Predictive maintenance of aircraft bleed air systems : an approach based on machine learning and variational autoencoders

William Todo

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *06/10/2023* **par :**

William TODO

Maintenance prédictive des systèmes d'air d'avion : une approche basée sur l'apprentissage automatique et les autoencodeurs variationnels

JURY

Jean-François DUPUY	Professeur d'Université	Président du Jury
Christophe BIERNACKI	Professeur d'Université	Rapporteur
Mathilde MOUGEOT	Professeur d'Université	Rapporteur
Béatrice LAURENT	Professeur d'Université	Directrice de thèse
Jean-Michel LOUBES	Professeur d'Université	Directeur de thèse
Nicolas CANOUET	Dir. Datalab Liebherr Aerospace Toulouse	Invité

École doctorale et spécialité :
MITT : Domaine Mathématiques : Mathématiques appliquées
Unité de Recherche :
Institut de mathématiques de Toulouse
Directeur(s) de Thèse :
Béatrice LAURENT et *Jean-Michel LOUBES*
Rapporteurs :
Christophe BIERNACKI et *Mathilde MOUGEOT*

Remerciements

En premier lieu, je tiens à exprimer ma gratitude envers mes directeurs de thèse, Béatrice et Jean-Michel. Votre confiance, vos conseils éclairés et votre disponibilité constante ont grandement facilité le déroulement de cette thèse. Je vous remercie pour votre patience infinie et votre capacité à me stimuler dans la rédaction de ce travail. Cette aventure était challengeante mais aussi passionnante et je suis extrêmement heureux d'avoir eu l'opportunité de travailler à vos côtés.

Par ailleurs, je souhaite sincèrement remercier Nicolas Canouet. Tu as su me faire confiance pour m'intégrer à l'équipe du datalab et conduire cette thèse. Tu m'as accordé une grande liberté dans le déroulement de ce travail, tout en fournissant un soutien et des conseils indispensables. Je ne saurais trop souligner à quel point j'ai apprécié travailler avec toi.

Il me tient aussi à cœur de remercier l'équipe du datalab. Merci à toi Merwann, ta maîtrise des signaux des systèmes d'air est impressionnante, et tu as su m'initier à ces problématiques avec un humour certain. Merci aussi à toi Fabien sans qui l'accès aux données et aux machines aurait été très compliqué, merci pour ta disponibilité. Je remercie aussi toute l'équipe du datalab qui s'agrandit à vue d'œil dans la bonne humeur.

Ces années ont été longues et j'ai eu le privilège de pouvoir compter sur le soutien indéfectible de mes amis et de ma famille. Un merci tout particulier à Hugo, dont les visites impromptues du jeudi soir ont souvent égayé ma semaine ; à Matthieu, toujours présent, même lorsqu'il s'agissait de perdre aux cartes avec un certain panache il faut le dire.Lucas, je te remercie pour ces précieux moments de répit que nous avons partagés autour d'un verre ou d'un repas. Et évidemment, un immense merci à Luna pour les rires, les encouragements et le reste. Bien sûr, il m'est impossible de citer tout le monde ici, mais sachez que ma gratitude vous est acquise à tous.

Je souhaite aussi remercier Mathilde Mougeot et Christophe Biernacki d'avoir bien voulu rapporter ma thèse.

Contents

List of Figures

List of Tables

Introduction

En français

L'industrie aéronautique, secteur exigeant une fiabilité et une précision rigoureuses, produit des avions capables de résister à des environnements hostiles. Ils doivent en particulier faire face à des températures fluctuant entre -60°C lors des croisières en altitude, et dépassant aisément les 500°C en sortie des moteurs. Liebherr-Aerospace se positionne au cœur de ce secteur, concevant, développant et fabriquant une variété de systèmes essentiels à l'avion : systèmes de climatisation, commandes de vol, trains d'atterrissage, sans oublier engrenages, boîtes de transmission et éléments électroniques.

En tant que constructeur d'équipements d'origine (Original Equipment Manufacturer - OEM), la responsabilité de Liebherr-Aerospace est d'assurer une production d'équipements hautement qualitatifs et fiables. Cette responsabilité s'étend non seulement à la garantie de la qualité de leurs produits, mais également à l'assurance d'un approvisionnement en pièces pour la production et la maintenance, ainsi qu'à la fourniture d'un support technique pour leurs produits. Mais les responsabilités de Liebherr-Aerospace ne se limitent pas à la production. En plus de son rôle d'OEM, l'entreprise fournit des services complets de Maintenance, Réparation et Opérations (MRO). En d'autres termes, elle s'implique dans l'entretien régulier des avions, la réparation des composants, et le support aux opérations de vol, assurant ainsi la sécurité et l'efficacité des opérations aériennes en maintenant les avions en bon état de fonctionnement. Un autre aspect crucial de son activité réside dans la fourniture de services Aircraft On Ground (AOG). Ces services, devenant essentiels lors de l'immobilisation d'un avion pour des raisons techniques ou mécaniques, peuvent permettre d'économiser des coûts substantiels pour les compagnies aériennes, car chaque minute d'immobilisation au sol a un coût financier conséquent. Les services AOG ont pour objectif de résoudre rapidement ces incidents, en assurant des réparations d'urgence, des pièces de rechange et une coordination des services afin de minimiser le temps d'immobilisation de l'appareil.

En résumé, Liebherr-Aerospace est une entreprise qui joue un rôle clé à plusieurs niveaux de l'industrie aéronautique, endossant plusieurs rôles majeurs : elle est fabricant d'équipements, assure des services de maintenance et fournit des services AOG. De manière plus spécifique, cette thèse portera sur une des spécialité de Liebherr-Aerospace Toulouse : les systèmes d'air, également appelés "bleed air systems". Dans ce contexte le développement d'outils de maintenance prédictive est essentiel, car ils permettent d'abord de réduire le temps d'immobilisation au sol des avions. En effet, si proposer une solution suite a une panne est important, pouvoir anticiper les pannes et en réduire le nombre est également crucial.

Cette thèse vise donc à élaborer de nouvelles méthodes de maintenance prédictive performantes, acceptables et explicables. Fruit d'une collaboration entre plusieurs entités sous le format CIFRE, ce travail de recherche implique l'entreprise Liebherr Aerospace Toulouse, l'Institut de Mathématiques de Toulouse, ainsi que ANITI (Artificial and Natural Intelligence Toulouse Institute).

Le premier chapitre, intitulé "**État de l'art de la maintenance prédictive**", servira d'introduction à la maintenance prédictive en soulignant à la fois ses similitudes et ses différences avec la détection d'anomalies. Nous effectuerons ensuite dans une revue de la littérature, où nous découvrirons comment les techniques de machine learning et de deep learning sont mises à profit dans le domaine de la maintenance prédictive. Cette partie du chapitre mettra en lumière l'importance des ensembles de données publiques pour la comparaison et l'évaluation des techniques de maintenance prédictive. De plus, nous mettrons en évidence comment ces datasets publics permettent aux chercheurs de développer de nouvelles méthodes dans des conditions optimales, favorisant ainsi l'innovation et l'avancement du domaine. Cependant, nous ne manquerons pas de souligner les contraintes actuelles liées à la disponibilité et à la qualité des datasets, tout en mettant en avant les efforts continus et les progrès réalisés pour améliorer ces ressources essentielles.

Le deuxième chapitre, "**Bleed Air Systems: Data & Challenges**" abordera tout d'abord le fonctionnement global des systèmes d'air, également connus sous le nom de systèmes bleed, ainsi que les types de signaux qu'ils peuvent générer. Cette exploration détaillée établira le contexte spécifique de cette thèse et favorisera une meilleure compréhension des types de signaux qui seront analysés et exploités tout au long de notre travail. Après avoir cadré notre champ d'études, nous présenterons les premières tentatives de maintenance prédictive, en commençant par l'extraction de features simples suivis de la prédiction de la durée de vie restante (Remaining Useful Life - RUL) avec des techniques d'apprentissage automatiques traditionnels et la prédiction de la durée de vie restante directement à l'aide de réseaux de convolutions.

Cette approche initiale, en dépit de sa simplicité apparente, mettra en lumière les difficultés fondamentales liées à la maintenance prédictive dans le contexte des systèmes bleed. Les défis rencontrés, tant sur le plan de la capture de la complexité des signaux que sur la précision des prédictions de durée de vie restante, mettront en évidence l'importance de l'approche méthodologique et motiveront nos choix pour les phases ultérieures de la thèse. Ces constatations préliminaires nous permettront de mieux orienter notre travail de recherche et de souligner la nécessité d'adopter des méthodes plus sophistiquées et précises pour la maintenance prédictive des systèmes bleed.

Le troisième chapitre, "**Dimension Reduction for time series with Variational Autoencoders**" mettra en lumière le rôle crucial des auto encodeurs variationnels (Variational Autoencoders - VAE) dans le domaine de la réduction de dimension. Nous examinerons la performance des VAE par rapport à d'autres techniques de réduction de dimension largement utilisées, notamment la transformée en ondelettes et l'analyse en composantes principales fonctionnelle (Functional Principal Component Analysis - FPCA). Notre analyse montrera que les VAE se distinguent par leur capacité à obtenir des taux de compression élevés, surpassant même la transformée en ondelettes dans des scénarios hypothétiques où l'on retiendrait les coefficients d'ondelettes optimaux. Ces résultats impressionnants sont validés lorsque nous appliquons le VAE à différents jeux de données d'électrocardiogrammes (ECG), soulignant ainsi sa capacité à généraliser et à être efficace sur une variété de cas d'utilisation. Au-delà de la performance brute du VAE, ce chapitre explorera l'impact de différentes architectures VAE sur la capacité de réduction de dimension. Nous examinerons comment des variations dans la conception des VAE peuvent influencer leurs performances, apportant des informations précieuses pour le choix de l'architecture appropriée dans des applications spécifiques. De plus, nous évaluerons la robustesse des VAE face au bruit et leur capacité à maintenir une réduction de dimension efficace dans des conditions non idéales. En résumé, ce chapitre mettra en avant les performanaces des VAE en matière de réduction de dimension, leur efficacité, leur flexibilité et leur robustesse, illustrant ainsi leur pertinence pour notre travail de thèse.

Le quatrième chapitre, "**Counterfactual explanation for multivariate times series using a contrastive variational autoencoder**" présentera une version étendue d'un article publié à conférence ICASSP 2023. Ce chapitre aborde une question essentielle : comment comprendre les comportements anormaux des séries temporelles multivariées ? Cette problématique est centrale dans notre étude étant donné la nature des données que nous manipulons, à savoir des séries temporelles

multivariées. En effet, il est crucial non seulement de pouvoir prédire la dégradation des données et l'évolution des modèles, mais aussi de comprendre les mécanismes sous-jacents à ces prédictions. Toutefois, peu de méthodes disponibles offrent une explication contrefactuelle pour les séries temporelles et les rares existantes ne sont pas adaptées ou scalables pour les types de jeux de données qui nous intéressent. Face à cette lacune, nous avons développé une nouvelle approche. Nous montrerons comment, grâce à une séparation astucieuse de l'espace latent d'un VAE à l'aide d'une contrainte contrastive, nous parvenons à générer des espaces latents partiellement ordonnés. Ces derniers nous permettent de concevoir des exemples contrefactuels avec une grande efficacité. Pour attester de l'efficacité de cette méthode, nous présenterons son application à un ensemble de données publiques de signaux ECG. Cette validation sur des données réelles soulignera le potentiel de notre approche pour améliorer la compréhension des prédictions d'anomalies pour des séries temporelles multivariées.

Le cinquième et dernier chapitre, "**Explainable Predictive Maintenance: Revealing Degradation Factors with Contrastive Semi-Supervised VAE**", sera dédié à l'adaptation de la méthode CVAE (Contrastive Variational Autoencoder) pour résoudre des problèmes de maintenance prédictive. Nous nous intéresserons notamment à l'intégration de la notion de voisinage dans le cycle de vie d'un équipement, ce qui correspond aux vols précédant et suivant le vol cible. Cette approche part de l'hypothèse que le niveau de dégradation devrait être similaire entre ces vols voisins. Par ailleurs, nous explorerons la possibilité d'entrainer le CVAE de manière semi-supervisée. Nous illustrerons comment, grâce à cette stratégie, le modèle peut être formé de manière efficace même en présence de données censurées, offrant ainsi une grande flexibilité en termes de gestion des données. Ce chapitre mettra également en évidence l'efficacité remarquable de ce modèle en le comparant à des modèles classiques de classification de séries temporelles, ainsi qu'à un VAE entraîné avec des données saines, complété par un algorithme de détection d'anomalies. Nos résultats démontreront que notre approche surpasse ces méthodes de base. Enfin, nous adapterons une technique connue de la classification d'images, appelée "selective kernels", à la classification de séries temporelles.

In english

The aeronautics industry, with its rigorous demands for reliability and precision, produces aircraft that must withstand harsh environments. These machines must be able to withstand temperatures ranging from -60°C during high-altitude flights to over 500°C when the air exits the engines. Liebherr-Aerospace is at the heart of this sector, designing, developing and manufacturing a wide range of systems essential to the aircraft: air conditioning, flight controls, landing gear, not to mention gears, gearboxes and electronic elements.

As an Original Equipment Manufacturer (OEM), Liebherr-Aerospace is responsible for the production of high quality and reliable equipment. This responsibility extends not only to guaranteeing the quality of its products, but also to ensuring the supply of parts for production and maintenance, as well as providing technical support for its products. But Liebherr-Aerospace's responsibility is not limited to production. In addition to its OEM role, the company provides comprehensive maintenance, repair and operations (MRO) services. In other words, it is involved in regular aircraft maintenance, component repair and flight operations support, ensuring safe and efficient flight operations by keeping aircraft in good working order. Another important aspect of its business is the provision of Aircraft On Ground (AOG) services. These services, which are essential when an aircraft is grounded for technical reasons, can save airlines significant costs, as every minute spent on the ground represents a significant financial expense. AOG services aim to resolve these incidents quickly by providing emergency repairs, spare parts and service coordination to minimize aircraft downtime.

In conclusion, Liebherr-Aerospace is a company that plays a key role in several levels of the aeronautical industry, taking on several important roles: it is an equipment manufacturer, a maintenance provider, and an AOG provider. More specifically, this thesis will focus on one of the specialties of Liebherr-Aerospace Toulouse: air systems, also called "bleed air systems". In this context, the development of predictive maintenance tools is essential, as they allow us to reduce aircraft downtime. Indeed, while it is important to provide a solution after a breakdown, it is also crucial to be able to anticipate breakdowns and reduce their number.

The aim of this thesis is to develop new predictive maintenance methods that are efficient, acceptable and explainable. This research is the result of a collaboration between several entities under the CIFRE format, involving Liebherr Aerospace

Toulouse, the Toulouse Mathematics Institute as well as ANITI (Artificial and Natural Intelligence Toulouse Institute).

The first chapter, entitled "**État de l'art de la maintenance prédictive**", will serve as an introduction to predictive maintenance by highlighting its similarities and differences with anomaly detection. We will then dive into a literature review where we will discover how machine learning and deep learning techniques are being used in the field of predictive maintenance. This part of the chapter will highlight the importance of public datasets for comparing and evaluating predictive maintenance techniques. In addition, we will show how these public datasets allow researchers to develop new methods under optimal conditions, thus promoting innovation and advancement of the field. However, we will not fail to point out the current limitations related to the availability and quality of datasets, while highlighting the ongoing efforts and progress made to improve these essential resources.

The second chapter, "**Bleed Air Systems: Data & Challenges**" will first discuss the overall operation of air systems, also known as bleed systems, and the types of signals they can generate. This detailed examination will establish the specific context of this thesis and promote a better understanding of the types of signals that will be analyzed and exploited throughout our work. After defining our field of study, we will present the first attempts of predictive maintenance, starting with the extraction of simple features, followed by the prediction of Remaining Useful Life (RUL) using traditional machine learning techniques, and the prediction of RUL directly using convolutional networks. This initial approach, despite its apparent simplicity, will highlight the fundamental difficulties associated with predictive maintenance in the context of bleed systems. The challenges encountered, both in terms of capturing signal complexity and the accuracy of remaining life predictions, will highlight the importance of the methodological approach and motivate our choices for the later phases of the thesis. These preliminary findings will help us to better focus our research efforts and highlight the need for more sophisticated and accurate methods for predictive maintenance of bleed systems.

The third chapter, "**Dimension Reduction for Time Series with Variational Autoencoders**" will highlight the crucial role of Variational Autoencoders (VAE) in the field of dimension reduction. We will examine the performance of VAE in comparison to other widely used dimension reduction techniques, including the Wavelet Transform and Functional Principal Component Analysis (FPCA). Our analysis will show that VAE is characterized by its ability to achieve high compression ratios, even outperforming the wavelet transform in hypothetical scenarios where

optimal wavelet coefficients are retained. These impressive results are validated when we apply VAE to different ECG datasets, highlighting its ability to generalize and perform across a variety of use cases. Beyond the raw performance of VAE, this chapter will explore the impact of different VAE architectures on dimension reduction capability. We will examine how variations in VAE design can affect its performance, providing valuable information for selecting the appropriate architecture in specific applications. In addition, we will evaluate the robustness of VAEs to noise and their ability to maintain effective dimension reduction under non-ideal conditions. In summary, this chapter will highlight the dimension reduction characteristics of VAEs, their efficiency, flexibility, and robustness, and illustrate their relevance to our thesis work.

The fourth chapter, "**Counterfactual explanation for multivariate times series using a contrastive variational autoencoder**" presents an extended version of a paper published at the ICASSP 2023 conference. This chapter addresses a key question: how to understand anomalous behavior in multivariate time series. This problem is central to our study given the nature of the data we are dealing with, multivariate time series. Indeed, it is crucial not only to be able to predict data degradation and model evolution, but also to understand the mechanisms underlying these predictions. However, few available methods provide a counterfactual explanation for time series, and the few that do exist are not suitable or scalable for the types of datasets we are interested in. Faced with this shortcoming, we have developed a new approach. We will show how, thanks to a strategic separation of the latent space of a VAE using a contrastive constraint, we manage to generate partially ordered latent spaces. These allow us to design counterfactual examples with great efficiency. To prove the effectiveness of this method, we will present its application on a public dataset of ECG signals. This validation on real data will highlight the potential of our approach to improve the understanding of multivariate time series prediction.

The fifth and final chapter, "**Explainable Predictive Maintenance: Revealing Degradation Factors with Contrastive Semi-Supervised VAE**", is dedicated to the adaptation of the CVAE (Contrastive Variational Autoencoder) method to solve the predictive maintenance problem. In particular, we will focus on the integration of the notion of neighborhood in the life cycle of an equipment, which corresponds to the flights before and after the target flight. This approach assumes that the level of degradation should be similar between these neighboring flights. We will also explore the possibility of semi-supervised training of the CVAE. We will illustrate how this strategy allows the model to be efficiently trained even in

the presence of censored data, thus providing great flexibility in terms of data management. This chapter will also highlight the remarkable efficiency of this model by comparing it to classical time series classification models, as well as to a VAE trained on healthy data and complemented by an anomaly detection algorithm. Our results will show that our approach outperforms these conventional methods. Finally, we will adapt a well-known image classification technique known as "selective kernels" to the classification of time series.

Chapter 1

État de l'art de la maintenance prédictive

1.1 Les maintenances dans l'industrie aéronautique

Il existe généralement trois stratégies de maintenance. La première, assez intuitive, est la maintenance réactive. Cette approche consiste à agir suite à une panne ou une alerte. Bien que parfois indispensable, cette forme de maintenance est imprévue et peut par conséquent entraîner des immobilisations d'appareils au sol - une situation que l'on cherche à éviter dans l'industrie aéronautique.

La deuxième forme de maintenance, largement adoptée dans le secteur aéronautique, est appelée la maintenance préventive. Cette stratégie de maintenance se focalise sur la prévention des pannes et l'amélioration de la fiabilité des équipements. Cela est réalisé grâce à des inspections régulières, des réparations, des mises à jour, des nettoyages et des remplacements de pièces. L'objectif principal de cette approche est d'identifier et de résoudre les problèmes potentiels avant qu'ils ne se transforment en défaillances d'équipement. Plusieurs caractéristiques essentielles distinguent la maintenance préventive. Tout d'abord, elle est planifiée : les tâches de maintenance préventive sont organisées en amont, selon un calendrier prédéterminé, basé sur le temps (par exemple, mensuellement) ou sur l'utilisation de l'équipement (par exemple, après un nombre précis d'heures de fonctionnement). De plus, la maintenance préventive est systématiquement effectuée en suivant des procédures standard pour chaque type d'équipement, garantissant ainsi que toutes les étapes nécessaires sont accomplies.

Enfin, contrairement à la maintenance réactive qui intervient après une panne, la maintenance préventive est proactive. Son but est d'anticiper et de prévenir les problèmes, plutôt que de simplement réagir à ceux-ci. En bref, la maintenance

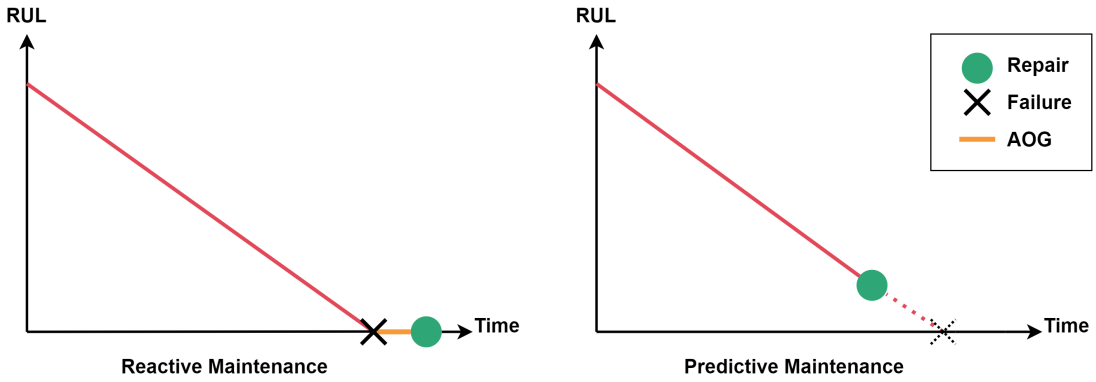


Figure 1.1: Comparaison entre la maintenance réactive et prédictive : impact sur le temps d'immobilisation au sol (AOG)

préventive est une approche de maintenance structurée et proactive, visant à optimiser la fiabilité et la longévité des équipements. Dans le contexte de l'industrie aéronautique, la maintenance préventive est essentielle pour assurer la sécurité et la fiabilité des avions. Elle comprend des activités telles que les inspections régulières des avions, le remplacement des pièces usées, le contrôle des systèmes de navigation et de communication, l'entretien des moteurs, et bien d'autres tâches.

Toutefois, la maintenance préventive a aussi des inconvénients. Elle peut être coûteuse et nécessiter beaucoup de temps, car elle implique des inspections et des travaux de maintenance même si aucun problème n'est apparent. De plus, elle peut parfois entraîner des réparations inutiles si les pièces sont remplacées avant la fin de leur durée de vie utile. Pour ces raisons, de nombreuses entreprises combinent la maintenance préventive avec d'autres approches, comme la maintenance prédictive, pour optimiser leurs stratégies de maintenance. La maintenance prédictive est une méthode de maintenance qui utilise des outils de suivi de l'état et des techniques d'analyse de données pour surveiller l'état des équipements et des systèmes en temps quasi-réel. L'objectif est de prédire quand une défaillance est susceptible de se produire, afin que la maintenance puisse être planifiée juste avant la panne afin de limiter le temps d'immobilisation des avions au sol, comme le montre la Figure 1.1. La maintenance prédictive repose fortement sur l'analyse des données collectées à partir des équipements. Ces données peuvent comprendre des mesures de températures, de pression, de vibrations, de débit, de bruit, et autres paramètres qui peuvent indiquer l'état de l'équipement. Ces données sont collectés durant le vol et transmises de manière régulière d'où le fait que ce soit, dans notre cas, en temps quasi réel. Comme la maintenance préventive, la maintenance prédictive est une approche proactive qui vise à anticiper et à prévenir les défaillances avant qu'elles ne se produisent.

Cependant, elle est généralement plus ciblée que la maintenance préventive, car elle se concentre sur les problèmes spécifiques qui sont les plus susceptibles de se produire en fonction de l'analyse des données. Enfin la maintenance prédictive utilise souvent des technologies avancées, telles que l'Internet des objets (IoT), l'apprentissage automatique, l'intelligence artificielle (IA), et l'analyse prédictive. Ces technologies permettent de collecter et d'analyser une grande quantité de données en temps réel, ce qui rend possible la détection précise des anomalies et la prédiction des défaillances.

L'implémentation de la maintenance prédictive dans le secteur aéronautique offre plusieurs avantages substantiels. Cette méthode offre l'avantage d'identifier les défaillances potentielles à l'avance, ce qui facilite la planification des opérations de maintenance avant l'apparition effective des problèmes. Ainsi, elle aide à prévenir les arrêts de production ou d'exploitation imprévus et potentiellement coûteux. En identifiant et en rectifiant les problèmes avant qu'ils n'atteignent un stade critique, la maintenance prédictive contribue à prolonger la durée de vie des équipements. Dans le cadre d'un système de maintenance prédictive généralisé, il est possible d'optimiser la gestion des stocks, évitant les ruptures de stock susceptibles de retarder les réparations. De plus, cela permet d'améliorer la planification des activités de maintenance, ce qui peut favoriser une meilleure allocation des ressources et renforcer l'efficacité opérationnelle. Enfin, en minimisant les pannes et les retards, la maintenance prédictive améliore la satisfaction des clients, qu'il s'agisse de compagnies aériennes ou d'autres utilisateurs d'équipements. En somme, la maintenance prédictive s'avère être un outil précieux, contribuant à la performance et à l'efficacité globales de l'industrie aéronautique.

1.2 La maintenance prédictive

La maintenance prédictive, sujet principal de cette section, représente une approche proactive novatrice dans le domaine de la maintenance industrielle. Cette méthode s'appuie sur l'exploitation de données, l'analyse statistique, l'apprentissage automatique et l'application d'algorithmes pour anticiper les défaillances d'équipements ou prévoir les nécessités de maintenance. Bien que ces techniques soient l'objet d'intenses recherches, leur complexité intrinsèque a jusqu'à présent limité leur déploiement dans l'industrie en général. Cependant, leur intégration semble particulièrement prometteuse dans le contexte de l'industrie aéronautique. En effet, la nature sophistiquée des composants d'un avion, souvent dotés de nombreux capteurs pour assurer leur performance optimale, génère une quantité impressionnante de données

sur les équipements vitaux de l'appareil. Cet afflux de données crée une opportunité unique pour l'application de la maintenance prédictive, offrant un cadre propice à l'expansion de ces techniques au sein de cette industrie.

La philosophie fondamentale de la maintenance prédictive repose sur la conviction qu'il est plus efficace, plus sûr et économiquement plus rentable de prévenir les pannes plutôt que d'y remédier. En anticipant les problèmes avant qu'ils ne se manifestent, nous pouvons non seulement prolonger la durée de vie des composants de l'avion, mais aussi prévenir les pannes imprévues qui pourraient immobiliser un avion au sol. En outre, cette démarche proactive contribue à renforcer la sécurité globale des vols en minimisant les risques associés aux défaillances inattendues. Ainsi, la maintenance prédictive se présente comme une stratégie d'avenir pour l'industrie aéronautique, promettant une optimisation des procédures de maintenance et une amélioration de la sécurité des vols.

Maintenance prédictive et détection d'anomalies

La maintenance prédictive et la détection d'anomalies sont étroitement liées, toutes deux visant à identifier des comportements anormaux dans les signaux de données. La détection d'anomalies se concentre sur l'identification de points de données aberrants qui s'écartent d'un schéma de données couramment observé, qui peuvent signaler des erreurs dans la collecte ou l'enregistrement des données ou des événements inhabituels. En revanche, la maintenance prédictive se focalise davantage sur l'identification de signes de dégradation avant une éventuelle panne.

Dans leur étude, Chandola et al., 2009 examinent diverses techniques de détection d'anomalies. Ils les répartissent en trois grandes catégories : les techniques basées sur la classification, celles basées sur le voisinage, et les techniques statistiques. Les techniques de classification s'appuient sur un modèle formé à partir d'un ensemble de données dont la classe de chaque observation est connue à l'avance (anomalie ou non-anomalie), ce modèle étant par la suite utilisé pour classer de nouvelles observations comme normales ou anormales. Les techniques basées sur le voisinage, comme leur nom l'indique, étudient l'environnement immédiat de chaque point de données, et les points dont le voisinage diffère significativement de la majorité sont considérés comme des anomalies. Enfin, les techniques statistiques partent de l'hypothèse que les données normales suivent une distribution statistique donnée. Les points ne correspondant pas à cette distribution sont alors identifiés comme des anomalies.

La maintenance prédictive va au-delà de la simple détection d'anomalies en prévoyant le moment où une intervention de maintenance sera nécessaire sur un équipement spécifique. Cela est rendu possible grâce à l'exploitation de données historiques, de mesures de capteurs et d'analyses avancées. La principale différence entre ces deux approches réside dans le fait que, tandis que la détection d'anomalies se base sur des classes clairement définies, en maintenance prédictive, la distinction entre les données normales et anormales peut être moins évidente. En effet, la dégradation de l'équipement se produit généralement de manière progressive tout au long de sa durée de vie, marquée par des épisodes sporadiques de dégradation plus sévère. Ainsi, il est difficile d'établir un critère objectif pour distinguer ces phases de dégradation.

La détection d'anomalies peut souvent servir d'étape préliminaire à la maintenance prédictive. L'identification d'une anomalie peut signaler qu'un équipement commence à présenter des dysfonctionnements et qu'il pourrait nécessiter une maintenance prochaine. Par conséquent, la détection d'anomalies permet d'alerter les opérateurs à un stade précoce, favorisant une intervention proactive pour prévenir les pannes d'équipement. Cependant, il est crucial de souligner que, bien que la détection d'anomalies puisse enrichir la maintenance prédictive, elle ne saurait la remplacer. La maintenance prédictive requiert des modèles plus sophistiqués, capables non seulement de prédire la probabilité d'une panne d'équipement, mais aussi d'estimer le moment de survenue de cette panne.

La mise en œuvre de la maintenance prédictive est souvent entravée par le manque de données pertinentes. En effet, l'état de l'équipement ne peut généralement être déterminé qu'à deux moments critiques : au moment de la défaillance, signalant une dégradation ou une anomalie, et lors de l'installation ou de la réparation de l'équipement, témoignant d'un état sain ou normal. De surcroît, la rareté des défaillances amplifie la pénurie de données labellisées. Une autre difficulté inhérente à la maintenance prédictive est la dégradation graduelle de l'état de l'équipement au fil du temps, impliquant qu'il ne devrait pas y avoir de différence notable dans le niveau de dégradation ou l'indice de santé entre deux instances rapprochées dans le temps.

L'objectif premier en maintenance prédictive est souvent de déterminer la durée de vie utile restante, ou Remaining Useful Life (RUL), comme souligné par Jardine et al., 2006. Le concept de RUL est fondamental dans le domaine de la maintenance prédictive et du pronostic. Il désigne le temps estimé restant avant qu'une machine,

un composant ou un système atteigne la fin de sa vie utile ou nécessite une intervention de maintenance. En d'autres termes, la RUL prédit le temps de fonctionnement restant avant qu'une défaillance prévue ou une dégradation de la performance de la machine ne survienne. Dans le cas des vannes de bleed, les alarmes qui signalent un dysfonctionnement sont activées directement dans l'avion. On considère qu'il y a une défaillance lorsque ces alarmes nécessitent une intervention de maintenance. Le déclenchement des alarmes peut être influencé par divers facteurs environnementaux, tels que la température des aéroports de départ et d'arrivée, ce qui peut introduire une variabilité supplémentaire dans le calcul de la RUL.

Une autre complexité notable dans l'estimation du RUL dans le secteur aéronautique provient de la diversité des mesures disponibles pour estimer le RUL. Ce dernier peut être estimé en termes d'heures d'utilisation ou de cycles de vol, par exemple. Chaque mesure présente des avantages et des inconvénients, et le choix le plus approprié dépend de nombreux facteurs.

L'estimation du RUL en termes d'heures d'utilisation est généralement précise puisqu'elle est basée sur le temps réel de fonctionnement de l'équipement. Cependant, elle ne tient pas compte de l'intensité de l'utilisation de l'équipement durant différentes phases de vol, comme au décollage ou à l'atterrissage, où les équipements peuvent être plus sollicités. Par exemple, certains équipements, tels que les vannes, ne fonctionnent pas en continu pendant le vol. Ces vannes peuvent rester entièrement ouvertes et donc être en quelque sorte au repos. Cela soulève la question de savoir si le temps de travail effectif d'une vanne doit être compté, ou s'il faut considérer le nombre de fois où elle est actionnée. D'un point de vue technique, ces questions sont difficiles à trancher. De plus, le temps de travail effectif d'une vanne est difficile à prévoir pour les vols futurs car il dépend des conditions de vol. L'objectif principal de la maintenance prédictive est de pouvoir avertir l'utilisateur d'une panne imminente. Par conséquent, les prédictions du RUL ne peuvent pas être directement utilisées pour prévenir d'une panne future. Il est nécessaire de transformer cette information en quelque chose de plus exploitable, comme un nombre de jours ou de vols avant une panne, informations utiles pour l'utilisateur.

L'estimation du RUL en matière de cycles de vol est particulièrement pertinente pour les équipements aéronautiques car elle prend en compte le nombre de décollages et d'atterrissages, cette mesure est aussi directement exploitable pour les compagnies aériennes. Cependant, cette mesure peut s'avérer difficile à estimer car elle ne tient pas compte de la durée des cycles. Même en se focalisant sur un modèle spécifique d'avion, on observe fréquemment une grande variabilité dans la durée des cycles. Cette variabilité est notable au sein d'une flotte d'avions, même au sein

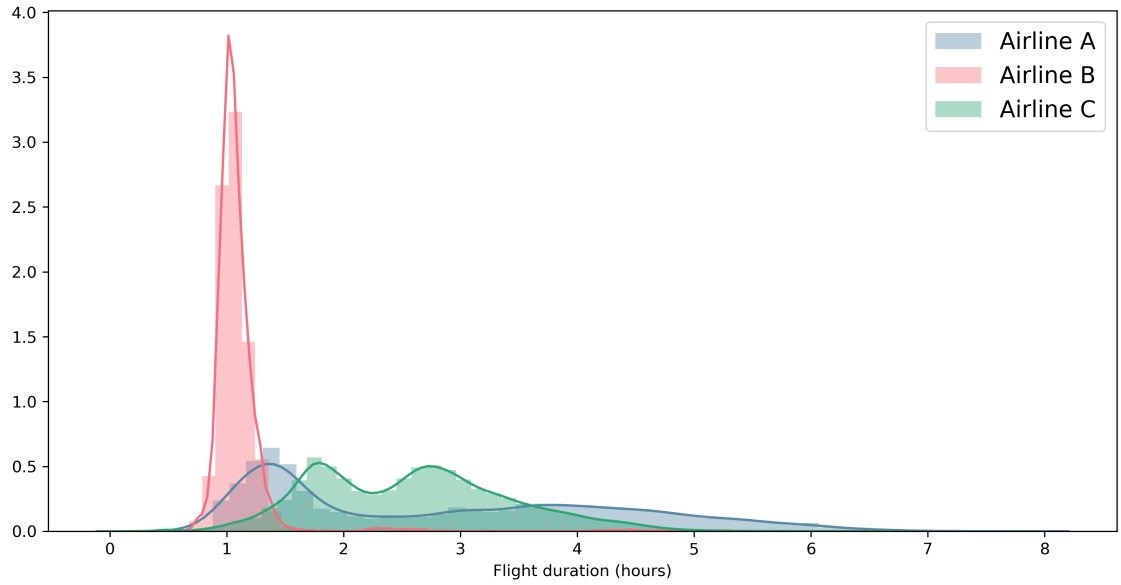


Figure 1.2: Comparaison de la durée des vols entre trois compagnies aériennes

d'une même compagnie. Les compagnies aériennes peuvent utiliser le même modèle d'avion de différentes manières selon leurs besoins spécifiques. Par exemple, une compagnie aérienne peut utiliser un avion pour des vols courts et fréquents, tandis qu'une autre compagnie aérienne peut utiliser le même modèle d'avion pour des vols longs et moins fréquents. La Figure 1.2 illustre ce phénomène. On observe que la compagnie A effectue des vols variant de 1 à 8 heures, tandis que la compagnie B se concentre principalement sur des vols d'une heure. Quant à la troisième compagnie, la majorité de ses vols se situe entre 1h30 et 4h. Ces variations mettent en lumière la grande disparité de durée des vols à la fois inter et intra-compagnies. Ces différences d'utilisation peuvent avoir un impact significatif sur la durée totale d'utilisation des équipements. Par conséquent, avec cette métrique, des avions ayant le même nombre de cycles peuvent avoir un nombre d'heures d'utilisation très différent.

Il convient de souligner que l'estimation de la RUL comporte une part d'incertitude intrinsèque, due à la complexité et à la dynamique du fonctionnement des machines, à la multitude de facteurs influents et aux éventuelles erreurs de mesure. Lorsqu'il s'agit de prévoir la durée de vie restante d'un élément intégré à un système, comme c'est le cas dans notre contexte, les autres composants du système peuvent affecter les données recueillies. On peut ainsi observer des phénomènes de compensation ou d'aggravation de la performance d'un élément en fonction des autres composants du système. Ces interactions complexifient davantage l'évaluation de l'état de santé de l'élément. L'estimation de la RUL peut se révéler être un processus complexe, étant donné que le taux de dégradation peut varier en fonction des pannes. Dans ce

contexte, l'indice de santé est fréquemment employé comme outil d'estimation de la RUL (Kang et al., 2021; Riad et al., 2010). Ce concept permet de quantifier de manière continue l'état de santé d'un équipement, fournissant ainsi un paramètre essentiel pour la prédiction de la durée de vie restante.

Généralement, l'estimation de la RUL est réalisée à l'aide d'un indice de santé (Health Index - HI), qui suppose une dégradation linéaire jusqu'à la défaillance ou, dans certains cas, adopte une fonction linéaire par morceaux (Jiang et al., 2020; Laredo et al., 2019; Teng et al., 2016). Ces indices de santé sont basés sur l'hypothèse que la dégradation est linéaire par morceaux avec une vitesse de dégradation identique pour toute la flotte ce qui peut être problématique si ces défaillances proviennent de causes différentes. De plus, dans le modèle linéaire par morceaux, le choix du taux de dégradation est souvent approximatif bien que déterminant pour l'entraînement du modèle.

Afin de mieux comprendre les données où la vitesse de dégradation n'est pas constante tout au long du cycle de vie d'un équipement, Kang et al., 2021 exploitent l'indice de santé pour estimer la RUL. Ils élaborent un modèle d'apprentissage automatique afin de prévoir l'HI d'un moteur turbo à chaque cycle. Étant donné que la RUL n'est pas incluse dans les jeux de données d'entraînement, une fonction polynomiale est adaptée aux HI, et le point d'intersection entre le polynôme et l'axe du cycle est considéré comme le point de défaillance. Ils supposent que les cycles initiaux présentent un HI = 1, et que les derniers cycles ont un HI = 0. Les données restantes sont ensuite estimées par interpolation. Cette méthode permet de modéliser des formats de dégradation plus complexes, mais ne prend pas en compte les éventuelles différences de vitesse de dégradation entre les cycles de vie.

Pour gérer des vitesses de dégradation différentes, Omshi et al., 2020 suggèrent une politique de maintenance prédictive qui s'ajuste en fonction des inspections effectuées au cours du cycle de vie. L'hypothèse clé est que les paramètres du processus de dégradation sont inconnus. Toutefois, contrairement aux méthodes précédentes, l'estimation de la RUL ici est effectuée à partir d'inspections et non d'un grand nombre de données recueillies tout au long du cycle de vie des systèmes. Cela permet de faire ces prédictions lorsque peu de données sont disponibles, mais cette approche n'est pas conçue pour un suivi permanent durant le cycle de vie.

Ces difficultés considérables associées à l'utilisation d'un indice de santé pour ajuster les algorithmes ont guidé notre attention vers des informations plus fiables : les données collectées juste avant une défaillance doivent être considérées comme anormales, tandis que les données provenant d'un nouvel équipement ou après des réparations doivent être considérées comme saines. Nous ne voulons pas faire d'hypothèses sur l'état de dégradation des équipements en dehors de certains mo-

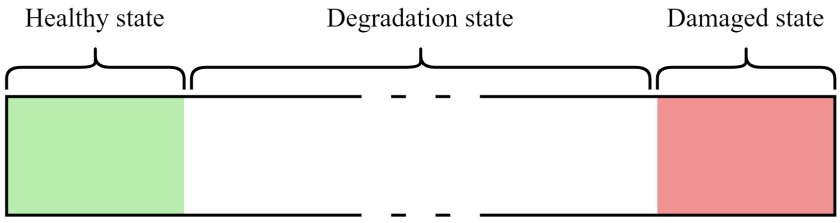


Figure 1.3: Périodes de dégradation dans le cycle de vie de l'équipement

ments précis. Comme illustré à la Figure 1.3, nous ne définirons que trois périodes : une où aucune dégradation n'est observée, et une autre où la dégradation est confirmée. Nous désignerons une troisième période comme celle pendant laquelle la dégradation se produit, mais sans prétendre connaître le moment exact de son début ni la rapidité de ce processus de dégradation.

En conclusion, si les techniques de détection de défauts et de maintenance prédictive contribuent toutes deux à la gestion globale de la santé des machines et des systèmes, la maintenance prédictive apporte une valeur supplémentaire en estimant l'état de santé et en optimisant les activités de maintenance. En développant des stratégies de maintenance prédictive, les organisations peuvent réaliser des améliorations significatives en matière de fiabilité des équipements, d'efficacité opérationnelle et d'économies de coûts.

Techniques de maintenance prédictive

La maintenance prédictive est un domaine en pleine expansion, avec de nombreuses techniques développées pour améliorer ses capacités dans une large gamme de cas d'utilisation. En conséquence, une variété de méthodes a émergé pour répondre aux défis posés par différentes applications. Selon Ran et al., 2019, nous pouvons classifier les techniques de maintenance prédictive en trois groupes distincts. Le premier groupe inclut les méthodes basées sur les connaissances, qui se fondent sur une expertise spécifique du système ou sur la modélisation physique. Le second regroupe les méthodes d'apprentissage automatique traditionnelles, tandis que la troisième et dernière catégorie comprend les approches axées sur l'apprentissage profond.

Approches basées sur les connaissances

Les approches basées sur les connaissances peuvent complémenter les méthodes axées sur les données en fournissant un contexte et un raisonnement humain pour aider à interpréter les résultats. On peut distinguer trois sous-catégories de ces méthodes.

Approches basées sur l'ontologie : L'ontologie exprime formellement les connaissances contextuelles à travers les concepts et les relations existant dans un domaine précis. Elle peut servir de base de connaissances pour différents systèmes de machines et peut être associée à divers algorithmes de raisonnement existants pour réaliser le diagnostic et le pronostic des pannes.

Approches basées sur les règles : Ces méthodes évaluent les données surveillées en temps réel selon un ensemble de règles préétablies à partir de l'expertise humaine, connues sous le nom de systèmes experts. Ces systèmes conservent le "savoir-faire" des experts humains sous forme de règles expertes.

Approches basées sur les modèles : Ces méthodes associent généralement des modèles mathématiques à des processus physiques ayant un impact direct ou indirect sur la santé des systèmes ou des composants concernés. Elles utilisent des résidus comme caractéristiques, en vérifiant la cohérence entre les résultats mesurés et le comportement attendu du processus au moyen d'un modèle analytique. Cette catégorie comprend notamment ce que l'on appelle couramment les "jumeaux numériques" ou "digital twins", très utilisés dans l'industrie.

Pour résumer, l'ontologie offre un moyen potentiel d'intégrer, de partager et de réutiliser les connaissances contextuelles d'un système, mais nécessite l'intégration d'autres méthodes de raisonnement pour réaliser la maintenance prédictive. Les approches basées sur les règles sont utiles lorsqu'il y a une grande expertise mais pas suffisamment de détails pour développer des modèles quantitatifs précis. Toutefois, un système basé sur des règles a souvent du mal à gérer les nouvelles pannes et à acquérir une connaissance exhaustive pour construire un système fiable. Les approches basées sur les modèles sont pertinentes lorsque des modèles mathématiques précis peuvent être construits à partir de systèmes physiques. Cependant, pour de nombreux systèmes complexes, les modèles mathématiques explicites peuvent être inaccessibles.

Modèles d'apprentissage automatique traditionnels

La maintenance basée sur les données est devenue la méthode la plus utilisée pour gérer la maintenance prédictive (PdM), notamment pour la surveillance de la santé des machines (par exemple le diagnostic des défaillances et l'évaluation de la durée de vie restante). Les algorithmes d'apprentissage automatique sont couramment utilisés pour analyser les données, en se concentrant particulièrement sur la précision de la prédiction. Selon W. Zhang et al., 2019, les signaux utilisés pour le diagnostic des

défaillances peuvent inclure les émissions acoustiques, les paramètres de signature électrique (courant et tension), la température, la pression, la vitesse de rotation et la vibration. Néanmoins, les signaux de vibration sont le plus souvent exploités, car ils fournissent des informations précieuses, notamment pour les systèmes rotatifs. Dans le cadre de cette thèse, nous nous concentrerons principalement sur des signaux physiques tels que la pression et la température, mais aussi sur des signatures électriques correspondant aux commandes envoyées aux instruments contrôlés.

Parmi les modèles d'apprentissage automatique traditionnels, on retrouve la régression logistique (LR), un modèle de classification bien connu pour sa faible complexité algorithmique. La revue de W. Zhang et al., 2019 énumère plusieurs applications de la régression logistique pour la maintenance prédictive. Par exemple, H. Li et al., 2015 ont proposé une méthode qui combine un modèle LR avec des signaux d'émission acoustique et de force de coupe pour surveiller le processus d'usure des outils de coupe et déterminer le moment optimal pour la maintenance. On peut citer également Yan and Lee, 2005 pour la surveillance de la santé des portes d'ascenseur. La régression logistique est une méthode précieuse pour la maintenance prédictive, offrant une précision de prédiction élevée dans certains cas, une faible complexité de modèle, et une interprétabilité appréciable pour les experts industriels. Cependant, la simplicité de ces méthodes peut limiter leur utilisation dans les cas plus complexes de dégradation, où des méthodes plus avancées se révéleront plus performantes.

On retrouve également les modèles basés sur les forêts aléatoires (RF). Les RF sont des collections d'arbres de décision formés à partir de sous-ensembles aléatoires de caractéristiques, dont les résultats sont ensuite agrégés par une moyenne. Les travaux de Prytz et al., 2015 ont utilisé RF comme algorithme de classification, en association avec deux méthodes de sélection de features, pour prédire les réparations de divers composants de véhicules commerciaux. De même, Canizo et al., 2017 ont fait appel à la méthode des RF pour générer des modèles prédictifs dynamiques destinés au suivi des éoliennes. Cette approche a permis une accélération significative du traitement des données, tout en garantissant la scalabilité et l'automatisation des prédictions. Cette technique, robuste face au surapprentissage et simple à mettre en place, est la méthode d'apprentissage automatique la plus utilisée et comparée dans les applications de maintenance prédictive Carvalho et al., 2019.

Les machines à vecteurs de support (SVM) et les k-plus proches voisins (KNN) sont également des méthodes largement utilisées. Les SVM sont couramment employées pour des tâches de classification et de régression en raison de leur précision élevée.

Elles permettent une séparation précise entre différentes classes de données. Les KNN sont des méthodes utilisées pour trouver des partitions (ou clusters) dans un ensemble de données. Faciles à mettre en œuvre, performants et capables de gérer de grands ensembles de données, ces méthodes ont été testées par Mathew et al., 2017 sur un jeu de données de la NASA dont nous discutons plus en détail dans la Section 1.3.

Ces modèles d'apprentissage automatique traditionnels nécessitent généralement l'utilisation de features qui représentent les données. Toutefois, dans des cas d'utilisation complexes, un ensemble simple de feature peut s'avérer insuffisant, nécessitant l'intervention d'un expert du système pour générer des features sur mesure qui peuvent capturer avec précision la dégradation de la santé des équipements pour un cas d'utilisation spécifique. Cette tâche peut être assez difficile en raison des connaissances expertes requises, ce qui pousse à l'utilisation de modèles plus complexes qui vont permettre d'extraire de manière automatique ces features.

Modèles d'apprentissage profond

La troisième et dernière catégorie regroupe les approches basées sur l'apprentissage profond. L'enquête approfondie de Serradilla et al., 2022 examine diverses techniques d'apprentissage profond utilisées pour la maintenance prédictive (PdM), des méthodes dont l'usage est en croissance constante.

De nombreuses études ont mis en avant l'usage des réseaux de neurones convolutifs (CNN) pour extraire des caractéristiques pertinentes en vue de la prédiction de pannes. Parmi ces travaux, Huuhtanen and Jung, 2018 se sont penchés sur l'application des CNN à la maintenance prédictive des panneaux solaires. En dépit de défis tels que les variations météorologiques et les ombres générées par les objets environnants, l'étude a souligné le potentiel des approches basées sur les CNN pour la maintenance prédictive des systèmes photovoltaïques. Une autre étude de Kiangala and Wang, 2020 a mis l'accent sur l'utilisation de la maintenance prédictive pour les moteurs de convoyeurs dans le cadre de l'industrie 4.0. Ce travail a montré comment mettre en place un cadre de maintenance prédictive en s'appuyant sur un modèle de classification basé sur les CNN. Les séries temporelles, représentant diverses observations enregistrées au fil du temps, sont alors prétraitées pour optimiser leur utilisation dans ces réseaux. Par ailleurs, Çınar et al., 2020 ont proposé une revue exhaustive des progrès récents en matière d'application des techniques d'apprentissage automatique à la maintenance prédictive dans le contexte de la fabrication intelligente (industrie 4.0). Ce travail met en exergue les nombreux avantages des applications

d'apprentissage automatique, comme la réduction des coûts de maintenance, la diminution des interruptions pour réparation, et l'allongement de la durée de vie des pièces de rechange. Néanmoins, malgré ces avancées et le potentiel de l'apprentissage automatique, une enquête de PwC révèle que seulement 11% des entreprises ont effectivement mis en œuvre une maintenance prédictive basée sur l'apprentissage automatique. En somme, ces études mettent en lumière l'importance et le potentiel des CNN dans le domaine de la maintenance prédictive, tout en soulignant les défis à relever pour en généraliser l'usage. Les recherches futures pourraient envisager de combiner différentes approches et techniques pour surmonter ces obstacles et optimiser davantage l'efficacité de la maintenance prédictive.

Les données utilisées pour la maintenance prédictive proviennent souvent de capteurs qui génèrent des séries temporelles. Les réseaux de neurones récurrents (RNN), spécialement conçus pour traiter ce type de données, sont par conséquent fréquemment employés dans ce domaine. Dans une étude menée par Q. Wang et al., 2020, les chercheurs ont proposé une approche innovante pour moderniser les méthodes de maintenance désuètes des équipements d'alimentation ferroviaire à grande vitesse. Ils ont mis en œuvre un réseau de neurones récurrents à mémoire à long terme (LSTM-RNN) pour prédire le moment idéal pour effectuer la maintenance, se basant sur les données historiques. Les essais réalisés sur un disjoncteur isolé au gaz ont démontré la possibilité de prédire avec précision le moment de la prochaine maintenance. Dans une autre publication, Rahhal and Abualnadi, 2020 ont collecté une grande quantité de données pour chaque dispositif. Ces données ont été ensuite transmises à un serveur central de traitement pour la construction d'un modèle mathématique. Deux types de RNN, le Vanilla-RNN et le LSTM-RNN, ont été mis en œuvre pour réaliser des prédictions. Le LSTM-RNN a démontré une meilleure performance prédictive et a été recommandé pour des équipements ou des dispositifs dont la défaillance ou l'interruption de fonctionnement aurait des conséquences significatives. En conclusion, les RNN, notamment les LSTM-RNN, ont prouvé leur potentiel pour améliorer la précision de la maintenance prédictive. En couplant ces modèles d'apprentissage profond à d'autres technologies, telle que l'Internet des Objets (IoT), chercheurs et professionnels sont en mesure de concevoir des systèmes de maintenance plus efficaces et plus précis.

Les auto-encodeurs sont couramment utilisés pour extraire des features pertinentes sans avoir besoin de labels (Davari et al., 2021; Jakubowski et al., 2021; Su et al., 2020), contrairement aux modèles RNN et CNN qui nécessitent un label cible pour l'apprentissage. Les auto-encodeurs (AE) sont des architectures de réseaux de neurones conçues pour réduire la dimensionnalité des données. Ils se composent de

deux réseaux distincts : un premier réseau qui encode les données en réduisant leur dimensionnalité, et un second qui reconstruit les données à partir de cette version compressée. L'un des avantages majeurs des AE réside dans leur non-linéarité, qui permet de compresser les données de manière souvent plus efficace que d'autres techniques telles que l'analyse en composantes principales (PCA). De plus, ces modèles s'entraînent de manière non supervisée, éliminant ainsi le besoin de labels. Les paramètres du modèle sont optimisés pour minimiser l'erreur de reconstruction entre la donnée d'entrée et la sortie du décodeur. Dans leur étude, Jia et al., 2018 ont exploré l'utilisation d'auto-encodeurs pour le diagnostic automatique des défauts de machines. Comparativement aux méthodes conventionnelles, qui requièrent une caractérisation manuelle des anomalies limitant ainsi leur capacité à automatiser le processus, les auto-encodeurs ont démontré une nette supériorité. En effet, ils ont atteint un taux de précision de 99,43%, bien supérieur à celui obtenu en utilisant une méthode combinant une réduction de dimension via une PCA suivie d'une classification endommagé/sain à l'aide de SVM, qui a seulement atteint 41,04% de précision.

Lu et al., 2015 ont exploité un AE basique comme extracteur de features pour obtenir une représentation en petite dimension à partir de signaux de roulement de grande dimension. Néanmoins, des données brutes de grande dimension peuvent entraîner un coût de calcul élevé et du surapprentissage. Par conséquent, des caractéristiques multi-domaines peuvent être préalablement extraites à partir des données brutes, puis introduites dans des modèles basés sur AE. En raison du manque de données historiques sur les défaillances ou de données correctement labélisées, les modèles basés sur AE s'avèrent être un outil pertinent pour l'estimation du processus de dégradation. Ces approches sont capables de mesurer l'état de santé du système et de distinguer les différents niveaux de gravité des pannes. Enfin, les modèles basés sur AE sont généralement combinés avec divers modèles de régression pour prédire la durée de vie restante (RUL) des équipements. Par exemple, Xia et al., 2018 ont développé une approche de pronostic en deux étapes. D'abord, un AE est utilisé pour classer les signaux en différentes étapes de dégradation. Ensuite, des modèles de régression sont construits pour chaque étape de santé afin de prédire la RUL. Cette combinaison de techniques permet une prédiction plus précise et adaptée à chaque étape de dégradation.

En réalité, une multitude de méthodes élaborées pour la classification, la régression ou même le clustering peuvent être employées pour la maintenance prédictive. Le choix d'un modèle spécifique sera généralement guidé par la quantité et le type de données disponibles. L'avantage majeur de ces techniques d'apprentissage profond réside dans leur capacité à reconnaître automatiquement les patterns précurseurs de

pannes, réduisant ainsi le besoin d'une expertise spécifique pour leur génération. Par conséquent, elles ouvrent la voie à la mise en œuvre de la maintenance prédictive dans des environnements où les connaissances spécialisées peuvent être restreintes. Cependant, ces méthodes requièrent généralement un volume conséquent de données pour l'entraînement. Aussi, W. Zhang et al., 2019 pointent du doigt le fait que, bien que ces algorithmes soient utilisables pour la plupart des applications industrielles, ils souffrent d'un manque d'interprétabilité et sont incapables d'expliquer les phénomènes spécifiques.

Conclusion

Il est important de souligner que ces méthodes ne sont pas mutuellement exclusives. En effet, elles peuvent souvent être utilisées conjointement dans un système de maintenance prédictive, comme l'illustrent les méthodes en plusieurs étapes mentionnées précédemment. Ainsi, les données collectées par les capteurs sur les machines pourraient être analysées à la fois avec des techniques de machine learning et d'apprentissage profond. Les résultats de ces analyses pourraient alors servir à mettre à jour ou à affiner les systèmes fondés sur des connaissances spécifiques et expertes.

Tandis que chaque méthode présente ses propres avantages et inconvénients, il est essentiel de comprendre que la meilleure approche pour une application donnée dépendra de nombreux facteurs. Parmi ceux-ci figurent la disponibilité des données, l'expertise disponible, ainsi que le format des données en question.

1.3 De l'importance des jeux de données

L'étude de W. Zhang et al., 2019 met également l'accent sur l'importance de la provenance des données. Les auteurs remarquent que la majorité des jeux de données proviennent de centres de données publics ou de plateformes expérimentales à l'échelle du laboratoire. Ils observent que seuls quelques articles utilisent des jeux de données collectés à partir d'équipements en service réel, mettant en évidence la nécessité d'approfondir la recherche dans des conditions d'exploitation concrètes. C'est dans cette optique que nous discuterons plus en détail des jeux de données et de leurs limites dans cette section.

Le jeu de données sur les roulements de la Case Western Reserve University (CWRU Bearing Dataset) représente une ressource largement utilisée et reconnue pour le diagnostic des pannes. Dès 2015, il a été qualifié de jeu de données de référence

dans l'étude de Smith and Randall, 2015, qui a réalisé un benchmark sur les méthodes de détection des défauts. Ce jeu de données est composé d'enregistrements de tests de roulements à billes dans diverses conditions, allant de l'état normal à plusieurs niveaux de défectuosité. Les défauts ont été introduits méthodiquement à l'aide d'une machine d'électro-érosion (EDM), offrant des tailles de défauts multiples. La collecte des données a été réalisée en utilisant des accéléromètres fixés à la fois au boîtier du roulement et, dans certains cas, à la plaque de base de soutien du moteur. Les signaux de vibration ont été enregistrés via un dispositif à 16 canaux et ensuite post-traités dans un environnement Matlab. Les données numériques ont été collectées à des fréquences allant de 12 000 à 48 000 échantillons par seconde, avec des données complémentaires sur la vitesse et la puissance également enregistrées. En raison de la prévalence de ce type de données dans l'industrie, le jeu de données CWRU est devenu une ressource précieuse pour les chercheurs et les ingénieurs travaillant dans le domaine de la maintenance prédictive, la détection de défauts et l'analyse des vibrations dans les systèmes de roulements à billes.

Cependant, il est important de noter certaines limitations inhérentes à ce jeu de données. En premier lieu, les données ont été recueillies sur un banc d'essai, ce qui peut différer des conditions d'utilisation réelles et avoir une incidence sur la capacité de généralisation des résultats. En outre, les défauts ont été introduits de manière expérimentale, ce qui peut ne pas correspondre exactement aux types de dégradations observées dans des machines à roulements en conditions d'utilisation réelles. De plus, le jeu de données ne permet pas une étude approfondie de l'évolution et de la dégradation des équipements au cours de leur cycle de vie, car les données sont présentées de manière binaire (soit saines, soit avec un défaut spécifique). Ainsi, malgré son utilité et son utilisation répandue, ces aspects doivent être pris en compte lors de l'exploitation du jeu de données CWRU pour le diagnostic des défauts.

Le jeu de données Nectoux et al., 2012 repose sur PRONOSTIA, une plateforme expérimentale conçue pour les tests de dégradation accélérée des roulements à billes. Cette plateforme a été élaborée pour authentifier et affiner les méthodes relatives à l'évaluation de l'état, au diagnostic et au pronostic des roulements, compte tenu du fait que la majorité des défaillances des machines tournantes sont associées à ces composants. L'objectif principal de PRONOSTIA est de fournir des données concrètes concernant la dégradation accélérée des roulements, réalisée sous des conditions de fonctionnement constantes et/ou variables, contrôlées en temps réel. De plus, l'article annonce l'organisation du "IEEE PHM 2012 Prognostic Challenge" lors de la conférence PHM, où un lien vers les données de dégradation sera accessible aux concurrents, leur permettant de tester et de vérifier leur méthodologie de

pronostic. Ce jeu de données offre une caractéristique intéressante : il propose des exécutions jusqu'à la défaillance complète (run-to-failure), ce qui permet d'observer l'évolution de la dégradation au cours de la durée de vie des équipements. Les auteurs soulignent que certaines dégradations surviennent soudainement, rendant leur prédiction difficile. Tout comme le jeu de données précédent, celui-ci est produit grace à un banc d'essai, ce qui entraîne des limitations similaires. Cependant, les dégradations sont réelles et recueillies tout au long du cycle de vie, ce qui apporte une valeur ajoutée notable à ce dataset.

Bien que les ensembles de données soient cruciaux pour la maintenance prédictive, les données de vibrations de roulement sont très spécifiques et peu pertinentes pour le sujet de cette thèse, qui se concentre principalement sur des mesures physiques enregistrées à basse fréquence (1 Hz). Pour de ce type de données, l'ensemble de données C-MAPSS est largement exploité, voir Saxena et al., 2008. Les chercheurs ont utilisé une simulation thermo-dynamique pour générer les réponses d'un grand nombre de capteurs en fonction des variations de flux et de l'efficacité des modules concernés. Ils ont imposé un taux de changement exponentiel pour la perte de flux et d'efficacité pour chaque ensemble de données, indiquant une défaillance non spécifiée avec des conséquences de plus en plus nuisibles. La progression des dommages est autorisée à se poursuivre jusqu'à l'atteinte d'un critère d'échec. Ils ont défini un indice de santé comme le minimum de plusieurs marges opérationnelles à un moment donné et le critère d'échec est atteint lorsque l'indice de santé atteint zéro. Cet ensemble de données est largement utilisé pour illustrer les nouvelles méthodes de maintenance prédictive, mais il présente néanmoins certaines limites. Premièrement, les données proviennent de simulations numériques. Les systèmes moteurs des avions évoluent dans des contextes très variés et sous de fortes contraintes. Dans ces conditions, la simulation numérique s'écarte rapidement des données que l'on peut collecter en vol. Deuxièmement, les dégradations présentes dans cet ensemble de données sont facilement identifiables, même sans l'extraction de caractéristiques d'intérêt, comme le montre l'évolution des données brutes dans la Figure 1.4. Sur cette figure, nous représentons l'évolution de deux séries de données brutes jusqu'au moment de la panne (temps 0). En superposant simplement les signaux du dataset, on peut déjà trouver une règle "d'expert" pour distinguer les moments proches d'une panne de ceux où le système est en bon état.

Cet ensemble de données, tout en étant utile, présente des dégradations relativement simples à identifier et ne tient pas compte de la variété des contextes dans lesquels les systèmes moteurs évoluent lors des vols. En raison de ces facteurs, il est particulièrement complexe d'estimer la performance réelle d'un modèle fondé sur cet

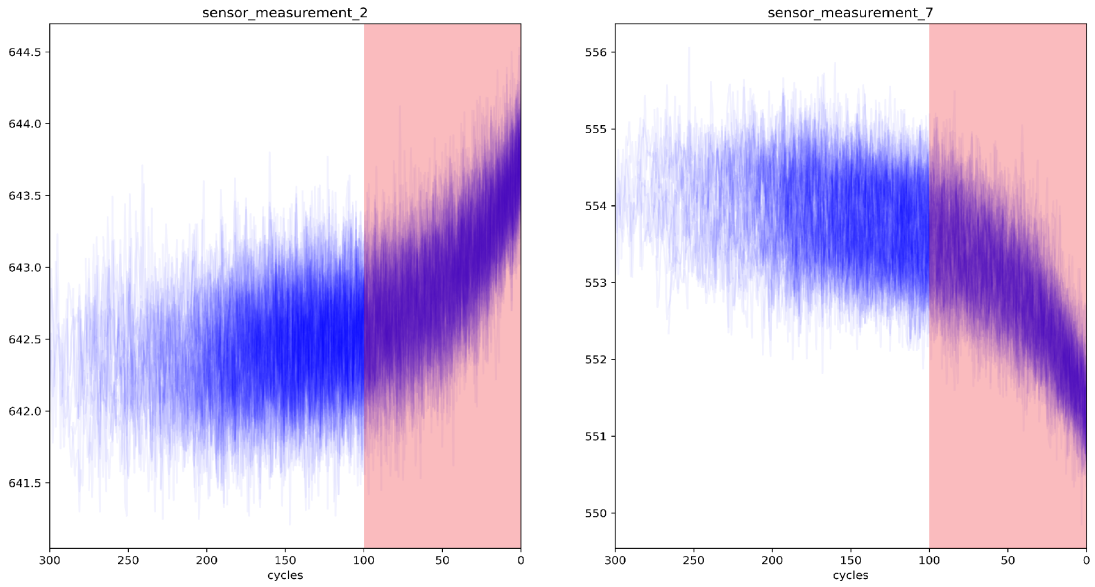


Figure 1.4: Évolution de donnés récoltés sur les capteurs du dataset C-MAPSS

ensemble de données. Il est donc crucial de relativiser les performances des méthodes uniquement testées sur cet ensemble. Une version améliorée de cet ensemble de données, appelée N-CMAPSS (Arias Chao et al., 2021), a été développée pour tenir compte des conditions de vol réelles, telles qu'elles sont enregistrées à bord des avions de ligne. Cette version élargit le spectre de la simulation et s'efforce ainsi de se rapprocher davantage de la réalité opérationnelle.

Récemment, des ensembles de données basés sur des informations recueillies à partir d'équipements en fonctionnement ont été mis à disposition, comme celui issu de la base de données NGAFID (National General Aviation Flight Information Database - Yan and Lee, 2005; Yang et al., 2022). Ce dataset constitue l'ensemble de données publiques non simulées le plus vaste, couvrant l'intégralité d'une flotte d'avions (des Cessna 172), les enregistrements de vols et les journaux de maintenance pour la prédiction des défaillances matérielles et des besoins en maintenance. Il contient 31 177 heures de données de vol sur 28 935 vols, correspondant à 2 111 événements de maintenance non prévus classés en 36 types de problèmes de maintenance. Cet ensemble de données revêt une importance capitale pour les recherches futures en matière de maintenance prédictive, car il offre la possibilité de tester des méthodes sur des données réelles généralement inaccessibles. La seule limitation que l'on pourrait signaler ici est l'absence de données sur le cycle de vie complet des équipements ; on ne trouve que des données sur les vols sains (après réparation) et les vols avec des équipements endommagés (vols réalisés 1 à 2 jours avant une panne). Ce format de données encourage des approches basées sur la classification des vols (sains ou

dégradés) mais ne permet pas vraiment de tester les prédictions de la durée de vie restante utile.

Un autre ensemble de données réelles est disponible, provenant du projet de maintenance prédictive mené avec le service de transport public urbain de métro à Porto, au Portugal (Veloso et al., 2022). Collectées en 2022, ces données englobent une diversité de signaux issus de capteurs analogiques (pression, température, consommation de courant), de signaux numériques (signaux de contrôle, signaux discrets) ainsi que des informations GPS (latitude, longitude, vitesse). L'accent est mis sur les défaillances de l'unité de production d'air, qui alimente différentes unités assurant diverses fonctions. Parmi ces unités, la suspension secondaire, qui maintient la hauteur du véhicule stable indépendamment du nombre de passagers à bord. Les données ont été recueillies de janvier à juin 2022, à partir d'un train en exploitation. Avec un taux d'acquisition de données de 1Hz, cet ensemble de données comprend plus de 3000 heures d'informations. Ce dataset offre un historique des données, mais il ne recense que trois pannes sur une période de six mois. Cette particularité pourrait limiter l'utilisation de cet ensemble de données aux méthodes de détection de défauts non supervisées.

Conclusion

Au terme de notre étude approfondie des divers jeux de données dans le domaine de la maintenance prédictive, il est clair que chaque ensemble de données a ses avantages distincts, mais aussi ses limites inhérentes. Les jeux de données traditionnels, tels que le CWRU Bearing Dataset et C-MAPSS, ont contribué de manière significative à l'avancement de la recherche dans ce domaine. Toutefois, il convient de reconnaître leurs limitations, notamment leur dépendance vis-à-vis des conditions d'essai en laboratoire ou des simulations numériques, qui ne reflètent pas toujours fidèlement la complexité des scénarios du monde réel.

D'autre part, nous saluons l'apparition d'ensembles de données plus réalistes et dynamiques, provenant d'équipements en fonctionnement, comme le dataset NGAFID et celui du service de métro de Porto. Ces ensembles de données nous rapprochent de l'objectif de la maintenance prédictive, à savoir prévoir et éviter les défaillances matérielles dans des conditions d'exploitation réelles. Cependant, il est également crucial de prendre en compte les limitations de ces datasets, comme l'absence de données sur le cycle de vie complet des équipements et le nombre limité de défaillances enregistrées.

En conclusion, l'exploration et l'évaluation rigoureuse de ces divers jeux de données mettent en lumière l'importance d'une approche pluraliste dans le domaine de la maintenance prédictive. Il est impératif de ne pas se reposer uniquement sur un type de données, mais plutôt de tirer le meilleur parti des caractéristiques spécifiques de chaque jeu de données, tout en tenant compte de leurs limites.

Un dataset de référence va devoir émerger pour permettre d'avancer dans le domaine. Mais il convient de souligner que bon nombre des méthodes de maintenance prédictive reposent généralement sur l'expertise spécialisée et sont souvent conçues pour s'adapter à des cas d'utilisation spécifiques. Cette réalité met en lumière la nécessité d'utiliser des données de nature diverse pour concevoir et évaluer des modèles. L'idée de recourir à un unique ensemble de données pour tester toutes les méthodes de maintenance prédictive semble donc irréaliste. Une variété de jeux de données est nécessaire pour embrasser les complexités et spécificités des scénarios de maintenance prédictive de la vie réelle et pour permettre de comparer sereinement les méthodes développées.

Chapter 2

Bleed Air Systems: Data & Challenges

In the context of predictive maintenance, particularly for critical aerospace components, this chapter lays a comprehensive foundation for our research by delving into the underlying systems, available data, and preliminary approaches. We begin with a detailed explanation of the bleed air system in the first section. As one of the key components in the functional operation of an aircraft, it is critical to understand the details of the bleed air system, its primary components, and the role it plays in the overall performance of the aircraft.

Once the technical landscape of the bleed air system is unraveled, we turn to the data available to us. Our research benefits from the rich data set provided by Liebherr Aerospace Toulouse, which furnishes us with a wealth of information, enabling us to analyze real-world functioning of bleed air systems and their failure patterns. The data is a crucial aspect of our research and the backbone of any machine learning project. In the second part of this chapter, we will provide an overview of the dataset's components, their interrelations, and the notations used to represent them. We will also discuss the data collection procedure, the steps involved in data pre-processing, and the splitting of data into training and testing sets.

In the third and final part of this chapter, we will evaluate some preliminary and admittedly naive approaches to predictive maintenance. Recognizing that a simple approach might not always yield the most accurate results, it serves as a critical stepping stone for our understanding of the problem at hand. It allows us to gauge the difficulty of the task, identify the limitations of straightforward methodologies, and lays the groundwork for more sophisticated and accurate models to be discussed

in the subsequent chapters.

This chapter sets the stage for the rest of the research, providing us with a deep understanding of the context in which we operate, the data at our disposal, and a starting point from which we can build upon.

2.1 Bleed air systems

A bleed air system is a critical component in the design and operation of an aircraft. It comprises a sophisticated network designed to deliver and regulate compressed air from an aircraft's engines. Within an aircraft, the air system performs several key functions that contribute to the safe and efficient operation of the aircraft. Primarily, it is used to provide air conditioning and manage cabin pressurization, which helps maintain a comfortable and safe environment for passengers and crew at high altitudes where the air is naturally thinner. In addition, the air system plays a critical role in de-icing and anti-icing the wings and engines, preventing the formation of ice that could degrade aircraft performance or even pose a hazard. The tapped air is also used to cool various avionic systems and heat-generating components. Consequently, the significance of the air system extends beyond ensuring passenger comfort. It plays a more substantial role in guaranteeing the safety and enhancing the operational efficiency of the aircraft. As such, the reliability of the air system becomes critical, positioning the maintenance of its components as a high priority within the aviation industry. Moreover, it's noteworthy that the bleed air system is one of the most energy-intensive systems on an airplane, surpassed only by the engines. This presents another compelling reason to monitor it closely, preventing potential losses in efficiency.

The air system operates by drawing air from the engine as needed, which can be taken from two different locations in the engine. The extracted air can easily reach temperatures exceeding 500°C, thereby imposing substantial stress on various system components. This mass of hot air is channeled, through a series of carefully orchestrated valves, to an initial radiator. This radiator, utilizing an additional influx of fresh air captured from outside, is tasked with moderating the temperature of the extracted air before its subsequent use. Depending on specific conditions, the air can then be directed towards the aircraft's wings to ensure a de-icing function. Finally, a terminal valve regulates the pressure of this treated air, making it suitable to be injected into the air conditioning system, commonly known as PACK. This sequence of operations ensures a constant supply of regulated air, crucial to the

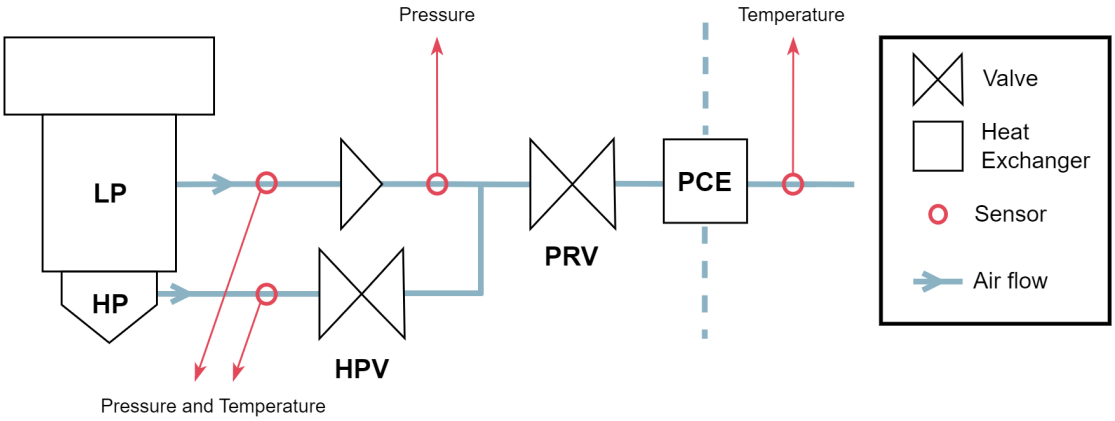


Figure 2.1: Simplified Overview of the Bleed Air System's Architecture

comfort and safety of the aircraft and its occupants. These first step of bleed air systems are detailed in the Figure 2.1.

This intricate system is composed of numerous components, among which the most critical are valves that can be controlled either pneumatically or electrically. These valves are primarily utilized to regulate output pressure, a function which is essential to the safety and efficiency of downstream components. Figure 2.2a provides an example of this type of valve. These pieces of equipment can be subjected to various types of degradation, including leaks in the valve's flap or the valve becoming stuck. These types of degradation can impact the valve's ability to maintain a consistent and accurate pressure.

Another key component is the heat exchanger, which uses ambient cool air to reduce the temperature of the hot air emanating from the turbines. This process enables downstream components to receive air at a moderately reduced temperature. An example of such a component is depicted in Figure 2.2b. Common degradation in this component often takes the form of leaks. While a minor leak might not pose a direct threat to the rest of the system, it can impact the overall efficiency of the aircraft. Therefore, it's crucial to ensure timely replacement of a leaking heat exchanger. The task of preventive maintenance in this case isn't straightforward because heat exchangers can be quite large, making their replacement a complex procedure. This makes heat exchangers ideal candidates for predictive maintenance.

Failures in the air system can have severe consequences from both safety and operational perspectives. From a safety viewpoint, a failure of this system could compromise cabin pressurization, endangering the health and comfort of passengers

(a) Valve (b) Heat Exchanger

Figure 2.2: Exemples of Key Components of the Bleed Air System

and crew, especially during high-altitude flights. Similarly, a failure in de-icing and anti-icing systems can cause ice accumulation on the wings and engines, a situation that can affect the aircraft's maneuverability and, in extreme cases, result in engine failures. However, it's essential to highlight that the design of these systems incorporates significant redundancy, with each engine equipped with its own air system. Thus, even in case of a failure on one side, the other can take over all vital functions, thereby minimizing risks and ensuring optimal safety.

From an operational perspective, air system failures can lead to flight delays or cancellations, bearing significant financial consequences for airlines. Moreover, if an aircraft is grounded for repairs, this can also impact flight scheduling and aircraft availability. For all these reasons, predictive maintenance of the air system is a major priority in the aviation industry, as it enables the detection of problems before they escalate and hence minimizes associated risks and costs.

2.2 Data collection

In the upcoming sections, we will delve deeper into the specifics of the data collection process, exploring the types of data primarily involved in our analyses and the rationale behind their selection. Additionally, we will also introduce the preprocessing steps and the evaluation strategy.

Data Acquisition and Analysis in Aircraft Bleed Air Systems

Aircraft, particularly in their complex functionalities and mechanical nuances, are equipped with an extensive network of sensors that ensure their optimal performance. These sensors, some of which are strategically placed around components of the bleed air systems (as illustrated in Figure 2.1), actively monitor and record crucial operational data during flights. The raw time-series data, constituting various measurements and status indicators, are transferred to our database post-flight, offering a comprehensive dataset spanning hundreds of thousands of flight hours across the fleet.

The primary focus of this thesis rests on physical measurements including, but not limited to, pressure and temperature values. Additionally, we consider control commands sent to the valves, flow measurements preceding the air conditioning component, and importantly, the altitude of the aircraft. The inclusion of altitude data plays a pivotal role in our analysis as it provides insights into the specific flight phase at the time of data collection. Aircraft operations are typically divided into distinct phases, each with their unique operational characteristics and requirements. A general overview of these phases includes: Ground before Departure (gbd): This phase marks the time leading up to the takeoff, encompassing pre-flight checks and preparatory procedures. Climb (clb): Initiating when the aircraft is off the ground, this phase extends until the aircraft reaches its cruising altitude. Cruise (crz): Starting when the aircraft attains its cruising altitude, the cruise phase continues until the initiation of the descent. Descent (des): This phase begins when the aircraft starts descending from the cruise altitude and concludes when the aircraft touches the ground. Ground after Arival (gaa): Following the landing, this phase encompasses the time until engine shutdown. Figure 2.3 illustrates a typical flight profile, highlighting the evolution and sequence of different phases throughout the course of a flight. The depicted altitude changes reflect standard patterns of ascent, cruising altitude, and descent, offering insights into key moments where the bleed air system may face varying operational demands. In fact, flight profiles vary significantly based on many factors. For example, the length of the flight, weather conditions, and air traffic control guidelines all play a significant role in shaping a flight profile.

Alarms, generated by the aircraft's monitoring systems, play a crucial role in safeguarding the aircraft's operational integrity. They continually oversee various parameters and operational metrics, working in synergy with avionic systems that handle navigation, communications, and display. Modern aircraft incorporate

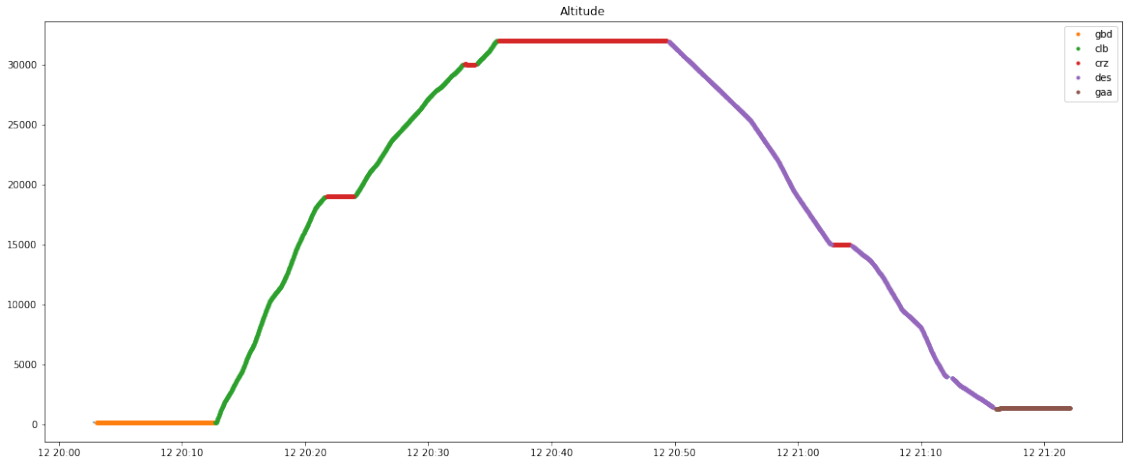


Figure 2.3: A Typical Flight Profile

advanced alarm management systems that help prioritize alarms and control 'alarm floods.' In the event of a cascade of triggered alarms, the system highlights the root cause, allowing the crew to address the primary issue first. These alarms are systematically logged, and any alarm concerning a component of the bleed system typically necessitates subsequent maintenance. Occasionally, maintenance may be conducted without a prior alarm, signifying a right-censored scenario in our data interpretation.

Our goal is to preemptively identify and resolve issues that may trigger these alarms. To this end, we define the last flight of a component's lifecycle as the flight preceding the alarm-triggering event. Therefore, the overall goal is to improve the accuracy of predicting potential maintenance needs, thereby promoting operational efficiency and safety.

Preprocessing

The data collection and processing pipeline for aircraft sensor data involve several critical steps. These steps ensure the integrity of the data and its suitability for the complex analyses required for predictive maintenance. One of the primary data sources in this regard are the numerous sensors installed throughout the aircraft, some of which focus on monitoring the bleed air systems. These sensors predominantly operate at a sampling rate of 1 Hz, ensuring a steady stream of data during each flight. The sensor data is gathered across two separate channels, which each capture slightly differing sets of values. This dual-channel setup is designed to enhance the robustness of the data collection process and provides an added level of redundancy to maintain data integrity. However, while most sensors adhere to the 1

Hz sampling rate, their data collection cycles might not be perfectly synchronized. This discrepancy in synchronization can potentially result in minor variations in the time-stamping of data points from different sensors. Even small inconsistencies could potentially affect the precision of certain analyses, particularly those requiring correlation between different types of measurements or intricate time-series analysis.

To address this challenge, the first step in our data preprocessing is the synchronization of all time series data. We achieve this via linear interpolation, a technique that fills in missing data points within a series based on the linear relationship between known data points. This process allows us to create a unified timeline for all data, ensuring accurate comparisons and correlations between different measurements. Following the synchronization of the time series data, we merge the data from the two channels associated with each sensor. While the channels collect slightly different values, they essentially capture the same physical measurements. Hence, they can be combined into a unified time series. This merging step simplifies our data, reducing the number of time series to be analyzed while preserving the comprehensiveness and richness of our dataset. These pre-processing steps are essential to ensure the accuracy and usability of our sensor data. Through effective synchronization and consolidation of our data, we lay the groundwork for robust subsequent analyses and the development of reliable predictive maintenance models.

Each flight is stored as a multivariate time series consisting of K parameters, each observed over $\check{T}_i$ time steps. We represent this time series as $\check{x} = \{\check{x}^{t,k}\}$, $t \in \check{\mathcal{T}} = \{1, \ldots, \check{T}_i\}$, $k \in \mathcal{K} = \{1, \ldots, K\}$. Here, t refers to the set of observation times $\check{\mathcal{T}}$, and k is a member of the set of parameters $\mathcal{K}$. Our dataset, of size n, consists of pairs, each including the i^{th} time series and its corresponding label (or class). We express these pairs as $\check{x}_i, \check{y}_i$, for each i in the range 1 to n. It's essential to note that the number of parameters, K, and the duration of each flight (which influences the length of each time series) can vary. This variability adds another level of complexity to the dataset. We denote these cropped time series as $x = \{x^{t,k}\}$, $t \in \mathcal{T} \subset \check{\mathcal{T}}_i$, $k \in \mathcal{K}$. The choice of $\mathcal{T}$ can depend on the flight phases we wish to investigate, as deriving meaningful features from each phase can enrich our analysis. Specifically, for deep learning models that require uniform input dimensions, we may need to standardize the lengths of our time series. In such scenarios, T can be selected to represent a 'window' within each flight of a specific duration. This approach allows us to accommodate the inherent variability in our time series lengths while still utilizing the advanced capabilities of deep learning models.

Data Partitioning for Robust Aircraft Predictive Maintenance

The implementation of a robust and unbiased testing strategy is a critical element in developing our predictive maintenance models. The aircraft bleed air system, characterized by its complexity and significant role in aircraft operations, is one such system that requires particular attention. An aircraft typically has two independent bleed air systems, each consisting of several interconnected components. To simplify the analysis, we treat each system independently in our model. However, because the components of these systems are interconnected, the health of one part can potentially affect others within the same system. As such, we need to carefully consider this interconnectedness when partitioning our dataset into training and testing subsets to avoid undue interference and bias.

The data partitioning approach we have adopted is driven by a fundamental principle: splitting by aircraft. This means, rather than taking a random sampling of individual flights or components to form the training and testing groups, we specifically assign all flights from particular aircraft to the test set. This technique plays a pivotal role in maintaining the integrity of the bleed air system under observation. As the system's components are inherently interconnected, it's essential to encapsulate the entire system's behavior in our test data. Allocating all flights from certain aircraft to the test set enables us to achieve this goal, thereby ensuring the relevance and validity of our testing procedures. This approach also acts as a safeguard against potential interference that could stem from overlapping health states within a system. When we consider the interconnected nature of different parts within a bleed air system, it becomes clear that the health state of one component could influence another. Therefore, by keeping the aircraft data distinct within the training and testing groups, we ensure a clear demarcation and an unbiased reflection of our model's predictive accuracy. Moreover, our data partitioning approach addresses the need to account for operational variability across different aircraft. By assigning all flights from specific aircraft to the test set, we encapsulate this variability within our test data. This strategy enables us to check our model's ability to generalize its predictions across varying operational conditions, thereby enhancing its practical applicability and strengthening its predictive capabilities. To implement this partitioning strategy, we randomly select n_{test} aircraft and assign all of their flights to form our test set. The remaining flights of the remaining aircraft form our training set.

2.3 First approach

In this section, we propose a comparative exploration of prevalent methodologies in the field of predictive maintenance. Our focus will primarily be centered around two key approaches: the application of classical machine learning techniques involving feature extraction in conjunction with tree-based machine learning models for predicting the Remaining Useful Life (RUL), and a deep learning-based strategy tailored specifically for the same purpose.

The metric of our primary interest throughout this investigation is the RUL, defined as the estimated number of operational hours that remain before an alarm triggers, signifying a potential failure or need for maintenance. We aim to provide an analytical overview of how these two diverse methodologies perform in predicting this critical parameter, thus enabling a more informed strategy for predictive maintenance.

The succeeding subsections will present the exploration of each of these approaches, their implementation, and the results of their performance in predicting the RUL. The objective of this section is not only to understand these approaches better but also to shed light on their strengths and potential areas for improvement within the context of predictive maintenance.

Feature Extraction Based Approach

This subsection delves into our methodological approach, grounded in the principles of feature extraction, where we concentrate on two pivotal flight phases: the ascent and descent. These flight segments have been identified as being vital for the bleed air system, primarily due to the strain placed on the engine and its components. For example, during climb, the engine operates at high throttle, requiring active valve control to maintain air balance. Conversely, the cruise phase typically exhibits stability, which could limit the informational value of related signals. The descent phase, nonetheless, provides another rich opportunity for data collection and analysis, potentially revealing unique system behaviors under varying throttle conditions.

As part of our feature extraction strategy, we collect an array of basic statistical features for each signal. This comprehensive collection, which results in a significant number of features per flight, includes not only the maximum and minimum values, but also the mean, median, variance, percentiles, interquartile range, skewness, and kurtosis. Each of these features offers unique insights into the bleed air system's performance during the critical ascent and descent phases of flight. To further enhance our feature set, we also compute the signal's autocorrelation, capturing

its self-similarity across different time lags and adding another layer of valuable data about its temporal behavior. We supplement these time-domain features by delving into the frequency domain. This involves performing a spectral analysis of the signal, identifying the frequency with the highest energy – a key factor that often reveals hidden patterns – and calculating the energy of 10 distinct frequency bands. Furthermore, we determine the skewness and kurtosis of these frequency values, providing information about the asymmetry and 'tailedness' of the frequency distribution. The choice of these features is informed by their potential to highlight critical system anomalies, or variations that could indicate an impending system failure. In summary, our feature extraction strategy encapsulates a comprehensive exploration of each signal from multiple analytical perspectives, aiming to maximize the potential to uncover meaningful patterns and correlations in our predictive maintenance task.

Following feature extraction, we pivot to the application of machine learning algorithms—specifically, Random Forest and Boosted Trees—to predict the Remaining Useful Life (RUL). Random Forest utilizes an ensemble of decision trees and operates by aggregating their predictions to enhance overall prediction accuracy. Boosted Trees, on the other hand, employs an iterative process to minimize errors in prediction, refining its model with each iteration for improved prediction accuracy. Through this approach, we intend to take advantage of the vast wealth of features that are collected from each flight and use this information to effectively predict RUL and facilitate proactive maintenance planning. These algorithms have been utilized in various fields and applications due to their robustness and capability to handle complex datasets, making them an ideal choice for our predictive maintenance framework.

This methodology offers several significant advantages, primarily in its computational efficiency and scalability, which are crucial when dealing with massive amounts of data. By extracting simple features and using tree-based models, computation and training are accelerated, allowing for faster data processing. This efficiency is particularly beneficial in a real-world scenario where it is necessary to make daily predictions for a large fleet. Furthermore, the use of tree-based models contributes to the robustness of predictions, handling various data distributions and complexities, thereby ensuring the applicability and effectiveness of this method in a dynamic operational environment.

Deep Learning Approach

In this portion of our investigation, we delve into the utilization of deep learning methodologies. As with the feature-based approach, we concentrate on the climb phase of the flight. For the deep learning models, the climb phase data are segmented into 256-length crops, with a 50% overlap between each crop, creating a comprehensive and interlinked representation of the phase. This data preparation strategy can be seen as a data augmentation strategy and helps the model find relevant patterns anywhere in the time series.

The duration variability of the climb phase is significantly lower than that of the cruise phase, making it technically feasible to present the entire climb phase to our deep learning models. However, addressing shorter phases by zero-padding could lead to an inflation of non-informative data, which could potentially dilute meaningful patterns and impact the learning capabilities of our models. Furthermore, using the entire climb phase as input can lead to significantly less training data points as each data point becomes longer. This reduction in the number of training examples might make the deep learning models harder to train, posing a challenge to achieving high prediction accuracy. Additionally, incorporating longer time series data necessitates larger model architectures. While these larger models might be capable of capturing more complex patterns, they also require more computational resources for both training and prediction. This increase in resource demand can be a barrier to operational applicability, particularly when predictions need to be made daily on a large fleet. In light of these considerations, we adopt a balanced approach of segmenting the climb phase into manageable lengths while ensuring we do not lose essential information. This strategy seeks to strike a balance between maximizing model performance and maintaining reasonable resource requirements for effective real-time prediction.

The final RUL prediction for each flight is computed by taking the mean of the predictions for each crop, providing a consolidated estimate. For this task, we employ a three-layer convolutional neural network (CNN). In our context, the CNN operates on our time-series sensor data, effectively identifying potentially complex patterns that could be indicative of future system failures.

Results

This section provides an in-depth examination of the outcomes of our experimentation with the feature extraction based approach and the deep learning approach. The analysis is rooted in several crucial aspects of predictive accuracy: capturing

degradation trends, predicting the remaining useful life (RUL) of the bleed air system, and the precision of predictions as the system approaches the end of its life.

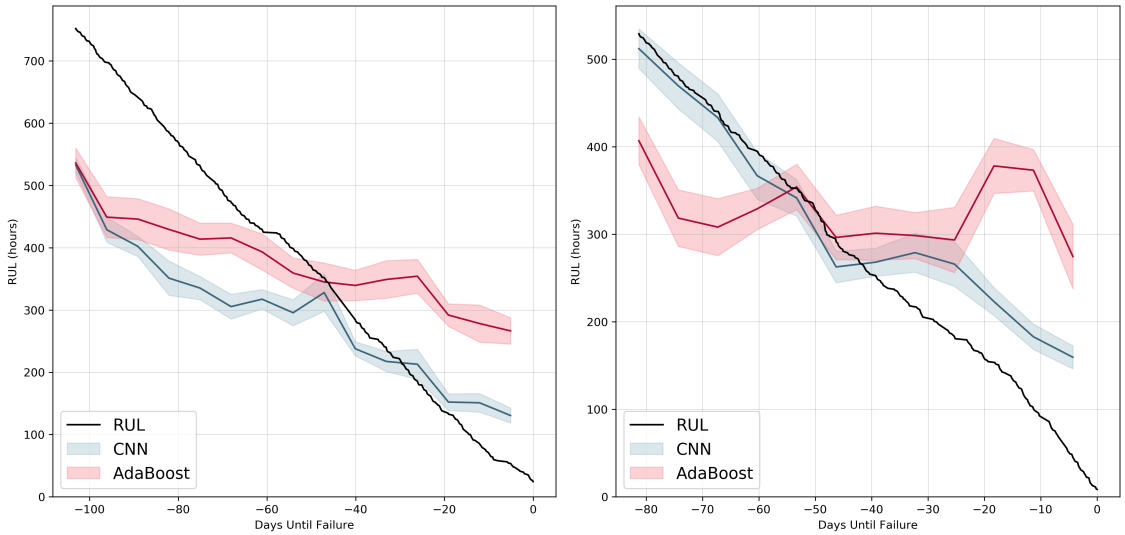
When comparing both methodologies, it's crucial to note that they were found to have different strengths and weaknesses. We have used three common metrics to evaluate the performance of these models: the Mean Absolute Error (MAE), the Mean Squared Error (MSE), and the median error. A summary of the comparative results is presented in Table 2.1. It can be observed that the deep learning model, more specifically the Convolutional Neural Network (CNN), has outperformed the AdaBoost model based on extracted features in all the metrics. The feature extraction based methodology struggles to effectively capture the degradation patterns of the bleed air system, and this is not surprising for two main reasons: First, the degradation patterns tend to manifest at specific moments during the flight. However, since the features are computed over large sections of the flight, these vital signals could be drowned out amidst the surrounding noise. Secondly, the bleed air system is a sophisticated device with no obvious degradation features. The lack of clear indicators that align with its component degradation makes the task of predicting its health status substantially more challenging than in comparatively simpler scenarios, such as the ones encountered in the C-MAPSS dataset. This complexity amplifies the difficulty of predictive maintenance, necessitating more robust and nuanced modeling strategies.

Models	MAE	MSE	Median error
AdaBoost	181 $\pm$12	206 $\pm$14	173 $\pm$12
CNN	162 $\pm$11	191 $\pm$12	146 $\pm$10

Table 2.1: First approach performances

For a more qualitative perspective on the results, we have charted the evolution of the RUL predictions over two lifecycles, as shown in Figure 2.4. In this figure, the solid lines depict the 7-day rolling means with their corresponding confidence intervals. The target RUL is indicated by the black line. In Figure 2.4a, both the AdaBoost and CNN models successfully capture the degradation trend, which bodes well for their predictive capabilities. However, the CNN model seems to deliver predictions that are closer to the target RUL. Figure 2.4b shows a similar pattern, with the CNN capturing the degradation trend while the AdaBoost model does not. Yet, there is a notable failing in the CNN model's predictions as the system approaches failure - it does not predict the RUL accurately during this crucial period. This issue is also apparent in the first lifecycle, which presents a significant problem.

The ability to predict with high precision as the system nears its end of life is crucial for scheduling timely maintenance. Hence, despite the better overall performance of the CNN model, this limitation presents an area needing further investigation and potential improvement.



(a) (b)

Figure 2.4: Comparative Analysis of RUL Predictions for Two Lifecycles

In summary, while the CNN-based method demonstrated superior capabilities in trend prediction compared to the AdaBoost model, its end-of-life prediction precision and higher computational demands present notable limitations. Future work could focus on refining these models further to enhance their prediction precision towards the end of life, improving their computational efficiency, and exploring hybrid methodologies that could leverage the strengths of both approaches. Despite the current shortcomings, these methodologies provide a strong foundation for the development of a more refined predictive maintenance framework.

2.4 Conclusion and Insights

This chapter marks the preliminary stages of our work on predictive maintenance for critical aircraft components, with a special focus on the bleed air system. Our discussion begins with a comprehensive exploration of the bleed air system, a major feature critical to the functionality of an aircraft. Understanding its complexity, primary components, and their impact on the overall performance of the aircraft is essential to our research. The final part of this chapter brings to light some fundamental approaches to predictive maintenance. Although relatively simple, these

methods allow us to understand the problem, appreciate the challenge involved, and identify the limitations of basic methodologies. It also lays the groundwork for the development and application of more sophisticated models discussed in subsequent chapters.

Building on this foundation, we have considered two prevalent methodologies for predicting the Remaining Useful Life (RUL) - one based on feature extraction with tree-based machine learning models, and the other rooted in deep learning strategies. Through our analyses, we highlighted that while the Convolutional Neural Network (CNN) demonstrated superior performance in capturing degradation trends, it failed in its predictive accuracy as the system approached end-of-life. Meanwhile, the AdaBoost model based on extracted features struggled to effectively capture degradation patterns due to the complexity and the sporadic nature of the degradation signals.

The results of our exploratory research highlight the complex nature of predictive maintenance in aviation, coupled with the existing limitations of the models we evaluated. However, these challenges should not discourage us, but rather serve as drivers for further investigation and refinement of our methods. A prominent result of our investigations is the demonstrated ability of deep learning models to detect pertinent features indicative of the degradation of the bleed air system. Despite this positive result, the issue of constructing an effective strategy for generating the Remaining Useful Life (RUL) target remains a significant hurdle. As we outlined in the first chapter, developing a robust RUL prediction strategy for our use cases is extremely challenging due to the intricacies of aircraft operations and maintenance schedules.

Another significant constraint we face is the lack of labeled data. In the aviation industry, preventive maintenance often precedes an actual system failure, making it impossible to establish an accurate RUL target in scenarios where there is no imminent failure or where preventive maintenance has been performed. This limitation reduces our ability to make the best use of available data and limits the application of more sophisticated deep learning models that could potentially detect more complex degradation features. However, we believe that there is unexploited potential in autoencoder models, particularly variational autoencoders (VAEs), for our use case. The ability of autoencoders to capture complex phenomena in high-dimensional data and effectively reduce dimensionality is widely recognized. Coupled with the need to distinguish between healthy and degraded system behavior, similar to anomaly detection, VAEs in particular appear to be an excellent fit for our needs.

In the next chapter, we will present the beneficial properties of VAE for dimensionality reduction, especially in the context of multivariate time series data. We will take the opportunity to compare different VAE architectures and explore their unique features and advantages. By taking this step, we expect to improve our understanding of predictive maintenance in aviation and identify better strategies to address the challenges we face.

Chapter 3

VAE as a feature extraction tool

In the next chapter, we will explore the beneficial properties of VAE for dimensionality reduction, especially with respect to multivariate time series data. We will engage in an extensive comparison of different VAE architectures, exploring their unique characteristics and advantages.

In order to do this, we will be using a data set that is different from the data from the bleed air system that we have been working with so far. Our choice was driven by two key considerations. First, we wanted to use a well-studied and publicly available dataset that would allow us to compare our results and findings more broadly. Second, we were looking for a dataset that represented multivariate time series data with intrinsic correlations between signals and that demonstrated the presence of anomalous behavior within these signals.

With these considerations in mind, we chose electrocardiogram (ECG) data as the subject of our investigation. ECG data not only meet our requirement for multivariate time series data, but also reflect the interplay observed in aircraft systems by showing robust correlations between signals. Furthermore, the presence of distinct pathological classes within the ECG data, representative of abnormal behavior, provides us with an ideal playground to explore the strengths of VAEs.

Dimension Reduction for time series with Variational Autoencoders

William Todo[1,2], **Merwann Selmani**[2], **Béatrice Laurent**[1,3], **Jean-Michel Loubes**[1]
[1] Institut de Mathématiques de Toulouse, Toulouse, France
[2] Liebherr Aerospace Toulouse
[3] INSA de Toulouse, Toulouse, France

Abstract

In this study, we investigate dimensionality reduction techniques for univariate and multivariate time series data, with a particular focus on comparing wavelet decomposition and convolutional variational autoencoders (VAEs) for this purpose. Our experiments demonstrate that VAEs are a promising option for reducing the dimensionality of high-dimensional data, such as electrocardiogram (ECG) signals. We conduct these comparisons on a real-world, publicly available ECG dataset characterized by substantial variability, using reconstruction error as the evaluation metric. Furthermore, we assess the robustness of these models in the presence of noisy data during both training and inference, reflecting the challenges commonly encountered in real-world time series data analysis. Our results indicate that the VAE exhibits remarkable resilience in both settings. Additionally, we explore the impact of different encoder and decoder architectures on the performance of VAEs. This comprehensive analysis provides valuable insights for practitioners seeking to employ dimensionality reduction techniques in the analysis of time series data.

3.1 Context

The curse of dimensionality has been at the heart of decades of research in statistics and machine learning, preventing the use of many methods as the dimension of the data increases. However, with the exponential growth of data collection, high-dimensional data such as images, time series, or functional data are being studied more and more. Traditional machine learning techniques are not a good fit for this kind of data, firstly due to the practical difficulty of handling this kind of data from a computational point of view, but also from a theoretical point of view since the

accuracy of the algorithms is hindered for high dimensional data. To overcome these problems, a well-studied way of preprocessing the data is to reduce its dimensionality. Principal Components Analysis (PCA) Wold et al., 1987 is a well-known general technique that has been widely studied together with Independent Components Analysis (ICA) Comon, 1994 or Factor Analysis (FA) Harman, 1976 and more. More recently, with the advent of AutoEncoders (AE) and Variational AutoEncoders (VAE) Kingma and Welling, 2013, more complex data-driven dimensionality reduction techniques have become possible and have been benchmarked (Mahmud et al., 2020 & Dai et al., 2017).

Time series and multivariate time series require a special attention. Actually classical dimensionality reduction techniques fail to capture the temporal aspect of the observations, therefore there are some specific techniques for that kind of data, in particular projection methods onto specific bases for instance onto generical wavelet basis as in Y. Liu, 2009 or data driven basis using for instance Functional Principal Component Analysis (FPCA) Di et al., 2009. Wavelet transform uses a low-pass filter to extract low frequency information and a high-pass filter to extract high frequency information. The advantage over the Fourier transform is that the positional information is conserved with the wavelet transform. FPCA decomposes functional data into basis functions that explain the variance. The outcome of such method is to discover features using such projections, that are expected to concentrate the information on a small number of coefficients. Dimension reduction plays a pivotal role in managing the voluminous information encapsulated in time series data. This technique, widely employed in tasks such as clustering Javed et al., 2020 and anomaly detection Barreyre et al., 2019, facilitates efficient data analysis by reducing computational requirements and mitigating the curse of dimensionality.

Very recently, Machine Learning methods using deep neural networks have been considered as alternatives to such methods. They enable to construct low dimension embedding by considering the features from the penultimate neural layer that are used to build the forecast. In the same vein, variational autoencoders map the data into a structured representation of lower dimension in a data driven way. However, the use of variational autoencoders on time series as a dimension reduction technique is not yet well studied or compared to other methodologies.

This paper shows the advantages of VAE regarding the dimensionality reduction power and robustness against more traditional methods like wavelet decomposition and FPCA on real world ECG datasets.

In this paper, we trained various convolutional variational autoencoders to obtain a lower dimensional representation of the dataset PTB-XL Wagner et al., 2020. We compare those results with the state of the art method of the wavelet decomposition. Then we stress the tests to highlight the good properties of VAEs. We also test these VAEs on other real world ECG datasets from the Physionet challenge Alday et al., 2020.

3.2 Background

Datasets

In this study, we utilize a real-world dataset of ECG measurements: the PTB-XL dataset. This dataset consists of samples that each contain 12 recordings of electrical activity on the body surface, measured over a duration of 10 seconds at a sampling frequency of 100 Hz. Consequently, each sample in this dataset is a 12,000-dimensional point, highlighting the necessity for dimensionality reduction. The PTB-XL dataset features a diverse range of ECG signals, providing a robust testbed for evaluating the effectiveness of our dimensionality reduction methods. To further investigate the performance of these techniques on univariate time series, we also employ a cropped version of the PTB-XL dataset. This cropped version retains all samples but includes only the first ECG lead, resulting in each sample consisting of 1,000 data points, as depicted in Figure 3.1.

By conducting experiments on both the original and cropped versions of the PTB-XL dataset, we can evaluate the efficacy of our proposed dimensionality reduction techniques on multivariate and univariate time series data, respectively. This comprehensive analysis will provide valuable insights into the most suitable methods for ECG signal compression and reconstruction.

In order to demonstrate the generalization capability of the VAE trained on the PTB-XL dataset, we evaluate its performance on two additional datasets from the PhysioNet Challenge Alday et al., 2020. The first dataset, referred to as "Georgia," consists of $10,344$ ECG recordings sourced from Georgia. The second dataset, named "China," contains $3,453$ ECG recordings of unused data from the CPSC2018 F. Liu et al., 2018. To ensure consistency across datasets and facilitate a fair comparison, we adopt the preprocessing steps outlined in Singstad and Tronstad, 2020. These steps include scaling the ECG signals and resampling them to a frequency of 100 Hz, which matches the sampling rate of the PTB-XL dataset. By evaluating the VAE's performance on these diverse datasets, we aim to demonstrate its potential for generalization and applicability to a wide range of ECG signal analysis tasks.

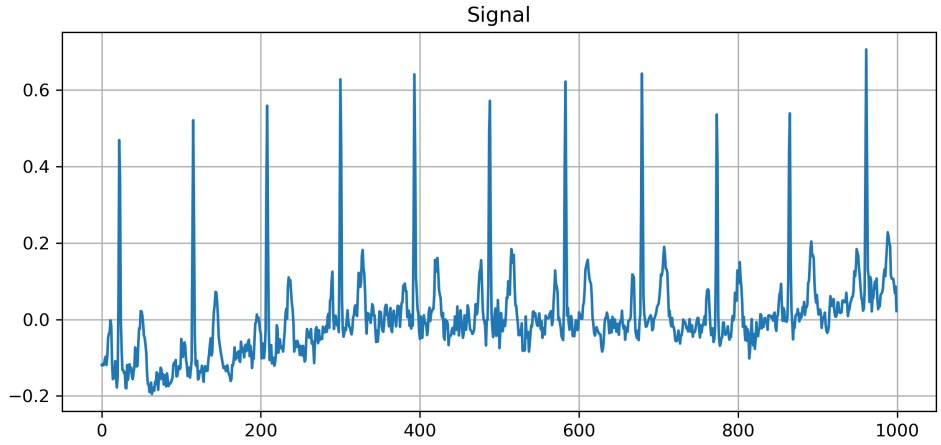


Figure 3.1: Example signal from the dataset PTB-XL

This comprehensive evaluation will provide valuable insights into the versatility and robustness of the VAE as a dimensionality reduction technique for ECG data.

For the purpose of the paper, we denote the length of the time series as T and the number of parameters or features in the data as K. For a given time series, each data point at time t can be represented as $x^{t,k}$, where k is the index of the feature dimension. Similarly, the corresponding reconstructed data point from the VAE is denoted as $\hat{x}^{t,k}$.

We define the sets of all time points and feature dimensions as $\mathcal{T} = 1, 2, ..., T$ and $\mathcal{K} = 1, 2, ..., K$ respectively.

VAE

Variational Autoencoder (VAE) is a powerful deep learning technique that has attracted considerable attention in the fields of machine learning and computer vision. VAEs are generative models that use a neural network architecture to learn the underlying probability distribution of a given data set. The main idea behind VAE is to find a latent representation of the data that captures the important features of the data while allowing us to sample from this latent space to generate new data points that are similar to the original data set.

VAEs were introduced by Kingma and Welling, 2013 and Rezende et al., 2014. They consist of two main components: an encoder and a decoder. The encoder, denoted as $q_\Phi(Z|X = x)$, takes an input x from a high-dimensional space and maps it to a J-dimensional Gaussian distribution with a mean vector $\mu(x) = (\mu_j(x))_{1 \leq j \leq J}$ and diagonal covariance matrix $\text{diag}\,(\sigma^2(x))$ with $\sigma^2(x) = \left(\sigma_j^2(x)\right)_{1 \leq j \leq J}$. The resulting

distribution can be expressed as $q_\Phi(Z|X = x) \sim \mathcal{N}_J(\mu(x), \text{diag}(\sigma^2(x)))$. The decoder, denoted as $p_\theta(X|Z = z)$, takes a sample z from the latent Gaussian distribution produced by the encoder as input and generates the associated element in the original high-dimensional space. This reconstructed data point is denoted by $\hat{x}$.

The VAE is trained by optimizing the evidence lower bound (ELBO), which is a lower bound on the log-likelihood of the data. The ELBO consists of two terms: the reconstruction loss and the KL-divergence regularization term. The reconstruction loss measures the difference between the original input data point x and its reconstructed version $\hat{x}$. This is typically calculated using the mean squared error (MSE) for continuous data. The KL-divergence regularization term, denoted as $D_{KL}(q_\Phi(Z|X = x)||p(Z))$, encourages the learned latent space to be close to a standard normal distribution, which is typically represented as $p(Z) \sim \mathcal{N}(0, Id)$, where Id is the identity matrix.

$$\begin{aligned}\mathcal{L}_{\text{VAE}}(\Phi, \theta; x) = &- \mathbb{E}_{Z\sim q_\Phi(Z|X=x)}\left[\log p_\theta(X|Z)\right] \\ &+ D_{KL}\left[q_\Phi(Z|X = x)||p(Z)\right]\end{aligned} \tag{3.1}$$

The conditional distribution, $p_\theta(X|Z)$, is specified by the decoder as a Gaussian distribution, denoted as $\mathcal{N}(\mu = \hat{x}, \Sigma)$. In the context of VAEs, it is common to assume a unit variance, i.e., $\Sigma = Id$, for simplicity. When optimizing the negative log-likelihood expectation, $-\mathbb{E}_{Z\sim q_\Phi(Z|X=x)}\left[\log p_\theta(X|Z)\right]$, under this assumption, it turns out to be equivalent to minimizing the mean squared error (MSE) between the input data x and the reconstructed data $\hat{x}$. This equivalence allows us to express the first term in Eq.(3.1) as the MSE within the VAE framework, providing an intuitive connection between the VAE's probabilistic formulation and familiar error metrics.

$$||\hat{x} - x||^2 = \frac{1}{T \times K} \sum_{t\in\mathcal{T}_x, ;k\in\mathcal{K}} \left(\hat{x}(t,k) - x(t,k)\right)^2 \tag{3.2}$$

The second term in Eq.(3.1) is the Kullback-Leibler (KL) divergence between $q_\Phi(Z|X = x)$ and $p(Z)$. Since the distributions are respectively $\mathcal{N}_J(\mu(x); \text{diag}(\sigma^2(x)))$ and $\mathcal{N}_J(0; Id_J)$, we can easily compute this term. The KL divergence is a measure of the difference between two probability distributions and is defined as:

$$D_{\text{KL}}[q_\Phi(Z|X = x)||p(Z)] = \mathbb{E}_{Z\sim q_\Phi(Z|X=x)} \log\left(\frac{q_\Phi(Z|X = x)}{p(Z)}\right) \tag{3.3}$$

For the given Gaussian distributions, the KL divergence can be calculated as:

$$D_{\mathrm{KL}}[q_\Phi(Z|X=x)||p(Z)] = \frac{1}{2}\sum_{j=1}^{J}\left[-1 - \log\left(\sigma_j^2(x)\right) + \mu_j^2(x) + \sigma_j^2(x)\right] \tag{3.4}$$

Thus, the optimization of the VAE involves minimizing the sum of the reconstruction loss, as represented by the MSE in Eq.(3.2), and the KL divergence regularization term in Eq.(3.4). By striking a balance between these two terms, VAEs provide a powerful framework for learning compact, meaningful representations of high-dimensional data through an unsupervised generative process. VAEs are particularly suitable for applications such as dimensionality reduction and anomaly detection.

Wavelets

Another dimensionality reduction technique studied is the wavelet decomposition technique, which is a popular method for signal compression and dimensionality reduction Hilton, 1997. Wavelet transform efficiently concentrates the signal information into a small number of coefficients, making it possible to perform lossy compression by eliminating coefficients with small magnitudes. This powerful concentration of information was in particular largely used in the standard of compression of images JPEG 2000 Rabbani and Joshi, 2002. To evaluate the effectiveness of wavelet-based compression on ECG signals, we adopt two different methods and compare the compression rate, which is computed by counting the number of kept coefficients.

The Global Approach, suitable for real-world settings, computes the wavelet transform across the entire dataset, retaining only the n coefficients with the highest energy. Here, the energy refers to the the absolute value of a wavelet coefficient, a measure indicative of its significance within the transformed signal. After calculating the energy of each coefficient across the entire dataset, we then select the n coefficients that possess the highest energy. This method is especially useful for handling non-synchronous time series, where the location of important features may vary across different series in the dataset.

In wavelet decomposition, the first wavelet coefficients typically arise from the highest level of decomposition and therefore possess the greatest energy. Hence, in practice, the Global Approach often boils down to retaining the first n wavelet coefficients. It's important to note that the mean energy of the wavelet coefficients throughout the dataset doesn't necessarily indicate the location of a specific feature,

but rather reflects the level of decomposition. This phenomenon is illustrated in Figure 3.2, which visualizes the wavelet decomposition process and the selection of high-energy coefficients.

The second method involves retaining the n largest coefficients for each individual signal. While this approach yields significantly better performance compared to the Global Approach, it is less practical in real-world applications. The reason for this limitation is that we need to keep track of which coefficients were retained in order to reconstruct the signal accurately. Essentially, this method serves as an oracle, demonstrating the best-case scenario for wavelet decomposition if we had perfect knowledge of which coefficients to keep for each signal.

By examining both the Global Approach and this oracle method, we can gain a comprehensive understanding of the potential performance of wavelet decomposition in compressing ECG signals.

By comparing the VAE compression and wavelet decomposition techniques, we aim to identify the most effective method for compressing ECG signals. Wavelet decomposition is a widely used technique in this context, as shown in Addison, 2005 and C. Li et al., 1995. We take into account factors such as compression rate, signal reconstruction quality, and computational efficiency in our comparison. This comparison will provide valuable insights for researchers and practitioners working on multivariate time series and related applications.

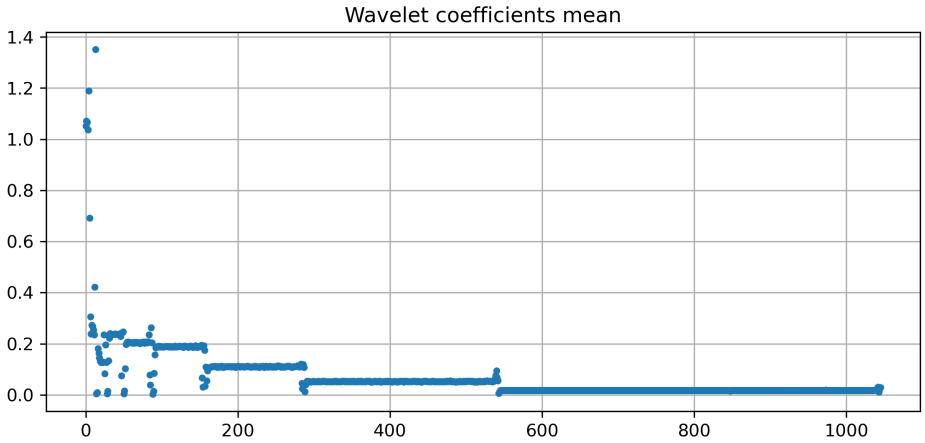


Figure 3.2: Mean throughout the dataset of the energy of the wavelet decomposition

Functional PCA

Functional Principal Component Analysis (FPCA) introduced in Di et al., 2009 is a widely-used technique for dimensionality reduction in functional data, where the observations are functions or curves measured over a continuous domain, such as time or space. FPCA is particularly effective at capturing the main patterns or structures in smooth and regular functional data with a strong temporal structure. The method involves several key steps, including the computation of the mean function $m(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t)$, which represents the average value of the functional data at each time point t.

The covariance function, $C(s,t) = \frac{1}{n}\sum_{i=1}^{n} [x_i(s) - m(s)]\,[x_i(t) - m(t)]$, measures the similarity between observations at different time points s and t. The eigenanalysis of the covariance function involves solving the eigenvalue-eigenfunction equation $\int C(s,t)\phi_k(t)dt = \lambda_k\phi_k(s)$, where λ_k are the eigenvalues and $\phi_k(s)$ are the corresponding eigenfunctions.

The functional principal components (FPCs) are obtained as the projections of the original data onto the eigenfunctions, given by $\alpha_{ik} = \int [x_i(t) - m(t)]\,\phi_k(t)dt$. Finally, the original functional data can be approximated by using a subset of the most significant FPCs, as shown in the formula $\hat{x}_i(t) = m(t) + \sum_{k=1}^{K} \alpha_{ik}\phi_k(t)$, where K is the number of selected FPCs and $\hat{x}_i(t)$ is the reconstructed version of the original data $x_i(t)$. By choosing an appropriate number of FPCs, FPCA can reduce the dimensionality of the data while preserving most of its information content as shown in Muelas et al., 2017.

In some cases, however, FPCA may not be the most suitable method for dimensionality reduction, particularly when dealing with non-synchronous or highly diverse functional data. Non-synchronous data refers to time series that are not aligned, meaning that significant features or events occur at different time points across the dataset. Applying FPCA to non-synchronous or highly diverse data often requires extensive preprocessing, such as resynchronizing the time series in the dataset. This can be a complex and time-consuming process, potentially introducing inaccuracies or biases in the data.

Due to these challenges and limitations, we have chosen not to use FPCA for comparison purposes in this paper. Instead, we will focus on alternative methods that are more suitable for our specific dataset and research objectives.

3.3 One dimensional time series

Method

In this section, we outline the architecture, training process, and implementation details of the VAEs used for one-dimensional time series experiments. We employ the PyTorch library Paszke et al., 2019 to implement the VAE, as it offers significant flexibility for developing these models.

Our experiments utilize a CNN VAE with symmetrical encoder and decoder architectures, each comprising a 3-layer deep convolutional neural network. Throughout the experiments, the number of filters and their sizes remain constant at 256, 512, and 512 for the number of filters, and a filter size of 5.

During the training phase, we incorporate 256-length crops of the signal, a data augmentation technique inspired by the methods presented in Strodthoff et al., 2020. In their work, they benchmark various deep learning models for ECG analysis. They demonstrated the efficacy of using shorter signal crops for training these models, which improved model generalization by allowing the model to learn from different parts of the ECG signals. Following a similar approach, we employ these crops in our training process to account for the potential variability in the location of regions of interest within the signal, especially given that the heartbeats are not synchronized. To further enhance the robustness of our model to different heart rates present in the dataset, we introduce additional variability into the training data. Specifically, we artificially modulate the cardiac rhythm by up-sampling and down-sampling the input signal. This procedure ensures that our model is exposed to a wider range of heart rate patterns during training.

The reconstruction error is computed as the mean of four VAE reconstruction errors, ensuring that the error measurement encompasses the entire signal rather than just the cropped portion. This approach provides a more robust assessment of the model's performance in reconstructing the original ECG signals.

For each experiment, we employ 10-fold cross-validation using the pre-defined folds within the dataset, which helps to mitigate the risk of overfitting and provides a more reliable estimation of the model's performance. In the accompanying figures, we present confidence intervals to provide a visual representation of the variability in the model's performance across different folds.

We investigate various compression ratios for both the VAE technique and the wavelet approaches described in Section 3.2. We utilize the mean squared error (MSE) between the original and reconstructed signals as our evaluation metric. The

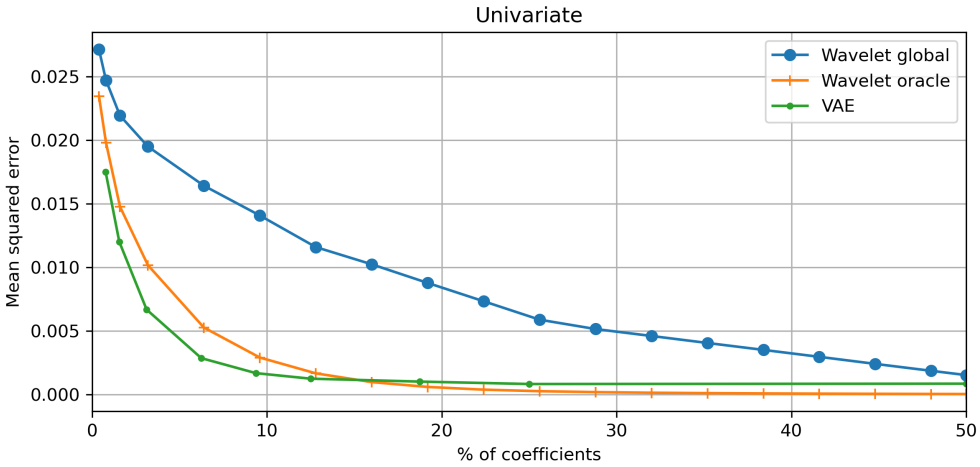


Figure 3.3: Performance comparison between VAE and wavelets giving the percentage of kept coefficients

performance of these methods is compared in Figure 3.3.

Our results demonstrate that variational autoencoders outperform both the standard wavelet decomposition method and the oracle when compressing data to a very low dimension (compression rate of 10 or higher). This outcome can be attributed to the fact that VAEs employ tailored convolutional filters, which provide a better fit for the data compared to the non-data-driven wavelet basis. Consequently, VAEs achieve a lower reconstruction error than wavelet decomposition, maintaining competitive performance even with a reduced compression rate.

Variational autoencoders offer several advantages over wavelet transform in high compression rate scenarios, including a lower error rate and a well-structured latent space that can be readily employed for further processing. Although VAEs also possess powerful feature extraction capabilities, we do not delve into this aspect in the current analysis. Instead, our focus remains on demonstrating the superior performance of VAEs in reconstructing ECG signals from highly compressed representations.

3.4 Multidimensional time series

Reconstruction

In the previous section, we established that the Oracle wavelet method is the only approach that can closely compare to the performance of the VAE at high levels of compression, which is our primary focus in this paper. Consequently, in this section,

we will only consider the comparison between the wavelet oracle and the VAE.

The oracle method retains the n largest coefficients from the decomposed signals, implying that we may not necessarily maintain the same number of coefficients for each dimension. In some cases, we might even completely ignore one dimension if no coefficients are retained. This aspect is crucial in understanding the performance difference between the oracle wavelet method and the VAE, especially when dealing with highly correlated multivariate time series data.

The CNN VAE employs filters that can combine information from all dimensions of the time series. The 12-lead ECG data are highly correlated since they all capture the same heartbeat from different positions. This inherent correlation provides an opportunity for the CNN VAE to exploit this relationship and effectively reduce dimensionality for this type of data. In practice, this holds true when retaining a small number of coefficients for reconstruction, as shown in Figure 3.4. When less than 4% of the coefficients are retained, the reconstruction error is significantly better than that of the wavelet oracle. However, beyond that point, the model does not exhibit substantial improvement despite an increase in the size of the latent space.

Several factors can contribute to the VAE's limited ability to reconstruct input with high precision, even with a large latent space. One primary reason is the regularization imposed on the VAE, which can sometimes create a trade-off between accurate reconstruction and a well-structured latent space. As a consequence, the VAE might struggle to reconstruct the input with high precision, even when provided with a larger latent space. In the context of image data, VAEs often produce blurry images due to this regularization effect. Similarly, for time series data, the regularization can result in the loss of high-frequency components during reconstruction. Another potential issue arises when the training set is not representative of the test set, leading to suboptimal learned decompositions. In such cases, the model may not perform well on unseen data as it has not adequately captured the underlying data distribution. In our particular case, it is most likely the regularization that is primarily responsible for the limited reconstruction performance of our models.

The results are summarized in Table 3.1, where the numbers are bolded when the VAE results surpass the global wavelet method and underlined when the VAE outperforms the wavelet oracle. The performance on the Georgia and China datasets are computed using the VAE trained on the PTB dataset. The results are similar on the Georgia dataset and only slightly inferior on the China dataset, demonstrating the impressive generalization capabilities of the VAE. We observe the same trends

across all three datasets: VAEs consistently perform better than the wavelet method at high compression ratios, even when compared to the oracle. This outcome highlights the VAE's ability to capture and reconstruct the important features of the ECG signals more effectively than wavelet methods, particularly when working with multivariate time series data. In addition to better reconstruction performance, the VAE's structured latent space can be easily used for further processing and analysis, offering numerous advantages over the wavelet transform.

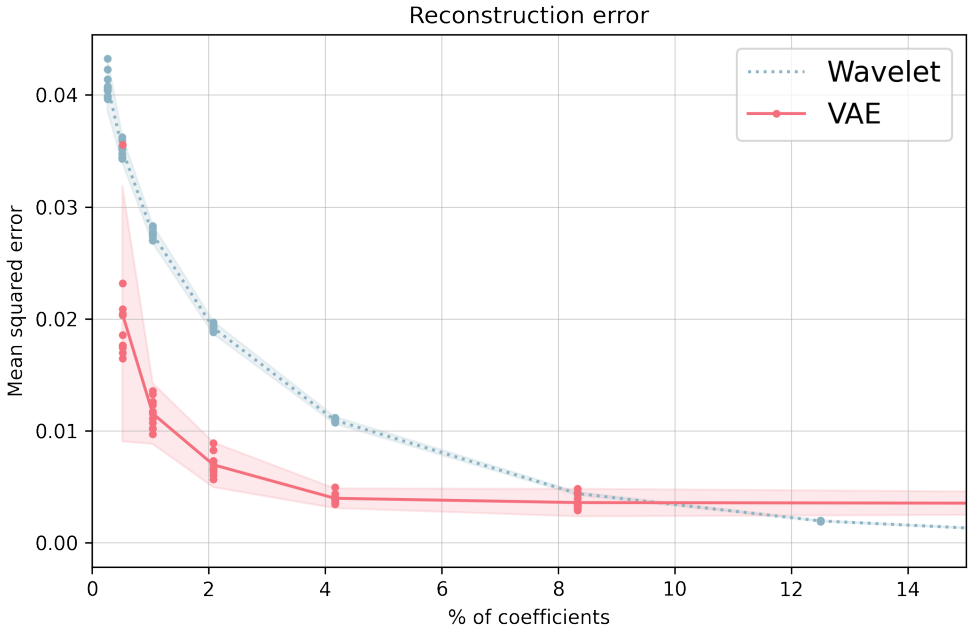


Figure 3.4: Performance comparison between VAE and wavelets giving the percentage of kept coefficients on PTBXL dataset, Wavelet correspond to the oracle

Prediction after reconstruction

The primary goal of dimensionality reduction is to enable efficient processing of multivariate time series data while mitigating the curse of dimensionality. In this context, we aim to assess how dimension reduction techniques impact the class-related characteristics of the signals. The PTB-XL dataset comprises samples labeled by one or multiple experts based on the associated pathologies. The dataset contains four pathological classes and one healthy class.

For each test fold, we train a convolutional neural network (CNN) classifier on raw signals and then predict the classes using various reconstructed signals. The classifier consists of a standard 4-layer CNN architecture, with each convolutional

Dataset	Method	Mean Squared Error		
PTB	VAE	**0.02201**	**0.00723**	**0.00487**
	Global	0.05074	0.03815	0.00603
	Oracle	0.03624	0.02002	0.00008
Georgia	VAE	**0.01759**	**0.00723**	**0.00531**
	Global	0.04057	0.03220	0.00576
	Oracle	0.02871	0.01612	0.00008
China	VAE	**0.03632**	**0.01972**	0.01424
	Global	0.06356	0.05313	0.00710
	Oracle	0.04053	0.02175	0.00009
Coefficients %		0.5	2	33

Table 3.1: Dataset Comparison

block containing a convolutional layer, batch normalization, leaky ReLU activation, and max pooling layer to increase the receptive field throughout the model. Dropout layers are placed between each convolutional block. The classifier employs 64, 128, 256, and 512 convolutional filters with sizes of 5, 5, 3, and 3, respectively. Overall, this architecture performs similarly to the classifiers presented in the benchmark from Strodthoff et al., 2020, achieving an AUC of approximately 0.92.

We compare the AUC scores of the raw data with the results from the reconstructed signals obtained using VAE and oracle wavelet methods, as presented in Figure 3.5. As expected, when the wavelet method provides better reconstruction, it also yields better prediction results, starting with more than 8% of the coefficients retained. Interestingly, at high compression ratios of 50 or more (i.e., less than 2% of the coefficients), the performance of the VAE reconstruction is significantly better. Notably, the disparity in prediction performance exceeds the difference in MSE scores. This observation implies that, despite the relatively small difference in MSE scores between the two methods, the VAE's ability to preserve class-related features at high compression ratios has a more substantial impact on the classifier's performance. Consequently, this highlights the importance of considering not only reconstruction errors, but also the preservation of meaningful features in the context of dimensionality reduction for classification tasks. It appears that, at high compression ratios, class-related features are largely absent from the wavelet reconstruction, leading to the classifier's inability to differentiate classes in the reconstructions. As previously explained, the oracle method can entirely miss some signals at high compression rates, which impairs the classifier's ability to distinguish between classes.

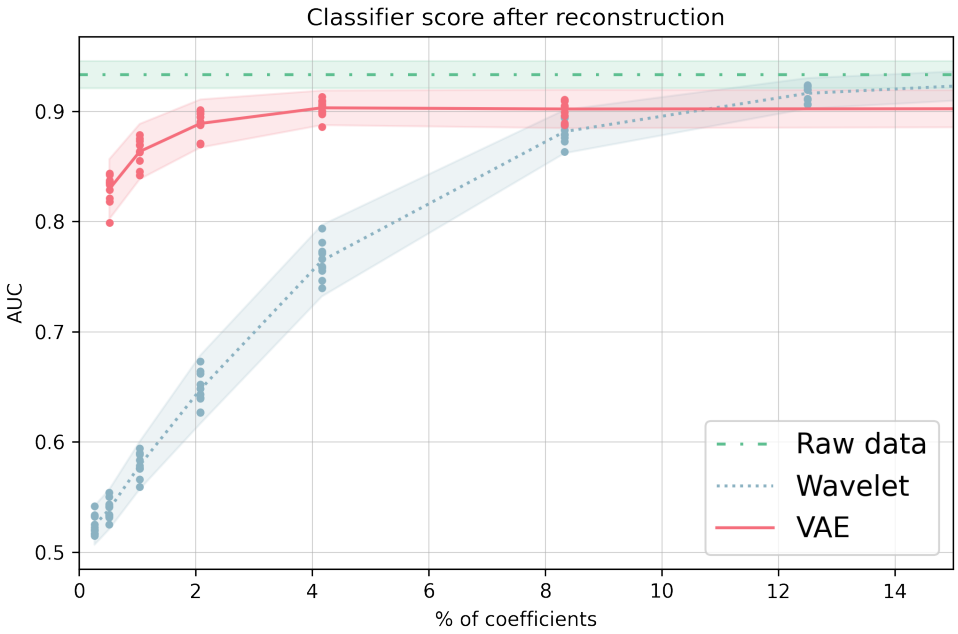


Figure 3.5: Prediction AUC after reconstruction

Noise Robustness

In this section, we present experiments conducted to assess the robustness of VAE feature extraction on noisy data, given that wavelet decomposition is frequently employed as a denoising tool. By eliminating small coefficients in the wavelet decomposition using hard or soft thresholds, noise in the reconstructed signal is reduced (Donoho and Johnstone, 1994, Chang et al., 2000). We introduce white noise to our time series, as illustrated in Figure 3.6, accounting for 20% of the signal variance.

We assess the robustness of the VAE to noise through two distinct experimental scenarios. The first scenario, labeled as 'noisy VAE noisy input' in the figure, simulates real-world conditions where data collection is often influenced by noise. Here, we train the VAE on noisy data and evaluate its performance by comparing the reconstruction of a noisy time series to the original clean time series. This allows us to gauge the impact of noise on the VAE's training and its ability to reconstruct clean signals from noisy inputs. The second scenario, termed 'VAE noisy input', tests the VAE's resilience to unexpected noise during inference. In this case, the VAE is trained on clean data, but it encounters noisy data during the testing phase. This scenario lets us explore the model's robustness when faced with noise that was not present during training.

As depicted in Figure 3.7, a VAE trained on clean data and tested on noisy data exhibits strong performance, with only a slight difference compared to the VAE tested on clean data. This result demonstrates the VAE's robustness to noise. Moreover, the figure indicates that the VAE trained on noisy data maintains good performance despite some information loss.

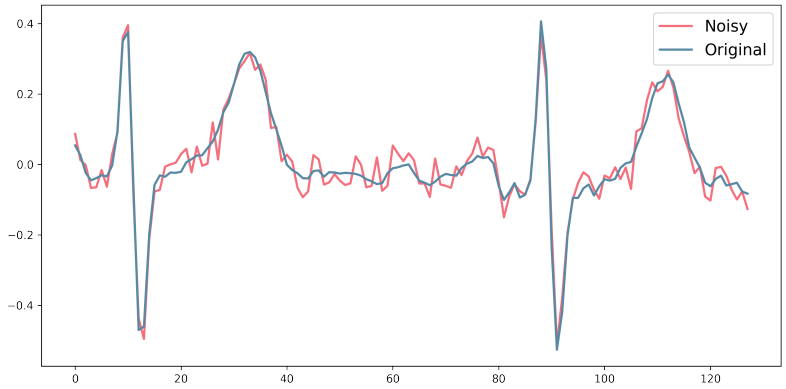


Figure 3.6: Effect of the added noise

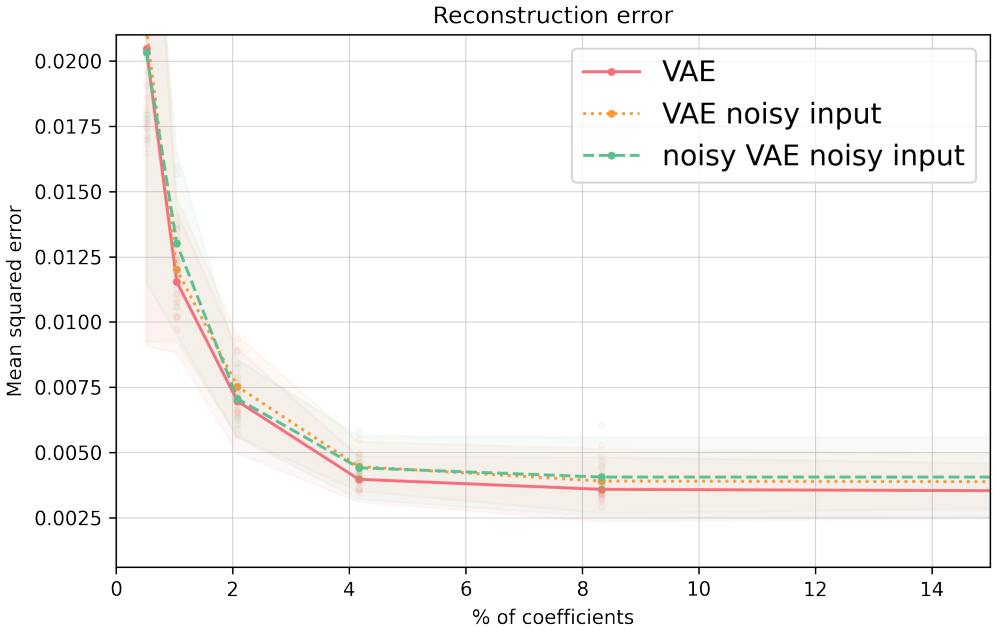


Figure 3.7: Performance comparison for VAE trained on various kinds of data

Anomaly detection

In this section, we present an experiment that evaluates the anomaly detection capabilities of a Variational Autoencoder (VAE) on noisy time series data. The primary

		Predicted label	
		Anomaly	Normal
True Label	Anomaly	55%	45%
	Normal	45%	55%

Table 3.2: Confusion matrix for anomaly detection using wavelets reconstruction errors

		Predicted label	
		Anomaly	Normal
True Label	Anomaly	69%	31%
	Normal	31%	69%

Table 3.3: Confusion matrix for anomaly detection using VAE reconstruction errors

advantage of VAEs over wavelet decomposition is their ability to learn specialized convolutional filters tailored to the data, as opposed to wavelet decomposition's generic, non-data-driven approach. This characteristic enables VAEs to maintain a relatively low reconstruction error even at high compression ratios, as illustrated in Section 3.4.

Variational Autoencoders (VAEs) are often employed in anomaly detection tasks. In this study, we investigate the utility of a VAE, originally trained for dimensionality reduction, in anomaly detection without requiring any modifications to its training process. To this end, we train a VAE exclusively on data without any detected diseases, utilizing the same parameters and hyperparameters as in previously trained VAE models. Our anomaly detection strategy is straightforward: we posit that a high reconstruction error indicates an anomaly. To operationalize this, we set a threshold on the reconstruction error. This threshold is determined by maximizing the accuracy on the training set. Following this, we implement a decision rule: a reconstruction error surpassing the set threshold is classified as an anomaly. The results of our approach are presented in confusion matrices in Tables 3.2 and 3.3. Remarkably, due to its adaptability, the VAE outperforms wavelet decomposition in this task

3.5 Other architectures of VAE

The results presented in this paper were obtained using simple variational convolutional autoencoders, where both the encoder and decoder consist of three layers of convolution and deconvolution, respectively. In order to examine the impact of different encoder and decoder architectures on VAE performance, we conducted

experiments with various configurations. This section provides a detailed overview of the experiments performed to investigate the influence of the encoder and decoder architecture in the VAE. The performance analysis of these configurations is summarized in Table 3.1, with all experiments conducted using a latent space of 32 dimensions.

The first type of autoencoder employed in this study is based on Convolutional Neural Networks (CNN), as detailed in Section 3.2. CNN-based autoencoders offer several advantages, including an easy-to-implement symmetric structure and a highly parallelizable architecture that allows for rapid training on GPUs.

Given that our research focuses on time series data, assessing recurrent neural network architectures, such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), is crucial. We tested both LSTM and GRU, observing no significant differences in reconstruction error for this particular dataset. This result is consistent with the findings of Greff et al., 2016, which suggest that LSTM variants perform similarly to the original LSTM, provided they maintain the forget gate and output activation function.

Incorporating LSTM or GRU architectures into the encoder is relatively simple, as we use the values from the last hidden layer of the RNN to feed a neural network that generates the Gaussian's mean and variance. However, the decoder implementation is more intricate due to multiple possible configurations. To reconstruct a multivariate time series, we first obtain a vector with the size of the latent space, sampled from the Gaussian output of the encoder. We then explored various approaches using LSTM.

In all of our approaches, we first employ a single-layer neural network to reshape the input vector to the desired length, which is subsequently used by the LSTM layers. We can either use a single-layer neural network to initialize the "cell states" (commonly denoted as C_0) and "hidden states" (H_0) with the input vector, or initialize them with zeros. Next, we need to generate a reconstruction with the shape $(nb_params, length_ts)$. This can be accomplished by either using an LSTM with the number of features in the hidden state h equal to nb_params, or by employing another single-layer neural network to reshape the LSTM output to the desired size. The method initializing C_0 and H_0 with a simple neural network and using a final neural network to reshape the output to the correct size performed better, so it was selected as the benchmark method.

Additionally, we experimented with a variation of the CNN-based VAE incorporating squeeze-and-excitation (SE) layers after the CNN layers, which we refer to as CNNSE. SE layers are a type of self-attention mechanism specifically designed for

convolutional neural networks (CNNs). They were introduced by Hu et al., 2018, the primary goal of SE layers is to adaptively recalibrate channel-wise feature responses in a CNN by explicitly modeling the interdependencies between channels.

In a traditional CNN, each convolutional filter operates independently, and the output feature map is a combination of these filters' responses. However, not all filters contribute equally to the final representation for every input sample. SE layers address this issue by learning to assign different weights to each filter depending on the input. This enables the model to emphasize the most relevant features and suppress less important ones, resulting in improved performance without significantly increasing the model's complexity. Our experiments demonstrate that this variation performs well for the encoder portion of the VAE.

Furthermore, we tested more sophisticated models inspired by the MLSTM-FCN classifier proposed by Karim et al., 2019, which represents the state of the art for multiple multivariate time series classification datasets. This model combines both LSTM and CNNSE components, with the outputs of the two models concatenated to form the final prediction. For the encoder, the implementation is straightforward; for the decoder, we added a neural network to reshape the input vector for each part of the model. In the CNN component, we employed deconvolution with squeeze-and-excitation. Instead of concatenating the results, we averaged the outputs of the two parts to obtain the final reconstruction. Utilizing this method for both the encoder and the decoder yields the best results for that specific latent space dimension.

Finally, we tested a new model architecture featuring an MLSTMFCN encoder and a CNNSE decoder with various latent space dimensions to compare with the basic CNN VAE tested throughout the paper. These results are illustrated in Figure 3.5 and indicate that employing more complex architectures for the VAE is beneficial at very high compression ratios but may lead to overfitting when using a larger latent space, resulting in compromised performance. Therefore, it is recommended to test different architectures depending on the latent space size and the dataset.

3.6 Conclusion

In this study, we conducted a comprehensive comparison between the data-driven dimension reduction technique, VAE, and the classical wavelet decomposition for analyzing real-world ECG datasets. Our results revealed that the CNN-VAE is a suitable architecture for addressing this specific problem, providing good performances in terms of reconstruction fidelity and in preserving important features of

Encoders \ Decoders	CNN	CNNSE	LSTM	MLSTMFCN	Mean
CNN	0.0131	0.0130	0.0165	0.0125	0.0137
CNNSE	0.0121	0.0137	0.0160	0.0124	0.0136
GRU	0.0123	0.0155	0.0206	<u>0.0113</u>	0.0149
LSTM	0.0127	0.0258	0.0233	0.0172	0.0198
MLSTMFCN	0.0120	0.0117	0.0144	**0.0108**	0.0122
Mean	0.0125	0.0159	0.0181	0.0128	0.0148

Table 3.4: Mean squared error depending on the architectures of the encpders and decoders

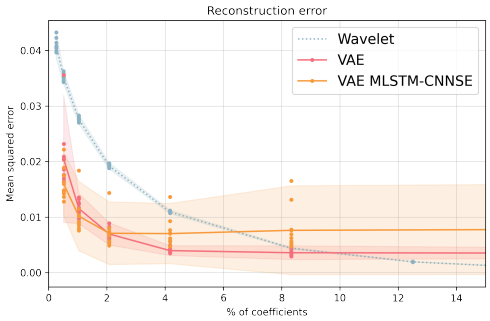


Figure 3.8: figure
Reconstruction error

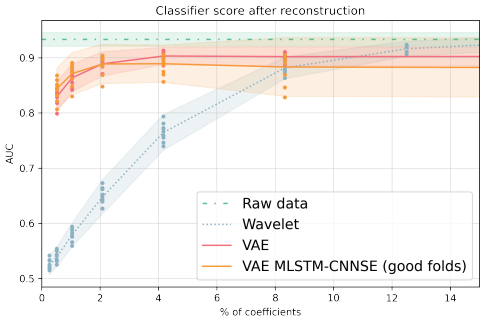


Figure 3.9: figure
Prediction after reconstruction

the time series, particularly when reducing dimensions to smaller spaces. Moreover, we demonstrated the robustness of the CNN-VAE to noise during both the training and inference phases and underscored its potential for anomaly detection in time series data.

Apart from these findings, our research contributes valuable insights to the field of dimensionality reduction in multivariate time series data, thereby paving the way for further investigation. We conducted an exploration of various encoder and decoder architectures, including LSTM and GRU-based models, and examined the impact of different latent space sizes on model performance. This exploration provided a deeper understanding of the factors that influence the efficacy of VAEs in the context of time series data analysis, as well as their potential limitations.

Future work may delve into the disentanglement of variational autoencoders' latent spaces to further improve performance in this domain. Additionally, researchers may consider investigating more sophisticated architectures, such as those that combine CNN, LSTM, and self-attention mechanisms, to achieve better adaptability and reconstruction quality. Exploring hybrid models and new architectures inspired by state-of-the-art classification techniques may also yield improvements in the analysis of time series data.

Chapter 4

Counterfactual explanation for MTS

Building on the insights gathered in the previous chapter, we now move to an even more crucial issue in the realm of multivariate time series data - the challenge of explainability. In this next chapter, we will focus specifically on developing a method for generating counterfactual explanations for this type of data.

As before, we continue with the ECG dataset for several reasons. First and foremost, it provides us with a familiar and well-studied context that allows for a more seamless comparison with other methods. However, another important motivation is the need to evaluate the performance of our newly developed method. To achieve this, we need a dataset that allows the application of a high performance classifier, a condition that is not always feasible in predictive maintenance scenarios.

The upcoming chapter expands and elaborates on a paper we presented at the ICASSP 2023 conference: Counterfactual Explanation for Multivariate Times Series Using A Contrastive Variational Autoencoder Todo et al., 2023. While the conference paper provides a high-level overview, this chapter provides a more comprehensive and detailed examination of the same topic.

Counterfactual explanation for multivariate times series using a contrastive variational autoencoder

William Todo[1,2], **Merwann Selmani**[2], **Béatrice Laurent**[1,3], **Jean-Michel Loubes**[1]
[1] Institut de Mathématiques de Toulouse, Toulouse, France
[2] Liebherr Aerospace Toulouse
[3] INSA de Toulouse, Toulouse, France

Abstract

We tackle the important problem of anomaly detection for multivariate functional data in a supervised setting, which has become increasingly important in medical applications such as electrocardiogram (ECG) analysis. While deep learning has shown great promise in this area, there are few techniques that provide explainability for multivariate time series. In this paper, we propose a novel approach to understand abnormal class features on multivariate time series by dividing the latent space generated by a variational autoencoder (VAE) into general and class-based features using contrastive learning. The resulting Contrastive VAE provides a well-organized latent space that enables us to modify only the class-based features and generate counterfactual examples. Our method is able to produce plausible counterfactual observations that highlight the differences between pathological and non-pathological data. We demonstrate the superiority of our approach over other counterfactual methods through a thorough evaluation that shows significant improvements in both validity and performance.

4.1 Introduction

Anomaly detection in time series data is a fundamental problem in many real-world applications, including healthcare, aerospace, and cybersecurity. Because of the complexity of defining what constitutes an anomaly, standard methods aim to extract key behaviors and understand the observations that deviate from those patterns. In recent years, many research works have focused on building functional features that characterize the normal behavior of time series data. These features can be

extracted using reduction techniques to overcome the high-dimensional aspect of the problem. For instance, we refer to Antoniadis et al., 2013, Jacques and Preda, 2014 and Tarpey and Kinateder, 2003 Barreyre et al., 2019.

Numerous methods for detecting outliers exist, ranging from probabilistic and parametric to non-parametric approaches, as reviewed in Pimentel et al., 2014 and Markou and Singh, 2003. Density-based methods, which rely on the fact that an outlier can be an individual situated in low-density regions, have also been proposed. The One-Class Support Vector Machine (OCSVM) Schölkopf et al., 1999 is a reference method for density level set, while the Local Outlier Factor (LOF) Breunig et al., 2000 was introduced to identify density-based local outliers. However, most existing methods in Explanable AI fail to generate explanations for time-series data, which is often present in critical systems and applications.

This can be a problem when dealing with such applications and can hinder the adoption of machine learning based techniques. One of the main challenges in generating explainable algorithms for time series is that they are high-dimensional objects composed of values observed at different observation times. Hence, using classical algorithms that consider explainability with respect to the input variables fail since they face variables which suffer from the curse of dimensionality and also from the fact that these values are highly dependent variables.

First, time series are very high-dimensional objects composed of values observed at different observation times. Thus, the use of classical algorithms that consider explainability with respect to the input variables fail because they are faced with variables that suffer from the curse of dimensionality and also from the fact that these values are highly dependent variables. Hence, feature based methods have to be considered. The usual features considered for time series are often either a low-dimensional representation such as its projection onto a proper basis (for instance wavelet basis as in Mallat, 1999) or a data driven basis as in Shang, 2014, or features obtained by the embedding using a deep neural network or a variational auto-encoder. Yet explanations that are based on such features carry little *explainability* for the user who only observes the initial time series. Indeed the relationship between the variability of the features and the corresponding variability of the time series is difficult to understand.

For this reason, counterfactual explanations are well-suited to highlight small changes in the time series that lead to a change in the predicted class. Counterfactual explanations can be generated by manipulating the original time series data in a controlled way to generate a new time series that is similar to the original data but belongs to a different class. This approach has the advantage of providing an

explanation that is based on the original data and is therefore easier for the end user to interpret.

Time series data often involve multiple explanatory factors, and disentangling these factors is a critical step in developing accurate models. Recent research has shown that a good representation of time series data should be able to separate these multiple explanatory sources (Bengio et al., 2013; Woo et al., 2022). In this paper, we consider time series that are generated from general explanatory factors labelled as general behaviour $\mathcal{F}_g$ and deviations around this general behaviour depending on the label $\mathcal{F}_s$. The latter are generated with salient features that represent alterations in the explanatory factors of the time series.

Specifically, we study electrocardiogram (ECG) data, where the general behavior is due to common characteristics of ECGs, and the deviations are due to cardiac pathologies. We use the encoder part of a variational autoencoder (VAE), denoted as $q_\theta(Z|X)$, to represent the functional data $\mathcal{F}$ in a low-dimensional latent space $\mathcal{Z}$, which can be manipulated more easily than the original functional data.

The encoder part of the VAE, $q_\theta(Z|X)$, maps the high-dimensional functional data $\mathcal{F}$ to a lower-dimensional latent space $\mathcal{Z}$. This latent representation captures the underlying structure of the data in a more compact and manipulable form. However, in many cases, the latent space may still be entangled and not fully disentangle the explanatory factors of the data. To overcome this, we use a contrastive loss that separates the features of the general shape of the signals and the features resulting from anomalies in the latent space. Note that the supervised contrastive loss can be trained either by using y the label or by using a prediction $\hat{y}$ of a classifier to explain.

In our proposed method, we use the latent space $\mathcal{Z}$ to represent the time series data, where $\mathcal{Z}_g$ captures the general features of the signal, such as the positions of the peaks, heart rate, and the overall shape of the ECG, and $\mathcal{Z}_s$ captures the salient features that differentiate the classes, such as the small alterations characteristic of pathologies. We use a latent prototype from healthy data in the salient space $\mathcal{Z}_s$ to transform the latent representation of the signal, and then use the decoder part of the VAE, $p_\theta(X|Z)$, to construct the counterfactual explanation. The prototype can be seen as a reference point that represents the ideal healthy salient features.

The contributions of this paper are firstly a new method of training VAEs that separates salient features from features shared across all classes, by using a contrastive loss to untangle salient features in the latent space. This partially untangled latent space is used to find latent prototypes that generate counterfactual

examples. These examples show good properties compared to other techniques with an independent classifier.

4.2 Background and related work

Dataset

This article is centered on analyzing multivariate ECG data, which is common in modern ECG datasets collected through the 12-lead procedure. We therefore use the PTB-XL dataset (Wagner et al., 2020), which contains 21,801 12-lead ECG records from 18,869 patients. Each ECG record comprises a 10-second time series of 12 channels (I, II, III, AVL, AVR, AVF, V1, ..., V6), with each channel representing a distinct measure of the heart's electrical activity at a given time, and has been annotated by one or two cardiologists.

These time series are divided in 5 classes : 'CD' for conduction disturbance, "HYP" for hypertrophy, "MI" for myocardial infarction, "NORM" for normal ECG, "STTC" for ST/T change, with normal data representing 42% of the dataset. We are interested in distinguishing between time series with pathology (i.e., those in the CD, HYP, MI, and STTC classes) and those without (i.e., the NORM class). Let K be the dimension of the multivariate time series and $\check{T}$ be the number of timesteps. $\check{x} = \{\check{x}^{t,k}\},\ t \in \check{\mathcal{T}} = \{1, \dots, \check{T}\},\ k \in \mathcal{K} = \{1, \dots, K\}$ denote the K-dimensional time series, $\check{\mathcal{T}}$ is the set of observation times. We note n the size of the dataset and $\{\check{x}_i, \check{y}_i\},\ i \in \{1, \dots, n\}$ a view of the i^{th} element of the dataset with the corresponding label (or class). In the use case, we have $\check{T} = 1000$, $K = 12$. We use data augmentations to generate different views of the time series by cropping them to T-length multivariate time series along with some slight transformations, T is fixed to 256. The whole process is detailed in Section 4.4. Let $x = \{x^{t,k},\ t \in \mathcal{T}_x,\ k \in \mathcal{K}\}$ be that view, and $\mathcal{T}_x$ the timesteps corresponding to the new x.

Variational Auto Encoders

Variational Autoencoder (VAE) is a powerful deep learning technique that has attracted considerable attention in the fields of machine learning and computer vision. VAEs are generative models that use a neural network architecture to learn the underlying probability distribution of a given data set. The main idea behind VAE is to find a latent representation of the data that captures the important features of the data, while at the same time allowing us to sample from this latent space to generate new data points that are similar to the original data set.

VAE was introduced by Kingma and Welling, 2013 and Rezende et al., 2014, they are generative models composed of an encoder $q_\Phi(Z|X=x)$ which takes the entry x from a high-dimensional space and maps it to a J-dimensional Gaussian with mean vector $\mu(x)$ and diagonal covariance matrix $\text{diag}(\sigma^2(x))$, $q_\Phi(Z|X=x) \sim \mathcal{N}_J\left(\mu(x), \text{diag}\left(\sigma^2(x)\right)\right)$. The decoder $p_\theta(X|Z=z)$ takes a sample z of that distribution as input and generates the associated element in the starting high-dimensional space, we note this reconstruction $\hat{x}$.

The VAE is trained by optimizing the evidence lower bound (ELBO), which is a lower bound on the log-likelihood of the data. The ELBO consists of two terms: the reconstruction loss and the KL-divergence regularization term. The reconstruction loss measures the difference between the original input data point x and its reconstructed version $\hat{x}$. The KL-divergence regularization term encourages the learned latent space to be close to a standard normal distribution.

$$\begin{aligned}\mathcal{L}_{\text{VAE}}(\Phi, \theta; x) = &- \mathbb{E}_{Z\sim q_\Phi(Z|X=x)}\left[\log\, p_\theta(X|Z)\right] \\ &+ D_{KL}[q_\Phi(Z|X=x)||p(Z)]\end{aligned} \tag{4.1}$$

The conditional distribution, $p_\theta(X|Z)$, is specified by the decoder as $\mathcal{N}(\mu = \hat{x}, \Sigma)$. When optimizing the negative log-likelihood expectation, $-\mathbb{E}_{Z\sim q_\Phi(Z|X=x)}\left[\log\, p_\theta(X|Z)\right]$, it turns out to be equivalent to minimizing the mean squared error (MSE) between the input data x and the reconstructed data $\hat{x}$. As a result, it is reasonable to express the first term in Eq.(4.1) as the MSE within the VAE framework.

$$||\hat{x} - x||^2 = \frac{1}{T \times K} \sum_{t\in\mathcal{T}_x,\, k\in\mathcal{K}} \left(\hat{x}^{t,k} - x^{t,k}\right)^2 \tag{4.2}$$

The second term in Eq.(4.1) is the Kullback-Leibler divergence between $q_\Phi(Z|X=x)$ and $p(Z)$. Since the distributions are respectively $\mathcal{N}_J\left(\mu(x),\ \text{diag}\left(\sigma^2(x)\right)\right)$ and $\mathcal{N}_J(0,\ Id_J)$, we can easily compute this term. Namely, we have :

$$\begin{aligned}D_{\text{KL}}[q_\Phi(Z|X=x)||p(Z)] &= \mathbb{E}_{Z\sim q_\Phi(Z|X=x)} \log\left(\frac{q_\Phi(Z|X=x)}{p(Z)}\right) \\ &= \frac{1}{2}\sum_{j=1}^{J}\left[-1 - \log\left(\sigma_j^2(x)\right) + \mu_j^2(x) + \sigma_j^2(x)\right]\end{aligned} \tag{4.3}$$

Although VAEs are typically trained in an unsupervised way, it is often possible to identify clusters in the latent space that are correlated with the input data labels, especially on well-known datasets such as MNIST or FashionMNIST. This is typically observed when differences between data points are largely driven by differences in the input labels. In the case of ECG data, variations in the data

are mainly caused by general features such as frequency, amplitude, and position. However, pathologies induce local variations around the general shape of the ECG, which have a minimal effect on the reconstruction loss and thus the shape of the latent space. To ensure that the VAE can effectively capture these salient features, it is necessary to incorporate additional mechanisms that encourage the encoder to encode and preserve these features.

Contrastive representation learning

Semi-supervised learning has proven to be an important part of training large image classifiers. In this vein, Chen, Kornblith, Norouzi, et al., 2020, inspired by Berthelot et al., 2019, propose a new method for training deep neural networks to learn useful visual representations in an unsupervised manner, which is called contrastive learning.

Contrastive learning is a method of training neural networks in which the network is trained to learn representations that bring similar inputs closer together and dissimilar inputs further apart in the representational space. The idea is to train the network to identify which images are "positive" (i.e., similar) and which images are "negative" (i.e., dissimilar), based on a set of "anchor" images. The network takes pairs of images as input, and each image is passed through the network to obtain two feature vectors. These feature vectors are then compared using a contrastive loss function that encourages similar images to have feature vectors that are close together and dissimilar images to have feature vectors that are far apart. The supervised version of this technique is proposed in Khosla et al., 2020.

We take a set of N randomly sampled sample/label pairs, $\{\check{x}_l, \check{y}_l\},\ l = 1, \ldots, N$. The training batch consists of $2N$ pairs $\{x_i, y_i\},\ i = 1, \ldots, 2N$ where x_{2i} and x_{2i-1} are 2 different views (random data augmentations, see Section 4.4) of $\check{x}_l$ and $\check{y}_l = y_{2i} = y_{2i-1}$ is the corresponding label. Let $I = \{1, \ldots, 2N\}$ be the set of indices of the different views. $A(i) = I \setminus i$ are all the indices of the batch except i, $P(i) = \{p \in A(i) : y_p = y_i\}$ is the set of indices of all positives in the batch, distinct from i and τ a positive parameter, we follow the original implementation to set it. Note that using 2 views of each time series ensure that $|P(i)| > 0,\ \forall i \in I$.

Let $j(i)$ be the index of the other augmented sample from the same source sample.

We note $z_i = \text{Enc}(x_i)$ the projection of x_i into a low-dimensional space using an encoder network Enc(.) and $< ., . >$ the usual scalar product in $\mathbb{R}^J$.

In Chen, Kornblith, Norouzi, et al., 2020, the unsupervised contrastive loss takes this form.

$$L_{\text{Con}}(\{z_i\}_{i\in I}) = -\sum_{i\in I} \log \frac{\exp\left(< z_i, z_{j(i)} > /\tau\right)}{\sum_{a\in A(i)} \exp\left(< z_i, z_a > /\tau\right)} \tag{4.4}$$

Here, we encounter a potential issue highlighted by Khosla et al., 2020. If two elements within the same batch belong to the same class, the unsupervised contrastive loss will not pull them closer together; instead, it may drive them apart. This observation is what led to the development of the supervised version of the contrastive loss, which addresses this limitation by incorporating label information. The supervised contrastive loss is formulated as follows:

$$L_{\text{supCon}}(\{z_i\}_{i\in I}) = \sum_{i\in I} \frac{-1}{|P(i)|} \sum_{p\in P(i)} \log \frac{\exp\left(< z_i, z_p > /\tau\right)}{\sum_{a\in A(i)} \exp\left(< z_i, z_a > /\tau\right)}. \tag{4.5}$$

In this case, incorporating the positive information $P(i)$ from the batch is expected to facilitate the formation of even more refined and meaningful representations, as it takes into account the similarity of data points belonging to the same class.

Counterfactual explanations

Counterfactual explanations are a way of explaining why a particular model makes certain decisions by generating hypothetical examples of what might have been different, answering the question of what it would take to change the classification of that particular input. In other words, counterfactual explanations help to identify the characteristics of an input that contributed most to the output. In Karimi et al., 2020, they point out that this type of explanation is easy for an individual to understand. By providing a clear and interpretable explanation of why a particular decision was made, counterfactual explanations can help improve the transparency and accountability of AI systems. This is particularly important in situations where the consequences of a decision can have significant real-world impacts, such as in aerospace, criminal justice, or in our use case: healthcare. There are many techniques developed for this purpose, and we will present some of them in this section.

Delaney et al., 2021 propose a novel model agnostic technique that generates counterfactual explanations for time series classifiers. The method is a case-based technique that retrieves existing instances from the training data and adapts them to generate counterfactual explanations. Parts of the signal are replaced by the signal of the target class using class activation mapping (CAM) (Zhou et al., 2016) to select the parts to be replaced.

Although this method gives impressive results, it requires the time series to be aligned, which is not usually the case with multivariate series. Moreover, using CAM on multivariate time series is not a straightforward process, as it requires a specific architecture, as shown in Assaf et al., 2019 and Fauvel et al., 2021.

In Ates et al., 2021 (CoMTE), a method for generating counterfactual explanations for multivariate time series, a nearby multivariate time series in the training dataset with the target label is selected using a KD-Tree in a feature space. This nearby time series, called a distractor, is then used to substitute parts of the signal into the original time series to generate a counterfactual explanation. Two algorithms are used to make substitutions in the time series. The Sequential Greedy Algorithm replaces each variable in the time series with the corresponding variable in the distractor time series that results in the largest increase in the probability predicted by the classifier. This process is repeated iteratively until the predicted probability exceeds a given threshold. The Random-Restart Hill Climbing algorithm is also applied to minimise the loss function, which involves adding or removing a variable to the set of substituted variables. This algorithm starts with a random initialization point for the set of variables and evaluates the loss function for random neighbours until it finds a better neighbour. However, hill climbing can settle into local minima, so random restart is used to explore more of the search space. In the rare cases where all variables can be pruned, the greedy search algorithm is used to find a viable solution.

Replacing parts of the signal in a multivariate framework would affect the inter-signal consistency, which is not desirable, as there are often significant correlations between the signals.

In Bahri et al., 2022, they propose a model-agnostic, instance-based explanation algorithm called Shapelet Explainer for Time Series (SETS). SETS uses a dictionary of shapelets extracted from the dataset to represent the time series data in an efficient and meaningful way. The algorithm generates counterfactual instances by replacing segments of the input signal with shapelets representing the target class. The process involves computing the distances between the shapelets and the dataset instances, selecting class-specific shapelets, and calculating the occurrence distributions. Then, for a given instance of class A, the algorithm generates a counterfactual instance of class B by replacing the shapelets of class A with the closest shapelets of class B according to the occurrence distributions. The resulting counterfactual instance is visually interpretable and can help explain the decision process of black box models.

The approach presented in Wachter et al., 2017 proposes constructing an objective function that modifies the classifier's output by making small changes to the original data using an optimization algorithm. The suggested loss function is defined as follows:

$$L(x, x', y') = (f(x') - y')^2 + d(x, x')$$

where $f()$ is the trained classifier, x is the original data point, x' is the counterfactual, y' is the target class, and $d(.,.)$ is a distance metric such as the Manhattan distance used to keep x and x' close to each other. Specialized techniques for time series have been developed in Dhurandhar et al., 2018 and Karlsson et al., 2018. However, these methods have two main drawbacks. Firstly, the optimization process can be costly and time-consuming. Secondly, the quality of the counterfactual is dependent on the quality of the classifier f, which can limit the method's practical applicability.

The counterfactual generation method proposed in Balasubramanian et al., 2020 works by searching for counterfactuals in the latent space of an autoencoder (AE). An autoencoder is a neural network architecture that learns a compressed representation (latent space) of an input data point. When generating counterfactuals, the autoencoder is trained on the original data set and learns to map each data point to a compressed representation in the latent space. To generate a counterfactual for a given data point x, the method first encodes the data point into its latent representation z using the trained autoencoder. It then searches near z to find a new latent representation z' that is close to z and produces a different classification output for the classifier. This search is done using gradient descent optimization on a loss function that aims to minimize the distance between z and z' while maximizing the difference in classifier outputs between x and the counterfactual x'. Once a suitable z' is found, it is decoded back into the original data space using the decoder part of the autoencoder, producing the counterfactual x'. Adding more constraints to the optimization problem and using Adam optimizer showed good results in Z. Wang et al., 2021.

These methods can produce plausible counterfactuals if the model is well trained, but they do not provide good explanations if the latent space is entangled. Unfortunately, the latent space produced by VAE and AE models on the ECG dataset does not separate the classes, and therefore these techniques do not work well (see table 4.4).

Desirable counterfactual properties are discussed in the paper Verma et al., 2020. Indeed, not all counterfactuals are created equal. To be truly useful and informative, a good counterfactual explanation should satisfy certain desirable

properties. The authors outline the properties that a good counterfactual should verify and explain why these properties are important. In this article, we will focus on some of the most important properties highlighted in the review and explore why they are crucial for creating effective and trustworthy counterfactual explanations.

One key aspect of this evaluation is *validity*, which measures the extent to which the generated counterfactuals actually lead to a change in the predicted class of the signal relative to the target. This can be quantified as the percentage of counterfactuals that cause an independent classifier f to change its predictions from y to y_{target}. A higher validity score indicates that the generated counterfactuals are more likely to be useful and informative for understanding the model, while a lower score suggests that the method may be struggling to generate effective counterfactuals. Thus, validity is an important metric for evaluating the overall quality and utility of a counterfactual generation method.

Another important metric for evaluating counterfactual explanations is *sparsity*, which measures the extent to which the generated counterfactuals change the smallest number of features necessary to achieve the desired outcome. For tabular data, sparsity is easily computed as the number of features changed in the counterfactual, while for time series data it is more complex. In our case, we use a parameter ϵ to control the possible deviation between the original time series and the counterfactual, allowing a deviation of up to 25% of the variance.

$$\text{Sparsity}(x, x_{\text{cf}}, \epsilon) = \frac{1}{T \times K} \sum_{t \in \mathcal{T}_x,\, k \in \mathcal{K}} \left| x^{t,k} - x_{\text{cf}}^{t,k} \right| < \epsilon \tag{4.6}$$

A sparser explanation should make it easier to understand the underlying differences between the classes.

The *plausibility* refers to the ability of the generated counterfactuals to resemble realistic, plausible data points that could have been observed in the original dataset. In our case, since we use the decoder part of a Variational Autoencoder (VAE) to construct the counterfactuals, this property is automatically verified by design. A well-trained VAE will produce plausible data points that are consistent with the underlying structure and patterns of the original dataset. By ensuring plausibility, we can increase the reliability and usefulness of the generated counterfactuals and gain a better understanding of the underlying factors and drivers that influence the model's predictions.

Scalability refers to the ability of the model to provide counterfactuals for large datasets in a timely and efficient manner. This is particularly important in real-world settings where large datasets are common and timely decision making is critical. For most counterfactual generation techniques, we need to consider both training time and inference time, as these can have a significant impact on the overall scalability and feasibility of the approach. While a longer training time may be acceptable if it leads to better performance and more accurate counterfactuals, it is important to ensure that the inference time remains reasonable and feasible for real-world applications.

4.3 Method

Contrastive variational autoencoder

We want a VAE where the latent space is separated into two distinct parts, one part coding for phenomena related to normal behaviour $\mathcal{Z}_g$ and another part coding for variations due to a change in behaviour $\mathcal{Z}_s$ (in our case pathologies). Unlike previous work (Cai et al., 2019, Zheng and Sun, 2019, Poels and Menkovski, 2022) we do not want to use an additional classifier, adversarial training or multiple encoders / decoders.

The representation of time series data in a latent space is a powerful approach that can reveal the underlying structure and relationships within the data. In particular, it is important to capture the disentangled sources of variation contributing to the data, as this can enable more effective downstream analysis and modeling. To this end, we aim to construct a latent space that reflects both the general and salient sources of the data. The general sources are the common features shared by all signals, such as the overall shape or rhythm, while the salient sources represent the specific patterns or variations that distinguish a healthy signal from a pathological one.

To achieve this disentanglement, we partition the latent space into two subspaces, $\mathcal{Z}_g$ and $\mathcal{Z}_s$, representing the general and salient features, respectively. Specifically, we define the dimensions of the general part of the latent space, J_g, and the salient part of the latent space, J_s, where $J = J_g + J_s$, and we use the sets of indices $\mathcal{J}_g = 1, \ldots, J_g$ and $\mathcal{J}_s = J_g + 1, \ldots, J$ to differentiate between the two subspaces. Within this framework, our goal is to ensure that signals with similar general features are close together in $\mathcal{Z}_g$, while those with different salient features are far apart in

$\mathcal{Z}_s$. This allows us to more effectively group signals that share common properties, while also emphasizing the differences that are most relevant for downstream tasks.

In order to achieve this goal, we make use of a contrastive loss that is specifically focused on the salient part of the latent space, $\mathcal{Z}_s$. This loss can be unsupervised, as in Eq.(4.4), or supervised, as in Eq.(4.5), depending on whether we have access to labeled data. By applying this loss to the salient features, we can better capture the differences between signals belonging to different classes or categories, which in turn can improve the discriminative power of the representation. Overall, this approach provides a powerful framework for representing time series data in a way that captures both the general and salient features, and enables effective downstream analysis and modeling. Through the use of contrastive loss on the salient features, we can more effectively capture the differences between signals and create a more powerful representation for a variety of tasks.

We sample z^x from the Gaussian output of the encoder $q_\Phi(Z|X=x) \sim \mathcal{N}_J\left(\mu(x),\ \text{diag}\left(\sigma^2(x)\right)\right)$, we apply the contrastive loss only on the salient dimensions $\{z_j^x,\ j \in \mathcal{J}_s\}$. This part of the latent space will be forced to disentangle classes and thus, encode class-specific features. The choice of the dimension of the salient space is not straightforward. This is why the experiments show a variable size of salient dimensions. The strategy chosen to select this dimension is discussed in Section 4.4. We finally define our contrastive variational loss two ways.

Initially, we define the unsupervised contrastive variational loss as follows:

$$
\begin{aligned}
L_{\text{CtrVAE}}(\Phi, \theta; \{x_i\}_{i\in I}) &= \frac{1}{|I|}\sum_{i\in I} ||\hat{x}_i - x_i||^2 + \frac{1}{|I|}\sum_{i\in I} D_{\text{KL}}[q_\Phi(Z|X=x_i)||p(Z)] \\
&+ L_{\text{Con}}(\{z^{x_i}\}_{i\in I}) \\
&= \frac{1}{|I|}\sum_{i\in I}\left(\frac{1}{T\times K}\sum_{t\in\mathcal{T}_x,\, k\in\mathcal{K}}\left(\hat{x}_i^{t,k} - x_i^{t,k}\right)^2\right) \\
&+ \frac{1}{|I|}\sum_{i\in I}\left(\frac{1}{2}\sum_{j=1}^{J}\left[-1 - \log\left(\sigma_j^2(x_i)\right) + \mu_j^2(x_i) + \sigma_j^2(x_i)\right]\right) \\
&- \sum_{i\in I}\log\frac{\exp\left(\sum_{j\in\mathcal{J}_s}\left(z_j^{x_i}.z_j^{x_{j(i)}}\right)/\tau\right)}{\sum_{a\in A(i)}\exp\left(\sum_{j\in\mathcal{J}_s}\left(z_j^{x_i}.z_j^{x_a}\right)/\tau\right)}
\end{aligned}
$$

Subsequently, we introduce the supervised variant of the contrastive variational loss, which incorporates class information:

$$
\begin{aligned}
L_{\text{SupCtrVAE}}(\Phi, \theta; \{x_i\}_{i\in I}) &= \frac{1}{|I|}\sum_{i\in I} ||\hat{x}_i - x_i||^2 + \frac{1}{|I|}\sum_{i\in I} D_{\text{KL}}[q_\Phi(Z|X = x_i)||p(Z)] \\
&\quad + L_{\text{supCon}}(\{z^{x_i}\}_{i\in I}) \\
&= \frac{1}{|I|}\sum_{i\in I}\left(\frac{1}{T\times K}\sum_{t\in\mathcal{T}_x,\, k\in\mathcal{K}}\left(\hat{x}_i^{t,k} - x_i^{t,k}\right)^2\right) \\
&\quad + \frac{1}{|I|}\sum_{i\in I}\left(\frac{1}{2}\sum_{j=1}^{J}\left[-1 - \log\left(\sigma_j^2(x_i)\right) + \mu_j^2(x_i) + \sigma_j^2(x_i)\right]\right) \\
&\quad + \sum_{i\in I}\frac{-1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp\left(\sum_{j\in\mathcal{J}_s}\left(z_j^{x_i}.z_j^{x_p}\right)/\tau\right)}{\sum_{a\in A(i)}\exp\left(\sum_{j\in\mathcal{J}_s}\left(z_j^{x_i}.z_j^{x_a}\right)/\tau\right)}
\end{aligned}
$$

These two formulations allow us to compare the effectiveness of contrastive learning in generating meaningful latent representations for time series data.

One way to look at the partially constrained latent space is to think of an encoder/decoder pair as being able to compress only a certain amount of information. During training, salient features are represented in the part of the latent space constrained by the contrastive loss to help form clusters and thus minimize the loss. Non-salient features should remain in the general latent space to avoid cluttering the salient space.

To better understand and visualize the structure of this latent space, we use a dimensionality reduction technique called UMAP introduced by McInnes et al., 2018. UMAP allows us to project the high-dimensional latent space onto a 2D plane, which we can then visualize and analyze. Fig. 4.1 shows the UMAP projections of the latent space for two different encodings: one for general features $\mathcal{Z}_g$, and one for pathological features $\mathcal{Z}_s$. In the latent space encoding for general features $\mathcal{Z}_g$ (located at the top of Fig. 4.1), we observe that the data are uniformly distributed across the 2D space. This indicates that there is no clear clustering or separation between the different classes of data. On the other hand, in the latent space encoding for pathological features $\mathcal{Z}_s$, we see a clear separation between normal data and anomalies. This means that the encoding has successfully captured the key features that distinguish between normal and anomalous data, and has represented these features in a way that allows for easy differentiation and classification.

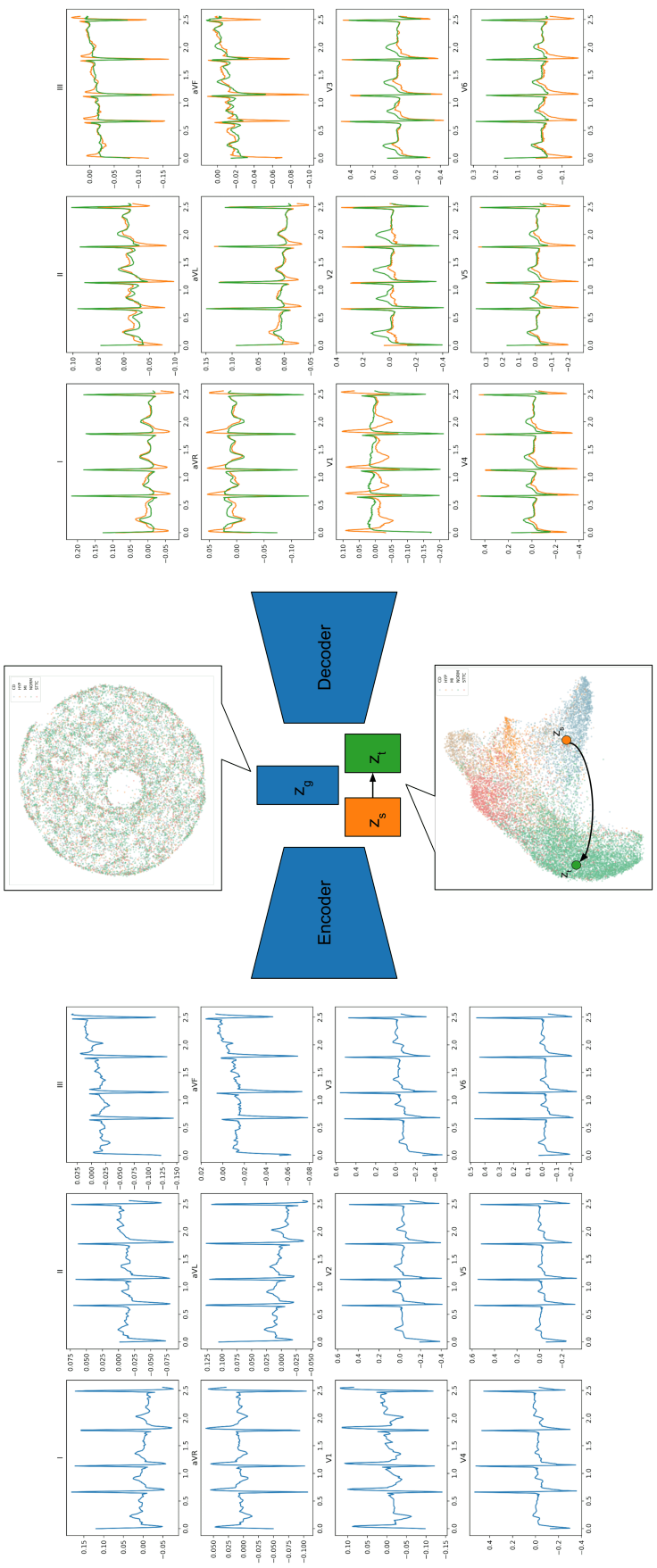


Figure 4.1: Framework of the contrastive VAE, each color in the latent space corresponds to a pathological class. The blue time series on the left is the raw data input to the model, the orange and green signals on the right represent the reconstructed signal and the counterfactual respectively.

Counterfactual explanation

The contrastive variational autoencoder is a method for disentangling the latent space of data into general and salient features. This disentanglement enables us to selectively modify the salient features in order to produce counterfactual explanations for pathological data. Specifically, we aim to generate the time series that a pathological instance would have exhibited if it had been healthy. To achieve this, we employ a "healthy latent prototype" z_t, which is representative of healthy data in the salient space $\mathcal{Z}_s$. By replacing the salient features of a pathological instance with those of z_t, we obtain a counterfactual instance that is expected to be healthy.

Our goal is to change only the salient part of the signal and thus to change only the salient part of the latent space. We use a "healthy latent prototype" to make this change, this prototype is a representative of the healthy data in $\mathcal{Z}_s$ and since this is our target, we call this prototype z_t. The transformation is done by replacing z_s by z_t.

Given a time series x, we have $z(x) = \{z_g(x), z_s(x)\}$ its latent representation and $z_{\text{cf}}(x) = \{z_g(x), z_t\}$ the latent counterfactual, let $x_{\text{cf}} = p_\theta\left(X|Z = z_{\text{cf}}(x)\right)$ be the counterfactual explanation for x.

To determine the best method for selecting a prototype, we evaluated four approaches. The first method uses the mode of the target class distribution, which is estimated by applying kernel density estimation to the target class in the training set of $\mathcal{Z}_s$. The second method selects the median of the target class, while the third method chooses the mean. The fourth method, called the sub-method, applies kernel density estimation to both the target and other classes and selects the point that maximizes the difference between the two estimated densities.

After evaluating the four methods, we found that the sub-method had slightly better validity performance than the other three methods. This is because it considers the density of both the target and other classes, which provides a more comprehensive understanding of the overall distribution.

Although the clusters produced by the different methods gave similar positions for z_t, the sub-method was more robust in terms of prototype selection. The counterfactuals generated by the three methods were similar because the salient part of the latent space produced good class-related clusters. Overall, the slightly superior performance and robustness of the sub-method makes it the preferred approach for selecting prototypes. These results are summarized in Table 4.1.

Table 4.1: Prototype comparison

		CVAE $(16, 4)$	CVAE $(32, 8)$	CVAE $(64, 32)$
Validity	Mode	0.984 ±0.006	0.938 ±0.016	0.895 ±0.008
	Median	0.982 ±0.007	0.933 ±0.017	0.895 ±0.007
	Sub	**0.985** ±0.006	**0.945** ±0.014	**0.908** ±0.009
Sparsity	Mode	0.514 ±0.016	0.579 ±0.018	**0.591** ±0.019
	Median	**0.519** ±0.014	**0.581** ±0.018	**0.591** ±0.019
	Sub	0.512 ±0.019	0.576 ±0.019	0.586 ±0.020

When using the unsupervised contrastive loss function in CVAE, it is still necessary to define a healthy latent prototype for generating counterfactual examples. This requires access to labeled healthy data to estimate $\mathcal{Z}_t$. However, the well-disentangled salient latent space of the CVAE ensures that the estimated $\mathcal{Z}_t$ produces high quality counterfactuals, as demonstrated by our experimental results, see Table (4.1).

In Fig. 4.1, we demonstrate an example of the transformation process using our proposed method, and the resulting counterfactual closely resembles the reconstructed signal. In particular, we observe that the cavities of the original signal are reduced after each "peak" in the counterfactual examples generated for *I*, *aVL*, and *V1*. This observation suggests that these types of cavities are representative of the pathological class and highlights the relevance of the features highlighted by our method. To further confirm the plausibility of our counterfactuals, we consulted a cardiologist who confirmed the significance of the features highlighted by our method. This validation supports the effectiveness of our approach in generating accurate counterfactuals for medical diagnosis.

One of the major advantages of our method is the speed with which we can generate counterfactuals using prototypical salient targets. Unlike optimization-based methods, our approach requires only the inference time of the VAE. This is a significant advantage, as optimization-based methods can require a significant amount of computation time to generate a counterfactual, which can be impractical in many real-world scenarios. Overall, the use of prototypical salient targets allows us to provide fast and accurate counterfactual explanations, which could have important implications for medical diagnosis and treatment planning.

4.4 Experiments

To mitigate the randomness in the formation of different architectures, we performed multiple training runs with varying random seeds. Confidence intervals of 2σ are

represented in the figures by colors around the mean.

Training details

The training was conducted on one Nvidia RTX A6000, and we trained a total of 270 models for the experiments, with a total training time of approximately 300 hours. We estimated total emissions to be 5.22 kgCO_2eq using the Lacoste et al., 2019 method and the electricitymap.org website.

To achieve our goal of developing a new way of generating counterfactual explanations, we used a standard architecture for the encoder and a reversed version for the decoder. The architecture consisted of three convolutional blocks with filters of size 5 and 256, 512, and 512 neurons, respectively. During training, we used a batch size of 256 and added coefficients to the KL and contrastive losses to balance them with the reconstruction loss.

We use the PyTorch implementation of the supervised contrastive loss from Tian, 2020 and used it as the basis for both the supervised and unsupervised contrastive losses. We employed a learning rate that reduces on plateau with a patience of 45 epochs and early stopping after 100 epochs without improvement. We also used the Adam optimizer to optimize the model parameters.

In addition, we used 20% dropout on the convolutional blocks to reduce overfitting during training. We also included leaky ReLU activation layers with a negative slope of 0.01 and batch normalization to improve the stability of the training process. Overall, our training process was designed to optimize the performance of our deep learning model for generating counterfactual explanations. The use of multiple training runs with varying seeds helped to mitigate the impact of randomness, while the addition of coefficients to the loss functions helped to balance their contributions to the overall objective.

Data augmentation

Data augmentation is an essential component of training deep learning models to improve their ability to generalize to new data and achieve good performance. While there are many techniques for image models to add diversity to the dataset, such as adding different types of noise, flipping and rotating images, and more advanced techniques such as Mixup and CutMix (Yun et al., 2019, H. Zhang et al., 2017), it can be more challenging to apply such transformations to multivariate time series data, as it may not always make sense to modify the time series in this way, and it could therefore harm the model. In our case, we did not add noise or other complex time series-specific augmentations, such as time wrapping, for two reasons. First,

because they must be done on the fly during training, these augmentations can be quite time-consuming. In addition, some of these techniques may not make sense for the type of data we are using. For example, heart rate is mostly stable during recording, so techniques like time wrapping would not be appropriate.

Instead, our data augmentation strategy focuses mainly on random cropping, which is widely used in time series data whenever possible, and is notably used in the benchmark of the Strodthoff et al., 2020 dataset. We also employ an augmentation technique that extends or condenses the time series time-wise, using a random length ($\mathcal{U}(256 - p_{\text{scale}}, 256 + p_{\text{scale}})$, $p_{\text{scale}} = \lfloor 0.3 \times 256 \rfloor$) crop of the time series and then linearly interpolating to make it 256-length.. This artificially alters the cardiac rhythm of the samples and is similar to the randomly resized cropping technique used in image data. Overall, our data augmentation strategy is designed to increase the diversity of our training data and improve the generalization performance of our deep learning model. While we did not use more complex augmentations specialized for time series, our approach still provides a simple and effective way to augment the data and train our model to better detect and respond to different cardiac rhythms and pathologies.

Contrastive VAE

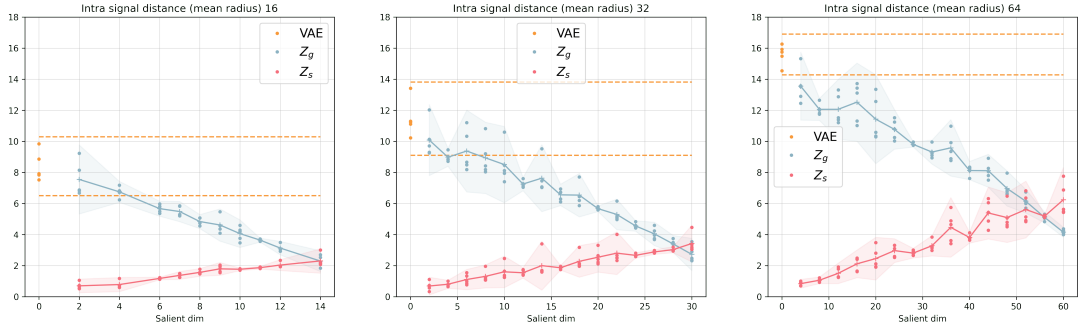


Figure 4.2: Mean dispersion radius. The mean dispersion radius over the latent space of the classical VAE is shown for comparison, the dotted lines represent $\pm 2\sigma$.

To evaluate the performance of our proposed contrastive VAE, we conducted several experiments to assess the quality of the disentangled representations learned by the model, as well as its ability to generate plausible data from the decoder. Our results demonstrate the ability of the CVAE model to produce high quality synthetic data from a well disentangled latent space.

A key metric for evaluating the disentanglement performance of our CVAE model is the average radius of dispersion in the two parts of the latent space ($\mathcal{Z}_g$ and $\mathcal{Z}_s$) for different views of the same time series. For each data sample, we take seven different views (crops) and compute the dispersion on the two parts of the latent space. Since each cropped signal is generated by the same underlying phenomenon (i.e., pathology), it should appear close together in $\mathcal{Z}_s$. As shown in Figure 4.2, the dispersion in $\mathcal{Z}_s$ is small, indicating that this part of the latent space is highly dependent on the pathology rather than the general shape of the signal.

This is a promising result that confirms the effectiveness of our disentanglement approach, which separates the signal's pathology from other factors that may influence its shape. Another important measure of the model's performance is its reconstruction capability, which determines how accurately the decoder can generate plausible data from latent space. We have compared the mean square error between the input signal and the reconstructed signal for both the traditional VAE and CVAE models in Figure 4.3. Our results indicate that the addition of the contrastive loss does not degrade the reconstruction compared to the traditional VAE, and may even slightly improve the quality when using a sufficiently large latent space.

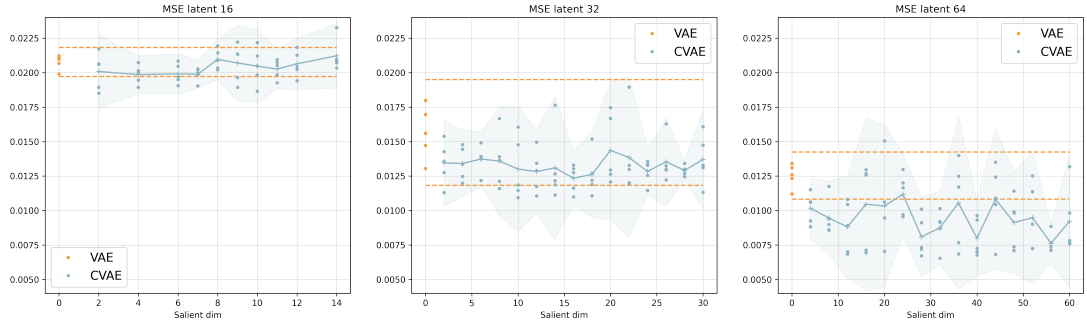


Figure 4.3: Reconstruction error, the dotted line represents the $\pm 2\sigma$ intervals of classical VAE

Counterfactual explanations

To thoroughly evaluate the validity and effectiveness of our counterfactual explanations, we used an independent classifier trained according to the recommended practices of the Strodthoff et al., 2020 benchmark. Our classifier, a convolutional neural network, achieved an overall AUC of 0.928 on the test set when its predictions were averaged over the entire signal. This high level of performance is similar to

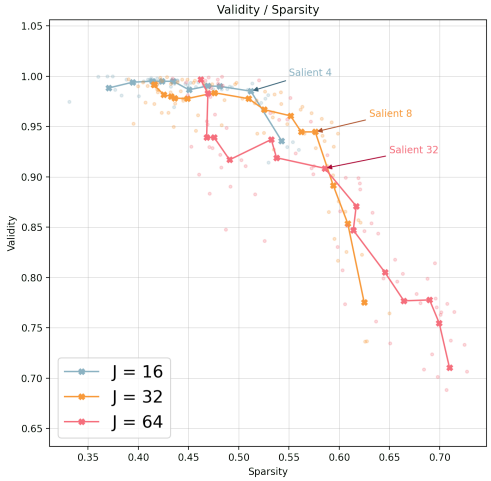


Figure 4.4: Validity as a function of sparsity, the higher the better. The arrows indicate the architectures that are a good compromise.

that of the benchmark, indicating that our classifier is a trustworthy and reliable tool for assessing the accuracy of our counterfactual explanations.

It is important to note, however, that our classifier is completely independent of the VAE and counterfactual generation process. This is a critical consideration because many methods that generate optimization-based counterfactuals rely on the same classifier to assess the validity of the explanations as to construct them, which can lead to confusion and misinterpretation. By using an independent classifier, we eliminate this potential problem and ensure that our counterfactuals are both credible and reliable. This added level of assurance is critical when using counterfactuals in high-stakes real-world applications, such as medical diagnosis or treatment planning.

To strike a balance between validity and sparsity in the generated counterfactuals, we must carefully select the salient dimensions of our contrastive VAE model. A smaller number of salient dimensions will result in smaller changes in the signal, leading to better sparsity, but it may also make it more challenging to change classes. Thus, we aim to choose the smallest number of salient dimensions that still produce acceptable validity scores. To aid in this decision-making process, we plot validity as a function of sparsity in Figure 4.4. This plot is also useful for comparing the effects of different latent space dimensions on the performance of the proposed method. As we can see from the figure, when the latent space is relatively large (e.g., 128 dimensions), the generation of counterfactual examples becomes more difficult.

Furthermore, we have achieved high validity scores without significantly altering the original signal, as confirmed by an independent classifier. Comparing the curves, we observe that the contrastive VAE model with a latent dimension of 16 can generate better counterfactual examples. However, it is crucial to note that the reconstruction error is also a crucial metric for evaluating the quality of the CVAE model. Based on our evaluation, we conclude that the preferred model has a latent dimension of 32 and uses 8 salient variables, for the model with a latent dimension of 64 we keep the one with 32 salient dimensions, and for the model with a latent space of 16 we keep the one with 4 salient dimensions.

Overall, the choice of salient dimensions is a crucial aspect of our model and affects the trade-off between validity and sparsity of the generated counterfactuals. Our approach leverages the validity-sparsity plot to determine the optimal number of salient dimensions, and our results show that our contrastive VAE model can effectively generate counterfactuals that are both valid and sparse.

The results of our experiments highlight the importance of partitioning the latent space of the CVAE into two distinct subspaces, $\mathcal{Z}_g$ and $\mathcal{Z}_s$. The contrastive loss function can be applied to the entire latent space to produce high validity scores, but at the cost of generating counterfactual examples with low sparsity and substantial deviation from the original sample. In contrast, applying the contrastive loss only to the subspace $\mathcal{Z}_s$, which captures the salient factors of the input data, results in more sparse counterfactual examples that preserve the essential features of the original sample. This approach provides a better balance between validity and sparsity, and generates counterfactuals that are more interpretable and actionable for end users. Our experiments show that this partitioning of latent space into salient and non-salient dimensions leads to better performance in terms of both sparsity and validity, and provides a valuable tool for generating counterfactuals that are reliable and accurate.

Another way to adapt the sparsity score to user needs is to use linear interpolation between z_s and z_t. This method involves replacing the prototype z_t with $z\text{proj} = z_s + \lambda(z_t - z_s), ; \lambda \in [0, 1]$. By varying the value of λ, we can adjust the tradeoff between sparsity and validity in the generated counterfactuals. As shown in Figure 4.5, small values of λ can increase sparsity while sacrificing some degree of validity, while larger values of λ result in more valid counterfactuals but with less sparsity. It is important to note that this parameter is optional and should be used at the discretion of the user based on their specific needs and preferences.These results demonstrate the trade-off between sparsity and validity in our counterfactual generation process and highlight the flexibility of our approach in meeting the needs of different users.

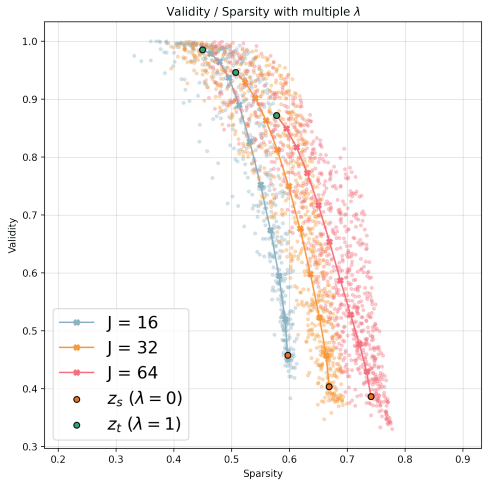


Figure 4.5: Evolution of sparsity and validity given different parameters λ

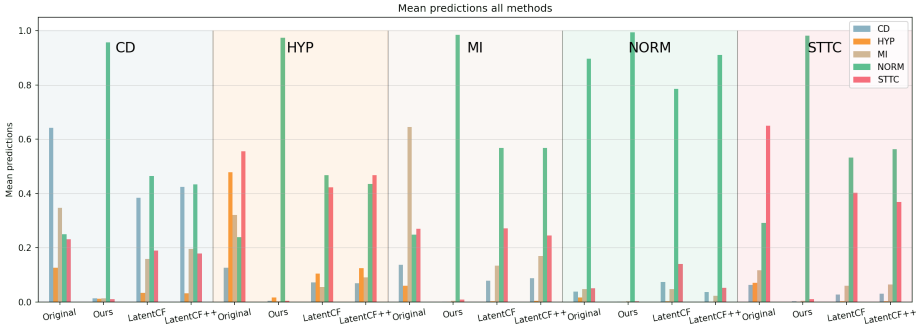


Figure 4.6: Average predictions of the baseline classifier on the different classes

The performance of the contrastive VAE with a latent space of 16 dimensions, four of which are dedicated to salient features, is shown in Figure 4.6. To evaluate its effectiveness, we compared the mean predictions of the baseline model between the input and counterfactual time series for each class. The results show a clear change in classifier prediction, highlighting the usefulness of our approach in generating realistic and meaningful counterfactual explanations for ECG data. This improvement is particularly important given the difficulty of making accurate predictions on ECG data, which is a complex and nuanced domain that requires specialized knowledge and techniques. The results demonstrate the potential of contrastive VAE as a powerful tool for generating meaningful counterfactual explanations in the context of ECG data analysis.

In summary, our study highlights the potential of using contrastive VAEs for

generating valid and sparse counterfactual explanations. By carefully selecting the salient dimensions and partitioning the latent space, our approach provides a valuable tool for generating actionable explanations that are both reliable and interpretable.

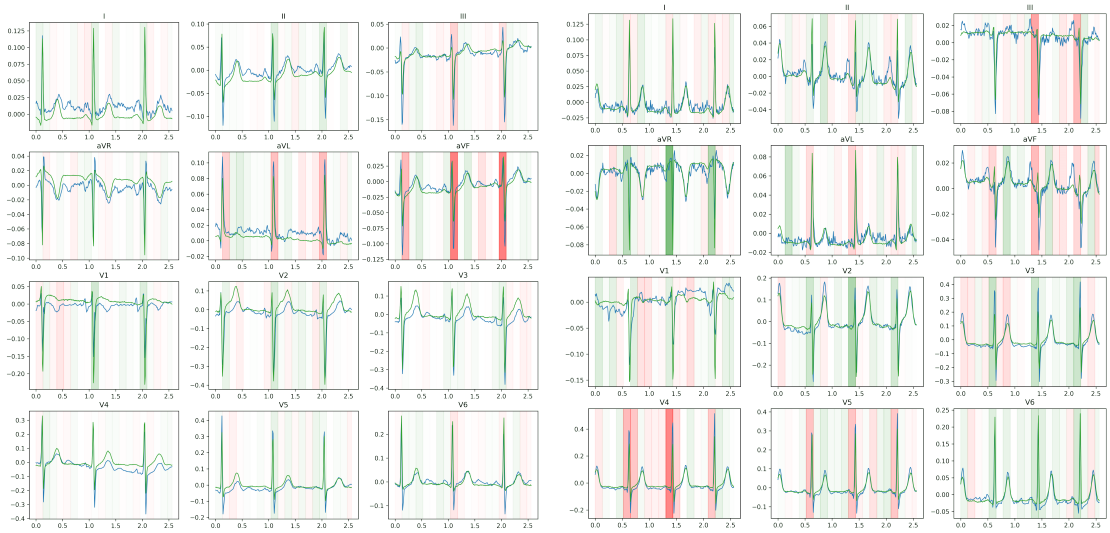
Comparison with Lime

In Ribeiro et al., 2016, the authors argue that transparency and interpretability of these models are critical for building trust and acceptance of their predictions. Their method, called LIME (Local Interpretable Model-Agnostic Explanations), provides a general framework for generating explanations for any type of classifier, both in quantitative and visual forms. The effectiveness of LIME is demonstrated through several experiments with different datasets and classifiers, including text classification, image recognition, and credit scoring. The paper presents a promising approach to addressing the challenges of interpreting and trusting the predictions of machine learning models.

To evaluate the effectiveness of our proposed counterfactual explanation method, we compared it with an adaptation of the LIME technique presented in Hering et al., 2020. In this technique, the signal is perturbed multiple times by replacing certain parts with the mean, and then fed to a basic classification model. The resulting predictions are then used to feed a linear model that estimates the influence of certain parts of the signal on the prediction. While this method assumes that a simple linear model is sufficient to provide a good local explanation, our approach aims to provide more accurate and localized explanations.

Our experiments show that the parts of the signal significantly changed by our CVAE method are more important than the rest of the signal according to the LIME method. However, in addition to providing more localized explanations, our approach also shows the form needed to change the class, which can be particularly useful for understanding the underlying causes of the change in classification.

As shown in Figure 4.7, we provide two examples of counterfactual explanations and saliency maps given by the LIME method. The LIME method gives great importance mainly to the peaks of the signal, and there is no clear difference between the explanations of the MI and CD classes. In contrast, our proposed method provides more detailed and accurate explanations, as shown in Figure 4.7a. In signals V2 and V3, the bumps after the peaks are larger in the counterfactual generated by our method, which is not the case for the MI class. On the other hand, the bumps before the peaks are smoothed in the III and aVF signals in the MI class



(a) Original label : CD (b) Original label : MI

Figure 4.7: Two counterfactual explanations with the original signal in blue, the counterfactual one in green and the background is colored given the LIME explanations

(Figure 4.7b), but not in the CD class, demonstrating that our method produces well-fitted counterfactual explanations that capture the key features of the data.

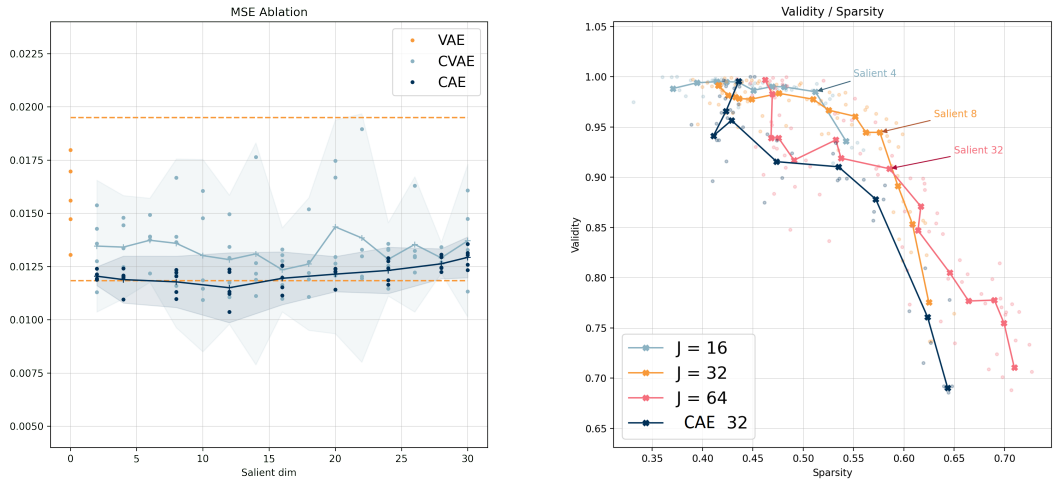
Ablation study

The ablation study presented in Figure 4.8 and in Table 4.2 provides important insights into the design of effective counterfactual generation models. One of the key findings is that the choice to use a variational autoencoder instead of a simple autoencoder has a significant impact on the quality of the generated counterfactuals. While the contrastive autoencoder performs slightly better in terms of reconstruction error, its performance on the sparsity and validity metrics is lower than that of the contrastive VAE. This suggests that the Kullback-Leibler divergence constraint, which is an essential component of the VAE, is crucial for generating good counterfactuals that are both sparse and valid. One possible reason why the VAE outperforms the AE in generating counterfactuals is that the VAE can learn a probabilistic distribution over latent space. This means that it can generate a diverse set of counterfactuals, rather than just reproducing the same set of input data. In contrast, the AE is limited in its ability to generate new data because it can only reconstruct the input data deterministically.

Overall, the results of the ablation study highlight the importance of using a variational autoencoder to generate counterfactuals. By leveraging the VAE's ability

Table 4.2: Ablation study depending on the salient dim J_s

	J_s	2	8	12	16	20	30
Validity	CVAE	0.75 ±0.09	0.93 ±0.03	0.95 ±0.02	0.97 ±0.01	0.95 ±0.09	0.99 ±0.01
	CAE	0.68 ±0.01	0.87 ±0.02	0.91 ±0.03	0.90 ±0.07	0.95 ±0.05	0.99 ±0.01
Sparsity	CVAE	0.89 ±0.02	0.78 ±0.03	0.73 ±0.04	0.65 ±0.05	0.57 ±0.09	0.51 ±0.04
	CAE	0.89 ±0.02	0.74 ±0.02	0.68 ±0.01	0.59 ±0.03	0.52 ±0.02	0.53 ±0.04



(a) Reconstruction MSE for latent space 32 (b) Sparsity / validity plot for Contrastive AE

Figure 4.8: Contrastive AE compared and contrastive VAE

to learn a probabilistic distribution over the data it is possible to generate high-quality counterfactuals that are both valid and sparse.

An explanable classifier ?

We claim that the salient space produced by our particular training of the VAE is disentangled, i.e., it allows us to distinguish between data samples belonging to different labels in the latent space. To further validate our claim, we trained a weak classifier, a logistic regression, on the salient space. We then compared the performance of this classifier with the baseline CNN model on the PTB-XL dataset. As shown in Table 4.3, the logistic regression classifier is not as accurate as the CNN model. However, the results also confirm that the salient space generated by our contrastive VAE is indeed disentangled, which is a promising finding for the development of explainable AI models.

Table 4.3: Classification performances

	AUC
CVAE(16,4) + LogReg	0.797 ± 0.010
CVAE(32,8) + LogReg	0.801 ± 0.020
CVAE(64,32) + LogReg	0.827 ± 0.003
Baseline CNN	0.928

Moreover, the use of logistic regression on the salient latent space allows us to have an explainable classifier that can provide valuable insights into the decision-making process of the CNN model. With this simple addition, we can now not only generate counterfactual examples, but also explain the classification decisions in a transparent and interpretable way, which is an essential step towards the development of trustworthy and reliable AI models.

Methods comparisons

In this section, we compare the effectiveness of our proposed counterfactual VAE method with other existing counterfactual techniques. The results are summarized in Table 4.4. In this table, training and inference time are in seconds per sample.

We first tested a simple baseline method composed of a rolling mean of the time series data. The goal of this test was to determine whether the smoothing effect resulting from the counterfactual VAE was responsible for the improved counterfactual properties of our model. We used a window of length 8 for the rolling mean and found that simply smoothing the signal was not enough to make it look like a healthy time series.

LatentCF (Balasubramanian et al., 2020) and LatentCF++ (Z. Wang et al., 2021) are counterfactual methods that use an autoencoder and a classifier to modify the latent representations in a direction that will change the prediction of the classifier. However, as stated in the introduction, the latent space produced by VAEs or AEs trained on this dataset does not correspond to the different classes present in the data. This leads to the inability of LatentCF and LatentCF++ to produce good counterfactuals. As a result, these methods have a low validity score, which is a critical metric for assessing the quality of counterfactual explanations. The inability of these methods to capture the different classes in the latent representation is a major limitation, and underscores the importance of developing techniques that can produce disentangled latent spaces that can better capture the underlying structure of the data.

Furthermore, these optimisation-based techniques have another drawback, as they

Table 4.4: Methods comparisons

	Validity	Sparsity	Training	Inference
CVAE ($J = 16, J_s = 4$)	**98.50**%	51.21 %	0.196 s	1.7e-3 s
CVAE ($J = 32, J_s = 8$)	94.46 %	57.61 %	0.241 s	1.7e-3 s
CVAE ($J = 64, J_s = 32$)	90.84 %	58.56 %	0.281 s	1.8e-3 s
RollMean	2.86 %	85.6 %	**0** s	**1e-4** s
LatentCF Balasubramanian et al., 2020	57.22 %	58.11 %	0.0935 s	7.3 s
LatentCF++ Z. Wang et al., 2021	52.97 %	59.29 %	0.0935 s	10.8 s
CoMTE Ates et al., 2021	34.45 %	**93.00** %	0.0146 s	84.07s

use gradient descent at the time of inference, the generation of the counterfactual takes about 7 and 10 seconds respectively. Making these techniques unable to produce counterfactual examples in large quantities.

While the CoMTE method has good sparsity performance, it has not worked well for several reasons. One of the key features of the method is the ability to replace only a certain part of the signal while leaving the rest unchanged, which explains the good sparsity score. The feature extraction and the classifier used for training are fast. However, the classifier performs poorly on the dataset. This is a major drawback since it is used for the optimization part, which is critical for generating effective counterfactuals. In addition, the optimization time during inference is significant, making it difficult to generate large numbers of counterfactuals. In summary, the CoMTE method has some useful features, but the poor performance of the classifier and the long optimization time during inference limit its effectiveness.

We attempted to use the SETS method (Bahri et al., 2022) presented in Section 4.2. However, we encountered significant challenges when attempting to apply this method to our dataset. In fact, the computational cost of generating the Shapelet dictionary grows with the length of the time series, making it prohibitively expensive to generate the dictionary for our dataset, which contains over 21,000 samples of 1000 time steps; the dataset used in the paper is relatively small (1354 samples of 60 time steps). As a result, we were unable to test the effectiveness of SETS for generating counterfactual explanations on our dataset.

Limitations

While our proposed counterfactual VAE method provides promising results, there are some limitations to consider. One of the main limitations is the smoothing effect of VAE on the generated data. This can make it difficult to understand and generate counterfactual explanations when pathologies cause only subtle changes in the signal.

In addition, the effectiveness of our method depends on the performance of the VAE, and out-of-distribution data may not be well encoded by the VAE, resulting in less relevant counterfactual explanations. Furthermore, while our model performs well with simple convolutional variational autoencoders, further optimization of the hyperparameters and architecture could potentially improve the model's performance. Finally, it's worth noting that our experiments were conducted on a specific dataset, and further research is needed to determine the generalizability of our method to other datasets and domains.

4.5 Conclusion

In conclusion, our proposed contrastive VAE method provides a promising approach for generating valid and sparse counterfactual explanations in multivariate time series data. Our approach successfully partitions the latent space into general and class-based features, allowing us to modify only the class-based features and generate interpretable and actionable explanations that are both reliable and effective. The validity-sparsity plot provides a useful tool for determining the optimal number of salient dimensions, and our experiments demonstrate the trade-off between sparsity and validity in our counterfactual generation process. Furthermore, our independent classifier showed a high validity score, indicating that our counterfactuals are credible and reliable. While our proposed method shows promising results, it is important to note its limitations, including the smoothing effect of the VAE on the generated data, the dependence on the performance of the VAE, and the need for further optimization of the hyperparameters and architecture. Some interesting research around diffusion models and VAE, as in Pandey et al., 2022, could help generate even better counterfactuals. In addition, further research is needed to determine the generalizability of our method to other datasets and domains.

In summary, our proposed approach has the potential to advance the field of anomaly detection in medical applications and other domains by providing a powerful tool for generating valid and interpretable counterfactuals. Our results suggest that our method could be used to improve the understanding of complex time series data, which could have significant implications for medical diagnosis and treatment planning.

4.6 Acknowledgements

The work is supported by the AI Interdisciplinary Institute ANITI, which is funded by the French 'Investing for the Future – PIA3' program under the Grant agreement ANR-19-PI3A-0004.

Chapter 5

VAE trained for predictive maintenance

In the following chapter, we present a refined version of the previously discussed method, specifically adapted to the unique requirements of predictive maintenance. This updated technique is rigorously tested and validated using three different datasets, giving us a broader understanding of its performance and applicability in different scenarios. The corresponding research paper is currently undergoing peer review for publication in the journal Reliability Engineering and System Safety.

Explainable Predictive Maintenance: Revealing Degradation Factors with Contrastive Semi-Supervised VAE

William Todo[1,2], **Merwann Selmani**[2], **Béatrice Laurent**[1,3], **Jean-Michel Loubes**[1]
[1] Institut de Mathématiques de Toulouse, Toulouse, France
[2] Liebherr Aerospace Toulouse
[3] INSA de Toulouse, Toulouse, France

Abstract

In this paper, we address the challenge of understanding degradation processes in multivariate time series data. Our primary goal is to identify the key parameters that influence the degradation process, while maintaining a good degradation estimate. We employ counterfactual explanations and develop a novel contrastive semi-supervised loss function for training a counterfactual variational autoencoder (CVAE), effectively leveraging censored data that remains inaccessible to traditional approaches. We evaluate our CVAE method on three datasets and against two state-of-the-art time series classification models - Inception Time and MLSTM FCN - as well as a standard predictive maintenance method using a variational autoencoder (VAE). The counterfactuals generated by our method reveal the critical role of specific parameters in the degradation process and demonstrate the effectiveness of counterfactual explanations in highlighting disparities between healthy and degraded time series. Our approach enables domain experts and decision-makers to concentrate on the most critical factors contributing to degradation, paving the way for the development of effective mitigation strategies.

5.1 Introduction

Predictive maintenance plays a crucial role in the aerospace industry, where ensuring the safety, reliability, and cost efficiency of aircraft is of paramount importance Mobley, 2002. As the adoption of artificial intelligence (AI) and machine learning (ML) techniques in predictive maintenance continues to grow the need for explainable

AI (XAI) becomes increasingly significant Shukla et al., 2020. Understanding the underlying reasons behind maintenance predictions enables engineers and decision-makers in the aerospace industry to gain valuable insights, refine designs, and improve the overall robustness of their products.

To address the challenges associated with explainable predictive maintenance in the aerospace industry, we propose a novel approach that leverages counterfactual explanations for in-flight data. This study is situated at the intersection of predictive maintenance and explainable AI, it seeks to advance the state of the art in these domains.

Our methodology is based on Variational Autoencoders (VAEs) and contrastive learning, a combination that has not been previously applied to predictive maintenance in the aerospace industry. We specifically tailor the contrastive loss to better fit the unique characteristics of predictive maintenance data, which allows for more effective learning of meaningful representations from time series data. The proposed approach demonstrates an improvement in the performance of predictive maintenance systems in the aerospace industry. The use of counterfactual explanations not only enhances the model's interpretability but also enables engineers and decision-makers to better understand the factors contributing to potential failures Goyal et al., 2019; Guidotti, 2022.

The remainder of this article is organized as follows: Section 5.2 provides an overview of related work on predictive maintenance, counterfactual explanations for time series and contrastive representation learning; Section 5.3 details our methodology, including the development of tailoring of the VAE-based contrastive learning approach; Section 5.4 presents our experimental results, comparing the performance of our proposed method with existing techniques and the explanations analysis.

In this paper, we make several significant contributions to the field of predictive maintenance, with a primary focus on introducing a novel technique that provides a higher level of explainability. Our key contributions are as follows:

We present a new approach to training predictive maintenance models by leveraging contrastive learning, an area that has not been extensively explored in this domain. This innovative methodology allows us to extract more meaningful representations from time series data, ultimately improving the performance of predictive maintenance systems in the aerospace industry.

In Figure 5.1, we visualize the effects of the novel contrastive loss developed in this paper (detailed in Section 5.3). The red arrows illustrate the attraction forces induced by the loss within the representational space. The grey circle represents

the time series data's representational space. In this example, we analyze a batch containing three flight IDs i, j, and k—each representing a distinct equipment lifecycle stage. The index i corresponds to the mid-lifecycle stage, at which the equipment's degradation state is unknown. Consequently, the loss treats i as unlabeled, and the pulling force is only applied between different views of this flight (i.e., time series from its neighborhood, see Section 5.3). In contrast, j and k both represent early stages of their respective lifecycles, and we assume that the equipment is healthy in both cases. Thus, the loss aims to pull j and k closer together within the representational space.

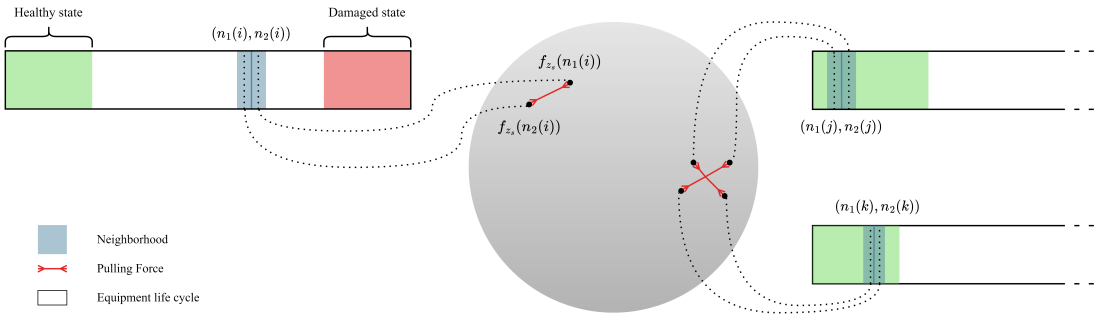


Figure 5.1: Illustration of the forces of attraction in representational space induced by the novel contrastive loss

Our proposed technique offers enhanced explainability, enabling a better understanding of the factors influencing equipment health. This increased interpretability can facilitate more informed decision-making and lead to more effective maintenance strategies for aerospace engineers and decision-makers. Additionally, the improved explainability provided by our approach can support retrofit programs in the aerospace industry, helping to identify areas for improvement and optimization in existing aircraft systems. This in turn can contribute to enhanced safety, reliability, and cost efficiency of aircraft operations.

5.2 Preliminaries

Predictive maintenance and anomaly detection

Fault detection methods, such as statistical process control, machine learning algorithms, and model-based techniques, aim to identify anomalies in the operation of machines or systems by classifying instances as either normal or anomalous Chandola et al., 2009. These techniques play a crucial role in monitoring the performance and health of equipment and inform maintenance activities. Predictive maintenance, on the other hand, goes beyond mere anomaly detection by predicting when mainte-

nance is required on a particular machine using historical data, sensor measurements, and advanced analytics. The primary difference between the two is that while anomaly detection deals with clearly defined classes (anomaly or non-anomaly).

In predictive maintenance, distinguishing between normal and abnormal data can be very difficult because equipment degradation typically occurs gradually over its lifetime, interspersed with sporadic episodes of more severe degradation. Predictive maintenance is hampered by the lack of data because the condition of the equipment can only be determined at two critical points: when a failure occurs, indicating degradation or abnormality, and when the equipment is installed or repaired, indicating a healthy or normal condition. In addition, the rarity of failures exacerbates the scarcity of labels. Another consideration in predictive maintenance is the gradual degradation of equipment condition over time, which means that there should not be a significant difference in the level of degradation or health index between two instances that are very close in time.

The objective function in predictive maintenance is frequently identified as the Remaining Useful Life (RUL) Jardine et al., 2006. Remaining Useful Life is a pivotal concept in the domain of predictive maintenance and prognostics. It denotes the estimated duration remaining before a machine, component, or system approaches the end of its useful life or necessitates maintenance intervention. Essentially, RUL is a prediction of the remaining operational time prior to an anticipated machine failure or performance degradation.

It is vital to acknowledge that RUL estimation is inherently uncertain due to the complex and dynamic nature of machine operation, the presence of numerous influencing factors, and potential measurement errors. Estimating RUL can be a convoluted process, as the degradation rate may fluctuate depending on the failures. In this context, the health index concept is often employed to estimate the RUL Kang et al., 2021 & Riad et al., 2010.

Commonly, RUL estimation is achieved through the use of a health index, which typically involves assuming linear degradation to failure or, in certain instances, adopting a piecewise constant linear function (Teng et al., 2016, Jiang et al., 2020, Laredo et al., 2019). These health indexes are based on the assumption that degradation is predominantly linear and consistent across different failures, which can be problematic when diverse failures stem from unique causes. In the piecewise linear model, determining the appropriate degradation rate is a critical factor.

In Kang et al., 2021, the authors estimate the RUL by fitting a polynomial function to the health index of the dataset. The degradation of the health index can be estimated using expert knowledge, as demonstrated in El Mejdoubi et al.,

2017. However, estimating the health index of equipment is a complex task, and predicting the RUL based on this index can be even more challenging.

These considerable difficulties associated with utilizing a health index for fitting algorithms have guided our attention towards more reliable information: data collected just before a failure should be regarded as abnormal, while data from new equipment or after repairs ought to be considered healthy.

In conclusion, while both fault detection and predictive maintenance techniques contribute to the overall health management of machines and systems, predictive maintenance provides additional value by estimating the health state and optimizing maintenance activities. By embracing the complexity and unique challenges of predictive maintenance, organizations can achieve significant improvements in equipment reliability, operational efficiency, and cost savings.

Predictive maintenance techniques

Predictive maintenance is a rapidly growing field, with numerous techniques being developed to enhance its capabilities across a wide range of use cases. As a result, a diverse array of methods has emerged to address the unique challenges posed by different applications.

In the survey by Ran et al., 2019, predictive maintenance techniques are categorized into three distinct groups. The first category consists of knowledge-based methods, which rely on expert knowledge about the system or physical modeling. The second category encompasses traditional machine learning methods, such as tree-based models, Support Vector Machines (SVM), and k-Nearest Neighbors (KNN), see Mathew et al., 2017. Machine learning-based methods require the use of features that adequately represent the data. However, for complex use cases, a simple set of features may not suffice, necessitating the involvement of a system expert to generate tailored features that can accurately capture health degradation for a specific use case. This task can be quite challenging, and due to the expert knowledge required, these features can also be considered part of the knowledge-based category.

The third and final category encompasses deep learning-based approaches that leverage deep learning models for predictive maintenance tasks. In their extensive survey, Serradilla et al., 2022 explore various deep learning techniques applied to predictive maintenance (PdM), such as Self-Organizing Maps (SOMs) for clustering and anomaly detection, and One-Class Neural Networks (OC-NNs) for identifying deviations from normal behavior. Deep learning techniques also enable the automatic

recognition of degradation indicators within raw data, streamlining the process of predictive maintenance. By facilitating the scalable deployment of such models, these approaches have the potential to transform the field of predictive maintenance, effectively eliminating the need for labor-intensive, custom feature engineering tailored to individual use cases. However, deep learning techniques also present their own challenges, as they can be difficult to train, computationally demanding, and often viewed as black boxes. This category includes Recurrent Neural Networks (RNN) models, convolutional neural networks (CNN) models, and autoencoder models.

Autoencoders are often employed to extract meaningful features without requiring labels (Davari et al., 2021; Jakubowski et al., 2021; Su et al., 2020), whereas RNN and CNN models necessitate a target label for training, which frequently takes the form of a piecewise linear RUL function. As previously mentioned, the choice of the RUL function is crucial. Despite the inherent challenges, deep learning-based approaches present promising advancements in the realm of predictive maintenance and prognostics.

In this study, we aim to tackle a critical challenge often encountered in the predictive maintenance field: the issue of censored data. Censored data is a prevalent concern in real-world predictive maintenance scenarios, as it represents incomplete information about the time to failure or the condition of a system or component. This incompleteness can be attributed to various reasons, such as the termination of observation before the occurrence of a failure event, planned maintenance actions, or replacement of a component before failure.

There are two main types of censored data: right-censored data and left-censored data. Right-censored data occurs when equipment failure has not yet happened within the monitoring period, implying that the exact time of failure remains unknown. On the other hand, left-censored data arises when observations on equipment are conducted after its early operational stage, leading to uncertainties in estimating the time to failure. While left-censored data can be employed for Remaining Useful Life (RUL) predictions, right-censored data presents a more significant challenge, as it cannot be utilized directly due to the absence of failure points. This limitation is particularly problematic in certain predictive maintenance application domains, such as aeronautics, where failure events are infrequent, resulting in a considerable amount of right-censored data.

To address this challenge, our paper proposes the use of a semi-supervised loss function capable of handling censored data effectively. By incorporating this loss

function into our predictive maintenance model, we aim to leverage the information contained in both left and right-censored data to improve model accuracy and reliability. This approach offers significant advantages in enhancing the performance and applicability of predictive maintenance models in real-world situations.

Contrastive VAE

In Todo et al., 2023, the critical problem of supervised anomaly detection in multivariate functional data is addressed, with a particular emphasis on medical applications such as electrocardiogram (ECG) analysis. While deep learning has demonstrated significant potential in this field, there is a notable scarcity of techniques that provide explainability for multivariate time series.

To tackle this challenge, and inspired by Khosla et al., 2020, an innovative approach that combines the power of variational autoencoders (VAEs) and contrastive learning is introduced. The method separates the latent space generated by the VAE into general features, which are common to all instances, and class-based features that distinguish between normal and abnormal instances. This partitioning of the latent space allows for the generation of counterfactual examples that showcase the differences between pathological and non-pathological data.

The paper emphasizes the importance of disentangling the multiple explanatory factors in time series data to develop accurate models. To address this issue, the authors introduce a contrastive loss function that effectively distinguishes between the general features of the time series and the salient features arising from anomalies. This distinction is achieved by applying the contrastive loss exclusively to a portion of the latent space, referred to as the salient part. Consequently, the salient part of the latent space becomes disentangled, while the remaining portion encodes the general features of the ECG. This approach ensures a more structured and interpretable latent space, facilitating the identification of meaningful relationships within the data.

To explicit the loss we introduce let introduce N randomly sampled sample/label pairs, $\{\check{x}_l, \check{y}_l\},\ l = 1, \ldots, N$. The training batch consists of $2N$ pairs $\{x_i, y_i\},\ i = 1, \ldots, 2N$ where x_{2i} and x_{2i-1} are 2 different views (random data augmentations, see Section 5.3) of $\check{x}_l$ and $\check{y}_l = y_{2i} = y_{2i-1}$ is the corresponding label. Let $I = \{1, \ldots, 2N\}$ be the set of indices of the different views. $P(i) = \{p \in I \setminus \{i\} : y_p = y_i\}$ is the set of indices of all positives in the batch, distinct from i and τ a positive parameter,

we follow the original implementation to set it. Note that using 2 views of each time series ensure that $|P(i)| > 0,\ \forall i \in I$.

With z^x a sample from the Gaussian output of the encoder $q_\Phi(Z|X = x) \sim \mathcal{N}_J\left(\mu(x),\ \text{diag}\left(\sigma^2(x)\right)\right)$, we apply the contrastive loss only on the salient dimensions $\{z_j^x,\ j \in \mathcal{J}_s\}$.

The supervised contrastive loss for VAE takes this from :

$$\begin{aligned} L_{\text{SupCtrVAE}}(\Phi, \theta; \{x_i\}_{i\in I}) &= \frac{1}{|I|}\sum_{i\in I} ||\hat{x}_i - x_i||^2 + \frac{1}{|I|}\sum_{i\in I} D_{\text{KL}}[q_\Phi(Z|X = x_i)||p(Z)] \\ &\quad + L_{\text{supCon}}(\{z^{x_i}\}_{i\in I}) \\ &= \frac{1}{|I|}\sum_{i\in I}\left(\frac{1}{T\times K}\sum_{t\in\mathcal{T}_x,\ k\in\mathcal{K}}\left(\hat{x}_i^{t,k} - x_i^{t,k}\right)^2\right) \\ &\quad + \frac{1}{|I|}\sum_{i\in I}\left(\frac{1}{2}\sum_{j=1}^{J}\left[-1 - \log\left(\sigma_j^2(x_i)\right) + \mu_j^2(x_i) + \sigma_j^2(x_i)\right]\right) \\ &\quad + \sum_{i\in I}\frac{-1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp\left(\sum_{j\in\mathcal{J}_s}\left(z_j^{x_i}.z_j^{x_p}\right)/\tau\right)}{\sum_{a\in I\setminus\{i\}}\exp\left(\sum_{j\in\mathcal{J}_s}\left(z_j^{x_i}.z_j^{x_a}\right)/\tau\right)} \end{aligned} \tag{5.1}$$

The study specifically focuses on ECG data, where the general behavior is attributed to common characteristics of ECGs, and deviations are due to cardiac pathologies. The authors use the encoder part of a VAE to represent the functional data in a low-dimensional latent space, which can be manipulated more easily than the original data.

Their proposed method employs the latent space to represent the time series data, capturing both the general features of the signal and the salient features that differentiate the classes. By using a latent prototype from healthy data in the salient space, the method transforms the latent representation of the signal only on the salient space and constructs counterfactual explanations through the decoder part of the VAE.

Datasets

This study is situated in the dynamic and complex aerospace industry, where the proper functioning and maintenance of equipment is of paramount importance. During each flight, sensors continuously monitor various equipment, capturing multivariate time series data. In this study, the time series are sampled at 1 Hz, which is the standard for this type of data. This data can be grouped with other time-series data collected throughout the equipment's lifetime, from the moment

of installation to its eventual failure. The result is a comprehensive collection of time series data for a wide range of equipment, providing invaluable insight into its performance and maintenance requirements.

Aircraft are equipped with a variety of components that require close monitoring. However, for the purposes of this paper, we will focus specifically on two predictive maintenance use cases, both derived from proprietary, undisclosed industrial datasets. The first use case involves the Heat Exchanger (HE) dataset, which is centered around the detection of leaks in a heat exchanger. These leaks can potentially impair the overall performance of the system. The data for this particular use case is gathered primarily from temperature sensors and the command of a valve adjacent to the heat exchanger, presenting an effective way to monitor the integrity of the heat exchanger.

Our second dataset pertains to a flow regulation valve (FV). This component plays a pivotal role in the overall functionality of the aircraft, as malfunctions can disrupt the operation of subsequent valves or equipment, thereby reducing the overall efficiency of the bleed system. In this case, the dataset is formed using data from sensors that measure upstream and downstream pressure, in addition to the commands sent to the downstream valve and the pressure after the downstream valve. In total, five parameters are utilized for this dataset, providing comprehensive insights into the operational status of the flow regulation valve.

Both these datasets provide valuable insights into the operation of key aircraft components. Moreover we can note that these two datasets present different kind of data, one is focused on temperature sensors and the other is focused on pressure sensors. Temperature changes generally occur more gradually because of the thermal inertia or heat capacity of materials, which is their ability to absorb and store heat. When a component or system is heated or cooled, it often takes time for that energy to be fully absorbed or released, leading to a more gradual change in temperature.

On the other hand, pressure changes can happen rapidly due to the compressible nature of gases. In systems like an aircraft's bleed air system, changes in pressure can occur almost instantaneously in response to alterations in flow rate or valve position. This is because gases are highly responsive to changes in volume or temperature, and their pressure can adjust almost immediately to balance with the new conditions.

In the context of the two datasets mentioned, these differences in behavior mean that the temperature data from the HE dataset is likely to present smoother trends over time, while the pressure data from the FV dataset might show more immediate responses to changes in system conditions.

Having access to these two distinct types of datasets, each with a different primary focus and rate of fluctuation, provides a robust platform to assess the performance of our predictive maintenance model.

Moreover, the diversity in these datasets represents a wide range of real-world scenarios, enhancing the generalizability and versatility of our model. By performing well across both these datasets, the model demonstrates its robustness and reliability, regardless of the nature of the data it encounters. This is particularly valuable in industrial applications where sensor data can vary greatly depending on the specific components and systems being monitored.

By examining these use cases, we aim to shed light on the broader context of equipment maintenance within the aerospace industry. To address the unique challenges of each use case, we will select portions of the flights that are most relevant to the problems we are trying to solve. This selection can be very simple, as in the FV dataset, where it is sufficient to select the part of the flight where the valve is active. Unfortunately, due to confidentiality constraints, these in-house datasets cannot be shared publicly.

In order to ensure the repeatability of our findings and to allow for independent validation, we also test our methodology on a public dataset known as NGAFID, Yang and Desell, 2022. The HE dataset is composed of 149 equipment lifetimes, referred to as periods, and includes data from 144 repaired or new equipment, only 5 faults, and a total of 168,000 flights, resulting in an impressive 12,000 hours of flight time. The FV dataset boasts 663 periods, encompassing data from 493 repaired or new equipment, 350 faults, and a grand total of 379,000 flights, which translates to 42,000 hours of flight time.

The publicly available NGAFID dataset, on the other hand, primarily consists of flights that take place immediately before equipment failure (damaged flights) or immediately after a repair (repaired or healthy flights). Interestingly, some flights in the dataset are labeled as having been repaired during the flight, which suggests that they are most likely test flights. In order to maintain the integrity of our analysis, we have chosen to exclude these flights from the dataset, along with any flights shorter than 1024 seconds. After applying these filters, the NGAFID dataset contains a total of 15,298 flights, with 7,987 damaged and 7,311 healthy flights, amounting to 19,000 hours of flight time. The table 5.1 summarizes the dimensions of the dataset presented here.

By utilizing these diverse datasets, our research endeavors to enhance the understanding and application of predictive maintenance within the aerospace industry, contributing to the development of more reliable, efficient, and safe aviation systems.

	Lifetimes	Faults	Flights	Datapoints
HE	149	5	12 000	43.2 M
FV	663	350	379 000	151.2 M
NGAFID	-	-	15 298	68.4 M

Table 5.1: Datasets size

5.3 Method

Contrastive Drift Loss

As mentioned earlier, accurately determining the condition of equipment in predictive maintenance data is only possible when the equipment is near the beginning (healthy) or end (damaged) of its life. Condition assessment becomes increasingly difficult when the data is far from these two points in time. Given that a majority of flights fall into this intermediate category, it is of utmost importance to incorporate these instances into the training dataset.

In the unsupervised contrastive learning framework, presented by Chen, Kornblith, Norouzi, et al., 2020, it is assumed that two views of the same sample should be situated in close proximity to each other in the feature space while remaining distant from other samples. When processing a batch of data, the method allows for the separation of two representations belonging to the same class, which is the primary reason behind the development of the supervised method.

In light of these considerations, the natural next step is to combine the two versions of the contrastive loss (supervised and unsupervised) to create a semi-supervised approach. When a label is available, the goal is to draw other representations of that same label closer together. However, if a label is not available, the focus shifts to bringing other views of the same sample closer to one another. This can be achieved by referring to Eq. (5.1) and modifying the definition of $P(i)$ to accommodate both supervised and unsupervised loss components:

$$P(i) = \begin{cases} \{p \in A(i) : y_p = y_i\} & \text{if } y_i \text{ exist} \\ j(i) & \text{if } x_i \text{ unlabeled} \end{cases}$$

We name that loss the semi supervised contrastive loss or $L_{\text{ssContrastive}}$. and it effectively leverage the strengths of both approaches to create a more robust and efficient learning framework.

An other significant challenge that may arise is ensuring the reliability of the labels assigned to equipment. Identifying the root cause of a failure can be a complex task, as failures often stem from multiple contributing factors. Even with an extensive

examination of the deteriorating equipment, pinpointing the exact origin of the failure may prove difficult. As a result, two equipment units nearing failure may display dissimilar behaviors, despite being in a similar state of disrepair.

Given the inherent complexity of determining the root cause of equipment failures, it has been decided to treat damaged samples differently. Rather than grouping them together, damaged samples will be considered as unlabeled data. This approach acknowledges the possibility that each damaged sample may be unique in its failure characteristics. Conversely, healthy data samples are expected to exhibit similar behavior and, therefore, should be grouped together within the representation space. To accomplish this goal, the definition of $P(i)$ in Eq. (5.1) can be adapted as follows :

$$P(i) = \begin{cases} \{p \in A(i) : y_p = y_i\} & \text{if } y_i = \text{healthy} \\ j(i) & \text{else} \end{cases}$$

Views in predictive maintenance

In the context of contrastive learning, generating multiple views of the same data point is a crucial aspect. This technique was initially developed for image models. In the case of image data, a variety of data augmentation techniques, such as rotation, random cropping, resizing, flipping, random solarization, and noise addition, are employed to generate different views. These techniques have been extensively discussed in various papers on contrastive learning, including Chen, Kornblith, Norouzi, et al., 2020, Chen, Kornblith, Swersky, et al., 2020, He et al., 2020, and Grill et al., 2020.

For time series data, the available data augmentation options are somewhat more limited. Nevertheless, some straightforward augmentations like random cropping and noise addition can be applied. More complex techniques, such as time warping-based augmentations, can also be utilized, as discussed in Ismail Fawaz et al., 2019 and Um et al., 2017. It is important to note, however, that the computational complexity of these advanced augmentation techniques may constrain their practical use, especially for large-scale datasets or real-time applications.

In this paper, we use simple data augmentation techniques for time series data. This includes the addition of random noise, slight modifications to the mean and trend of the data, and the application of random cropping to generate alternative views of the same data point.

The generation of random views for contrastive learning can be approached differently from traditional data augmentation methods. The main goal of using multiple views of the same data point in contrastive learning is to train the model to represent similar data (e.g., the same image) close to each other in the representation space. In the context of our study, we aim for data representing the same health level of a device to be situated close together in the representation space. As a result, it is reasonable to extend the concept of randomly sampling within a time series to randomly sampling from data points that are temporally close within a device's lifetime.

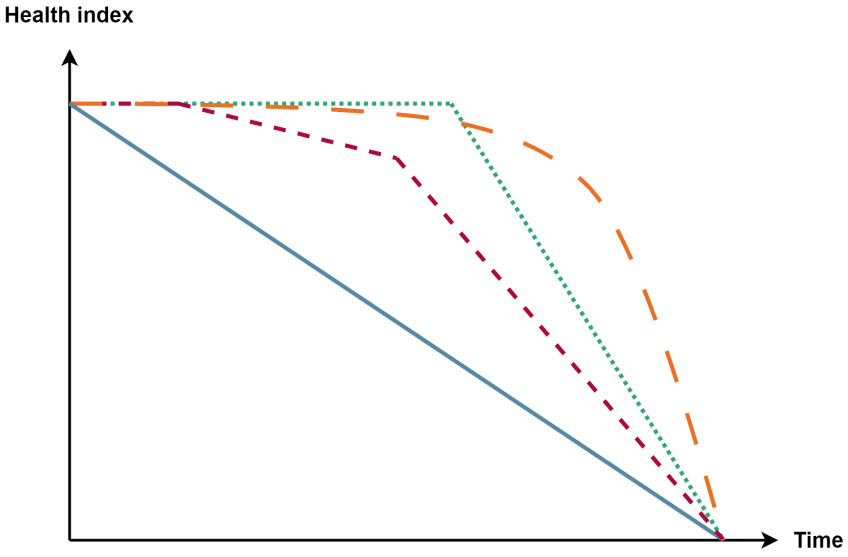


Figure 5.2: Graphical Comparison of RUL Modeling Approaches

Our proposed methodology effectively addresses the complexities inherent in the estimation of Remaining Useful Life (RUL). Regardless of the RUL modeling strategy chosen - be it a conventional linear approach, a piecewise linear approach, or a strategy that incorporates exponential decay - neighboring flights are expected to exhibit a similar health index. This is illustrated in Figure 5.2. This key assumption holds even in the face of different degradation rates over individual lifecycles. The power of our approach lies in its ability to detect and learn from these different degradation patterns, leading to a more nuanced understanding and more accurate prediction of device health over time.

By adopting this approach, we can not only leverage a more extensive range of data augmentation for use in contrastive loss, but also account for the temporal ordering within a series of time series. This consideration is also relevant for the analysis of

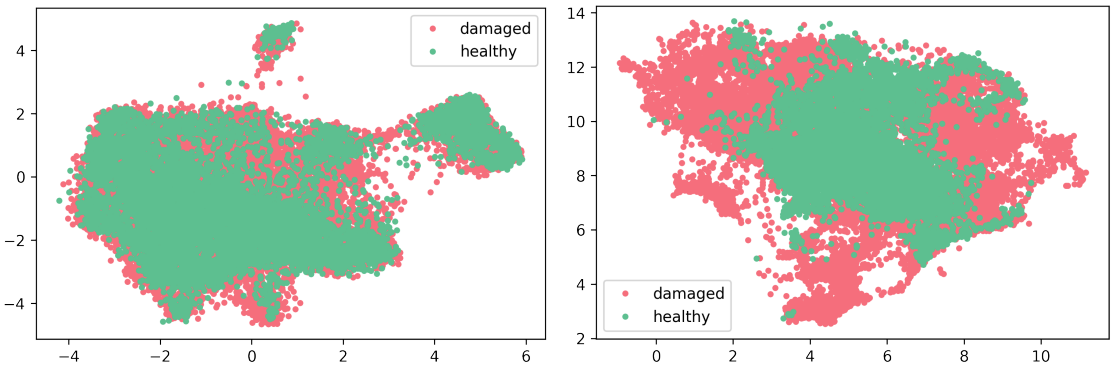
long time series data in predictive maintenance settings, where continuous sensor recordings are common. By taking random crops within a time window surrounding a given time series, we can effectively incorporate the temporal relationships between different data points, thereby enhancing the model's ability to learn meaningful representations of the underlying data.

Contrastive VAE training

The contrastive VAE for predictive maintenance is a two-phase approach that combines the advantages of unsupervised learning with specialized techniques designed for predictive maintenance tasks. The first phase involves training a contrastive VAE using the predictive maintenance-tailored loss function, referred to as the contrastive drift loss, which is detailed in 5.3. This loss function leverages different views of the predictive maintenance data.

As a result of this specialized training, the salient features within the CVAE are those that can effectively discriminate between healthy and faulty time series. This structured and informative latent space is then utilized in the second phase of the method, focusing on anomaly detection. The aim of this phase is to maintain simplicity; hence, an anomaly detection model is trained on the salient space representation of the healthy flights. v The differences between the representations of the two classes in each part of the latent space are illustrated in Figure 5.3 for the FV dataset. To effectively visualize the differences between these two components of the latent space, we use the Uniform Manifold Approximation and Projection (UMAP, from McInnes et al., 2018) technique to reduce the dimensionality to two dimensions. As shown in the figure, the general latent space does not show noticeable differences between the two classes. However, the salient space allows for a more noticeable separation between the classes. This finding highlights the effectiveness of the salient space representation in training the detection model for identifying anomalies in flight data.

Various methods have been explored for this purpose, including isolation forest (IF) F. T. Liu et al., 2008, one-class SVM (OCSVM) K.-L. Li et al., 2003, and local outlier factor (LOF) Breunig et al., 2000. Each of these approaches can provide an anomaly score, which is subsequently used to assess the health of a device. By integrating the strengths of the CVAE's discriminatory features and the simplicity of the anomaly detection models, this two-phase method offers a robust and efficient solution for identifying potential issues in the maintenance of equipment, ultimately enhancing the reliability and longevity of these systems.



(a) General part of the latent space : Z_g (b) Salient part of the latent space : Z_s

Figure 5.3: Comparison of General and Salient Latent Spaces for the FV Dataset

Explanation

In this study, we employ VAE based method to address the challenges associated with the analysis of time series data, particularly in the context of detecting and correcting abnormalities. One key advantage of utilizing this VAE-based technique is that we can exploit the decoder component of the VAE architecture to generate novel samples. This is achieved by partitioning the latent space into two distinct segments: one which encodes the general shape information of the time series, and another that captures the specific characteristics associated with abnormalities.

To effectively "remove" abnormalities from the signal, we manipulate the salient portion of the latent space corresponding to these anomalous features. We adopt the approach outlined in Todo et al., 2023, which entails identifying a healthy prototype within the latent space z_s, characterized as a representative element that embodies typical, non-anomalous behavior. To compute the healthy representative z_t, we use the mean of the healthy data in z_s. By substituting z_s with z_t in a damaged time series, we can then simulate the appearance of the time series in the absence of any damage.

This process facilitates a comparative analysis between the original degraded data and the counterfactual corrected version, which in turn enables users to gain a deeper understanding of the degradation process. Furthermore, this comparison makes it easier for users to discern the differences between healthy and damaged data. As such, this VAE-based method provides a powerful tool for detecting, isolating, and visualizing abnormalities in time series data. Additionally, the proposed VAE-based approach allows for the extraction of valuable insights from time series data. By enabling the identification of abnormal patterns or behaviours and subsequent

correction, practitioners can make more informed decisions and take appropriate actions to prevent or mitigate potential problems.

In summary, the VAE-based method proposed in this paper offers a versatile and effective means of addressing the complexities associated with the analysis of time series data. By leveraging the decoder component of the VAE architecture to generate new samples and partitioning the latent space to isolate and correct abnormalities, this technique presents a powerful tool for enhancing our understanding of the underlying processes and improving decision-making in a variety of applications.

5.4 Results

Experimental setup

We have maintained minimal preprocessing across all the datasets utilized in our study. For HE and FV datasets, we restricted the time series data to specific moments during the flights, as determined by domain experts. This selection process is crucial for constructing accurate and reliable datasets. However, we were unable to apply a similar selection process to the NGAFID dataset due to the absence of expert knowledge on these particular aircraft. Consequently, we want to emphasize the importance of selecting specific moments in flight data for generating reliable datasets. To ensure consistency, we applied standard scaling to the time series data for each dataset.

In each dataset, our models are designed to operate on short segments of 128 seconds. Throughout the training process, we use random cropping to provide input data to our models. For the final prediction, we compute the average score across all 128-second segments of the flight, yielding the overall classification. This approach proves effective in representing flights, as evidenced by the significant performance improvement observed on the NGAFID dataset compared to the results reported in the referenced study. This highlights the potential of our method to address complex predictive maintenance tasks, particularly for challenging datasets.

The Variational Autoencoder (VAE) models employed for both encoding and decoding purposes consist of three convolutional layers, with attention layers adopting the squeeze and excitation layers from Hu et al., 2018. These layers have been previously utilized in time series analysis, as demonstrated by Karim et al., 2019.

Regarding the training parameters, certain aspects will vary depending on the dataset being used. For instance, the dimensions of the latent space and the salient

features can be adjusted according to the specific dataset requirements. Furthermore, scaling parameters have been introduced to account for the composition of the loss function, which includes reconstruction loss, Kullback-Leibler (KL) divergence, and contrastive loss. The scaling parameters, named *kl_scaler* and *contrastive_scaler*, are designed to facilitate the optimization process by maintaining a similar scale across the three loss components.

The optimization process in our study employs the Adam optimizer, a widely-used optimization algorithm. We use the default parameters for the Adam optimizer, ensuring that our optimization approach aligns with established practices. To further refine the optimization process, we incorporate a learning rate schedule that combines Cosine Annealing with Warm Restarts and Linear Warmup. The learning rate schedule begins with a linear warmup phase, during which the learning rate gradually increases from an initial value to a specified maximum learning rate. In our case, the maximum learning rate is set at 0.0005. The linear warm-up phase helps stabilize the training process and prevents the occurrence of undesirable large weight updates in the early stages of training.

Following the linear warmup phase, the learning rate is adjusted using the Cosine Annealing technique. Cosine Annealing allows the learning rate to decrease smoothly, enabling the model to explore the optimization landscape effectively and avoid local minima. Warm Restarts after each Cosine Annealing cycle reset the learning rate to the maximum value, which subsequently follows the cosine annealing schedule once again. This technique promotes exploration of the optimization landscape, reducing the risk of the model becoming trapped in local minima and facilitating the discovery of better optima. In our experimental setup, the first Cosine Annealing cycle lasts for 100 epochs. With each subsequent cycle, we double the number of epochs.

To assess the performance of our models, we implemented cross-validation techniques, and the results presented are the average of the outcomes obtained from each test set. This approach ensures a robust evaluation of our models, providing a comprehensive understanding of their performance across different datasets and lending credibility to the conclusions drawn from this study.

Evaluation metrics

In this paper, we address two key objectives that are critical to the advancement of predictive maintenance in the aviation industry. The first objective is to create a robust predictive maintenance model that is able to discriminate between flights

Hyperparameter	Value
n_filters_conv1	256
n_filters_conv2	512
n_filters_conv3	512
size_filters	5
reduction_ratio (squeeze and excite)	16
kl_scaler	0.0005
contrastive_scaler	0.2

Table 5.2: Hyperparameters

in optimal operating condition and those on the verge of failure. To effectively assess this aspect, we adopt balanced accuracy as our performance metric, which accounts for the class imbalance commonly encountered in such problems. Balanced accuracy is defined as the average of the true positive rate (sensitivity) and the true negative rate (specificity), ensuring equal importance is given to both classes. Mathematically, balanced accuracy can be calculated as follows:

$$\text{BalancedAccuracy} = \frac{1}{2}\left(\frac{\text{TP}}{\text{P}} + \frac{\text{TN}}{\text{N}}\right)$$

where TP denotes the number of true positives, P represents the total number of positives, TN is the number of true negatives, and N signifies the total number of negatives. This metric provides a more insightful evaluation of the model's performance, particularly in scenarios where the class distribution is imbalanced, as it prevents the dominance of the majority class in the evaluation.

Recognizing the potential problem of false positives, we include an additional simple metric to address this concern. We set an acceptable false positive rate of 2% by determining the corresponding threshold on the training set. We then use this threshold to calculate the false positive rate (FP) on the test set, which includes the number of healthy flights misclassified as damaged, as well as the true positive rate (TP) , which is the proportion of correctly identified flights with impending failures.

The second objective involves generating counterfactual explanations for damaged time series data, thereby providing actionable insights for system improvement. To assess the effectiveness of our model in generating counterfactual explanations, we utilize a validity metric that quantifies the degree to which the generated counterfactuals result in a shift in the predicted class of the signal relative to the target. Mathematically, this is represented as the proportion of counterfactuals causing an independent classifier, denoted by f, to alter its predictions from y to y_{target}.

However, the computation of this metric necessitates the availability of a reliable independent classifier, which may not always be accessible within the context of predictive use cases.

Classification

In order to benchmark our proposed method, we selected two state-of-the-art time series models for comparison: InceptionTime Ismail Fawaz et al., 2020 and MLSTM-FCN Karim et al., 2019. InceptionTime employs various filter sizes (10, 20, and 40) to capture features at different time scales, making it a powerful tool for time series classification. On the other hand, MLSTM-FCN combines the strengths of convolutional neural networks (CNN) with squeeze-and-excitation layers, as well as the long short-term memory (LSTM) architecture, to effectively handle multivariate time series data. These two methods have demonstrated state-of-the-art performance in numerous time series classification tasks, and as such, they serve as strong benchmarks for our proposed model.

We also use a modified version of MLSTM-FCN with selective kernels called MLSTM-FCN(SK), this model is described in detail in the appendix 5.6.

Additionally, we explored the use of a classical Variational Autoencoder (VAE) to extract meaningful features from the time series data, as this approach has been commonly employed to address similar problems. By evaluating our model against these established techniques, we aim to demonstrate the efficacy and competitiveness of our proposed method in the context of predictive maintenance.

The performance metrics for all methods are summarized in Table 5.3. Upon examining the simpler use case (HE dataset), all techniques achieve satisfactory performance. However, InceptionTime and MLSTM-FCN exhibit slightly lower performance, likely due to the limited number of damaged flight examples within this specific dataset. As the analysis advances to the more challenging FV dataset, the CVAE method notably surpasses the MLSTM-FCN and InceptionTime models in terms of performance. This significant improvement highlights the potential of the CVAE approach to effectively tackle complex predictive maintenance problems in various application areas. By leveraging its strengths, practitioners can gain valuable insights and develop more efficient maintenance strategies for diverse systems.

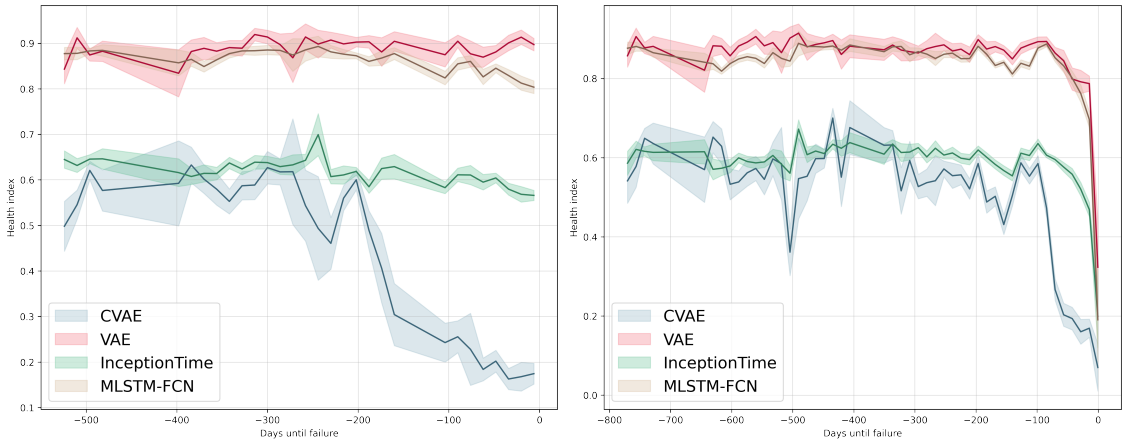
Figure 5.4 presents the anticipated degradation of the health index for two separate equipment lifetimes within the FV dataset. The visualized data have been smoothed using a 7-day rolling mean average to reduce random fluctuations and highlight overall trends. The colored bands encapsulating the mean lines correspond to

the standard deviations, providing a measure of variability around the average. It's important to note that the output from the anomaly detection models is not restricted to the range between 0 and 1. To facilitate more meaningful comparisons, these anomaly scores are appropriately scaled.

In this study, the CVAE approach consistently outperforms the other three techniques in terms of early detection of degradation across various scenarios. As illustrated in Figure 5.4a, the CVAE method stands out as the sole approach capable of identifying the degradation effectively, whereas the other techniques fall short in providing timely detection.

Meanwhile, in Figure 5.4b, the alternative methods exhibit a delay in detecting degradation compared to the CVAE method. This observation underscores the superior performance of the CVAE method in terms of degradation detection.

Overall, the CVAE method outperforms its counterparts in these specific lifetimes, which aligns with the comprehensive results presented in Table 5.3. This observation underscores the effectiveness of the CVAE approach for degradation detection in the context of the FV dataset.



(a) (b)

Figure 5.4: Examples of Health Index evolution across two equipment lifetimes

For the NGAFID dataset, deep learning-based classifiers outperform the CVAE method. This can be attributed to the lack of flight history in the dataset, which prevents the effective exploitation of the semi supervised contrastive loss applied to the neighborhood, limiting it to the crops of the same flight. Additionally, the NGAFID dataset is well-suited for time series classification models, as it comprises flights with clearly defined labels.

It is noteworthy that the performance of our methods surpasses that reported in the study by Yang and Desell, 2022 (achieving a maximum accuracy of 76.1%). The

Dataset	Model		FP	TP	Balanced Accuracy
HE	InceptionTime		0.0247 ±0.003	0.9821 ±0.034	0.8727 ±0.158
	MLSTM_FCN		0.0238 ±0.004	0.9992 ±0.001	0.9141 ±0.095
	MLSTM_FCN(SK)		0.0208	1.0	0.9144
	VAE (3)	IF	0.0134 ±0.003	0.7520 ±0.121	0.9728 ±0.013
		LOF	0.0320 ±0.007	0.9959 ±0.006	0.9944 ±0.001
		OCSVM	0.0147 ±0.009	0.9837 ±0.023	0.9897 ±0.007
	CVAE (16,8)	IF	0.0163 ±0.129	1.0 ±0.001	0.9930 ±0.005
		LOF	0.0382 ±0.028	0.9959 ±0.006	**0.9980** ±0.003
		OCSVM	0.0159 ±0.014	1.0 ±0.001	0.9152 ±0.119
FV	InceptionTime		0.0218 ±0.003	0.1031 ±0.020	0.5870±0.023
	MLSTM_FCN		0.0212±0.004	0.1036 ±0.037	0.5847 ±0.018
	MLSTM_FCN(SK)		0.0215	0.0983	0.5798
	VAE (8)	IF	0.0187 ±0.03	0.0530 ±0.02	0.5492 ±0.03
		LOF	0.0329 ±0.02	0.1170 ±0.015	0.5768 ±0.04
		OCSVM	0.0196 ±0.03	0.0674 ±0.02	0.5510 ±0.03
	CVAE (64,8)	IF	0.0261 ±0.02	0.1961 ±0.05	0.7249±0.04
		LOF	0.2077 ±0.06	0.6707 ±0.03	0.7013 ±0.03
		OCSVM	0.0670 ±0.05	0.4238 ±0.06	**0.7392**±0.04
NGAFID	InceptionTime		0.0263 ±0.004	0.4615 ±0.109	0.7794 ±0.031
	MLSTM_FCN		0.0265 ±0.005	0.4549±0.139	0.7713 ±0.028
	MLSTM_FCN(SK)		0.0306	0.6239	**0.7856**
	VAE (64)	IF	0.0202 ±0.003	0.0207±0.006	0.5004 ±0.006
	CVAE (256,128)	IF	0.1183 ±0.021	0.5459 ±0.062	0.7454 ±0.015

Table 5.3: Comparative Performance Metrics

performance improvement can be attributed to two factors: the averaging of scores across different crops and the utilization of smaller windows for our classifiers. This demonstrates the effectiveness of our approach in handling predictive maintenance tasks, particularly for challenging datasets such as FV and NGAFID.

Explanation

In Figure 5.5, we present a counterfactual explanation for a multivariate time series from the dataset FV, to identify factors contributing to degradation. Counterfactuals serve as valuable tools for identifying the differences between healthy and degraded time series. By examining these disparities, we can effectively investigate the parameters that influence the degradation process. This analysis allows us to

identify critical factors that contribute to degradation, providing insights for domain experts and decision makers to develop targeted strategies to mitigate and prevent degradation in various systems.

A closer look reveals that there are minimal differences between the counterfactual explanation and the original time series for the final pressure and the upstream pressures. This suggests that these parameters have a relatively low impact on the degradation process. In contrast, more substantial differences are observed for the downstream valve command and the downstream pressure, indicating their higher influence on degradation.

Regarding the downstream pressure, the counterfactual emphasizes the presence of spikes indicating degradation. In addition, a relatively high value for the downstream pressure seems to indicate a near failure, which is not surprising since the role of the valve is to regulate this pressure.

For the downstream valve command, the counterfactual explanation shows that a healthy time series should have lower values than those observed in the original time series. It is crucial to note that the overall shape of the time series is preserved in the counterfactual explanation, further reinforcing the importance of the downstream valve command in the degradation process. We can explain this because the downstream valve is trying to compensate for the high pressure from the faulty valve, so it is working harder than it normally would. These results are consistent with those reported by domain experts for this specific use case, lending credibility to our analysis.

In conclusion, our exploration of counterfactual explanations for the multivariate time series from the FV dataset has enabled us to identify key parameters that play a significant role in the degradation process. Importantly, these findings are consistent with observations made by domain experts for this specific use case, demonstrating the validity of our counterfactual explanations. This information is valuable to domain experts and decision makers, allowing them to focus on the most critical factors contributing to degradation and develop effective strategies to address them. By aligning our analysis with the expertise of domain experts, we can ensure that our approach is both accurate and relevant to the challenges faced in real-world applications.

Robustness

A major challenge we face when using VAE methods to detect degradation in time series data is the heavy dependence on the quality of the label annotations. The

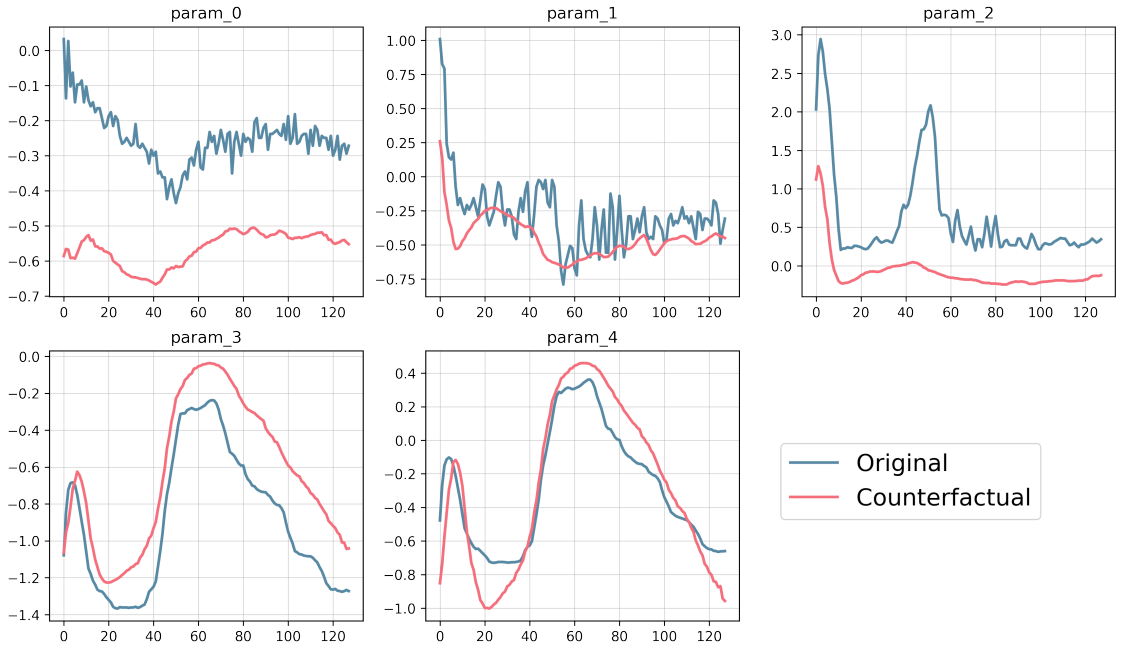


Figure 5.5: Counterfactual explanation of multivariate time series for dataset FV highlighting parameters impacting degradation process

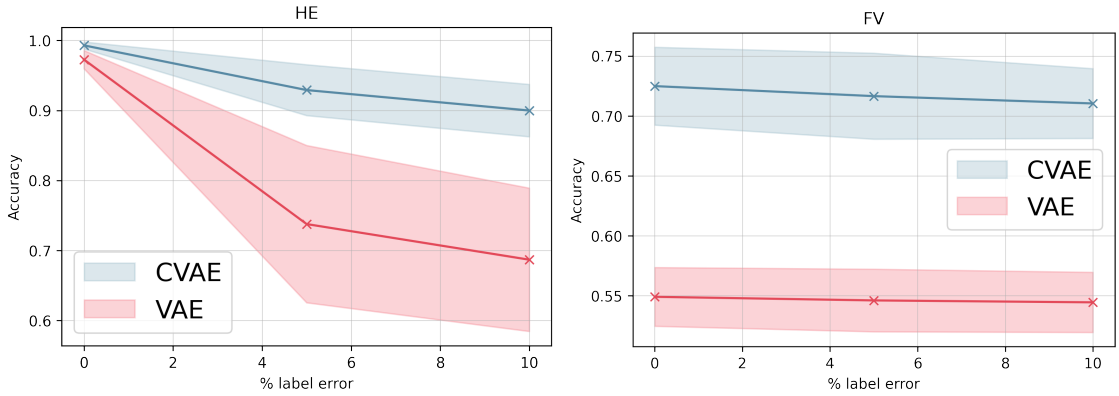
accuracy of these labels is of paramount importance, as the VAE may inadvertently learn the characteristics of mildly degraded or degraded time series during the training phase when they are fed into the model. As a result, the VAE may have difficulty distinguishing between truly degraded time series and those that are healthy during the testing phase.

In real-world predictive maintenance scenarios, the data labeling process is often complex and challenging. Several issues can arise during the generation of the data set, such as anomalies in other parts of the system that cause the selected equipment to fail, or other components within the system that compensate for the degradation of the target equipment. These factors can further complicate the task of accurately labeling data. In addition, it has been observed that VAE-based techniques underperform when applied to datasets with a high degree of diversity in both healthy and degraded time series. For example, in complex systems, a single piece of equipment may not consistently affect the accessible sensors, which can lead to misleading signals. In addition, other components within the system may introduce noise into the time series, further exacerbating the challenge of accurately detecting degradation.

In contrast to VAE, we found the CVAE method to be more robust to these challenges, prompting the present experiment. Our goal was to evaluate the per-

formance of CVAE and VAE under the worst-case scenario of swapping time series instances between the healthy and damaged classes in the training set and analyze the impact on accuracy. Figure 5.6 shows the results for the FV dataset (Figure 5.6b). In this case, the performance of the VAE is already relatively low, indicating that the deterioration of the training dataset did not significantly affect its accuracy. Conversely, for the HE dataset (Figure 5.6a), where the two methods show comparable results, the degradation of the dataset significantly affects the performance of the VAE, while only minimally affecting the CVAE.

This study demonstrates the superior suitability of the CVAE method for real-world predictive maintenance problems. These problems often involve a diverse dataset and the possibility of imperfectly classified time series. CVAE's robustness to noise and diverse data sets positions it as a reliable choice in scenarios where perfect labeling is unlikely. In contrast to traditional VAE approaches, CVAE maintains its accuracy in the presence of label noise and multiple data sources. This distinct advantage enables the CVAE methodology to effectively address the complexities inherent in real-world predictive maintenance scenarios.



(a) (b)

Figure 5.6: Evolution of the accuracy with bad labels

5.5 Conclusion

In this study, we addressed the challenge of understanding degradation processes in multivariate time series data by employing counterfactual explanations, focusing on three datasets: FV, HE, and the publicly available NGAFID dataset. Our primary objective was to identify key parameters that significantly impact the degradation process, providing valuable insights for domain experts and decision-makers, while

also assessing degradation estimation performance through the classification of healthy and degraded flights.

To train our counterfactual variational autoencoder (CVAE) method, we developed a novel contrastive loss function, which encourages the model to produce more meaningful and discriminative representations in the latent space. This loss function effectively captures the equipment lifecycle stages and improves the interpretability of the counterfactual explanations.

We compared our CVAE method against two state-of-the-art time series classification methods, Inception Time and MLSTM-FCN, as well as a standard predictive maintenance method using a variational autoencoder (VAE). Our method outperformed the deep learning models on the FV and HE datasets and demonstrated competitive performance on the NGAFID dataset.

The insights gained from this research contribute to the predictive maintenance field and degradation estimation. By pinpointing critical factors contributing to degradation and accurately classifying healthy and degraded flights, we can enable domain experts and decision-makers to focus their efforts on addressing these factors and devising effective strategies for mitigation.

In conclusion, our research demonstrates the potential of counterfactual explanations, CVAE, and the novel contrastive loss function in uncovering crucial parameters that influence the degradation process in multivariate time series data and accurately estimating degradation. We believe that continued exploration of this approach and its applications in predictive maintenance will yield even more impactful findings and contribute to the development of targeted, effective strategies to address complex challenges across various domains.

Appendix

5.6 Selective kernels

Selective kernels, as described in the work of X. Li et al., 2019, play an important role in image classification, primarily due to their adaptability and ability to exploit information at multiple scales as illustrated in the Figure 5.7. This method is widely used in image classification, providing a more dynamic and versatile model that competently captures complex patterns in the data.

A noteworthy application of this technique is found in the SimCLRv2 algorithm, specifically within the ResNet architecture, as explored by Chen, Kornblith, Swersky, et al., 2020. The inclusion of selective kernels in this context led to an enhancement in the model's performance, demonstrating the technique's effectiveness in deep learning architectures.

Moreover, the utility of selective kernels extends to time series data, particularly in human activity recognition (HAR) tasks. According to Gao et al., 2021, the incorporation of selective kernels into a relatively straightforward convolutional neural network (ConvNet) architecture resulted in performance improvements. This attests to the technique's potential not only in intricate models but also in simpler, more streamlined architectures. Thus, the broad applicability and adaptability of selective kernels.

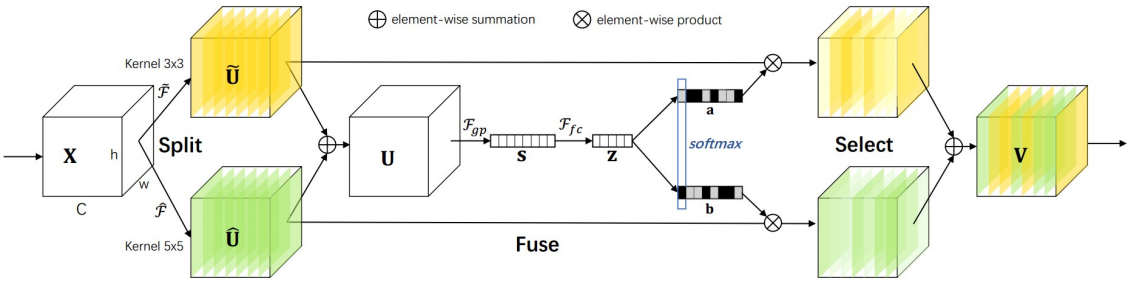


Figure 5.7: The Diagram of a Selective Kernel Convolution module from the paper X. Li et al., 2019

In terms of its mechanism, selective kernels share common ground with the 'squeeze-and-excitation' technique. Both methods adaptively modulate the prominence of various convolutional filters, guided by a compact representation of the model's current state. These techniques fundamentally aim to augment the model's discriminative prowess, honing in on the most informative features at a given juncture.

However, the key distinction between selective kernels and the 'squeeze-and-excitation' approach arises in the diversity of filters deployed. While the squeeze-and-excite technique primarily adjusts the weights of filters of the same size, selective kernels increase the adaptability of the model. They dynamically select from convolutional filters of different sizes, thus accommodating a spectrum of receptive field sizes within the model. This feature facilitates the capture and integration of multiscale features from the image or time series, enhancing the model's ability to handle complex classification tasks. The ability of selective kernels to adapt their receptive field in response to input features sets them apart and provides a more flexible and robust approach to image classification.

One can argue that InceptionTime is competitive on a lot of benchmark dataset for time series classification because of its multi scale feature extraction capabilities, indeed it uses different sizes of convolutional filter to capture those patterns of different scale. The selective kernels allows the model to have convolutional filters of different size thus being able to capture differnt scaled features. More importantly , as per the squeeze and excite networks, it can choose which filters to use. This make the selective kernel a key asset to learn complex patterns from time series. Also our experiences shows that it can in certain times surpass previous method for time series classification.

The efficacy of InceptionTime across numerous benchmark datasets for time series classification can be largely attributed to its multiscale feature extraction capabilities. It employs convolutional filters of varying sizes to capture patterns at different scales, thus providing a comprehensive view of the data.

To have better performances they used in the paper one 3x3 convolution and instead of using the 5x5 convolution they used a 3x3 convolution with a dilation rate of 2. Dilated convolution, a technique that allows CNNs to augment their receptive field without an increase in computational complexity or parameter count. The dilation rate, indicating the spacing between kernel values, is a crucial factor in this process. For instance, a dilated convolution with a rate of two and a kernel size of 3 will achieve the same receptive field as a 5 convolution, but without the added parameters and complexity. This efficient expansion of the receptive field is particularly advantageous in tasks requiring larger contextual information.

Crucially, similar to squeeze-and-excitation networks, selective kernels have the unique ability to modulate the use of filters. This ability provides the model with a high degree of learning adaptability, allowing it to focus on the most informative filters based on the given context. As a result, selective kernels emerge as a critical component in decoding complex patterns from time series data.

We have integrated selective kernels into the convolutional layers of the MLSTM-FCN network, giving rise to a variant named MLSTM-FCN(SK). Specifically, we employed kernel sizes of 5, 7, and 17 with a dilation rate of 2. This configuration results in receptive fields of 9, 13, and 33, respectively. We chose these particular parameters to align with the well-known receptive fields of InceptionTime, providing a familiar and proven framework as the basis for our innovative approach.

Our empirical evaluations confirm the theoretical advantages posited earlier. In several scenarios, we observed that the inclusion of selective kernels led to models that outperformed traditional methods for time series classification. These results highlight the potency of selective kernels in improving overall classification performance, thereby demonstrating their value as a powerful tool in the time series analysis toolkit. These results suggest that selective kernels could bring substantial improvements to time series classification tasks, further reinforcing their potential for wider adoption in the field.

Chapter 6

Conclusion

This Ph.D. project is an exploration and advancement of applied artificial intelligence, specifically in the field of predictive maintenance. The project departs from traditional, often complex Remaining Useful Life (RUL) estimation techniques and instead focuses on developing a highly scalable and novel explanatory tool for multivariate time series, a task that has been relatively unexplored and difficult to implement effectively. At the beginning of the thesis, we highlighted the lack of standard methodologies in the field of predictive maintenance and the limitations of currently available public datasets. Our research also revealed the limitations associated with traditional RUL prediction approaches and emphasized the importance of effective feature extraction. By shifting our focus in this more focused and promising direction, we were able to lay the foundation for a new wave of research in predictive maintenance.

In the first focus of our research - dimension reduction for time series - we established the superiority and utility of Variational Autoencoders (VAEs) over other popular dimension reduction methods, including the wavelet transform and Functional Principal Component Analysis. Our results highlighted the robustness and high compression capabilities of VAEs, especially those with convolutional neural network architectures, when dealing with different ECG data sets, even under noisy conditions. In addition, we explored the impact of different VAE architectures on their dimension reduction capabilities, providing invaluable guidance for selecting appropriate architectures for specific applications. In our second area of investigation-counterfactual explanation for multivariate time series-we developed an innovative method that strategically bifurcates the VAE latent space using a contrastive constraint. This approach resulted in the generation of partially ordered latent spaces and consequently produced efficient and sparsely valid counterfactual examples. This contributed to a deeper understanding of anomalies within multivariate time

series. Our validation on real-world ECG datasets underscored the potential of this new technique to improve the interpretability of multivariate time series predictions. This part of our research goes beyond predictive maintenance and introduces a new method for generating counterfactuals in an area where previous methods have struggled to operate efficiently on large, highly diverse datasets. This novel approach could prove critical for multivariate time series datasets, particularly in cases where expert knowledge is scarce. It helps users to better understand complex interactions within the dataset.

In our third and final research area - Explainable Predictive Maintenance - we adapted the CVAE approach to the complexities of predictive maintenance. By emphasizing the concept of "neighboring" lifecycle stages within an asset's lifetime, our strategy proved effective even in the presence of censored data, providing greater flexibility in data management. Our comparison with conventional time series classification models and a standard CVAE augmented with an anomaly detection algorithm demonstrated the superior performance of our adapted CVAE method in several predictive maintenance tasks. Indeed, the dual benefits of this new method are noteworthy. First, it introduces a robust technique for predicting future failures, especially under challenging conditions with censored data. Second, the method provides explainable results that not only gain wider acceptance, but also foster a deeper understanding of how equipment evolves over time. This knowledge can potentially accelerate improvement cycles, which is of great benefit to the organization. As a result, our research not only advances the field of predictive maintenance, but also provides practical, actionable strategies for organizations striving for efficiency and excellence in equipment maintenance.

While the use of CVAE for early failure prediction has yielded commendable results, it has also revealed several avenues for further research. The method, as efficient as it is, has a significant computational burden as the model's "energy" or resources are used for the explanatory component. In essence, the method requires the training of two models - the encoder and the decoder - which may not always be necessary or feasible. For example, in situations where explainability is not a mandatory requirement, or when a more complex model that cannot be configured as a VAE needs to be trained, alternative strategies may be preferable. In such scenarios, it would be interesting to explore the use of contrastive loss as a pre-training loss for more complex models. This approach could still take advantage of CVAE training: the use of censored data and the elimination of the need to compute the RUL at this stage. However, it would avoid the need to regenerate the time series, thereby saving computational resources.

It would be interesting to determine the applicability of the contrastive loss function as a pre-training step for more sophisticated models. Exploring such an avenue could provide valuable insights into the potential reduction in computational complexity and the impact on predictive performance. The feasibility of retaining the benefits of the current approach, such as the use of censored data and the avoidance of RUL computation at this stage, in this new methodology is also an exciting area of study. Solving this puzzle could reveal potential trade-offs and highlight an optimal balance between retaining these benefits and moving to a more complex model. We also envision the potential for significant advances in computational efficiency through the proposed technique. The factors that would influence this efficiency gain, and the methods to effectively manage them, could form the basis of interesting explorations in machine learning and predictive maintenance.

The use of contrastive pre-training is not yet commonplace in predictive maintenance, primarily because it requires extensive training data to achieve optimal performance. However, as industries increasingly adopt data collection for predictive maintenance, the use of this method may become more widespread. This transition may be accelerated by the emergence of large, publicly available data sets in the future. It's a clear trend that points to a potential norm in the coming years as industries recognize the undeniable benefits of using comprehensive data to improve their predictive maintenance capabilities. Beyond these specific areas of interest, optimizing the technique across different data sets and domains holds great promise. Given that the effectiveness of a model is closely tied to its adaptability to a wide range of scenarios, this line of inquiry could shed light on how our approach can be tailored and fine-tuned for datasets with different characteristics and levels of complexity.

This research journey unfolded in the highly competitive landscape of predictive maintenance, maintaining its exploratory ethos throughout. In a field teeming with diverse techniques to address predictive maintenance issues, our study has been characterized by a persistent pursuit of novel and results-oriented approaches. Such an exploratory path may not have been the path of least resistance, but it allowed us to revisit a well-established problem with a fresh, innovative perspective. In summary, this research, while deeply rooted in the evolving context of predictive maintenance, ventured into uncharted territory in an effort to uncover innovative solutions. By choosing to deviate from the well-trodden path, we were able to approach a familiar problem with a fresh lens. This approach not only led to the development of new, more efficient and explainable methods, but also opened up promising avenues for further exploration and innovation. The journey may not

have been the easiest, but it has certainly been enlightening and rewarding.

Bibliography

Addison, P. S. (2005). Wavelet transforms and the ecg: A review. *Physiological measurement*, *26*(5), R155.

Alday, E. A. P., Gu, A., Shah, A. J., Robichaux, C., Wong, A.-K. I., Liu, C., Liu, F., Rad, A. B., Elola, A., Seyedi, S., et al. (2020). Classification of 12-lead ecgs: The physionet/computing in cardiology challenge 2020. *Physiological measurement*, *41*(12), 124003.

Antoniadis, A., Brossat, X., Cugliari, J., & Poggi, J.-M. (2013). Clustering functional data using wavelets. *International Journal of Wavelets, Multiresolution and Information Processing*, *11*(01), 1350003.

Arias Chao, M., Kulkarni, C., Goebel, K., & Fink, O. (2021). Aircraft engine run-to-failure dataset under real flight conditions for prognostics and diagnostics. *Data*, *6*(1), 5.

Assaf, R., Giurgiu, I., Bagehorn, F., & Schumann, A. (2019). Mtex-cnn: Multivariate time series explanations for predictions with convolutional neural networks. *2019 IEEE International Conference on Data Mining (ICDM)*, 952–957.

Ates, E., Aksar, B., Leung, V. J., & Coskun, A. K. (2021). Counterfactual explanations for multivariate time series. *2021 International Conference on Applied Artificial Intelligence (ICAPAI)*, 1–8.

Bahri, O., Boubrahimi, S. F., & Hamdi, S. M. (2022). Shapelet-based counterfactual explanations for multivariate time series. *arXiv preprint arXiv:2208.10462*.

Balasubramanian, R., Sharpe, S., Barr, B., Wittenbach, J., & Bruss, C. B. (2020). Latent-cf: A simple baseline for reverse counterfactual explanations. *arXiv preprint arXiv:2012.09301*.

Barreyre, C., Laurent, B., Loubes, J.-M., Boussouf, L., & Cabon, B. (2019). Multiple testing for outlier detection in space telemetries. *IEEE Transactions on Big Data*, *6*(3), 443–451.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1798–1828.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., & Raffel, C. A. (2019). Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, *32*.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). Lof: Identifying density-based local outliers. *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 93–104.

Cai, R., Li, Z., Wei, P., Qiao, J., Zhang, K., & Hao, Z. (2019). Learning disentangled semantic representation for domain adaptation. *IJCAI: proceedings of the conference*, *2019*, 2060.

Canizo, M., Onieva, E., Conde, A., Charramendieta, S., & Trujillo, S. (2017). Real-time predictive maintenance for wind turbines using big data frameworks. *2017 ieee international conference on prognostics and health management (icphm)*, 70–77.

Carvalho, T. P., Soares, F. A., Vita, R., Francisco, R. d. P., Basto, J. P., & Alcalá, S. G. (2019). A systematic literature review of machine learning methods applied to predictive maintenance. *Computers & Industrial Engineering*, *137*, 106024.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, *41*(3), 1–58.

Chang, S. G., Yu, B., & Vetterli, M. (2000). Adaptive wavelet thresholding for image denoising and compression. *IEEE transactions on image processing*, *9*(9), 1532–1546.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *International conference on machine learning*, 1597–1607.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., & Hinton, G. E. (2020). Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, *33*, 22243–22255.

Çınar, Z. M., Abdussalam Nuhu, A., Zeeshan, Q., Korhan, O., Asmael, M., & Safaei, B. (2020). Machine learning in predictive maintenance towards sustainable smart manufacturing in industry 4.0. *Sustainability*, *12*(19), 8211.

Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*, *36*(3), 287–314.

Dai, B., Wang, Y., Aston, J., Hua, G., & Wipf, D. (2017). Hidden talents of the variational autoencoder. *arXiv preprint arXiv:1706.05148*.

Davari, N., Veloso, B., Ribeiro, R. P., Pereira, P. M., & Gama, J. (2021). Predictive maintenance based on anomaly detection using deep learning for air production unit in the railway industry. *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, 1–10.

Delaney, E., Greene, D., & Keane, M. T. (2021). Instance-based counterfactual explanations for time series classification. *International Conference on Case-Based Reasoning*, 32–47.

Dhurandhar, A., Chen, P.-Y., Luss, R., Tu, C.-C., Ting, P., Shanmugam, K., & Das, P. (2018). Explanations based on the missing: Towards contrastive explanations with pertinent negatives. *Advances in neural information processing systems*, *31*.

Di, C.-Z., Crainiceanu, C. M., Caffo, B. S., & Punjabi, N. M. (2009). Multilevel functional principal component analysis. *The annals of applied statistics*, *3*(1), 458.

Donoho, D. L., & Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *biometrika*, *81*(3), 425–455.

El Mejdoubi, A., Chaoui, H., Sabor, J., & Gualous, H. (2017). Remaining useful life prognosis of supercapacitors under temperature and voltage aging conditions. *IEEE Transactions on Industrial Electronics*, *65*(5), 4357–4367.

Fauvel, K., Lin, T., Masson, V., Fromont, É., & Termier, A. (2021). Xcm: An explainable convolutional neural network for multivariate time series classification. *Mathematics*, *9*(23), 3137.

Gao, W., Zhang, L., Huang, W., Min, F., He, J., & Song, A. (2021). Deep neural networks for sensor-based human activity recognition using selective kernel convolution. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1–13.

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., & Lee, S. (2019). Counterfactual visual explanations. *International Conference on Machine Learning*, 2376–2384.

Greff, K., Srivastava, R. K., Koutnık, J., Steunebrink, B. R., & Schmidhuber, J. (2016). Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, *28*(10), 2222–2232.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. (2020). Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, *33*, 21271–21284.

Guidotti, R. (2022). Counterfactual explanations and how to find them: Literature review and benchmarking. *Data Mining and Knowledge Discovery*, 1–55.

Harman, H. H. (1976). *Modern factor analysis*. University of Chicago press.

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Hering, J., Metzenthin, E., & Zenner, A. (2020). Lime for time. *https:// github.com/ emanuel-metzenthin/ Lime-For-Time*. https://github.com/HobbitLong/SupContrast.

Hilton, M. L. (1997). Wavelet and wavelet packet compression of electrocardiograms. *IEEE Transactions on Biomedical Engineering*, *44*(5), 394–402.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Huuhtanen, T., & Jung, A. (2018). Predictive maintenance of photovoltaic panels via deep learning. *2018 ieee data science workshop (dsw)*, 66–70.

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P.-A. (2019). Deep learning for time series classification: A review. *Data mining and knowledge discovery*, *33*(4), 917–963.

Ismail Fawaz, H., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D. F., Weber, J., Webb, G. I., Idoumghar, L., Muller, P.-A., & Petitjean, F. (2020). Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, *34*(6), 1936–1962.

Jacques, J., & Preda, C. (2014). Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis*, *71*, 92–106.

Jakubowski, J., Stanisz, P., Bobek, S., & Nalepa, G. J. (2021). Anomaly detection in asset degradation process using variational autoencoder and explanations. *Sensors*, *22*(1), 291.

Jardine, A. K., Lin, D., & Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical systems and signal processing*, *20*(7), 1483–1510.

Javed, A., Lee, B. S., & Rizzo, D. M. (2020). A benchmark study on time series clustering. *Machine Learning with Applications*, *1*, 100001.

Jia, F., Lei, Y., Guo, L., Lin, J., & Xing, S. (2018). A neural network constructed by deep learning technique and its application to intelligent fault diagnosis of machines. *Neurocomputing*, *272*, 619–628.

Jiang, Y., Lyu, Y., Wang, Y., & Wan, P. (2020). Fusion network combined with bidirectional lstm network and multiscale cnn for remaining useful life estimation. *2020 12th International Conference on Advanced Computational Intelligence (ICACI)*, 620–627.

Kang, Z., Catal, C., & Tekinerdogan, B. (2021). Remaining useful life (rul) prediction of equipment in production lines using artificial neural networks. *Sensors*, *21*(3), 932.

Karim, F., Majumdar, S., Darabi, H., & Harford, S. (2019). Multivariate lstm-fcns for time series classification. *Neural Networks*, *116*, 237–245.

Karimi, A.-H., Barthe, G., Schölkopf, B., & Valera, I. (2020). A survey of algorithmic recourse: Definitions, formulations, solutions, and prospects. *arXiv preprint arXiv:2010.04050*.

Karlsson, I., Rebane, J., Papapetrou, P., & Gionis, A. (2018). Explainable time series tweaking via irreversible and reversible temporal transformations. *2018 IEEE International Conference on Data Mining (ICDM)*, 207–216.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., & Krishnan, D. (2020). Supervised contrastive learning. *Advances in Neural Information Processing Systems*, *33*, 18661–18673.

Kiangala, K. S., & Wang, Z. (2020). An effective predictive maintenance framework for conveyor motors using dual time-series imaging and convolutional neural network in an industry 4.0 environment. *Ieee Access*, *8*, 121033–121049.

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

Lacoste, A., Luccioni, A., Schmidt, V., & Dandres, T. (2019). Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Laredo, D., Chen, Z., Schütze, O., & Sun, J.-Q. (2019). A neural network-evolutionary computational framework for remaining useful life estimation of mechanical systems. *Neural networks*, *116*, 178–187.

Li, C., Zheng, C., & Tai, C. (1995). Detection of ecg characteristic points using wavelet transforms. *IEEE Transactions on biomedical Engineering*, *42*(1), 21–28.

Li, H., Wang, Y., Zhao, P., Zhang, X., & Zhou, P. (2015). Cutting tool operational reliability prediction based on acoustic emission and logistic regression model. *Journal of Intelligent Manufacturing*, *26*, 923–931.

Li, K.-L., Huang, H.-K., Tian, S.-F., & Xu, W. (2003). Improving one-class svm for anomaly detection. *Proceedings of the 2003 international conference on machine learning and cybernetics (IEEE Cat. No. 03EX693)*, *5*, 3077–3081.

Li, X., Wang, W., Hu, X., & Yang, J. (2019). Selective kernel networks. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 510–519.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. *2008 eighth ieee international conference on data mining*, 413–422.

Liu, F., Liu, C., Zhao, L., Zhang, X., Wu, X., Xu, X., Liu, Y., Ma, C., Wei, S., He, Z., et al. (2018). An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, *8*(7), 1368–1373.

Liu, Y. (2009). Feature extraction and dimensionality reduction for mass spectrometry data. *Computers in Biology and Medicine*, *39*(9), 818–823.

Lu, W., Wang, X., Yang, C., & Zhang, T. (2015). A novel feature extraction method using deep neural network for rolling bearing fault diagnosis. *The 27th Chinese Control and Decision Conference (2015 CCDC)*, 2427–2431.

Mahmud, M. S., Huang, J. Z., & Fu, X. (2020). Variational autoencoder-based dimensionality reduction for high-dimensional small-sample data classification. *International Journal of Computational Intelligence and Applications*, *19*(01), 2050002.

Mallat, S. (1999). *A wavelet tour of signal processing*. Elsevier.

Markou, M., & Singh, S. (2003). Novelty detection: A review—part 1: Statistical approaches. *Signal processing*, *83*(12), 2481–2497.

Mathew, V., Toby, T., Singh, V., Rao, B. M., & Kumar, M. G. (2017). Prediction of remaining useful lifetime (rul) of turbofan engine using machine learning. *2017 IEEE international conference on circuits and systems (ICCS)*, 306–311.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Mobley, R. K. (2002). *An introduction to predictive maintenance*. Elsevier.

Muelas, D., Garcıa-Dorado, J. L., de Vergara, J. E. L., & Aracil, J. (2017). Application of functional feature extraction to the compression of network time series. *2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, 592–595.

Nectoux, P., Gouriveau, R., Medjaher, K., Ramasso, E., Chebel-Morello, B., Zerhouni, N., & Varnier, C. (2012). Pronostia: An experimental platform for bearings accelerated degradation tests. *IEEE International Conference on Prognostics and Health Management, PHM'12.*, 1–8.

Omshi, E. M., Grall, A., & Shemehsavar, S. (2020). A dynamic auto-adaptive predictive maintenance policy for degradation with unknown parameters. *European Journal of Operational Research*, *282*(1), 81–92.

Pandey, K., Mukherjee, A., Rai, P., & Kumar, A. (2022). Diffusevae: Efficient, controllable and high-fidelity generation from low-dimensional latents. *arXiv preprint arXiv:2201.00308*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, *32*, 8026–8037.

Pimentel, M. A., Clifton, D. A., Clifton, L., & Tarassenko, L. (2014). A review of novelty detection. *Signal Processing*, *99*, 215–249.

Poels, Y., & Menkovski, V. (2022). Vae-ce: Visual contrastive explanation using disentangled vaes. *International Symposium on Intelligent Data Analysis*, 237–250.

Prytz, R., Nowaczyk, S., Rögnvaldsson, T., & Byttner, S. (2015). Predicting the need for vehicle compressor repairs using maintenance records and logged vehicle data. *Engineering applications of artificial intelligence*, *41*, 139–150.

Rabbani, M., & Joshi, R. (2002). An overview of the jpeg 2000 still image compression standard. *Signal processing: Image communication*, *17*(1), 3–48.

Rahhal, J. S., & Abualnadi, D. (2020). Iot based predictive maintenance using lstm rnn estimator. *2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, 1–5.

Ran, Y., Zhou, X., Lin, P., Wen, Y., & Deng, R. (2019). A survey of predictive maintenance: Systems, purposes and approaches. *arXiv preprint arXiv:1912.07383*.

Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. *International conference on machine learning*, 1278–1286.

Riad, A., Elminir, H., & Elattar, H. (2010). Evaluation of neural networks in the subject of prognostics as compared to linear regression model. *International Journal of Engineering & Technology*, *10*(6), 52–58.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Saxena, A., Goebel, K., Simon, D., & Eklund, N. (2008). Damage propagation modeling for aircraft engine run-to-failure simulation. *2008 international conference on prognostics and health management*, 1–9.

Schölkopf, B., Williamson, R. C., Smola, A., Shawe-Taylor, J., & Platt, J. (1999). Support vector method for novelty detection. *Advances in neural information processing systems*, *12*.

Serradilla, O., Zugasti, E., Rodriguez, J., & Zurutuza, U. (2022). Deep learning models for predictive maintenance: A survey, comparison, challenges and prospects. *Applied Intelligence*, *52*(10), 10934–10964.

Shang, H. L. (2014). A survey of functional principal component analysis. *AStA Advances in Statistical Analysis*, *98*(2), 121–142.

Shukla, B., Fan, I.-S., & Jennions, I. (2020). Opportunities for explainable artificial intelligence in aerospace predictive maintenance. *PHM Society European Conference*, *5*(1), 11–11.

Singstad, B., & Tronstad, C. (2020). Convolutional neural network and rule-based algorithms for classifying 12-lead ecgs. *2020 Computing in Cardiology*, 1–4.

Smith, W. A., & Randall, R. B. (2015). Rolling element bearing diagnostics using the case western reserve university data: A benchmark study. *Mechanical systems and signal processing*, *64*, 100–131.

Strodthoff, N., Wagner, P., Schaeffter, T., & Samek, W. (2020). Deep learning for ecg analysis: Benchmarks and insights from ptb-xl. *arXiv preprint arXiv:2004.13701*.

Su, C., Li, L., & Wen, Z. (2020). Remaining useful life prediction via a variational autoencoder and a time-window-based sequence neural network. *Quality and Reliability Engineering International*, *36*(5), 1639–1656.

Tarpey, T., & Kinateder, K. K. (2003). Clustering functional data. *Journal of classification*, *20*(1), 093–114.

Teng, W., Zhang, X., Liu, Y., Kusiak, A., & Ma, Z. (2016). Prognosis of the remaining useful life of bearings in a wind turbine gearbox. *Energies*, *10*(1), 32.

Tian, Y. (2020). Supcontrast. *https://github.com/HobbitLong/SupContrast*. https://github.com/HobbitLong/SupContrast.

Todo, W., Selmani, M., Laurent, B., & Loubes, J.-M. (2023). Counterfactual explanation for multivariate times series using a contrastive variational autoencoder. *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Um, T. T., Pfister, F. M., Pichler, D., Endo, S., Lang, M., Hirche, S., Fietzek, U., & Kulić, D. (2017). Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. *Proceedings of the 19th ACM international conference on multimodal interaction*, 216–220.

Veloso, B., Ribeiro, R. P., Gama, J., & Pereira, P. M. (2022). The metropt dataset for predictive maintenance. *Scientific Data*, *9*(1), 764.

Verma, S., Boonsanong, V., Hoang, M., Hines, K. E., Dickerson, J. P., & Shah, C. (2020). Counterfactual explanations and algorithmic recourses for machine learning: A review. *arXiv preprint arXiv:2010.10596*.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, *31*, 841.

Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F. I., Samek, W., & Schaeffter, T. (2020). Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, *7*(1), 1–15.

Wang, Q., Bu, S., & He, Z. (2020). Achieving predictive and proactive maintenance for high-speed railway power equipment with lstm-rnn. *IEEE Transactions on Industrial Informatics*, *16*(10), 6509–6517.

Wang, Z., Samsten, I., Mochaourab, R., & Papapetrou, P. (2021). Learning time series counterfactuals via latent space representations. *International Conference on Discovery Science*, 369–384.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, *2*(1-3), 37–52.

Woo, G., Liu, C., Sahoo, D., Kumar, A., & Hoi, S. (2022). Cost: Contrastive learning of disentangled seasonal-trend representations for time series forecasting. *arXiv preprint arXiv:2202.01575*.

Xia, M., Li, T., Shu, T., Wan, J., De Silva, C. W., & Wang, Z. (2018). A two-stage approach for the remaining useful life prediction of bearings using deep neural networks. *IEEE Transactions on Industrial Informatics*, *15*(6), 3703–3711.

Yan, J., & Lee, J. (2005). Degradation assessment and fault modes classification using logistic regression.

Yang, H., & Desell, T. (2022). A large-scale annotated multivariate time series aviation maintenance dataset from the ngafid. *arXiv preprint arXiv:2210.07317*.

Yang, H., LaBella, A., & Desell, T. (2022). Predictive maintenance for general aviation using convolutional transformers. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(11), 12636–12642.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. *Proceedings of the IEEE/CVF international conference on computer vision*, 6023–6032.

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). Mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhang, W., Yang, D., & Wang, H. (2019). Data-driven methods for predictive maintenance of industrial equipment: A survey. *IEEE Systems Journal*, *13*(3), 2213–2227.

Zheng, Z., & Sun, L. (2019). Disentangling latent space for vae by label relevant/irrelevant dimensions. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12192–12201.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.