorie : "Systèmes de prédiction et d'analyse des défaillances", car notre méthode se concentre sur la réalisation de prédictions et l'analyse des éventuelles défaillances des composants ou sous-systèmes. En effet, le modèle XGBoost que nous avons implémenté est spécifiquement conçu pour identifier les écarts par rapport au comportement normal du système, signalant ainsi les potentielles défaillances.

L'étude référencée est celle de Kang et al. [29], dans laquelle un modèle de prédiction de défaillances pour les systèmes de trains autonomes basé sur LSTM est présentée.

Pour fournir une comparaison plus claire, nous présentons les métriques de performance des deux modèles dans la Figure 4.4 :

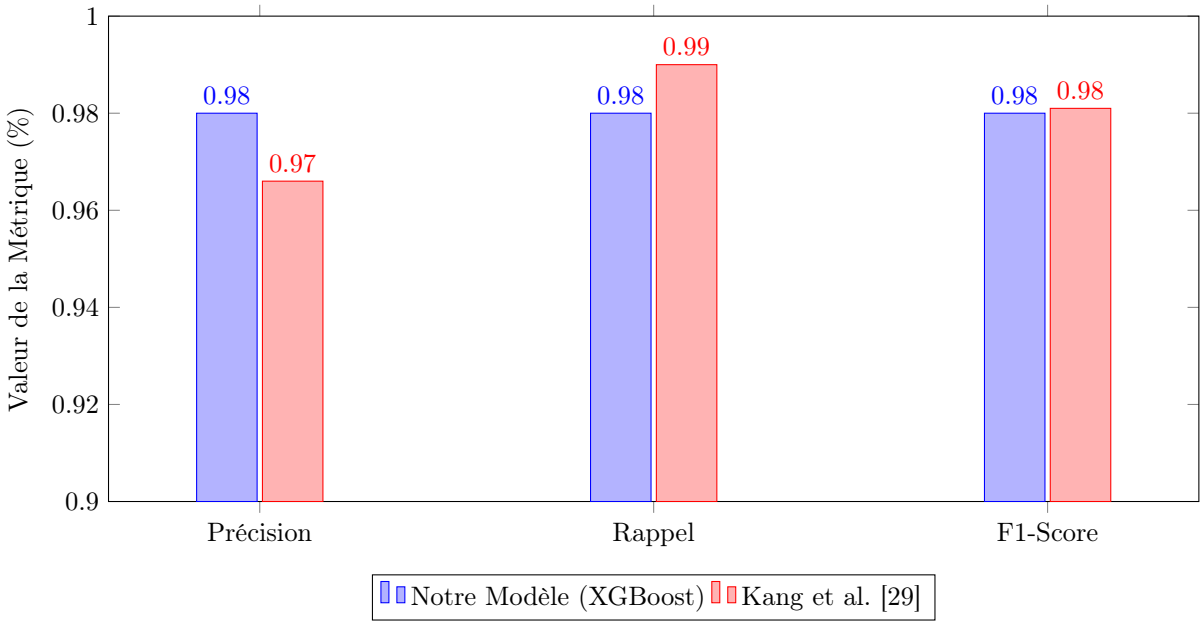

FIGURE 4.4 – Comparaison des métriques de performance entre notre modèle et le modèle référencé.

4.5.2 Discussion

Selon les résultats, les deux modèles sont particulièrement efficaces dans la prédiction des pannes des systèmes de trains autonomes. Toutefois, notre modèle présente une précision légèrement plus élevée, assurant un meilleur taux de prédictions positives vraies parmi toutes les prédictions positives. Bien que le

modèle LSTM-AP ait un rappel légèrement supérieur, ce qui suggère une meilleure capacité à repérer tous les cas pertinents, les F1-scores globaux sont presque identiques, ce qui suggère une efficacité similaire.

4.6 Analyse de la résilience et de la robustesse du système

Dans un système complexe tel que celui des trains autonomes, la résilience est essentiel pour assurer un fonctionnement fiable en cas de défaillances ou de conditions imprévues. Notre système de prédiction de défaillances doit être en mesure de résister aux pannes, de se remettre rapidement des perturbations et de continuer à fonctionner de manière efficace. Pour atteindre cet objectif, nous intégrons divers mécanismes de résilience à différents niveaux de notre architecture.

4.6.1 Redondance des capteurs

La première ligne de défense contre les défaillances réside dans la redondance des capteurs. Les trains autonomes sont équipés de multiples capteurs qui surveillent différents aspects du système, tels que la pression, la température, le courant du moteur, etc. En cas de défaillance d'un capteur, les autres capteurs redondants peuvent fournir des informations de secours, permettant au système de continuer à fonctionner correctement.

4.6.2 Surveillance en temps réel et gestion des alertes

Notre système de prédiction de défaillances est conçu pour fonctionner en temps réel, analysant continuellement les flux de données provenant des capteurs. Cela nous permet de détecter rapidement les anomalies et les tendances indiquant des pré-défaillances potentielles. Lorsqu'une pré-défaillance est détectée, notre système déclenche des alertes appropriées pour informer les opérateurs et les systèmes de contrôle.

4.7 Discussion des avantages et limites de la solution proposée

Le système de prédiction de défaillance proposé offre des bénéfices notables, tout en présentant certaines limites. Voici une analyse détaillée des avantages et des contraintes de cette solution.

4.7.1 Avantages

Notre approche de prédiction de défaillance basée sur XGBoost présente plusieurs avantages importants :

— **Détection précoce des défaillances** : Notre système permet de détecter les schémas précurseurs de défaillance jusqu'à 3 jours avant leur occurrence réelle. Cela offre un délai précieux pour intervenir et effectuer une maintenance préventive, évitant ainsi des temps d'arrêt coûteux et des risques pour la sécurité.

— **Performances élevées** : Le modèle proposé offre des performances élevées en termes de précision et de rapidité d'exécution, grâce aux optimisations internes de XGBoost.

— **Interprétabilité** : Notre système présente une bonne interprétabilité, ce qui facilite l'identification des éléments clés qui contribuent aux prédictions de défaillance. Cela peut faciliter la compréhension des causes sous-jacentes et guider les efforts de maintenance.

4.7.2 Limites

Malgré ses nombreux avantages, notre système présente également certaines limitations :

— **Dépendance à la qualité des données** : Bien que le système proposé intègre des techniques de gestion des données imparfaites, sa performance dépend fortement de la qualité et de la fiabilité des données d'entrée provenant des capteurs. Des erreurs systématiques ou des défaillances majeures de capteurs peuvent affecter la précision des prédictions.

— **Applicabilité restreinte** : Même si les résultats sont prometteurs sur l'ensemble de données utilisé, il n'est pas certain que ces performances se maintiennent sur d'autres jeux de données, ou dans des environnements opérationnels réels différents de celui simulé.

— **La complexité des données** : L'ajout de nouvelles fonctionnalités a également entraîné une augmentation de la complexité du modèle. Cette dernière demande une collecte de données plus avancées, ce qui pourrait potentiellement limiter l'efficacité du modèle.

— **Nécessité d'une expertise humaine** : Bien que notre système est largement automatisé, il sera toujours essentiel d'avoir une expertise humaine pour interpréter les résultats, valider les prédictions et prendre des décisions de maintenance éclairées.

4.8 Conclusion

Dans ce chapitre, nous avons présenté les résultats de simulation et d'évaluation des performances de notre système de prédiction de défaillance basé sur XGBoost. Les résultats obtenus démontrent des performances élevées avec une précision globale de 98%, ainsi qu'une amélioration significative du rappel pour les instances de pré-défaillance, passant de 31% à 85%, ce qui illustre une capacité améliorée à détecter les anomalies potentielles avant qu'elles ne se produisent. La robustesse du modèle, combinée aux mécanismes de résilience intégrés, offre une solution prometteuse pour améliorer la sécurité et la fiabilité des opérations ferroviaires autonomes. Cependant, les limites identifiées soulignent l'importance d'une approche prudente dans le déploiement et la maintenance continue du système.

CONCLUSION GÉNÉRALE ET PERSPECTIVES

Ce mémoire a présenté une approche innovante pour la prédiction de défaillances dans les trains autonomes. il met en évidence le potentiel important des techniques d'apprentissage automatique avancées, notamment l'algorithme XGBoost, pour améliorer la sécurité et l'efficacité des trains autonomes. L'étude a traversé plusieurs étapes cruciales, de l'analyse approfondie des défis posés par les trains autonomes à la conception et l'implémentation d'un système de prédiction de défaillances avancé.

Dans notre travail, nous avons posé les bases théoriques nécessaires pour comprendre les enjeux de la prédiction de défaillances, soulignant l'importance cruciale de ces systèmes pour la sécurité et l'efficacité des opérations ferroviaires modernes. L'apport central de ce mémoire réside dans la conception et le développement d'un système de prédiction de défaillances basé sur XGBoost. Cette approche innovante, capable de prévoir les pré-failures jusqu'à trois jours avant leur occurrence, a montré un potentiel significatif pour améliorer la maintenance prédictive ainsi que la sécurité opérationnelle. L'utilisation de XGBoost, combinée à des techniques avancées de prétraitement des données et d'ingénierie des caractéristiques, a permis d'obtenir des résultats remarquables en termes de précision et de rappel.

Les résultats obtenus ont confirmé l'efficacité de l'approche proposée, avec une amélioration notable des performances après l'ajout de nouvelles fonctionnalités basées sur des moyennes mobiles. La précision globale de 98% et le score ROC-AUC de 99% témoignent de la capacité du modèle à distinguer efficacement les instances de pré-défaillance des conditions normales de fonctionnement.

Bien que notre système de détection de défaillances présente de nombreux avantages, certaines limitations demeurent telles que la dépendance aux données historiques et les défis liés à la mise à jour continue du modèle. Ces limitations soulignent la nécessité de poursuivre les recherches et les développements dans ce domaine. L'amélioration de la gestion des données, en particulier en termes de qualité et de sécurité, est également cruciale pour garantir la fiabilité et la précision des prédictions.

Pour les recherches à venir, il serait possible d'explorer des méthodes d'apprentissage en ligne afin d'ajuster dynamiquement le modèle aux nouvelles conditions, l'intégration de sources de données supplémentaires pour améliorer la robustesse des prédictions, et le développement de méthodes plus avancées pour la gestion des fausses alertes. En outre, il serait judicieux d'améliorer la résistance des systèmes de prédiction de défaillance en incorporant des technologies avancées comme l'Internet des Objets (IoT) et le traitement en temps réel des données.

BIBLIOGRAPHIE

[1] R. Zhao et al. "Deep learning and its applications to machine health monitoring". In : *Mechanical Systems and Signal Processing* 115 (2019), p. 213-237.

[2] C. Zhang, L. Yao et Y. Sun. "Deep Learning-Based Approaches for Predictive Maintenance in Industrial IoT : A Comprehensive Review". In : *IEEE Internet of Things Journal* 9.11 (2022), p. 8046-8070.

[3] X. Xu et H. He. "Internet of Things in Industries : A Survey". In : *IEEE Transactions on Industrial Informatics* 10.4 (2014), p. 2233-2243.

[4] S. M. Asadzadeh et S. Buyaki. "Machine Failure Prognostics : A Systematic Review of Unsupervised and Transfer Learning Approaches". In : *Reliability Engineering & System Safety* 214 (2021), p. 107731.

[5] T. Chen et C. Guestrin. "Xgboost : A scalable tree boosting system". In : *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, p. 785-794.

[6] Bastian Simoni. *What is an autonomous train ?* `https://voie-libre.com/en/what-is-an-autonomous-train/`. (Consulté le 31 Mai 2024).

[7] `https://www.alstom.com/autonomous-mobility-future-rail-automated`. (Consulté le 15 Mars 2024).

[8] Observatory of Automated Metros Data. *WORLD REPORT ON METRO AUTOMATION*. `https://cms.uitp.org/wp/wp-content/uploads/2020/06/Statistics-Brief-Metro-automation_final_web03.pdf`. 2018.

[9] Pedro Manuel da Silva Oliveira Pinto. *Projeto de um sistema de aquisição de dados para a manutenção preditiva de uma unidade embarcada de produção de ar comprimido*. 2019.

[10] acerta. *What is anomaly detection in manufacturing ?* `https://acerta.ai/blog/anomaly-detection-in-manufacturing/`. (Dèrnière mise à jour le 21 Mars 2024).

[11] J. Han, M. Kamber et J. Pei. *Data Mining : Concepts and Techniques*. Morgan Kaufmann, 2012.

[12] N. Davari et al. "A Survey on Data-Driven Predictive Maintenance for the Railway Industry". In : *Sensors* 21 (2021), p. 5739. doi : `10.3390/s21175739`.

[13] R. K. Mobley. *An Introduction to Predictive Maintenance*. 2e éd. Elsevier, 2002, p. 437. isbn : 0080478697.

[14] M. Bengtsson et G. Lundström. "On the importance of combining "the new" with "the old" – One important prerequisite for maintenance in Industry 4.0". In : *8th Swedish Production Symposium, SPS 2018*. T. 25. Mälardalen University et Volvo Construction Equipment Operations. Stockholm, Sweden : Procedia Manufacturing, 2018, p. 118-125.

[15] J. H. Friedman. "Greedy function approximation : a gradient boosting machine". In : *The Annals of statistics* 29.5 (2001), p. 1189-1232.

[16] R. E. Schapire. "The boosting approach to machine learning : An overview". In : *Nonlinear estimation and classification*. Springer, New York, NY, 2003, p. 149-171.

[17] Adib Habbou. *Algorithmes de gradient boosting et introduction à XGBoost*. https://larevueia.fr/algorithmes-de-gradient-boosting-et-introduction-a-xgboost/. Consulté le 23-05-2024.

[18] M. Hossin et M.N. Sulaiman. "A REVIEW ON EVALUATION METRICS FOR DATA CLASSIFICATION EVALUATIONS". In : *International Journal of Data Mining Knowledge Management Process (IJDKP)* 5 (2015). doi : /10.5121/ijdkp.2015.5201.

[19] A. Ouadah, L. Zemmouchi-Ghomari et N. Salhi. "Selecting an appropriate supervised machine learning algorithm for predictive maintenance". In : *The International Journal of Advanced Manufacturing Technology* 119 (2022), p. 4277-4301. doi : 10.1007/s00170-021-08551-9.

[20] M. Sokolova et G. Lapalme. "A systematic analysis of performance measures for classification tasks". In : *Information Processing and Management* 45 (2009), p. 427-437. doi : /10.1016/j.ipm.2009.03.002.

[21] A. Jaramillo-Alcazar, J. Govea et W. Villegas-Ch. "Anomaly Detection in a Smart Industrial Machinery Plant Using IoT and Machine Learning". In : *Sensors* 23 (2023), Article number 8286. doi : /10.3390/s23198286.

[22] M. Maabreh et al. "The robustness of popular multiclass machine learning models against poisoning attacks : Lessons and insights". In : *International Journal of Distributed Sensor Networks* 18 (7 2022). doi : /10.1177/15501329221105159.

[23] A.Vadde. *Scalability*. https://www.modernismmodernity.org/forums/posts/scalability. (Consulté le 16 Mars 2024). 2018.

[24] Kobia. *F1-score, la synthèse entre precision et recall*. https://kobia.fr/classification-metrics-f1-score/. (Consulté le 14 Mars 2024).

[25] D.G. Altman et J. M. Bland. "Diagnostic tests. 1 : Sensitivity and specificity". In : *BMJ* 308.6943 (1994). Clinical research ed., p. 1552. doi : 10.1136/bmj.308.6943.1552.

[26] J. A. Hanley et B. J. McNeil. "The meaning and use of the area under a receiver operating characteristic (ROC) curve". In : *Radiology* 143.1 (1982), p. 29-36. doi : 10.1148/radiology.143.1.7063747.

[27] L. De Simone et al. "LSTM-based failure prediction for railway rolling stock equipment". In : *Expert Systems With Applications (ESWA)* 222 (2023). doi : /10.1016/j.eswa.2023.119767.

[28] S. Kauschke, J. Fürnkranz et F. Janssen. "Predicting Cargo Train Failures : A Machine Learning Approach for a Lightweight Proto-type". In : Lecture Notes in Artificial Intelligence 9956 (2016), p. 151-166. doi : /10.1007/978-3-319-46307-010.

[29] R. Kang et al. "A method of online anomaly perception and failure prediction for high-speed automatic train protection system". In : *Reliability Engineering and System Safety* 226 (2022). doi : /10.1016/j.ress.2022.108699.

[30] M.H. Le-Nguyen et al. "Real-time learning for real-time data : online machine learning for predictive maintenance of railway systems". In : *Transportation Research Procedia* 72 (2023), p. 171-178. doi : /10.1016/j.trpro.2023.11.391.

[31] Jie. Li et al. "A Novel Method for Aging Prediction of Railway Catenary Based on Improved Kalman Filter". In : *Structural Durability Health Monitoring (SDHM)* 18.1 (2024). doi : /10.32604/sdhm.2023.044023.

[32] P.C.L. Gerum, A. Altay et M. Baykal-Gürsoy. "Data-driven predictive maintenance scheduling policies for railways". In : *Transportation Research Part C* 107 (2019), p. 137-154. doi : /10.1016/j.trc.2019.07.020.

[33] H. Hadj-Mabrouk. "Analysis and prediction of railway accident risks using machine learning". In : *AIMS Electronics and Electrical Engineering* 4 (1 2019), p. 19-46. doi : /10.3934/ElectrEng.2020.1.19.

[34] M. de Almeida Costa, J. P. de A. P. Braga et A. R. Andrade. "A data-driven maintenance policy for railway wheelset based on survival analysis and Markov decision process". In : *Quality and Reliability Engineering International* (2020), p. 1-23. doi : /10.1002/qre.2729.

[35] N. Davari et al. "Predictive maintenance based on anomaly detection using deep learning for air production unit in the railway industry". In : *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*. 2021, p. 1-10.

[36] Scikit-learn developers. *MinMaxScaler*. https : / / scikit - learn . org / stable / modules / generated/sklearn.preprocessing.MinMaxScaler.html. Consulté le 21-05-2024.

[37] C. F. Dormann et al. "Collinearity : a review of methods to deal with it and a simulation study evaluating their performance". In : *Ecography* 36 (2013), p. 27-46. doi : 10.1111/j.1600-0587.2012.07348.x.

[38] Z. H. Ali et A. M. Burhan. "Hybrid machine learning approach for construction cost estimation : an evaluation of extreme gradient boosting model". In : *Asian Journal of Civil Engineering* 24 (2023), p. 1-16. doi : 10.1007/s42107-023-00651-z.

[39] S. M. Lundberg et al. "From local explanations to global understanding with explainable AI for trees". In : *Nature machine intelligence* 2.1 (2020), p. 56-67.

[40] *Kaggle : Your Machine Learning and Data Science Community*. https://www.kaggle.com/about. Consulté le 25-05-2024.

[41] Python Software Foundation. *Python*. https://www.python.org/. Consulté le 06-06-2024.

[42] Nina Reunes. *NumPy, Pandas, and Scikit-Learn Explained*. https : / / medium . com / personal - project / numpy - pandas - and - scikit - learn - explained - e7336baecedc. (Consulté le 07-06-2024).

[43] F. Pedregosa, G. Varoquaux et all. “Scikit-learn : Machine Learning in Python”. In : *Journal of Machine Learning Research* 12 (2011), p. 2825-2830.

[44] Matplotlib. *Matplotlib : Visualization with Python*. `https://matplotlib.org/`. Consulté le 07-06-2024.

[45] NumPy developers. *NumPy*. `https://numpy.org/`. Consulté le 10-06-2024.

RÉSUMÉ

Les systèmes de trains autonomes modernes posent des défis majeurs en matière de maintenance prédictive, en raison de leur complexité grandissante et de l'interdépendance de leurs composants. Cette situation se traduit par des arrêts imprévus, des coûts de maintenance élevés et des risques pour la sécurité des passagers. Pour relever ces défis, ce mémoire propose une approche innovante basée sur l'apprentissage automatique avancé, avec l'utilisation de l'algorithme XGBoost, afin de développer un système de détection précoce des défaillances.

La méthode employée combine des techniques de normalisation, de sélection de caractéristiques pertinentes et d'apprentissage automatique performant. Cela permet d'identifier les anomalies plusieurs jours avant leur survenue, offrant ainsi un délai précieux pour mettre en place des actions préventives. En détectant préventivement les défaillances potentielles, cette solution rend le système de trains autonomes plus robuste et résilient face aux imprévus. Grâce à cette détection précoce, le système peut mieux s'adapter aux perturbations, en minimisant les temps d'arrêt et en assurant une continuité de service accrue.

Les résultats obtenus sont très encourageants, avec des taux de précision et de rappel atteignant 98%. Cette approche contribue donc à améliorer la sécurité, la sûreté et la résilience des services de trains autonomes.

Mots clés : Maintenance prédictive, Trains autonomes, Apprentissage automatique, Détection d'anomalies, XGBoost, Résilience

ABSTRACT

Modern autonomous train systems pose significant challenges in predictive maintenance due to their increasing complexity and the interdependence of their components. This situation leads to unexpected shutdowns, high maintenance costs, and safety risks for passengers. To address these challenges, this thesis proposes an innovative approach based on advanced machine learning, using the XGBoost algorithm, to develop an early failure detection system.

The method employed combines normalization techniques, relevant feature selection, and high-performance machine learning. This enables the identification of anomalies several days before they occur, providing a crucial window for preventive actions. By preemptively detecting potential failures, this solution makes autonomous train systems more robust and resilient to unexpected issues. This early detection allows the system to better adapt to disruptions, minimizing downtime and ensuring increased service continuity.

The results obtained are very encouraging, with precision and recall rates reaching 98%. This approach thus helps to enhance the safety, security, and resilience of autonomous train services.

Keywords : Predictive Maintenance, Autonomous Trains, Machine Learning, Anomaly Detection, XGBoost, Resilience