République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Abderrahmane Mira de Béjaïa

Faculté des Sciences Exactes
Département d'Informatique

Mémoire de Master Recherche
Informatique
Option : Systèmes d'Information Avancés

Thème

Prédiction des prix au comptant et à terme du pétrole par l'intégration de méthodes non linéaires de réduction de la dimensionnalité et de techniques de machine learning

Présenté par :

HASSISSI Amir

Soutenu le 27 juin à 10h30 devant le jury composé de :

Présidente	Mme S. Ait Kaci Azzou	U. A/Mira Béjaïa
Promotrice	Mme D. Boukredera	U. A/Mira Béjaïa
Co-promoteur	M. S. Goudjil	U. A/Mira Béjaïa
Examinateur	M. N.E. Mouffok	U. A/Mira Béjaïa

Promotion 2023 – 2024

Remerciements

En premier lieu, je remercie le bon Dieu tout-puissant pour m'avoir donné la force et le courage qui m'ont permis de réaliser ce travail.

Je tiens à exprimer ma sincère reconnaissance et ma profonde gratitude envers les personnes suivantes :

À Madame **D. BOUKREDERA**, en tant que mon encadrante, pour son enthousiasme, sa patience, son soutien, ainsi que pour tous ses précieux conseils et informations tout au long de la réalisation de ce travail. Également, en tant qu'enseignante du module « Apprentissage Automatique », pour m'avoir fait découvrir et comprendre ce domaine. Grâce à ses excellentes séances de cours, j'ai non seulement développé une véritable passion pour le sujet, mais aussi l'ambition de m'y approfondir davantage.

À Monsieur **S. GOUDJIL**, mon co-encadrant, pour m'avoir fourni des informations précieuses et un solide bagage en économie, une partie essentielle et indispensable de mon thème, ainsi que pour son enthousiasme et son soutien tout au long de la réalisation de ce travail.

À Madame **S. AIT KACI AZZOU** et à Monsieur **N.E. MOUFFOK**, pour leur enseignement exceptionnel dans l'un de leurs modules que j'ai eu la chance de suivre, ainsi que pour avoir accepté d'examiner et d'évaluer ce mémoire.

À mes chers parents pour leur soutien inconditionnel. Grâce à eux, la réalisation de ce travail et de toutes mes réussites a été rendue possible.

À tous mes enseignants qui ont contribué à ma formation académique.

Enfin, un sincère remerciement à tous ceux qui ont contribué, directement ou indirectement, à la réussite de ce travail.

Dédicaces

Je dédie ce modeste travail :

À mes chers parents

À mes frères et à ma sœur

À mes enseignants

À mes amis

Table des Matières

Table des Figures

Liste des Tableaux

Liste des Algorithmes

Liste des Acronymes

AR	AutoRegressive
ARIMA	AutoRegressive Integrated Moving Average
ARMA	AutoRegressive Moving Average
DME	Dubaï Mercantile Exchange
GRU	Gated Recurrent Unit
ICE	Intercontinental Exchange
KPCA	Kernel Principal Component Analysis
LSTM	Long Short-Term Memory
MA	Moving Average
MAE	Mean Absolute Error
NYMEX	New York Mercantile Exchange
PCA	Principal Component Analysis
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
UMAP	Uniform Manifold Approximation and Projection
WTI	West Texas Intermediate

Introduction Générale

Le pétrole reste la pierre angulaire de l'économie mondiale, essentielle à la structure même de notre mode de vie moderne. En 2021, le pétrole représentait environ 33% de la consommation mondiale d'énergie primaire [39]. D'après le rapport sur le marché pétrolier publié en novembre 2023 par l'Agence Internationale de l'Énergie (AIE) [36], la demande mondiale quotidienne était d'environ 102 millions de barils cette année-là.

L'électricité, omniprésente dans nos activités quotidiennes, tire souvent son énergie des centrales thermiques, alimentées en grande partie par les énergies fossiles, dont le pétrole. Environ 59,9% de cette production mondiale était issue des énergies fossiles au premier semestre 2023 [40]. L'absence de pétrole risquerait de paralyser nos machines, de plonger nos maisons et nos infrastructures dans l'obscurité, et de rendre nos appareils électroniques muets et inactifs.

En outre, le pétrole est à l'origine de nombreux produits dont nous dépendons, tels que le plastique. En 2023, la production mondiale de plastique a atteint 460 millions de tonnes, un chiffre qui pourrait se multiplier par trois d'ici 2060 [34], une grande partie de cette production reposant sur les dérivés du pétrole. De nos emballages alimentaires à nos appareils médicaux sophistiqués, en passant par nos gadgets électroniques et même nos vêtements, le pétrole imprègne notre quotidien de manière indéniable.

Ajoutons à cela la fonction vitale qu'il remplit dans nos moyens de transport. Nos voitures, navires et avions ne peuvent avancer sans le carburant issu du pétrole. En 2021, le secteur des transports a utilisé environ 62% de la consommation mondiale de pétrole [43]. Il est le carburant qui alimente notre mobilité, reliant les gens et les marchandises à travers le monde.

Cette importance vitale confère au pétrole une place cruciale sur l'échiquier économique mondial. Les pays producteurs en tirent des revenus inestimables, avec des exportations pétrolières représentant jusqu'à 90% des revenus nationaux pour certains d'entre eux [31]. Tandis que les grandes puissances économiques s'appuient étroitement sur sa disponibilité constante.

Cependant, cette dépendance généralisée expose les pays et les sociétés à une fragilité inquiétante face aux fluctuations des prix du pétrole. Les chutes abruptes peuvent engendrer des répercussions économiques, sociales et politiques néfastes, telles que l'inflation et le chômage. Par exemple, une chute de 50% des prix du pétrole en 2014-2015 a entraîné une réduction significative des revenus pour de nombreux pays producteurs, provoquant des crises économiques.

De même, des hausses de prix peuvent réduire les revenus des grandes puissances économiques en raison de l'augmentation des coûts de production et de transport. Les prix du pétrole sont très exposés à des fluctuations en raison de divers facteurs, notamment l'offre et la demande. Tous les pays ne sont pas producteurs de pétrole, et ceux qui le sont,

comme les membres de l'OPEP, contrôlent les prix en gérant les quantités produites. Par exemple, si les prix chutent, les pays producteurs de pétrole, tels que ceux de l'OPEP, peuvent réduire la production pour diminuer l'offre et ainsi faire remonter les prix.

De même, si les pays consommateurs de pétrole connaissent une crise économique, ils réduiront leur production industrielle et leur consommation de pétrole, ce qui conduit à une baisse de la demande et, par conséquent, à une baisse des prix du pétrole.

D'autres événements imprévus peuvent également causer des fluctuations des prix du pétrole. Parmi ces événements, nous comptons les guerres, les conditions météorologiques extrêmes, et les tensions politiques. Ces facteurs peuvent perturber la production, le transport et la consommation de pétrole, influençant ainsi les prix sur les marchés mondiaux.

La prévision des prix du pétrole est donc essentielle pour anticiper les fluctuations du marché, informer les stratégies d'investissement et assurer la stabilité économique mondiale.

Avant le développement de l'intelligence artificielle et l'apparition des modèles de machine learning, des modèles économétriques tels que ARIMA étaient utilisés pour prédire les prix du pétrole. Cependant, en raison de la relation non linéaire entre les facteurs influençant les prix du pétrole et les prix eux-mêmes, ces modèles économétriques se sont révélés inefficaces.

L'apparition des modèles de machine learning a amélioré les prévisions des prix du pétrole. Notamment, les Réseaux de Neurones Récurrents (RNN) sont apparus comme des outils prometteurs. Les RNN se distinguent par leur capacité à exploiter les données séquentielles et à apprendre des tendances historiques des prix du pétrole, en intégrant ces connaissances dans leurs prévisions. De plus, les RNN sont particulièrement efficaces pour capturer les relations non linéaires complexes présentes dans les données de prix du pétrole, ce qui les rend particulièrement adaptés à cette tâche.

Cependant, les RNN rencontrent des difficultés avec les séquences longues, notamment les problèmes de « Exploding Gradient » et de « Vanishing Gradient ». Pour surmonter ces limitations, les modèles Long Short-Term Memory (LSTM) ont été développés. Les LSTM, avec leurs mécanismes de portes élaborés, sont particulièrement efficaces pour maintenir et mettre à jour la mémoire sur de longues séquences.

En outre, les Gated Recurrent Units (GRU), une variante optimisée des LSTM, offrent une alternative computationnellement plus efficace tout en maintenant de bonnes performances. Les LSTM et les GRU se sont ainsi avérés particulièrement efficaces pour la prédiction des prix du pétrole en exploitant des relations complexes dans les données temporelles.

Malgré les performances des modèles RNN et de leurs variantes LSTM et GRU, ils peuvent être renforcés afin d'obtenir des résultats encore meilleurs.

Parmi les approches proposées pour renforcer ces types de réseaux de neurones, Lei Yan et al. [47] ont démontré que certaines techniques de réduction de la dimensionnalité, telles que LLE, peuvent améliorer leurs performances.

Cependant, il est crucial de choisir judicieusement parmi les nombreuses techniques de réduction de la dimensionnalité disponibles afin de déterminer celles qui permettront aux modèles RNN et à leurs variantes de réaliser les meilleures prédictions. Dans ce mémoire, nous explorons deux de ces techniques, à savoir l'Analyse en Composantes Principales à Noyau (KPCA) et l'Approximation Uniforme et Projection de Manifold (UMAP), toutes

deux non linéaires. L'objectif est de déterminer quelles techniques de réduction de la dimensionnalité, telles que l'Analyse en Composantes Principales à Noyau (KPCA) et l'Approximation Uniforme et Projection de Manifold (UMAP), renforcent le plus efficacement les RNN et leurs variantes pour améliorer la prédiction des prix du pétrole et conduire ainsi à des résultats optimaux. Ensuite, nous comparerons les performances de nos modèles qui combinent les RNN et leurs variantes LSTM et GRU avec les deux techniques de réduction de la dimensionnalité KPCA et UMAP. Nous évaluerons également ces performances par rapport aux modèles RNN, LSTM et GRU sans réduction de la dimensionnalité, ainsi qu'au modèle le plus performant de Yan [47], qui combine LLE et RNN.

Ce mémoire est structuré en quatre chapitres :

- **Chapitre 1 :** Ce chapitre traite des généralités et des notions de base sur le marché pétrolier.
- **Chapitre 2 :** Il explore les notions de base des séries temporelles, des RNN et de leurs variantes, ainsi que des techniques de réduction de la dimensionnalité, notamment celles que nous allons aborder dans notre approche.
- **Chapitre 3 :** Il examine quelques travaux antérieurs pertinents et récents sur l'utilisation du machine learning dans la prédiction des prix du pétrole.
- **Chapitre 4 :** Ce chapitre présente en détail les étapes de notre méthodologie, ainsi que les résultats obtenus lors de l'évaluation des modèles proposés et de leur comparaison avec les modèles RNN et leurs variantes sans réduction de la dimensionnalité, ainsi qu'avec le modèle le plus performant de [47].

I

Généralités sur le marché pétrolier

I.1 Introduction

Ce chapitre explore les aspects essentiels du marché pétrolier, des produits dérivés et des modèles économétriques utilisés pour analyser ces domaines. Le marché pétrolier est d'une importance capitale pour l'économie mondiale, avec ses fluctuations de prix influençant divers secteurs. Les produits dérivés, tels que les contrats à terme et les options, jouent un rôle crucial dans la gestion des risques liés aux prix du pétrole. Enfin, les modèles économétriques fournissent un cadre analytique pour comprendre et prédire les tendances du marché pétrolier, bien que ces modèles présentent également des limites.

I.2 Pétrole

Le pétrole est une ressource naturelle non renouvelable essentielle dans de nombreux domaines. Il est utilisé comme matière première pour la production d'une variété de produits, notamment les plastiques, les carburants, les lubrifiants et les produits chimiques. Principal carburant pour les transports, il est également utilisé dans la production d'autres formes d'énergie, dont l'électricité. Son rôle crucial dans l'économie lui a valu le surnom d'« or noir ».

I.2.1 Types de pétrole

Il existe plusieurs types de pétrole brut, chacun présentant ses propres caractéristiques distinctes en termes de densité, de viscosité, de teneur en soufre et d'autres paramètres. Ces différents types de pétrole brut sont négociés sur des marchés spécifiques à travers le

monde. Parmi ces types, certains sont largement reconnus comme des références principales sur les marchés pétroliers internationaux et servent de benchmarks dans l'évaluation des prix du pétrole [42]. Parmi ces types, figurent :

- **West Texas Intermediate (WTI) :** est un type de pétrole brut léger et doux extrait principalement dans la région du bassin permien au Texas, aux États-Unis. Il est négocié sur le New York Mercantile Exchange (NYMEX) [3] et représente l'un des principaux indices de référence pour les prix du pétrole aux États-Unis et en Amérique du Nord. Le prix du WTI est souvent utilisé comme indicateur de la santé économique et énergétique des États-Unis.
- **Brent :** est un type de pétrole brut léger et doux extrait au milieu de la mer du Nord, entre les îles Shetland, en Écosse, et la Norvège. Il est négocié sur l'ICE, ainsi que sur le NYMEX et les marchés de Rotterdam. Le prix du Brent sert de référence pour les prix du pétrole dans une partie de l'Europe, de l'Afrique et des pays du pourtour méditerranéen [11].
- **Dubaï Light :** est un type de pétrole brut produit dans le Golfe Persique. Le Dubaï Light est un pétrole brut plus lourd que le WTI et le Brent. Il est principalement utilisé comme référence pour les prix du pétrole dans la région asiatique [33]. Il est négocié sur le marché de l'énergie à Dubaï, connu sous le nom de Dubaï Mercantile Exchange (DME).
- **Arabian Light :** est un type de pétrole brut produit en Arabie saoudite. Il fait partie de la gamme des pétroles bruts légers produits par l'Arabie saoudite. Il est est souvent utilisé comme référence pour les prix du pétrole dans la région du Moyen-Orient [10].

I.3 Marché pétrolier

« Le marché pétrolier est avant tout le lieu de rencontre entre l'offre des producteurs et la demande des consommateurs. Le rapport de forces entre ces derniers permet d'établir le prix du pétrole, le plus souvent exprimé en dollar par baril. » [30]

I.3.1 Types de marché pétrolier

D'après [8], il existe un marché pétrolier physique (au comptant et à terme) et un marché à terme.

I.3.1.1 Marché physique

Le marché physique est un marché de gré à gré, où les transactions s'effectuent directement entre deux parties sans l'intervention d'un intermédiaire. Ce marché implique des échanges concrets où les acheteurs acquièrent du pétrole brut directement des vendeurs, souvent au travers de contrats précisant la quantité, la qualité, le lieu de livraison et la date de livraison.
Il existe deux variantes principales de ce marché :

- **Marché physique au comptant (spot) :** Ce marché implique la livraison immédiate ou quasi immédiate des produits. Il fait intervenir les sociétés pétrolières en tant que vendeurs, les raffineurs en tant qu'acheteurs, ainsi que des négociants présents des deux côtés (vendeur et acheteur).
- **Marché physique à terme (forward) :** Dans ce marché, s'effectue l'échange de cargaisons de pétrole à une date ultérieure avec un prix prédéterminé. Il permet d'assurer la production future des producteurs et l'approvisionnement des consommateurs.

I.3.1.2 Marché pétrolier à terme (marché des « futures »)

Le marché pétrolier à terme est un marché organisé, ce qui signifie que les transactions ont lieu de manière structurée et réglementée à travers une plateforme d'échange officielle comme les bourses. À la différence du marché physique où le pétrole brut est vendu et acheté en tant que produit physique, dans le marché à terme, le pétrole est négocié sur papier sous forme de contrats appelés « contrats à terme ». Ces contrats permettent d'acheter du pétrole à une date future prédéterminée et à un prix convenu aujourd'hui. Il est utilisé pour optimiser son portefeuille en recherchant un équilibre entre la rentabilité potentielle et la sécurité contre les risques. Plusieurs bourses négocient des contrats à terme sur le pétrole, telles que :

- **New York Mercantile Exchange (NYMEX) :** était une importante bourse de marchandises basée à New York, spécialisée dans le commerce de contrats à terme et d'options sur divers produits de base, notamment l'énergie (pétrole, gaz naturel), les métaux précieux (or, argent) et les produits agricoles. Fondée en 1872 sous le nom « Butter and Cheese Exchange », elle a été transformée en NYMEX en 1882. Le NYMEX était particulièrement connu pour ses contrats à terme sur le pétrole brut, notamment le West Texas Intermediate (WTI).
- **Intercontinental Exchange (ICE) :** est une importante bourse mondiale basée à Atlanta, en Géorgie, spécialisée dans le commerce de contrats à terme et d'options sur une variété de produits financiers et de matières premières. Fondée en 2000, ICE est connue pour ses marchés dans les domaines de l'énergie (notamment le pétrole et le gaz naturel, y compris le pétrole Brent), des métaux précieux (or, argent), des produits agricoles et d'autres actifs financiers.
- **Dubaï Mercantile Exchange (DME) :** est une bourse basée à Dubaï, aux Émirats arabes unis, spécialisée dans le commerce de contrats à terme sur le pétrole brut. Lancée en Juin 2007, le DME propose des contrats à terme sur le pétrole Oman, une variété de pétrole brut produite dans la région du golfe Persique.

I.3.2 Facteurs influents sur les prix du pétrole

Les prix du pétrole sont influencés par divers facteurs, comme mentionné par [20], qui peuvent les faire fluctuer à la hausse ou à la baisse. Parmi ces facteurs :

- **L'offre :** L'offre de pétrole influence le prix en fonction de l'équilibre entre la quantité de pétrole disponible sur le marché et la demande mondiale. Lorsque l'offre excède la demande, les prix ont tendance à baisser car les producteurs doivent concurrencer pour vendre leur pétrole. À l'inverse, si l'offre est limitée par des facteurs tels que des réductions de production ou des perturbations géopolitiques, cela peut entraîner une hausse des prix car la demande excède l'offre disponible. Ainsi, des fluctuations dans l'offre de pétrole, influencées par la production mondiale, les politiques des pays producteurs, les stocks et les événements géopolitiques, ont un impact direct sur les prix du pétrole.

- **La demande :** La demande de pétrole exerce une influence significative sur son prix en fonction de l'équilibre entre l'offre disponible et les besoins mondiaux. Lorsque la demande de pétrole augmente, en raison par exemple de la croissance économique, de l'industrialisation ou de politiques favorisant l'utilisation d'énergies fossiles, cela peut entraîner une hausse des prix si l'offre ne parvient pas à suivre le rythme. En revanche, une diminution de la demande due à des facteurs économiques défavorables ou à des politiques favorisant les alternatives énergétiques peut exercer une pression à la baisse sur les prix du pétrole. Ainsi, les fluctuations de la demande mondiale de pétrole jouent un rôle déterminant dans la dynamique des prix du pétrole sur le marché mondial.

- **La géopolitique :** Le cours du pétrole est fortement influencé par les tensions géopolitiques. En effet, ces tensions affectent la possibilité de prévoir la production et le prix du baril de pétrole.
  Ces sources de tension incluent les conflits armés, les tensions diplomatiques, l'instabilité politique, les menaces d'attentats et les projets nucléaires [20].

- **La force du dollar :** Le pétrole est généralement négocié en dollars américains sur les marchés mondiaux. Par conséquent, lorsque le dollar américain s'affaiblit par rapport à d'autres devises, le coût du pétrole en termes de ces autres devises diminue, ce qui peut stimuler la demande de pétrole. En revanche, un dollar fort rend le pétrole plus cher pour les acheteurs dans d'autres devises, ce qui peut potentiellement réduire la demande et exercer une pression à la baisse sur les prix du pétrole.

- **Les décisions de l'Organisation des Pays Exportateurs de pétrole et leurs alliés (OPEP+)** : Les décisions de l'OPEP+ ont un impact direct sur l'offre de pétrole et, par conséquent, sur les prix du marché mondial. Lorsque l'OPEP+ décide d'augmenter ou de réduire sa production de pétrole, cela influence l'offre globale sur le marché. Par exemple, si l'OPEP+ décide de réduire sa production pour soutenir les prix en cas de surabondance, cela peut entraîner une diminution de l'offre disponible, ce qui peut potentiellement faire augmenter les prix du pétrole. À l'inverse, une augmentation de la production par l'OPEP+ peut accroître l'offre et exercer une pression à la baisse sur les prix. Les décisions de l'OPEP+ sont donc surveillées de près par les marchés pétroliers car elles peuvent avoir un impact significatif sur la dynamique de l'offre et de la demande de pétrole à l'échelle mondiale.

- **Les catastrophes naturelles :** Les catastrophes naturelles, telles que les ouragans, les tempêtes ou les tremblements de terre, peuvent également influencer l'offre et donc les prix du pétrole. Ces événements peuvent perturber les infrastructures pétrolières, telles que les plateformes offshore, les pipelines et les raffineries, ce qui entraîne souvent une réduction temporaire de la production ou une perturbation de la distribution. Une diminution de l'offre résultant de telles catastrophes peut potentiellement faire monter les prix du pétrole si la demande reste stable ou augmente. De plus, les craintes anticipées de perturbations dans l'approvisionnement en raison de catastrophes imminentes peuvent également provoquer une hausse des prix du pétrole sur les marchés à terme. En résumé, les catastrophes naturelles peuvent avoir un impact imprévisible mais potentiellement significatif sur l'offre de pétrole et, par conséquent, sur les prix mondiaux du pétrole.

I.4 Produits dérivés

« Les produits dérivés sont des instruments financiers reposant sur des valeurs mobilières ou sur des indices de marché appelés "sous-jacents". La valeur d'un produit dérivé dépend de celle de son sous-jacent au cours du temps. L'usage de ces produits permet aux investisseurs de se couvrir contre l'évolution défavorable d'un marché ou de spéculer en amplifiant la valorisation du sous-jacent grâce à l'effet de levier. » [6]

I.4.1 Catégories de produits dérivés

Les produits dérivés sont classés en deux catégories : les produits fermes, comprenant les contrats forwards, les contrats à terme et les contrats swaps, et les produits optionnels, incluant les options et les warrants [7].

I.4.1.1 Produits fermes

Ce sont des engagements contractuels pris entre deux parties qui les obligent à acheter ou à vendre une action. Les types de produits fermes incluent :

- **Les contrats forward :** Un contrat forward est un accord négocié de gré à gré entre un acheteur et un vendeur, dans lequel les parties s'engagent à livrer un actif spécifique (comme une marchandise, une devise ou un instrument financier) à une date future prédéterminée, à un prix convenu lors de la conclusion du contrat. Ce type de contrat permet de réduire la volatilité de certains marchés.

- **Les contrats à terme (futures) :** Les contrats futures sont similaires aux contrats forwards, mais la principale différence réside dans le fait que les contrats à terme sont négociés sur des marchés organisés. Sur ces marchés, les modalités telles que la taille du contrat,la date de livraison standardisée et les spécifications de l'actif sous-jacent sont déterminées par le marché lui même.

- **Les contrats swaps :** Un contrat swap est un accord financier entre deux parties négocié sur le marché gré à gré où elles conviennent d'échanger périodiquement des

flux financiers ou d'autres instruments financiers pendant une période déterminée, selon des modalités préétablies. Les swaps sont des instruments dérivés utilisés pour gérer les risques financiers tels que les variations de taux d'intérêt, de devises ou d'autres expositions.

I.4.1.2 Produits optionnels

Les principaux produits optionnels sont les suivants :

- **Les options :** Une option est un contrat entre deux parties qui donne au détenteur le droit, mais pas l'obligation, d'acheter ou de vendre un actif financier à un prix convenu à l'avance avant une date d'expiration spécifique. En échange de ce droit, l'acheteur paie au vendeur de l'option une somme d'argent appelée prime.
- **Les warrants :** Les warrants sont similaires aux options, à la seule différence que les warrants sont émis par les institutions financières.

I.4.2 Utilisation des produits dérivés

Selon [7], les produits dérivés peuvent être utilisés dans :

- **Le hedging (couverture) :** Un investisseur peut utiliser les produits dérivés comme stratégie visant à réduire ou compenser le risque financier potentiel lié à la fluctuation des prix ou des valeurs des actifs sous-jacents. Par exemple, un investisseur détient des actions dans une entreprise mais souhaite se protéger contre une baisse potentielle du prix de ces actions. Pour cela, il peut utiliser des contrats d'options pour couvrir sa position. S'il détient des actions, il peut acheter des options qui lui permettent de vendre ses actions à un prix prédéterminé à l'avenir, réduisant ainsi le risque de perte si le prix des actions diminue.
- **La spéculation :** Un investisseur peut définir une stratégie d'investissement où il prend des positions dans des produits dérivés dans le but de réaliser un profit en anticipant les futurs mouvements des prix ou des valeurs des actifs sous-jacents, sans nécessairement chercher à couvrir ou à réduire un risque existant. Par exemple, un spéculateur pourrait acheter des contrats à terme sur une matière première comme le pétrole en anticipant une hausse des prix à l'avenir, avec l'espoir de revendre ces contrats plus tard à un prix plus élevé pour réaliser un profit.
- **L'arbitrage :** L'arbitrage désigne une stratégie d'investissement visant à tirer profit des inefficiences temporaires ou des écarts de prix entre différents actifs ou marchés. L'objectif de l'arbitrage est de réaliser un profit sans risque ou presque, en exploitant ces différences de prix. Cette stratégie est généralement considérée comme à faible risque car elle repose sur l'idée que les écarts de prix devraient converger à terme.

I.5 Modèles économétriques pour la prédiction des prix du pétrole

Dans la littérature divers modèles économétriques ont été proposés et utilisés dans la prédiction des prix du pétrole et parmi eux :

- **Le modèle Auto Regressif (AR) :** Ce modèle utilise les valeurs passées d'une série temporelle pour prédire les valeurs futures. Un processus stationnaire y_t est dit autorégressif d'ordre p si l'on peut expliquer sa valeur à l'instant et en utilisant ses p termes précédents.

$$y_t = \mu + \sum_{i=1}^{p} \phi_{ii} y_{t-i} + \epsilon_t [13]$$

  où :

  - y_t est la valeur de la série temporelle à l'instant t.
  - μ est une constante.
  - ϕ_{ii} sont les coefficients auto-régressifs.
  - ϵ_t est le terme d'erreur à l'instant t, supposé être un bruit blanc avec une moyenne de zéro et une variance constante.

- **Le modèle Moving Average (MA) :** Ce modèle permet de prédire les valeurs futures à partir de moyennes pondérées des observations passées. Un processus y_t est considéré comme étant un processus MA d'ordre p si on peut exprimer sa valeur à l'instant t comme une combinaison linéaire d'erreur aléatoires (bruit blanc).

$$y_t = \mu + \epsilon_t + \sum_{k=1}^{q} \theta_{kk} \epsilon_{t-k} [13]$$

  - y_t est la valeur de la série temporelle à l'instant t.
  - μ est une constante.
  - θ_{ii} sont les coefficients du modèle MA.
  - ϵ_t est le terme d'erreur à l'instant t, supposé être un bruit blanc avec une moyenne de zéro et une variance constante.
  - ϵ_{t-k} sont les termes d'erreur passés.

- **Le modèle AutoRegressive Moving Average (ARMA) :** Le modèle ARMA est une combinaison de deux modèles : AR et MA. Il permet de modéliser des séries temporelles plus complexes, mais il est limité par le fait qu'il ne peut modéliser que des séries temporelles stationnaires. De plus, il n'est pas adapté pour modéliser une série temporelle présentant une tendance linéaire croissante [28].

$$y_t = \mu + \sum_{i=1}^{p} \phi_{ii} y_{t-i} + \epsilon_t + \sum_{k=1}^{q} \theta_{kk} \epsilon_{t-k}$$

- **Le modèle AutoRegressive Integrated Moving Average (ARIMA)** : Le modèle ARIMA est en effet une combinaison du modèle ARMA avec un processus de différenciation. Ce processus de différenciation permet d'éliminer la dépendance temporelle en produisant une série temporelle transformée qui est stationnaire, sur laquelle on peut ensuite appliquer le modèle ARMA.

I.6 Limites des modèles économétriques dans la prédiction des prix du pétrole

Ces méthodes dites traditionnelles ne donnent pas des meilleures prédictions en raison de :

- **La volatilité et le comportement non linéaire :** La volatilité élevée et le comportement non linéaire des prix du pétrole rendent souvent difficile la modélisation précise à l'aide de techniques économétriques traditionnelles qui supposent des relations linéaires entre les variables. Les chocs soudains, les crises géopolitiques, les changements de politiques et d'autres événements imprévus peuvent entraîner des mouvements brusques et non linéaires dans les prix du pétrole, ce qui rend difficile la capture de ces dynamiques avec des modèles linéaires.

- **Les problèmes de prévision à long terme :** Les modèles économétriques peuvent rencontrer des difficultés dans les prévisions à long terme du prix du pétrole. Les projections à plus longue échéance peuvent être sensibles à des facteurs difficiles à prévoir, comme les changements technologiques, les évolutions géopolitiques, ou les développements macroéconomiques mondiaux qui peuvent influencer de manière significative la demande et l'offre de pétrole. Les modèles économétriques peuvent avoir du mal à capturer ces influences complexes sur de longues périodes.

- **L'analyse multivariée :** L'analyse multivariée représente une limite des modèles économétriques utilisés pour prédire les prix du pétrole en raison de la complexité des relations entre le prix du pétrole et d'autres variables telles que l'offre, la demande et les conditions économiques mondiales. Les modèles économétriques peuvent rencontrer des difficultés à capturer efficacement ces interactions complexes en raison d'hypothèses simplificatrices sur la nature des relations entre les variables, de données limitées ou biaisées, ainsi que de la volatilité et du comportement non linéaire du marché pétrolier.

I.7 Conclusion

Les modèles économétriques ont été largement utilisés pour la prédiction des prix du pétrole. Cependant, ces modèles possèdent des limites, notamment en ce qui concerne la prise en compte de la volatilité et des comportements non linéaires des prix du pétrole, ainsi que la complexité des relations multivariées. Pour surmonter ces défis et améliorer les prévisions, une transition vers l'utilisation de techniques de machine learning peut être envisagée. Les approches de machine learning offrent la possibilité de capturer des

modèles plus complexes et non linéaires à partir de données massives, permettant ainsi une analyse plus approfondie et des prévisions plus précises dans le domaine du marché pétrolier .

II

Séries temporelles, réseaux de neurones récurrents et techniques de réduction de la dimensionnalité

II.1 Introduction

Les prix du pétrole, ainsi que les données qui influencent ces prix, sont des données séquentielles collectées quotidiennement, et elles évoluent d'un jour à l'autre. Cela démontre que le traitement et la prédiction des prix du pétrole relèvent du domaine des séries temporelles. Jusqu'à présent, les réseaux de neurones récurrents sont considérés comme l'un des outils les plus puissants pour la prédiction des séries temporelles. En effet, ils possèdent une mémoire interne qui leur permet de capturer l'historique des données passées, offrant ainsi la capacité de traiter des données séquentielles et de gérer les dépendances à court et à long terme en prenant en compte des événements passés susceptibles d'influencer les prédictions futures. Ils peuvent ainsi apprendre automatiquement à reconnaître les motifs temporels complexes dans les données, ce qui est essentiel pour la prédiction des séries temporelles où les tendances, les saisons et d'autres schémas peuvent être présents. De plus, ils sont capables de gérer des séquences de taille variable. Les techniques de réduction de la dimensionnalité renforcent les réseaux de neurones récurrents et permettent d'obtenir de meilleurs résultats tout en accélérant leur apprentissage, en éliminant le bruit et en prévenant le surapprentissage. Dans ce chapitre, nous aborderons brièvement quelques notions de base des séries temporelles, et nous nous pencherons sur les réseaux de neurones récurrents ainsi que sur quelques techniques de réduction de la dimensionnalité, notamment celles que nous utiliserons dans notre approche.

II.2 Séries temporelles

Les séries temporelles, également appelées séries chronologiques, sont des données séquentielles collectées dans le temps[19]. Elles sont utilisées dans divers domaines tels que la finance, la météorologie, la santé et l'IoT, ...etc. La Figure II.1 illustre l'évolution des prix au comptant (Spot Prices) du pétrole au fil du temps, extraite du dataset que nous avons utilisé.

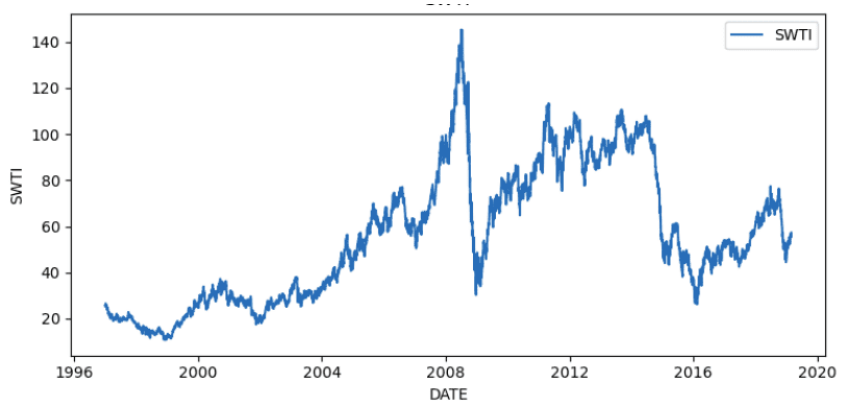

Fig. II.1 : Évolution des prix au comptant du pétrole au fil du temps

II.2.1 Composantes d'une série temporelle

Comme le montre la figure II.2 les séries temporelles sont composées de :

- **Tendance** : Il s'agit de l'évolution de la série temporelle au fil du temps, pouvant être à la hausse ou à la baisse.
- **Saisonnalité** : Elle représente les variations qui se produisent de manière périodique ou à intervalles réguliers dans le temps.
- **Résidu** : Le résidu est une composante aléatoire qui représente les fluctuations ou les variations imprévues et non expliquées par les modèles. Contrairement à la saisonnalité, le résidu ne peut pas être prédit à l'avance.

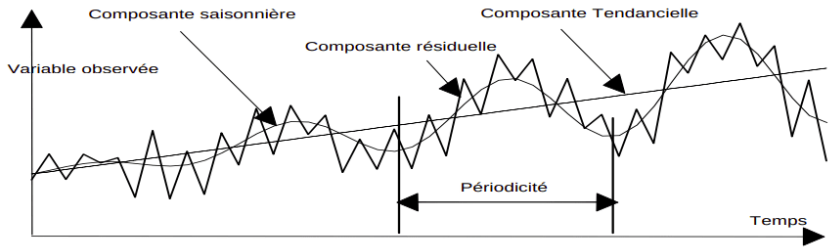

Fig. II.2 : Composantes d'une série temporelle [19]

Une autre composante qui peut être présente dans une série temporelle est la **composante cyclique**, qui présente un comportement périodique [19].

II.2.2 Stationnarité d'une série temporelle

Une série temporelle est dite stationnaire si et seulement si ses propriétés statistiques (moyenne, écart-type, variance) restent constantes au fil du temps. La figure II.3 extraite de [32] illustre une série temporelle stationnaire (en vert) et une série temporelle non stationnaire (en rouge). Il est observé qu'une série temporelle stationnaire ne présente pas de tendance.

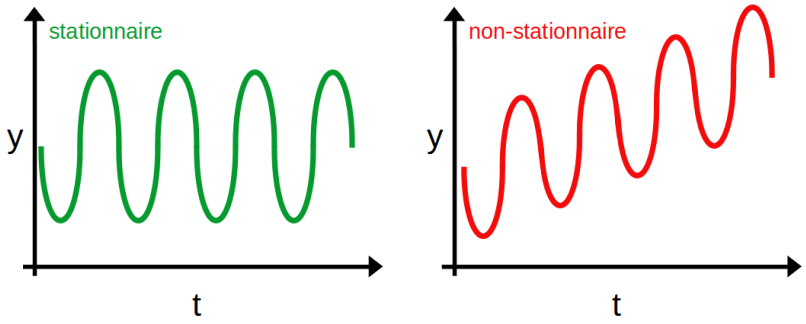

FIG. II.3 : Série temporelle : stationnaire vs. non stationnaire [32]

II.3 Réseaux de Neurones Récurrents (RNN)

Un réseau de neurones artificiels, inspiré du fonctionnement du cerveau humain, se compose de trois types de couches. La première couche, appelée couche d'entrée, est la première à recevoir les données d'entrée. Ensuite, les couches cachées, situées entre la couche d'entrée et la couche de sortie, effectuent des opérations de calcul intermédiaires. Enfin, la couche de sortie, qui est la dernière couche du réseau, fournit les prédictions.
Par rapport à ces réseaux de neurones classiques, un réseau de neurones récurrent est un type de réseau dans lequel les neurones sont connectés de manière à permettre le retour d'une sortie précédente comme entrée à des étapes ultérieures.

II.3.1 Architecture du RNN

Un RNN est composé d'une entrée, d'une sortie et d'un état caché. L'état caché est utilisé en tant qu'entrée pour le pas de temps suivant (t+1). Ainsi, l'état caché à l'instant t influe sur la sortie et l'état caché à l'instant t+1. L'état caché à un instant t est calculé comme suit :

$$h_t = \sigma_h \left(W_x x_t + W_h h_{t-1} + b_h \right) \text{[29]}$$

La sortie à l'instant t est déterminée par :

$$y_t = \sigma_y \left(W_y h_t + b_y \right) \text{ [29]}$$

Où :

- $\mathbf{x_t}$ est l'entrée à l'instant t.
- $\mathbf{h_t}$ est l'état caché à l'instant t.

- $\mathbf{y_t}$ est la sortie à l'instant t.
- $\mathbf{W_x}$ et $\mathbf{W_h}$ représentent les matrices de poids entre les unités cachées et à la fois l'entrée et la sortie respectivement et $\mathbf{W_y}$ représente la matrice de poids entre les pas de temps adjacents.
- $\mathbf{b_h}$ et $\mathbf{b_y}$ sont des vecteurs de biais.
- $\boldsymbol{\sigma_x}$ et $\boldsymbol{\sigma_y}$ sont des fonctions d'activation.

Les figures II.4 et II.5 illustrent respectivement le principe de fonctionnement du RNN et la structure d'une cellule RNN.

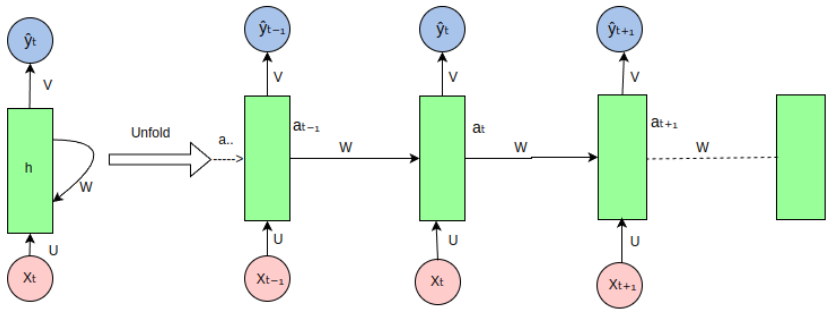

FIG. II.4 : RNN déplié [12]

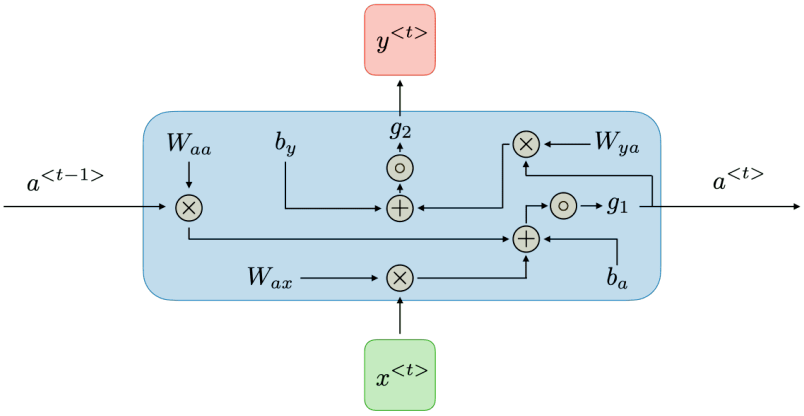

FIG. II.5 : Structure d'une cellule RNN [12]

II.3.2 Types de RNN

La nature de l'entrée et de la sortie des RNN (qu'il s'agisse de données individuelles ou de séquences) distingue les types de RNN suivants :

II.3.2.1 One-to-one

Il prend une seule entrée et produit une seule sortie, comme cela est montré dans la figure II.6.

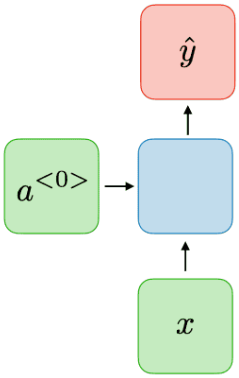

FIG. II.6 : RNN one-to-one [12]

II.3.2.2 One-to-many

Comme représenté dans la figure II.7, ce type de RNN prend une seule entrée et génère une séquence de sorties.
Il peut par exemple être utilisé pour générer une description textuelle à partir d'autres types de données , telles qu'une image radiographique [46], ainsi que pour la génération de musique [12] .

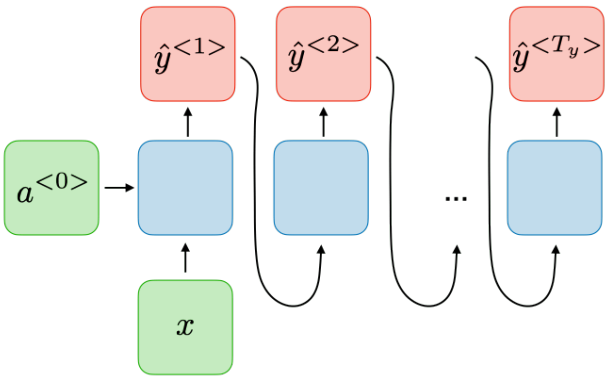

FIG. II.7 : RNN one-to-many [12]

II.3.2.3 Many-To-one

C'est l'inverse de « one-to-many ». Il prend une séquence d'entrées et génère une seule sortie à la dernière étape temporelle, comme illustré dans la figure II.8.
Il est utilisé par exemple dans la classification des sentiments [12].

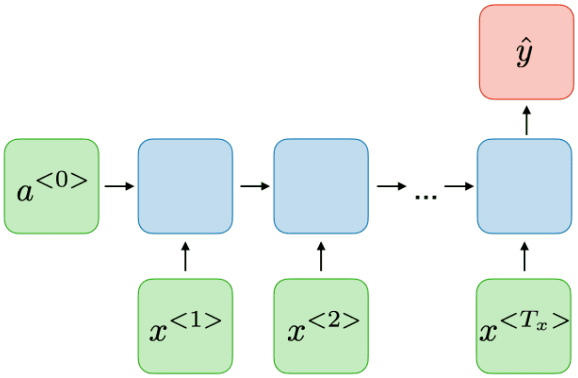

FIG. II.8 : RNN many-to-one [12]

II.3.2.4 Many-to-many

Ce type prend une séquence en entrée et génère également une séquence en sortie. Il existe deux types de ce modèle :

- **Many-to-many avec séquences d'entrée et de sortie identiques :** Comme le montre la figure II.9, pour chaque entrée de la séquence, une sortie est produite. Autrement dit pour chaque élément de la séquence d'entrée est associé à un élément correspondant dans la séquence de sortie à chaque étape de temps t.
  Parmi les exemples d'application de ce type, la prédiction d'événements cliniques de la prochaine visite à l'hôpital au fil du temps [46] et la reconnaissance d'entités nommées [12].

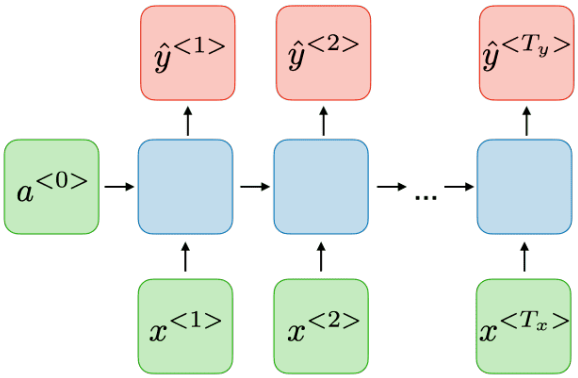

FIG. II.9 : RNN Many-to-many avec séquences d'entrée et de sortie identiques [12]

- **Many-to-many avec séquences d'entrée et de sortie différentes** : également connu sous le nom de Seq2Seq (Sequence to Sequence). Dans ce type, la séquence d'entrée est d'abord traitée par un processus de codage avant que le processus de décodage ne génère une séquence de sortie [46]. La figure II.10 illustre le principe de ce type. L'un des exemples de son utilisation est la traduction automatique [12, 46].

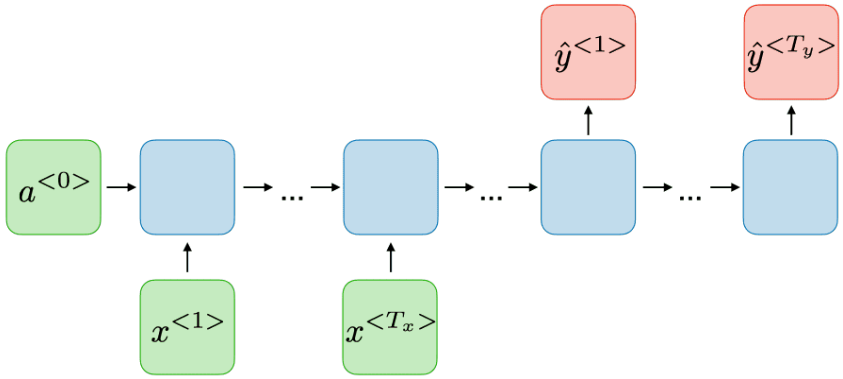

FIG. II.10 : RNN Many-to-many avec séquences d'entrée et de sortie différentes[12]

II.3.3 Problèmes de RNN

Comme indiqué dans [38, 47] le modèle RNN classique est confronté à deux problèmes qui sont abordés pour la première fois par [5] , à savoir :

- **Vanishing Gradient (Disparition du gradient)** : est un problème fréquent lors de l'entraînement des RNN. Au fur et à mesure que ces réseaux sont entraînés, le gradient devient de plus en plus petit en raison de la diminution rapide des composantes à long terme et de leur convergence vers 0. Cette diminution du gradient rend difficile la mise à jour des poids.
- **Exploding Gradient (Explosion du gradient)** : est un problème où le gradient augmente de manière exponentielle, en raison de l'augmentation excessive des composantes à long terme, devenant ainsi très grand. Cela rend l'apprentissage instable et difficile à converger vers la solution optimale.

Pour remédier à ces problèmes des RNN classiques, deux variantes de RNN dotées de mécanismes internes spéciaux, permettant de mieux gérer les dépendances à long terme dans les séquences, ont été développées : les LSTM et les GRU.

II.3.4 Long Short-Term Memory (LSTM)

Ce modèle a été proposé par Hochreiter et Schmidhuber [22]. Comme son nom l'indique, la cellule LSTM possède deux types de mémoires : une mémoire courte (à court terme) contenant des informations actuelles et récentes sur la séquence d'entrée, et une mémoire longue qui est capable de stocker et de maintenir des informations à long terme sur la séquence entrante [47].

Comme illustré dans la figure II.11, la cellule LSTM est constituée de trois portes (chacune étant associée à des équations spécifiques extraites de [18]), permettant ainsi de gérer les mémoires internes et de contrôler le flux d'informations, comme décrit dans [44].

- **Porte d'oubli :** Cette porte détermine quelles informations de la mémoire c_t doivent être oubliées (c'est-à-dire supprimées). Elle est représentée par l'équation :

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

- **Porte d'entrée** : Elle contrôle les informations à ajouter à l'état de la cellule, cela se traduit mathématiquement par l'équation suivante :

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

Également, l'état candidat est calculé par l'équation :

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$

Ensuite, l'état de la cellule est mis à jour par l'équation suivante :

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$$

- **Porte de sortie** : La porte de sortie contrôle quelles informations de l'état de la cellule seront transmises à la sortie de la cellule.
Elle est calculée par :

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

et le nouvel état caché obtenu par :

$$h_t = o_t \odot \tanh(C_t)$$

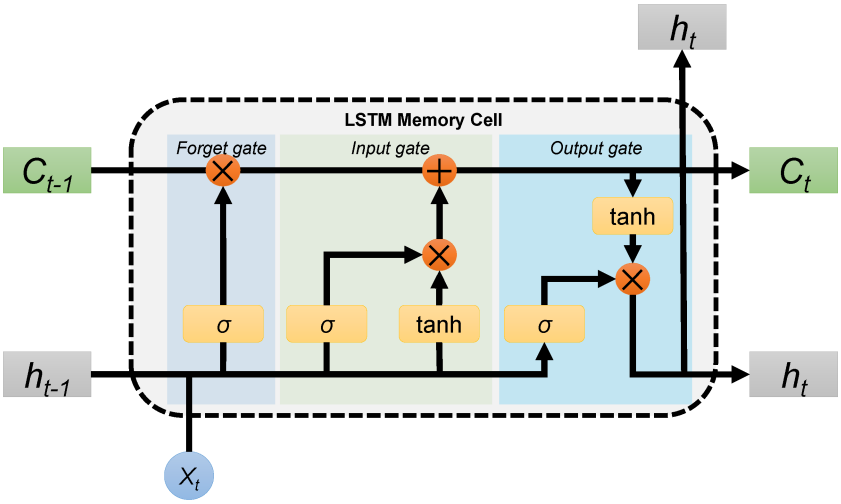

Fig. II.11 : Structure d'une cellule LSTM [18]

II.3.5 Gated Recurrent Unit (GRU)

Une autre variante développée pour résoudre les deux problèmes du gradient susmentionnées , introduite en 2014 par Kyunghyun Cho et al. [9]. Comme le montre la figure II.12, la cellule GRU dispose de deux portes, chacune avec des équations extraites de [48] : la « **Reset Gate** », qui contrôle la quantité d'informations à oublier, et la « **Update Gate** », qui détermine la quantité d'informations à intégrer dans le nouvel état caché h_t. Le fonctionnement du GRU est le suivant :

1. **Calcul de la porte de réinitialisation (Reset Gate)** : Cette opération est réalisée à l'aide de l'équation :
$$r_t = \sigma(W_{rx}x_t + W_{rh}h_{t-1} + b_r)$$
2. **Calcul de la porte de mise à jour (Update Gate)** : Cela s'effectue en utilisant cette équation :
$$z_t = \sigma(W_{zx}x_t + W_{zh}h_{t-1} + b_z)$$
3. **Calcul du candidat à l'état caché** : Le vecteur candidat à l'état caché est calculé par :
$$h_t^c = \tanh(W_{hx}x_t + W_{hh} \cdot (r_t \odot h_{t-1}) + b_h)$$
4. **Mise à jour de l'état caché** : L'état caché est mise à jour comme suit :
$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot h_t^c$$

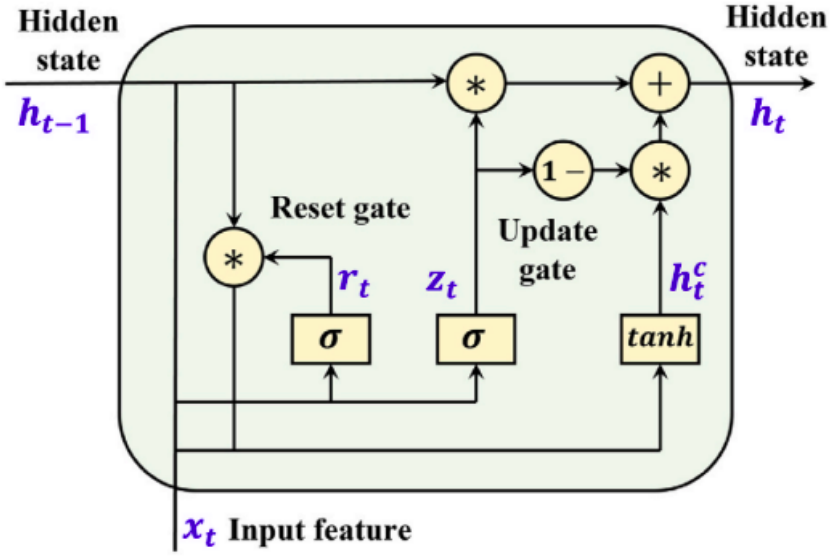

FIG. II.12 : Structure d'une cellule GRU [48]

II.4 Réduction de la dimensionnalité

La réduction de la dimensionnalité est une technique qui consiste à réduire le nombre de variables tout en conservant les informations les plus importantes. Il existe une multitude de techniques de réduction de la dimensionnalité qui peuvent être catégorisées en deux groupes : les techniques linéaires et les techniques non linéaires.

II.4.1 Techniques linéaires

Ce sont des techniques basées sur des opérations linéaires, souvent appliquées sur des données qui sont intrinsèquement linéaires. L'une des techniques linéaires les plus connues est PCA (Principal Component Analysis).

II.4.1.1 Principal Component Analysis (PCA)

Une technique statistique de réduction de la dimensionnalité très populaire. Ses étapes majeures sont selon Yan, Zhu et Wang [47] :

1. Calcul de la matrice variance-covariance.
2. Calcul des valeurs propres et des vecteurs propres de la matrice de variance-covariance.
3. Ordonner les valeurs propres dans l'ordre décroissant.
4. Construction de la matrice des vecteurs propres selon l'ordre des valeurs propres (la i-ème ligne de la matrice des vecteurs propres correspond au vecteur propre associé à la i-ème valeur propre de la liste des valeurs propres ordonnées dans l'ordre décroissant).
5. Sélection du nombre de composantes principales k (le nombre de composantes reflète le nombre de dimensions à conserver).
6. Construction d'une matrice ne contenant que les k premières lignes de la matrice des vecteurs propres.
7. Construction des nouvelles données en multipliant la matrice créée à l'étape 6 par l'ensemble de données initial (dont la réduction de dimension est souhaitée).

L'algorithme 1, extrait de l'article de Ezukwoke et Samaneh Zareian [17], illustre la procédure PCA.

Algorithm 1 PCA procedure [17]

Input : $x \in X$ where $x \in \mathbb{R}^D$
Output : $\hat{x} \in \mathbb{R}^k$ where $k \ll D$
begin
$\hat{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$;
$dx_i = x_i - \hat{x}$;
$\sum = \frac{1}{N} \sum_{i=1}^{N} dx_i dx_i^T$;
$\sum \mathbf{u}_k = \lambda_k \mathbf{u}_k$;
$\mathbf{u}_k = \arg\,\text{sort}(\mathbf{u}_k)$;
$\hat{x} = \mathbf{u}_k \cdot dx_i$;
end

II.4.2 Techniques non linéaires

Les technique de réduction de la dimensionnalité non linéaires sont des méthodes qui peuvent capturer des relations non linéaires entre les variables.

II.4.2.1 Kernel Principal Component Analysis (KPCA)

C'est une extension de la méthode PCA utilisée pour la réduction de dimension de caractéristiques non linéaires elle repose sur l'utilisation de fonctions du noyau pour qui permette de transformer les caractéristiques en un espace de dimension plus élevé pour détecter un sous espace réduit qui conserve les informations [17]. Ezukwoke et Samaneh Zareian [17] ont expliqué les équations et les étapes de la méthode KPCA, puis ont résumé ces étapes dans l'algorithme 2 , où $\phi(x)$ représente la transformation des données d'origine x dans un espace de dimension plus élevée. $k(x_i, x_j)$ exprime la mesure de similarité ou de distance entre les points x_i et x_j, capturée par la fonction noyau k. K désigne la matrice de noyau.

Algorithm 2 Kernel PCA Algorithm [17]

Input : $\phi(x) \in \kappa(x, x_j)$ where $\phi(x) \in \mathbb{R}^D$. Let $\mathbf{K} = \kappa(x_i, x_j)$

Output : $\hat{x} \in \mathbb{R}^k$ where $k \ll D$
begin
  Select a kernel κ ;
  Construct Gram matrix
    $\hat{\mathbf{K}} = \mathbf{K} - \mathbf{1}_{1/N}\mathbf{K} - \mathbf{K}\mathbf{1}_{1/N} + \mathbf{1}_{1/N}\mathbf{K}\mathbf{1}_{1/N}$;
  Solve eigen problem $\mathbf{K}\mathbf{u}_k = \lambda\mathbf{u}_k$;
  Project data in new space
    $\hat{x} = \sum_{i=1}^{N} u_k \kappa$;
end

II.4.2.2 Uniform Manifold Approximation and Projection (UMAP)

UMAP représente une autre technique non linéaire de la réduction de la dimensionnalité. Elle adopte une approche basée sur la topologie et la géométrie riemannienne, lui permettant de capturer efficacement des relations complexes entre les points de données, visant à préserver les structures à la fois locales et globales des données. UMAP est très apprécié pour sa capacité à capturer des structures de données complexes et elle se distingue par sa rapidité, surpassant souvent d'autres techniques non linéaires de réduction de la dimensionnalité comme l'incorporation de voisinage stochastique distribué en t (t-SNE) [35].
La figure II.13 extraite de [15] illustre les étapes d'UMAP.

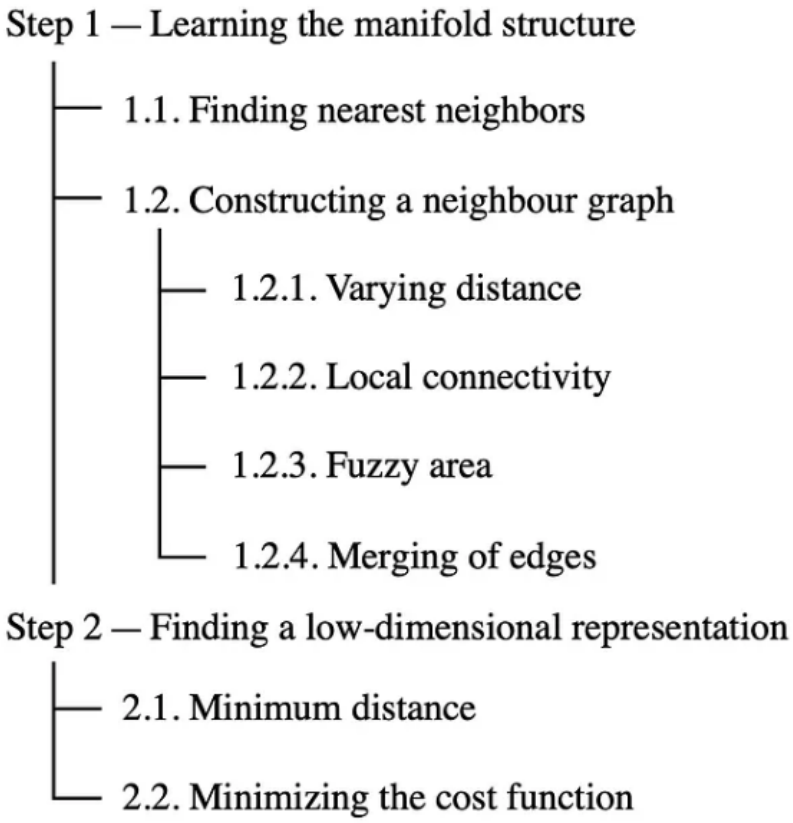

Fig. II.13 : Étapes d'UMAP [15]

II.5 Conclusion

Dans ce chapitre, nous avons exploré les notions de base des séries temporelles, ainsi que les principes de fonctionnement des Réseaux de Neurones Récurrents (RNN). Nous avons évoqué certains problèmes inhérents aux RNN et examiné les variantes qui offrent des solutions à ces problèmes. De plus, nous avons introduit quelques méthodes mathématiques de réduction de la dimensionnalité, en mettant en avant celle que nous allons utiliser dans notre approche. Étant donné l'importance cruciale de la prédiction du prix du pétrole pour l'économie mondiale et l'attention qu'elle suscite de la part de nombreux acteurs, de nombreux travaux antérieurs ont été réalisés et plusieurs approches basées sur des méthodes du machine learning ont été proposées. Dans le prochain chapitre, nous explorerons en détail ces différentes approches et méthodes utilisées pour la prédiction des prix du pétrole.

III

Travaux antérieurs sur l'utilisation du machine learning pour la prédiction des prix du pétrole

III.1 Introduction

Les prix du pétrole ont un impact socio-économique considérable, influençant les économies mondiales, les décisions politiques, et les stratégies d'investissement. Leur nature volatile rend leur prédiction complexe, nécessitant une prise en compte de multiples facteurs économiques, géopolitiques, et environnementaux. Avec l'avancement des technologies et l'essor du machine learning, de nombreuses recherches ont été menées pour améliorer la précision des prédictions des prix du pétrole. Dans ce chapitre nous allons présenter quelques articles récents et pertinents sur ce sujet.

III.2 Classification des travaux antérieurs

Parmi les travaux qui ont été réalisés et les approches qui ont été développées, ces derniers peuvent être classés selon les types de données utilisées :

- **Utilisation des prix antérieurs du pétrole** : Ces travaux se concentrent sur les prix passés du pétrole pour prédire les prix futurs.
- **Utilisation des prix antérieurs du pétrole avec données quantitatives hors prix du pétrole :** En plus des prix antérieurs du pétrole, ces approches utilisent

des données quantitatives passées telles que des données économiques et financières pour prédire les prix du pétrole.

- **Utilisation des prix antérieurs du pétrole avec données textuelles hors prix du pétrole :** Ces méthodes intègrent les prix antérieurs du pétrole ainsi que des données textuelles issues de techniques de text mining pour effectuer les prédictions.
- **Utilisation des prix antérieurs du pétrole avec données quantitatives et textuelles hors prix du pétrole :** Ces travaux combinent les prix antérieurs du pétrole, des données quantitatives et des données textuelles pour la prédiction des prix du pétrole.

La figure III.1 illustre la classification des travaux antérieurs selon les types de données utilisées.

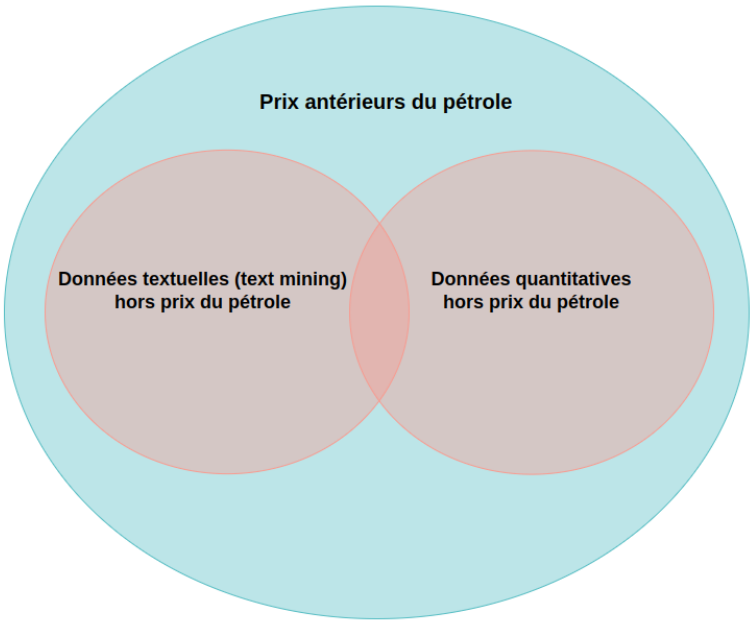

Fig. III.1 : Classification des travaux antérieures

III.3 Présentation des travaux antérieurs

Parmi les nombreux articles publiés sur la prédiction des prix du pétrole, nous avons choisi de présenter quelques articles pertinents et récents.

III.3.1 Travaux se limitant à l'utilisation des prix antérieurs du pétrole

Dans cette partie, nous allons présenter quelques articles proposant des approches de machine learning, mais se concentrant uniquement sur les prix historiques du pétrole.

III.3.1.1 Crude oil price prediction using deep reinforcement learning [1]

Ils ont développé un algorithme d'apprentissage par renforcement profond appelé DNPP (Dynamic Noise Proximal Policy), qui se démarque par deux améliorations principales : un mécanisme d'exploration de bruit dynamique et un mécanisme d'optimisation d'approximation dynamique. Ils l'ont utilisé pour prédire les prix du pétrole sur trois bourses différentes et ont comparé ses performances à celles d'autres modèles tels que DDPG (Deep Deterministic Policy Gradient), LSTM, DBN (Deep Belief Network), SVM et RF (Random Forest) en calculant les erreurs moyennes absolues (MAE), les erreurs quadratiques moyennes (MSE) et les erreurs moyennes absolues en pourcentage (MAPE) sur les données de (Brent, WTI, OMAN) selon différents ratios (60-40, 70-30, 80 :20, 90-10) et sur différentes étapes de prévision (1, 6, 12). Ils ont constaté que leur algorithme donnait les meilleurs résultats dans toutes les conditions et les comparaisons effectuées.

La figure III.2 illustre le pseudo code de leur algorithme :

The DNPP's training process.

Algorithm 1: Dynamic Noisy Proximal Policy

Initialize the weight of DNPP, including the actor and critic networks;
Set the memory buffer $M \neq \varnothing$;
for episode=[1,2,3,...,E_{max}] **do**
  Reset the agent and the simulation environment;
  Reset the initial state s_0;
  Reset the initial reward r_0;
  for $[t = 1, 2, 3, ..., T]$ **do**
    Initialize the OU process;
    According to the policy π, select a_t;
    Execute a_t from the environment,
    observe the immediate reward r_t and the next state s_{t+1};
    Push transition into M;
    Randomly sample N transition tuples as mini-batch training data;
    Calculate the gradient of the critic network;
    Update the parameters of the critic network;
    Calculate the gradient of the actor network;
    Update the parameters of the actor network;
  end for
end for

Fig. III.2 : Pseudo code DNPP [1]

III.3.1.2 Crude Oil Price Forecast Based on Deep Transfer Learning : Shanghai Crude Oil as an Example [14]

Les auteurs de cet article ont proposé une approche pour résoudre le problème du manque de données sur les contrats à terme du pétrole de Shanghai, en raison de leur cotation sur une période courte, remontant à 2018. Leur approche repose sur l'utilisation du LSTM avec le transfert d'apprentissage, où ils ont entraîné leur modèle sur un ensemble de données concernant le pétrole brut, puis affiné le modèle sur l'ensemble de données spécifique au pétrole de Shanghai. Après avoir comparé les résultats de leur modèle (T-LSTM) avec celui du modèle LSTM sur différents paramètres tels que la taille du lot, le nombre de neurones et la taille de la fenêtre, ainsi qu'avec ARIMA en utilisant des métriques de performance tels que MAE et RMSE, ils ont constaté que leur modèle

surpassait les autres en termes de précision et démontrait une grande capacité prédictive.

III.3.1.3 Forecasting crude oil price using LSTM neural networks [49]

Ils ont utilisé le modèle LSTM pour la prédiction du prix du pétrole en se basant uniquement sur les prix passés du pétrole Brent et du WTI. Ensuite, ils ont comparé les résultats du modèle LSTM avec ceux du modèle de réseau de neurones et du modèle ARIMA à court terme, moyen terme et long terme, en utilisant les trois mesures de performance suivantes : MSE, MAE et SDAPE. Ils ont trouvé que le modèle LSTM a une forte capacité de généralisation, surpassant les deux autres modèles à court et à long terme. Le modèle ANN, quant à lui, surpasse le LSTM (légèrement) et le modèle ARIMA à moyen terme.

Le tableau III.1 présente une synthèse des travaux se concentrant uniquement sur les prix historiques du pétrole.

Auteurs	Données collectées	Approche(s) proposée(s)	Résultat(s)	Critique(s)
Liang et al. [1]	Prix de clôture du pétrole des bourses WTI, Oman et Brent collectés de novembre 2009 à février 2022.	Dynamic Noise Proximal Policy (DNPP)	Le modèle DNPP surpasse le DDPG, LSTM, DBN, SVM et RF	Étant donné qu'ils n'ont pris en considération que les prix historiques du pétrole, il est difficile de juger si leur modèle est plus performant que le LSTM dans la prédiction des prix du pétrole.
Zhang et Hong [49]	Prix du pétrole WTI et du Brent collectés du 10 février 1986 au 17 mai 2021.	LSTM	-Le modèle LSTM a surpassé les modèles ARIMA et ANN en termes de précision et de stabilité des prévisions à la fois à court terme et à long terme. -Le modèle ANN a surpassé le LSTM (légèrement) et l'ARIMA à moyen terme.	Ils n'ont pas prétraité leurs données, par exemple en les normalisant, à l'exception du partitionnement de leur dataset.

Deng, Ma et Zeng [14]	- Prix à terme du pétrole de Shanghai collectés du 26 mars 2018 au 26 octobre 2021. - Prix de clôture du pétrole Brent collectés du 29 août 2000 au 31 décembre 2020.	T-LSTM (transfert learning avec LSTM)	T-LSTM à une bonne capacité de généralisation et surpasse le LSTM	Ils pourraient également tester d'autres méthodes réputées en matière de prédiction des séries temporelles, telles que le GRU, pour approfondir davantage leurs analyses.

Tab. III.1 : Tableau récapitulatif des travaux se limitant aux prix historiques du pétrole

- Le problème commun de ces travaux est qu'ils se sont limités uniquement aux données historiques du pétrole, alors que les prix du pétrole sont influencés par divers facteurs. Cela pourrait conduire à des conclusions biaisées et incomplètes.

III.3.2 Travaux se limitant à l'utilisation des prix antérieurs du pétrole et des données quantitatives

Les articles que nous allons présenter dans cette partie incluent et renforcent les données sur les prix historiques du pétrole avec d'autres données économiques.

III.3.2.1 Selection of Machine Learning Models for Oil Price Forecasting : Based on the Dual Attributes of Oil [47]

Les auteurs de cet article ont développé six modèles combinant les Réseaux de Neurones Récurrents (RNN) classiques et leurs variantes LSTM avec trois techniques de réduction de la dimensionnalité : PCA, MDS et LLE. Ces modèles ont été utilisés pour prédire les prix à terme (future prices) et les prix au comptant (spot prices) du pétrole, sur des fenêtres de taille respectivement 40, 60 et 80.
Leurs résultats montrent que les modèles utilisant les méthodes de réduction de la dimensionnalité PCA, MDS et LLE obtiennent de meilleurs résultats que les modèles RNN et LSTM qui ne font pas appel à ces techniques, à l'exception de la prédiction des prix à terme du pétrole dans une fenêtre de taille 80. Les modèles RNN-LLE donnent de meilleurs résultats que les modèles RNN-PCA, tandis que les modèles LSTM-LLE surpassent les modèles LSTM-MDS et LSTM-PCA. En outre, les modèles utilisant PCA

présentent de meilleures performances que ceux utilisant MDS, sauf dans la prédiction des prix au comptant du pétrole dans une fenêtre de taille 60, la prédiction des prix à terme dans une fenêtre de taille 80 avec RNN et MDS, ainsi que dans une fenêtre de taille 60 avec LSTM et MDS.

III.3.2.2 Oil Price Prediction using Deep Neural Network Technique Gated Recurrent Unit (GRU) and Multivariate Analysis [24]

Ils ont présenté un algorithme intitulé « HMOP-NCT MODEL », qu'ils ont appliqué à un ensemble de données comprenant des variables telles que le DJU, le US 10 YR BOND, le US DOLLAR INDEX, le SP 500, ainsi que les prix de l'or et du WTI sur une période de 6 ans. Après avoir prétraité les données en supprimant les lignes contenant des valeurs manquantes, ils ont réalisé une analyse multivariée. Ensuite, ils ont divisé l'ensemble de données en ensembles de test et d'entraînement, puis appliqué le modèle GRU. Enfin, l'évaluation du modèle s'est faite à l'aide de métriques de performance.

La figure III.3 illustre la prévision à l'aide du modèle GRU en utilisant la technique de suppression des valeurs aberrantes financières.

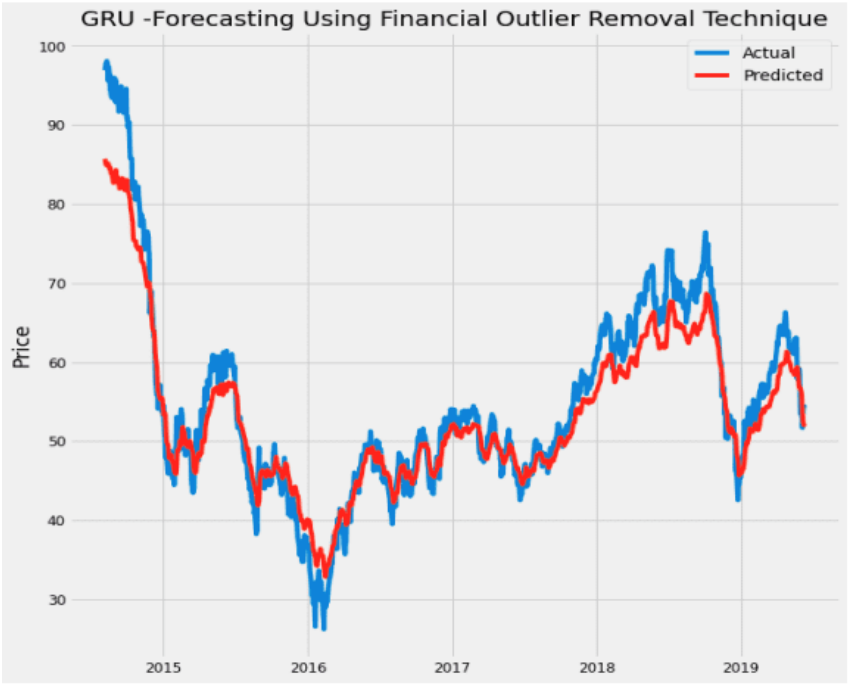

Fig. III.3 : GRU - Prévision à l'aide de la technique de suppression des valeurs aberrantes financières [24]

III.3.2.3 Forecasting crude oil futures price using machine learning methods : Evidence from China [21]

Les auteurs de cet article ont pour objectif de prédire les prix à terme du pétrole brut chinois. Ils ont collecté des données sur le pétrole chinois du 26 mars 2018 jusqu'au 28

février 2023, des données sur le marché monétaire (telles que le taux de change, le taux d'intérêt et le taux des bons du Trésor), des données sur le marché à terme (contrats à terme sur les matières premières et sur le pétrole brut (WTI et Brent)), des données sur le marché des actions et l'incertitude économique aux États-Unis. Ils ont appliqué des modèles de machine learning, à savoir RNN, LSTM, GRU, SVR, MLP, CNN et BP, sur leur ensemble de données sans l'inclusion des facteurs qui influent sur les prix du pétrole, ainsi qu'avec l'inclusion d'un facteur influant sur les prix du pétrole (inclusion uniquement du facteur de marché monétaire, inclusion uniquement du facteur du marché à terme, inclusion uniquement du facteur sur le marché des actions et inclusion uniquement du facteur d'incertitude économique aux États-Unis) et enfin l'inclusion de tous les facteurs à la fois. Ils ont prédit les prix à terme du pétrole chinois à 5 jours et 10 jours, puis ont évalué et comparé leurs modèles en utilisant les métriques MAE, MSE et RMSE. Ils ont trouvé que le modèle GRU surpasse les autres modèles dans toutes les situations, et que l'inclusion des facteurs externes influençant les prix du pétrole améliore considérablement la précision du modèle.

Le tableau III.2 synthétise les articles utilisant les prix historiques du pétrole et les données économiques pour la prédiction des prix du pétrole.

Auteurs	Données collectées	Approche(s) proposée(s)	Résultat(s)	Critiques(s)
Yan, Zhu et Wang [47]	Données collectées du 2 janvier 1997 au 28 février 2019 comprennent : - Données financières. - Données de marchandise. - Prix au comptant du pétrole WTI. - Prix à terme du pétrole WTI.	PCA-RNN, MDS-RNN, LLE-RNN, PCA-LSTM, MDS-LSTM, LLE-LSTM	LLE-RNN et LLE-LSTM surpasse les autres modèles.	Utilisation de la technique PCA bien que clairement inefficace, car les facteurs influençant les prix du pétrole sont non linéaires.
Al-Jasoor et Al-Janabi [24]	- Prix du pétrole WTI. - Données financières (prix de l'or, S&P500, indice du dollar américain, obligations à 10 ans américaines, DJU).	GRU et analyse multivariée	Leur modèle à une bonne capacité de généralisation.	-Les résultats des métriques de performance n'ont pas été présentés. -La durée de collecte de données est très courte (six ans).

Guo et al. [21]	- Prix à terme du pétrole chinois collectés du 26 mars 2018 jusqu'au 28 février 2023. - Données sur le marché monétaire (taux de change, taux d'intérêt et taux des bons du Trésor). -Données sur le marché à terme (contrats à termes sur les matières premières et sur le pétrole brut (WTI et Brent)). - Données sur le marché des actions. - Incertitude économique aux États-Unis	RNN, LSTM, GRU, SVR, MLP, CNN et BP	-Le modèle GRU surpasse les autres modèles en termes de prédiction des prix à terme du pétrole chinois. - L'inclusion des facteurs externes influençant les prix du pétrole améliore considérable-ment la précision des modèles.	La durée de collecte de données est très courte (2018-2023), une approche permettant de combler ce manque de données sur les prix à terme du pétrole de Shanghai est envisageable.

Tab. III.2 : Tableau récapitulatif des travaux incluant les données économiques

III.3.3 Travaux Incorporant le text mining

Il y a des travaux qui sont allés encore plus loin dans la collecte de données influençant les prix du pétrole, où ils ont utilisé des techniques de text mining afin de collecter des données textuelles à partir des actualités.

III.3.3.1 Crude oil price forecasting incorporating news text [4]

Ils ont collecté des données à partir des titres d'actualités et ont utilisé le modèle SeaNMF pour la catégorisation des documents, surpassant le modèle LDA pour les textes courts. Leurs caractéristiques comprennent des topics, l'indice de polarité et dprice (la différence de prix entre les observations consécutives). Ensuite, ils ont créé deux nouveaux indicateurs : l'intensité quotidienne des topics et l'intensité quotidienne des sentiments

pour chaque topic, représentant ainsi les valeurs de chaque caractéristique à chaque instant t. Ils ont utilisé l'algorithme AdaBoost.R pour prédire les prix du pétrole, constatant que leur approche surpassait les autres méthodes et avait également de bonnes performances dans la prédiction d'autres matières premières, comme les prix du gaz et de l'or. La figure ci-dessus illustre la méthodologie de leur approche.

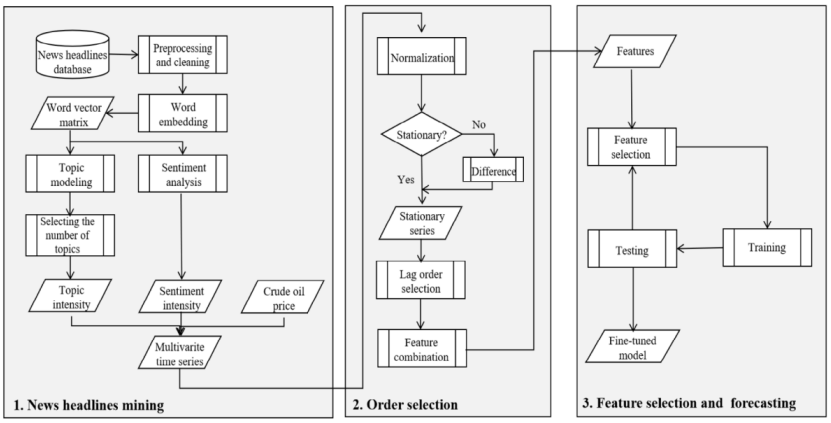

FIG. III.4 : The framework of crude oil price forecasting [4]

III.3.3.2 A novel hybrid model with two-layer multivariate decomposition for crude oil price forecasting [50]

Les auteurs ont proposé une approche combinant l'extraction de texte (text mining) et une décomposition multivariée en deux couches pour prédire les prix du pétrole (WTI). Ils ont utilisé des modèles de régression linéaire multiple (MLR) et des réseaux de neurones à propagation arrière (BPNN) dans leur méthode.

Après avoir testé et comparé leurs modèles, ils ont tiré les conclusions suivantes :

- Les modèles utilisant la deuxième décomposition multivariée surpassent ceux utilisant la première en simplifiant la prévision des composantes haute fréquence.
- La méthode de décomposition multivariée améliore significativement la précision des prévisions des prix du pétrole brut par rapport aux méthodes univariées en simplifiant les séries temporelles complexes.
- La performance de la Décomposition Empirique en Modes Multivariés (MVMD) est comparable à celle de la Décomposition Empirique en Modes (MEMD), probablement en raison du nombre de sous-composantes MVMD.
- Les techniques de prévision hybrides surpassent les techniques simples en utilisant différentes méthodes pour les sous-composantes aux caractéristiques distinctes.
- Leur modèle proposé démontre la meilleure performance en matière de prévision.

Le tableau III.3 présente un récapitulatif des articles proposant des approches du machine learning, incluant les données textuelles des actualités.

Auteurs	Données collectées	Approche(s) proposée(s)	Résultat(s)	Critique(s)
Bai et al. [4]	Prix quotidien du pétrole WTI collectés du 29 mars 2011 au 22 mars 2019, ainsi que les nouvelles couvrant cette période.	Proposition de deux nouveaux indicateurs de sujet et de sentiment pour les données textuelles courtes et éparses	AdaBoost.RT combiné avec l'approche proposée surpasse les autres modèles.	Les données collectées se limitent aux données textuelles, notamment les titres des articles d'actualités.
Zhao et al. [50]	- Prix au comptant de WTI. - Données financières (DJIA, S&P500, NASDAQ, BUSDX, USRUE). - Données textuelles issues des actualités (textmining).	Décomposition multivariée en deux couches (EMD+VMD)	La deuxième décomposition multivariée augmente la précision du modèle, et le modèle basé sur cette décomposition surpasse les autres modèles.	Les données sur les prix du pétrole ont été collectées sur une courte durée, de 5 septembre 2014, jusqu'au 26 novembre 2021.

Tab. III.3 : Tableau récapitulatif des travaux incluant les données textuelles (en utilisant le text mining)

- Le problème commun de ces études est qu'elles n'ont pas choisi les modèles de machine learning réputés pour la prédiction des séries temporelles, comme les RNN et leurs variantes, et ne les ont pas comparés avec les travaux utilisant ces modèles sans utiliser les données issues du text mining. Cela aurait permis de démontrer véritablement la force et la valeur ajoutée de leurs approches, notamment en utilisant des données issues du text mining.

III.4 Conclusion

En conclusion de ce chapitre, l'examen des articles pertinents sur l'utilisation du machine learning pour la prédiction des prix du pétrole nous a offert une vision approfondie de la complexité de ce domaine. Nous avons pu appréhender la dynamique des prix du pétrole ainsi que les multiples facteurs qui les influent, tout en explorant diverses techniques de prédiction. Cette exploration nous a permis de tirer des enseignements précieux et de nous inspirer pour proposer une approche innovante. Dans le prochain et dernier chapitre, nous détaillerons cette approche, fruit de nos réflexions et de nos analyses, dans le but de contribuer à l'avancement de la prédiction des prix du pétrole.

IV

Conception, implémentation, évaluation et comparaison des modèles de prédiction des prix du pétrole

IV.1 Introduction

Dans ce dernier chapitre, nous concrétisons notre approche en décrivant et en expliquant en détail les étapes que nous avons entreprises. Nous commençons par l'analyse et la description du dataset, puis passons à son prétraitement. Ensuite, nous présentons la réalisation de nos modèles, qui combinent deux techniques de réduction de la dimensionnalité (KPCA et UMAP) avec les RNN et leurs variantes LSTM et GRU. Enfin, nous évaluons et comparons nos modèles proposés avec les modèles RNN et leurs variantes sans réduction de la dimensionnalité, ainsi qu'avec le modèle LLE-RNN de [47].

IV.2 Langage, outils et bibliothèques utilisés

Afin d'analyser le dataset que nous avons utilisé, d'entraîner et d'évaluer nos modèles, nous avons eu recours à un langage de programmation ainsi qu'à certains outils et bibliothèques.

IV.2.1 Python

Python est un langage de programmation de haut niveau particulièrement apprécié dans les domaines du machine learning et de la data science. Sa syntaxe claire et sa large

panoplie de bibliothèques dédiées en font un outil puissant et accessible pour un large éventail de tâches liées au traitement et à l'analyse de données.

IV.2.2 Google Anaconda

Anaconda est un logiciel gratuit qui offre une interface graphique permettant d'accéder à plusieurs environnements de développement intégrés (IDE) comme Jupyter, Spyder,... etc. De plus, plusieurs bibliothèques de base sont déjà préinstallées, et il offre la facilité et la possibilité d'installer d'autres bibliothèques et fonctionnalités [41].

IV.2.3 Juyter Notebook

Jupyter est un IDE qui se lance sur un navigateur. Il est structuré sous forme de blocs contenant des lignes de code, et chaque bloc peut être exécuté séparément. Les sorties de chaque bloc, telles que les visualisations, peuvent être utilisées en un seul endroit, ce qui le rend efficace pour la présentation du code. De plus, il est facile de partager son notebook avec d'autres personnes [41].

IV.2.4 TensorFlow

TensorFlow est une bibliothèque développée par Google utilisée dans le machine learning et le réseaux de neurones. Elle fonctionne sur des CPUs, des GPUs et des processeurs IoT. Elle est apparue pour la première fois en 2015 [25].

IV.2.5 Keras

Keras est une API de réseau neuronal exécutée sur TensorFlow 2.0, sa dernière version étant la 2.3.0. Elle fonctionne en Python et se distingue par sa modularité, sa facilité d'extensibilité et sa convivialité pour l'utilisateur [25]

IV.2.6 NumPy

NumPy, abréviation de « Numerical Python », est une bibliothèque Python utilisée pour effectuer des opérations mathématiques telles que celles liées à l'algèbre linéaire sur des tableaux multidimensionnels à haute performance qu'elle fournit [27].

IV.2.7 Matplotlib

Matplotlib est une bibliothèque Python utilisée pour afficher une variété de graphiques et de tracés, y compris des graphiques linéaires, des graphiques de dispersion et des tracés en 3D. En plus de cela, elle est également utile pour créer des animations et des affichages interactifs, offrant ainsi une large gamme de fonctionnalités pour la visualisation des données en Python [23].

IV.2.8 Pandas

Pandas est une bibliothèque d'analyse de données open source sous licence BSD. Elle fournit des structures de données telles que les séries (Series) et les DataFrames, qui permettent d'effectuer la manipulation et l'analyse des données [16].

IV.2.9 Seaborn

Seaborn est une autre bibliothèque Python de visualisation de graphiques statistiques qui s'intègre étroitement avec les structures de données Pandas [45].

IV.2.10 Sickit learn

Scikit-learn est une bibliothèque Python open source, reposant sur NumPy, SciPy et Matplotlib. Elle offre une large sélection d'algorithmes tels que ceux de régression, de classification et de réduction de la dimensionnalité. C'est une ressource inestimable pour quiconque travaille dans le domaine de l'apprentissage automatique [26].

IV.3 Méthodologie de l'approche proposée

Notre approche repose sur la combinaison des techniques non linéaires de réduction de la dimensionnalité : KPCA et UMAP, avec les modèles de réseaux de neurones : RNN, LSTM et GRU. Comme illustré dans la figure IV.1, notre approche se divise en quatre phases : l'analyse des données, le prétraitement, la conception et l'entraînement des modèles, et enfin l'évaluation et la comparaison des modèles où nous avons évalué nos modèles et les avons comparés avec les modèles de réseaux de neurones récurrents et leurs variantes sans réduction de la dimensionnalité, ainsi qu'avec le modèle LLE-RNN de [47].

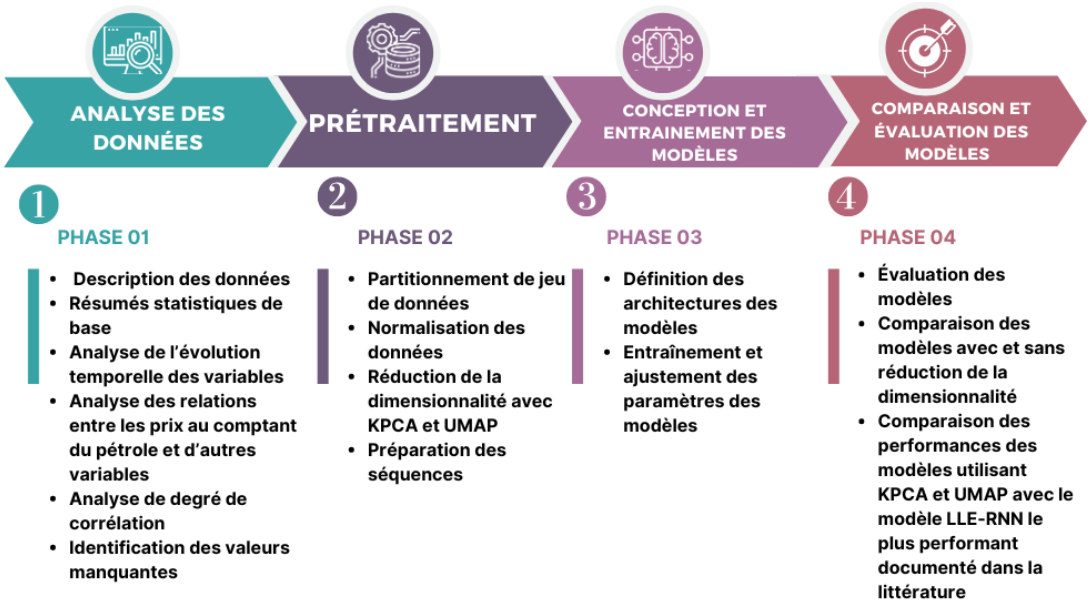

Fig. IV.1 : Méthodologie de l'approche proposée

IV.4 Analyse des données

Avant de passer au prétraitement du dataset et à la réalisation des modèles de prédiction, nous allons d'abord découvrir, comprendre et analyser le dataset.

IV.4.1 Description des données

Le dataset utilisé dans notre étude a été acquis auprès d'une source externe, plus précisément tiré de [47]. Il comporte 5548 observations et 10 variables, couvrant la période du 02/01/1997 au 28/02/2019. Ce dataset inclut les variables suivantes :

- **Date :** Représente la date.
- **SWTI :** Prix au comptant du pétrole brut de West Texas.
- **FWTI :** Prix à terme du pétrole brut de West Texas.
- **COM :** Représente la demande totale de pétrole et est calculée en utilisant la somme de la consommation des principaux pays, incluant les données pour le Canada, la France, l'Allemagne, l'Italie, le Japon, la Corée du Sud, le Royaume-Uni, les États-Unis et la Chine.
- **PRO :** Représente la production mondiale de pétrole.
- **GAS :** Prix à terme du gaz naturel.
- **GOLD :** Prix de clôture des contrats à terme sur l'or.
- **SIL :** Prix de clôture des contrats à terme sur l'argent.
- **RDL :** Taux de change du dollar converti en utilisant 2010 comme date de base.
- **RUB :** Taux de change du rouble converti en utilisant l'année 2010 comme période de référence.

IV.4.2 Résumés statistiques de base

Comprendre les propriétés statistiques de chaque variable est crucial pour une analyse de données précise. Elles nous aident à évaluer la dispersion et la distribution des données dans un ensemble, ainsi qu'à saisir leur tendance centrale et leur variabilité. La figure IV.2 illustre et résume les propriétés statistiques de chaque variable. Par exemple, pour les prix au comptant du pétrole (SWTI), la moyenne est de 56.131213. L'écart-type, qui indique la dispersion des valeurs autour de la moyenne, est de 29.131305, ce qui montre une dispersion significative des données. La médiane est de 52.24, différente de la moyenne, ce qui suggère que la distribution des données est asymétrique avec une tendance vers la droite. La valeur minimale est de 10.82, tandis que la valeur maximale est de 145.31, ce qui indique une grande variabilité dans l'ensemble des prix au comptant du pétrole.

```
[5]: df.describe()
```

[5]:

	FWTI	SWTI	PRO	COM	GAS	RDL	RUB	GOLD	SIL
count	5547.000000	5547.000000	5547.000000	5547.000000	5547.000000	5547.000000	5547.000000	5547.000000	5547.000000
mean	56.165008	56.131213	73674.682463	44172.156411	4.401282	97.177591	225.102028	832.224049	13.589400
std	29.139667	29.131305	5144.922765	2492.195800	2.251733	12.755228	537.402911	479.026704	8.773069
min	10.720000	10.820000	64307.900000	38030.840000	1.630000	68.970000	3.360000	253.000000	4.030000
25%	29.640000	29.730000	68772.780000	42397.330000	2.780000	91.345000	66.595000	334.550000	5.210000
50%	52.380000	52.240000	73835.870000	44159.280000	3.770000	98.100000	107.190000	803.000000	13.420000
75%	78.965000	78.860000	76756.490000	45712.300000	5.520000	103.795000	160.865000	1256.050000	17.530000
max	145.290000	145.310000	84305.480000	50988.580000	15.380000	128.070000	3414.400000	1873.700000	48.420000

Fig. IV.2 : Résumés statistiques

IV.4.3 Analyse de l'évolution temporelle des variables

La figure IV.3 illustre l'évolution des prix au comptant et à terme du pétrole ainsi que d'autres variables au fil du temps. Il est constaté qu'il y a eu une fluctuation des prix du pétrole, avec une forte baisse observée dans la période de 1997-1998 en raison de la crise financière qui a touché l'Asie, entraînant une diminution de la demande de pétrole dans cette région. En 2008, les prix ont atteint un niveau record, dépassant les 140 dollars le baril en juillet 2008, mais à la fin de cette même année, une nouvelle baisse des prix s'est produite, entraînant une diminution de la demande. Par la suite, les prix ont commencé à augmenter.

En 2011, les prix du pétrole brut ont régulièrement augmenté, atteignant des pics d'environ 113 dollars le baril. En 2014, les prix ont commencé à chuter jusqu'à atteindre leur valeur la plus basse le 11 février 2016, où le prix du WTI est tombé à environ 26 dollars le baril. Cette baisse s'explique par plusieurs facteurs, tels que la surproduction mondiale et le stockage élevé de pétrole dans certains pays. Par la suite, les prix ont commencé à augmenter en raison de la décision de l'Organisation des Pays Exportateurs de Pétrole (OPEP) de réduire la production de pétrole pour soutenir les prix.

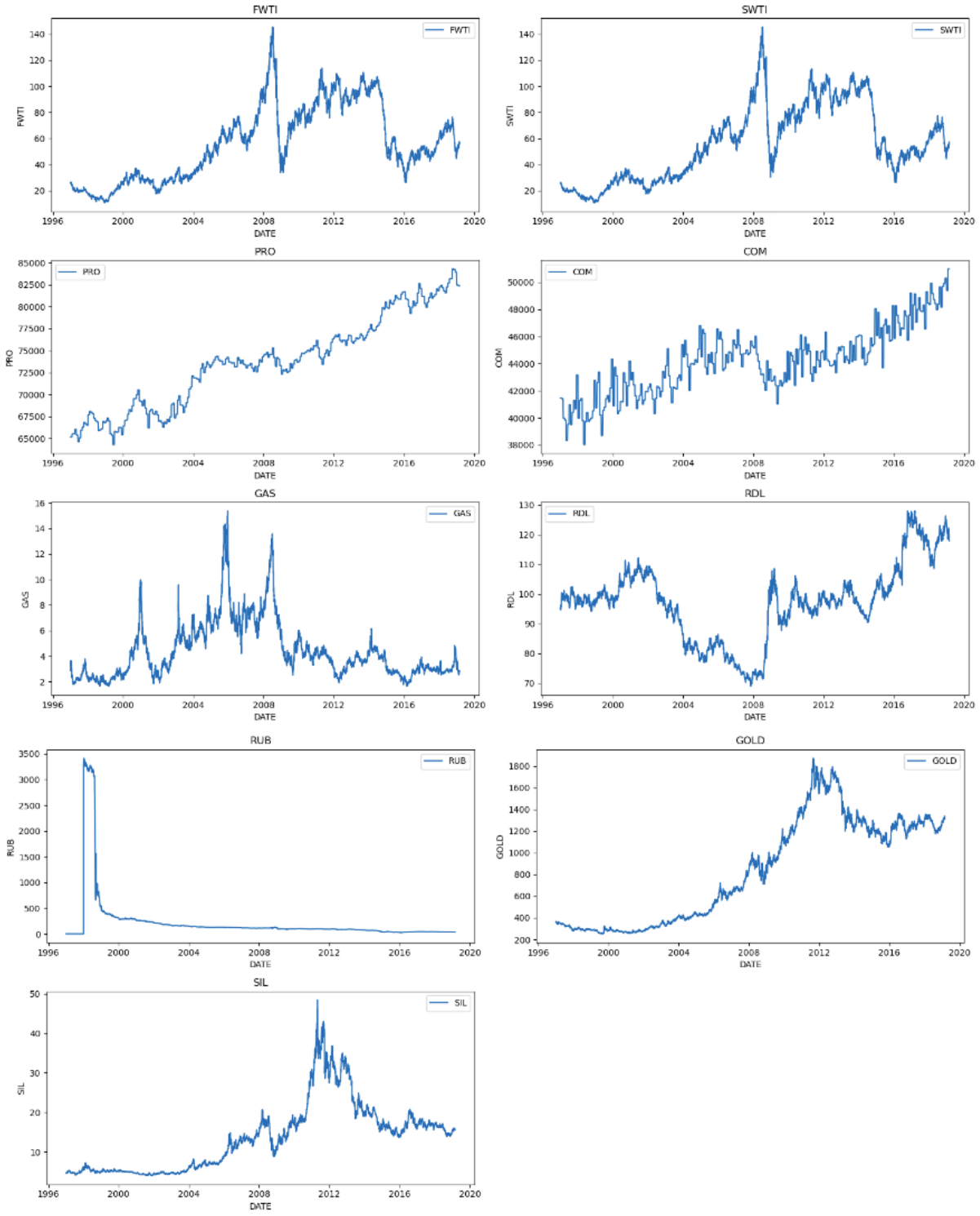

FIG. IV.3 : Évolution temporelle

IV.4.4 Analyse des relations entre les prix au comptant du pétrole et d'autres variables

D'après la figure IV.4, qui illustre la relation entre les prix au comptant du pétrole (SWTI) et les autres variables, nous observons des motifs complexes dans les nuages de points plutôt qu'une simple ligne droite, mettant ainsi en évidence une non-linéarité significative. Cette observation souligne la complexité des relations entre les prix du pétrole et les facteurs qui les influencent, notamment avec des variables telles que le taux de change du dollar (RDL), le taux de change du rouble (RUB) et les prix à terme du gaz naturel (GAS).

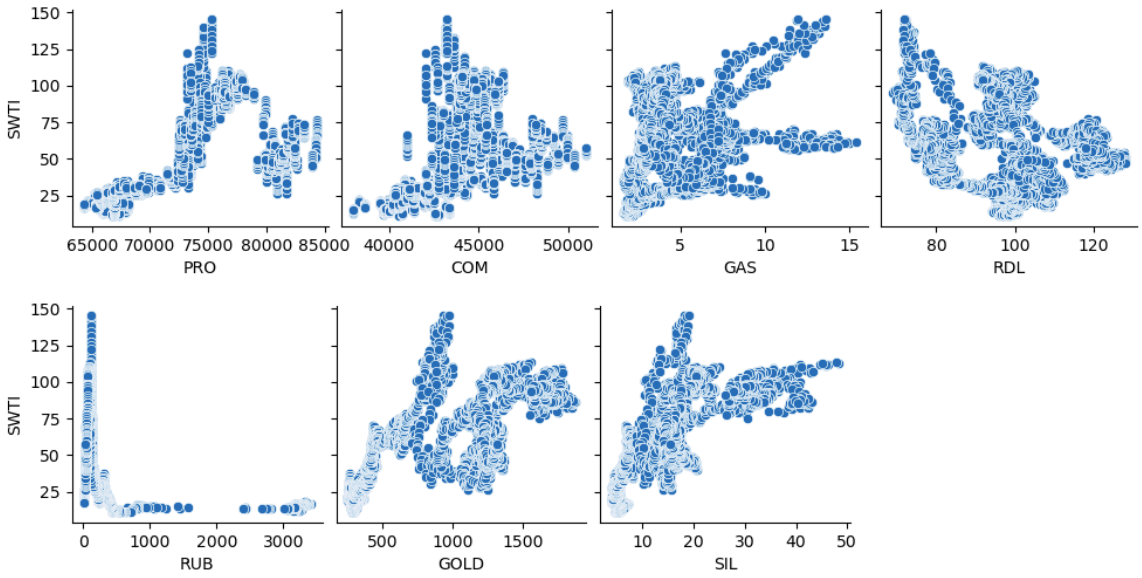

FIG. IV.4 : Relations entre les prix au comptant du pétrole et d'autres variables

IV.4.5 Analyse de degré de corrélation

Dans le contexte de ce dataset, comme le montre la figure IV.5, une forte corrélation existe entre les prix au comptant et les prix à terme du pétrole, les prix de clôture des contrats à terme sur l'argent et sur l'or, ainsi que la production mondiale de pétrole, avec des coefficients de corrélation de 0.79, 0.75 et 0.56 respectivement. Une corrélation modérée est également observée entre les prix au comptant et à terme du pétrole et la demande totale de pétrole, ainsi qu'entre les prix à terme du gaz naturel et le taux de change du rouble, avec des coefficients de corrélation de 0.4, 0.35 et -0.33 respectivement. Enfin, une faible corrélation existe entre les prix au comptant et à terme du pétrole et le taux de change du dollar, avec un coefficient de corrélation de -0.28.

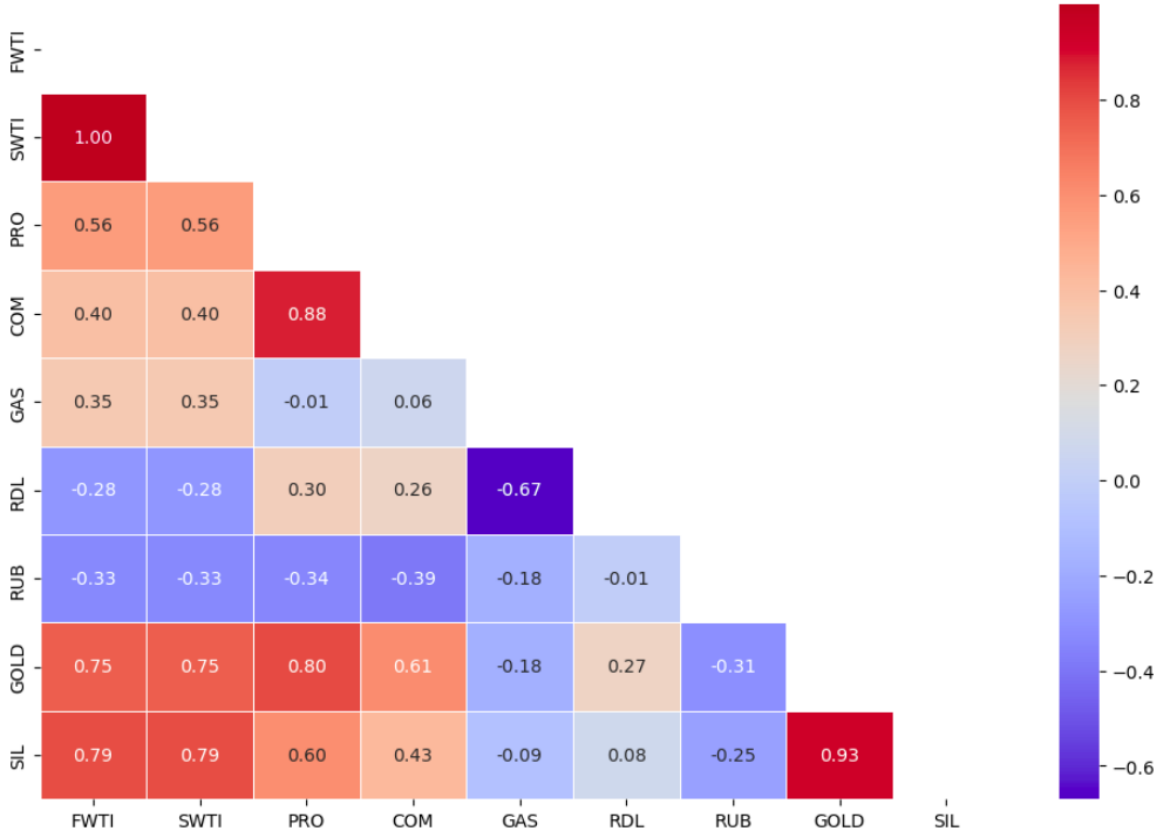

FIG. IV.5 : Matrice de corrélation

IV.4.6 Identification des valeurs manquantes

Comme le montre la figure IV.6, ce dataset est complet et ne contient aucune valeur manquante.

```
# Afficher le nombre de valeurs manquantes par variable
df.isnull().sum()
```

```
FWTI    0
SWTI    0
PRO     0
COM     0
GAS     0
RDL     0
RUB     0
GOLD    0
SIL     0
dtype: int64
```

Fig. IV.6 : Valeurs manquantes par variable

IV.5 Prétraitement des données

Le prétraitement des données est une autre étape indispensable pour la réalisation d'un modèle de prédiction.

IV.5.1 Partitionnement de jeu de données

Il est important de préciser que dans le cadre des séries temporelles, le partitionnement des données ne s'effectue pas de manière aléatoire. Il est judicieux de conserver l'ordre temporel des données, c'est-à-dire que les données utilisées pour l'entraînement du modèle précèdent chronologiquement les données utilisées pour le test. La figure IV.7 illustre ce principe de partitionnement des séries temporelles.

Fig. IV.7 : Principe de partitionnement des séries temporelles [37]

En ce qui concerne le jeu de données que nous avons utilisé, comme illustré dans les figures IV.8 et IV.9 nous avons choisi la période allant du 02/02/1997 jusqu'au 31/12/2014 pour l'entraînement et la période du 01/01/2015 jusqu'au 28/02/2019 pour le test et l'évaluation, comme ils l'ont fait dans [47], afin de pouvoir mener une comparaison plus fiable.

```
var_for_dividing=['SWTI','PRO','COM','GAS', 'RDL', 'RUB', 'GOLD', 'SIL']


train_start_date = pd.to_datetime('1997-01-01')
train_end_date = pd.to_datetime('2014-12-31')

df_train = df.loc[(df.index >= train_start_date) & (df.index <= train_end_date), var_for_dividing]

test_start_date = pd.to_datetime('2015-01-01')
test_end_date = pd.to_datetime('2019-02-28')
df_test = df.loc[(df.index >= test_start_date) & (df.index <= test_end_date), var_for_dividing]
```

FIG. IV.8 : Partitionnement de dataset pour la prédiction des prix au comptant (spot prices)

```
var_for_dividing=['FWTI','PRO','COM','GAS', 'RDL', 'RUB', 'GOLD', 'SIL']


train_start_date = pd.to_datetime('1997-01-01')
train_end_date = pd.to_datetime('2014-12-31')

df_train = df.loc[(df.index >= train_start_date) & (df.index <= train_end_date), var_for_dividing]

test_start_date = pd.to_datetime('2015-01-01')
test_end_date = pd.to_datetime('2019-02-28')
df_test = df.loc[(df.index >= test_start_date) & (df.index <= test_end_date), var_for_dividing]
```

FIG. IV.9 : Partitionnement de dataset pour la prédiction des prix à terme (future prices)

IV.5.2 Normalisation des données

La normalisation des données permet de mettre les valeurs sur la même échelle entre 0 et 1, ce qui permet au modèle de converger rapidement et d'améliorer sa stabilité numérique. Afin de s'assurer que l'ensemble de test reste séparé de l'ensemble d'entraînement, nous avons appliqué la normalisation après le partitionnement des données. Pour ce faire, comme l'illustre la figure IV.10 , nous avons opté pour la méthode de z-normalisation. Cette technique implique de soustraire la moyenne à chaque variable des deux ensembles et de diviser par son écart type, où la moyenne et l'écart type sont calculés à partir de l'ensemble d'entraînement.

```
mean_train=df_train.mean()
std_train=df_train.std()

df_train_normalise=(df_train-mean_train)/std_train
df_test_normalise=(df_test-mean_train)/std_train
```

FIG. IV.10 : Normalisation des données

IV.5.3 Réduction de la dimensionnalité

Dans cette étape, nous avons réduit la dimensionnalité du dataset en trois composantes en utilisant deux techniques : la KPCA et l'UMAP. Nous détaillerons l'application de ces deux techniques dans la section IV.6, dédiée à la réduction de la dimensionnalité.

IV.5.4 Préparation des séquences

L'analyse et la prédiction des séries temporelles, ainsi que l'utilisation du modèle RNN et de ses variantes pour traiter des données séquentielles, nécessitent la création de séquences de taille souhaitée contenant des observations passées pour prédire la sortie qui suit la dernière observation d'une séquence. Dans ce processus, un concept appelé « fenêtre glissante » est utilisé. Cela signifie qu'après la création d'une séquence, la fenêtre est décalée d'un pas de temps défini pour créer la séquence suivante. La figure IV.11 illustre le principe de préparation des séquences de taille 4 avec un pas de temps de taille 1. Ces paramètres sont donnés à titre d'exemple.

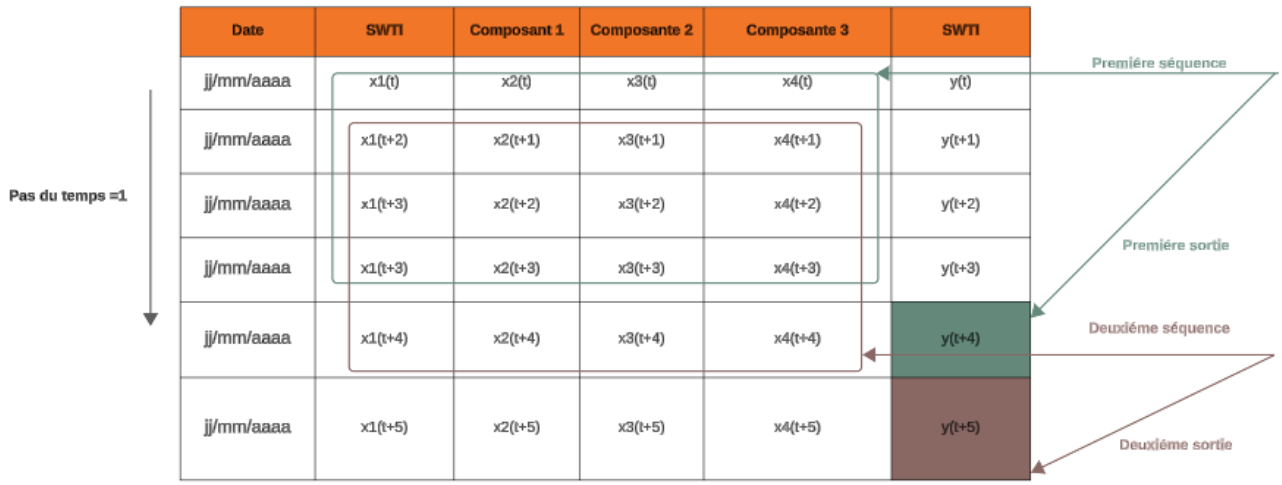

Fig. IV.11 : Principe de préparation des séquences

En ce qui concerne notre travail, nous avons créé des séquences de taille respectivement 40 (deux mois), 60 (trois mois) et 80 (quatre mois) pour la prédiction des prix au comptant et à terme du pétrole en utilisant une fonction TimeseriesGenerator de Keras. Cette fonction prend en paramètre les données d'entrée, la variable cible, le pas de temps et la taille de lot. Nous avons choisi ces tailles afin de comparer nos résultats avec ceux obtenus par [47].

La figure IV.12 illustre la création de séquences de taille 40 pour la prédiction des prix au comptant.

```
caracteristiques_kpca=['SWTI' ,'Composante_1','Composante_2','Composante_3']
cible=['SWTI']
caracteristiques_len=len(caracteristiques_kpca)
sequence_len=40
batch_size=32
epochs=20
```

```
train_generator = TimeseriesGenerator(kpca_df_train[caracteristiques_kpca].to_numpy().astype(np.float32),
                                      kpca_df_train[cible].to_numpy().astype(np.float32),
                                      length=sequence_len,
                                      stride=1,
                                      batch_size=batch_size)

test_generator=TimeseriesGenerator(kpca_df_test[caracteristiques_kpca].to_numpy().astype(np.float32),
                                   kpca_df_test[cible].to_numpy().astype(np.float32),
                                   length=sequence_len,
                                   stride=1,
                                   batch_size=batch_size)
```

Fig. IV.12 : Création des séquences de taille 40 pour la prédiction des prix au comptant

IV.6 Réduction de la dimensionnalité

Dans cette phase nous avons réduit la dimensionnalité du dataset en 3 en utilisant deux techniques non linéaires différentes.

IV.6.1 Réduction de la dimensionnalité avec KPCA

KPCA transforme les données dans un espace de dimension plus élevée afin de détecter les motifs non linéaires, puis réduit la dimensionnalité. Dans un premier temps, comme le montre la figure IV.13, nous laissons le nombre de composantes par défaut et calculons les valeurs propres des composantes. Nous déterminons le nombre de composantes au moment où la valeur propre commence à se stabiliser. Ensuite, nous utilisons ce nombre de dimensions pour re-sélectionner le nombre de composantes afin de réduire la dimension du jeu de données en utilisant le ratio de variance cumulative.

```
var_for_reducing=['PRO','COM','GAS', 'RDL', 'RUB', 'GOLD', 'SIL']
kpca = KernelPCA( kernel="rbf", gamma=0.01, fit_inverse_transform=True, alpha=0.1)
kpca_train = kpca.fit_transform(df_train_normalise[var_for_reducing])
```

Fig. IV.13 : Application de KPCA sans précision du nombre de composantes

En examinant la figure IV.14, qui illustre les valeurs propres de chaque composante, nous remarquons que les valeurs propres commencent à se stabiliser à partir de la 7ème composante. Nous utiliserons donc ce nombre de composantes pour réduire la dimension du jeu de données.

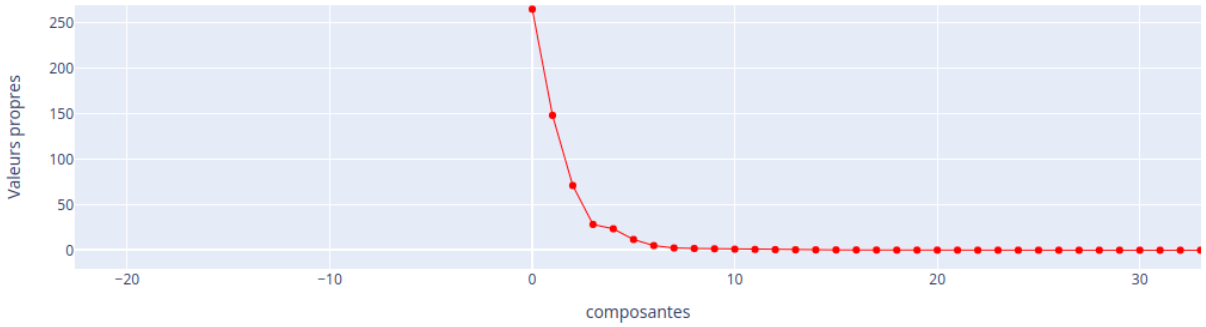

FIG. IV.14 : Valeurs propres des composantes principales

La figure IV.15 illustre l'application de KPCA en initialisant le paramètre du nombre de composantes à 7, à partir duquel nous avons mesuré et évalué la quantité d'information contenue dans les composantes principales.

```
kpca_test = KernelPCA(n_components=7, kernel="rbf", gamma=0.01, fit_inverse_transform=True, alpha=0.1)
kpca_test = kpca_test.fit_transform(df_train_normalise[caracteristiques_for_reduce])
```

FIG. IV.15 : Application de KPCA avec précision de nombre de composantes égal à 7

Pour mesurer et évaluer la quantité d'information contenue dans les composantes principales, deux mesures couramment utilisées sont :

- **Ratio de Variance Expliquée** qui est donnée par l'équation : $\frac{\lambda_{\text{comp}}}{\sum_{i=1}^{p} \lambda_i}$

  Où λ_{comp} représente la valeur propre de la composante principale spécifique, et $\sum_{i=1}^{p} \lambda_i$ est la somme totale des valeurs propres pour toutes les composantes principales.

- **Ratio de Variance Cumulative** qui est donnée par l'équation : $\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$

  Où k est le nombre de composantes principales considérées, et λ_i sont les valeurs propres (variances expliquées) des composantes principales.

Afin de trouver un compromis entre la réduction maximale de la dimensionnalité et la conservation de l'information, nous avons choisi de conserver les trois premières composantes principales, qui capturent une quantité significative d'information (87.5% de variance cumulative expliquée).

Le tableau IV.1 montre le résultat du ratio de variance cumulative ainsi que du ratio de variance expliquée des trois premières composantes principales que nous avons gardées.

Composante	Ratio de Variance Cumulative	Ratio de Variance Expliquée
1	0.479	0.479
2	0.747	0.2680
3	0.875	0.128

TAB. IV.1 : Ratio de variance cumulée et expliqué des trois premières composantes

La Figure IV.16 présente l'application de la KPCA à la fois sur l'ensemble d'entraînement et sur l'ensemble de test, réduisant ainsi ces ensembles à trois dimensions.

```
kpca_model = KernelPCA(n_components=3, kernel="rbf", gamma=0.01, fit_inverse_transform=True, alpha=0.1)
kpca_model.fit(df_train_normalise[caracteristiques_for_reduce])
```

```
kpca_train = kpca_model.transform(df_train_normalise[caracteristiques_for_reduce])
kpca_test = kpca_model.transform(df_test_normalise[caracteristiques_for_reduce])
```

Fig. IV.16 : Réduction de la dimensionnalité en 3 composantes principales avec KPCA

IV.6.1.1 Paramètres du modèle KPCA

Les paramètres que nous avons initialisés sont :

- **Kernel :** Ce paramètre spécifie le type de noyau à utiliser dans la transformation KPCA.
- **gamma :** Ce paramètre est spécifique au noyau RBF (gaussien). Il contrôle la largeur de la gaussienne et influence ainsi la flexibilité de la transformation.
- **alpha :** L'hyperparamètre alpha contrôle la régularisation dans KPCA. Il est utilisé pour éviter le surajustement (overfitting) en restreignant la complexité du modèle.
- **fit_inverse_transform** : Ce paramètre contrôle si la méthode fit_inverse_transform doit être disponible après l'ajustement. Si cela est défini sur True, cela permettra d'inverser la transformation KPCA et de récupérer les données originales (ou des approximations de celles-ci) à partir des composantes principales.

Les valeurs associées à chaque paramètre sont résumées dans le tableau IV.2.

Paramètre	Valeur
kernel	rbf
gamma	0.01
fit_inverse_transform	True
alpha	0.1

Tab. IV.2 : Paramètres du KPCA

IV.6.2 Réduction de la dimensionnalité avec UMAP

Nous avons également utilisé la technique UMAP pour la réduction de la dimensionnalité en 3. Nous avons initialisé les paramètres suivants :

- **n_component :** Représente le nombre de dimension de l'espace réduit.
- **n_neighbors :** Ce paramètre contrôle la taille du voisinage local que UMAP utilise pour l'apprentissage de la structure de la manifold.

- **min_dis :** Il permet de contrôler à quel point les points proches sont rassemblés dans l'espace latent.

  Les valeurs associées à chaque paramètre sont résumées dans le tableau IV.3.

Paramètre	Valeur
n_component	3
n_neighbors	10
min_dis	0.1

Tab. IV.3 : Paramètres d'UMAP

Remarque : Contrairement à la KPCA, après la réduction de la dimensionnalité avec UMAP, nous avons constaté que la moyenne des composantes était éloignée de 0 et que l'écart-type était éloigné de 1. Par conséquent, nous avons normalisé ces composantes afin de permettre aux modèles de réseaux de neurones récurrents de converger rapidement et d'améliorer leur stabilité.

IV.7 Architecture des modèles de réseaux de neurones

Dans la partie réseaux de neurones, nous avons utilisé trois modèles, à savoir RNN, GRU et LSTM, avec des architectures différentes.

IV.7.1 Architecture du modèle RNN

Le modèle RNN comporte une couche RNN et une couche dense , la figure IV.17 illustre l'architecture de ce modèle.

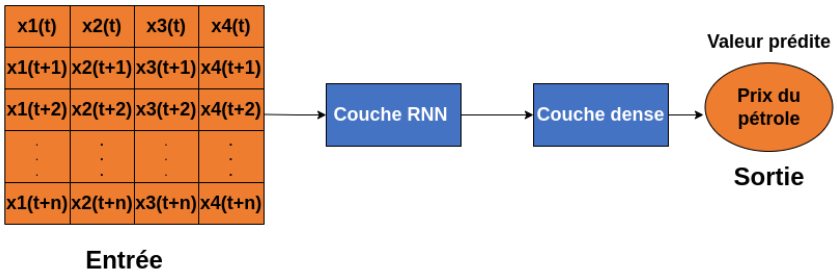

Fig. IV.17 : Architecture du modèle RNN

La figure IV.18 présente un résumé textuel de l'architecture du modèle RNN incluant certains paramètres.

```
modele_rnn.summary()
```

```
Model: "sequential_2"
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 simple_rnn_2 (SimpleRNN)     (None, 100)               10500

 dense_2 (Dense)              (None, 1)                 101

=================================================================
Total params: 10,601
Trainable params: 10,601
Non-trainable params: 0
_________________________________________________________________
```

FIG. IV.18 : Résumé textuel de l'architecture du modèle RNN

IV.7.2 Architecture du modèle GRU

Comme le montre la figure IV.19, le modèle GRU comporte deux couches GRU et une couche dense.

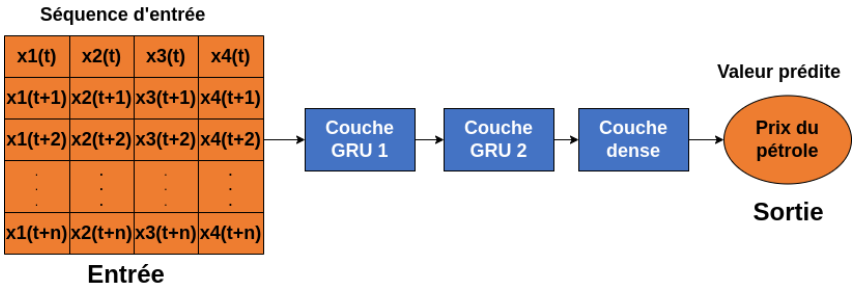

FIG. IV.19 : Architecture du modèle GRU

Le résumé textuel de l'architecture détaillé du modèle GRU est illustré dans la figure IV.20.

```
modele_gru.summary()
```

```
Model: "sequential_11"
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 gru_16 (GRU)                 (None, 40, 100)           31800

 gru_17 (GRU)                 (None, 100)               60600

 dense_11 (Dense)             (None, 1)                 101

=================================================================
Total params: 92,501
Trainable params: 92,501
Non-trainable params: 0
_________________________________________________________________
```

FIG. IV.20 : Résumé textuel de l'architecture du modèle GRU

IV.7.3 Architecture du modèle LSTM

Le modèle LSTM comprend deux couches LSTM et une couche dense, comme illustré dans la figure IV.21.

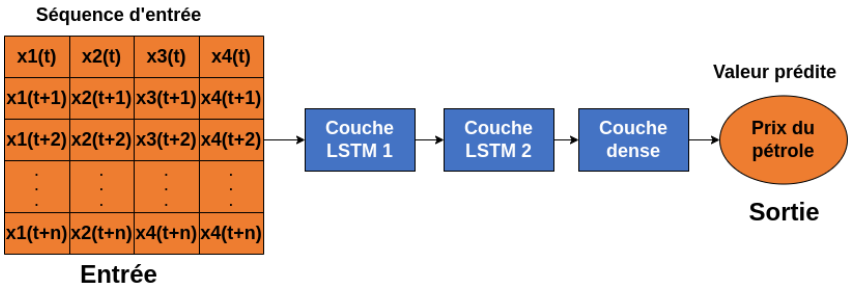

FIG. IV.21 : Architecture du modèle LSTM

La figure IV.22 illustre le résumé textuel de l'architecture du modèle LSTM.

```
modele_lstm.summary()
```

```
Model: "sequential_12"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, 40, 100)           42000

 lstm_1 (LSTM)               (None, 100)               80400

 dense_12 (Dense)            (None, 1)                 101

=================================================================
Total params: 122,501
Trainable params: 122,501
Non-trainable params: 0
```

FIG. IV.22 : Résumé textuel de l'architecture du modèle LSTM

IV.7.4 Paramètres des trois modèles (RNN, GRU et LSTM)

Les trois modèles comportent les paramètres suivants qui doivent être initialisés :

- **Optimizer (optimiseur)** : est un algorithme qui permet d'ajuster les poids du modèle afin de minimiser la fonction de perte.
- **Nombre d'époque (epochs)** : C'est le nombre d'itérations complètes sur l'ensemble des données d'entraînement.
- **Taille de lot (batch_size)** : Le nombre d'échantillons utilisés pour une seule mise à jour du modèle.
- **Fonction d'activation (activation)** : Elle détermine les sorties des cellules. Généralement, elle introduit la non-linéarité dans le modèle afin de capturer les motifs non linéaires et complexes.

- **Métriques de performance (loss et metrics)** : Elles permettent d'évaluer les performances et la qualité du modèle lors de l'entraînement et du test.

Le tableau IV.4 présente les paramètres relatifs aux trois modèles RNN, GRU et LSTM.

Paramètre	Valeur
optimizer	adam
epochs	20
batch_size	32
activation	- tanh (fonction d'activation par défaut) - Linear pour la couche dense (fonction par défaut)
loss	MAE
metrics	MSE

Tab. IV.4 : Paramètres des modèles RNN, GRU et LSTM

IV.8 Évaluation et comparaison des modèles

Après avoir entraîné les modèles, nous sommes passés à leur évaluation et à leur comparaison avec les modèles existants.

IV.8.1 Évaluation des modèles proposés

Nous avons évalué nos trois modèles sur des fenêtres de taille respectivement 40, 60 et 80, en utilisant les deux métriques de performances suivantes :

- **MAE (Mean Absolute Error)** : C'est l'une des métriques de performance les plus populaires. Elle permet de quantifier l'ampleur des erreurs sans distinction entre les surestimations et les sous-estimations, et elle est également robuste aux valeurs aberrantes [2]. Elle est calculée de la manière suivante : $\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$, où y_i représente les valeurs réelles et $\hat{y}_i$ représente les valeurs prédites.

- **RMSE (Root Mean Squared Error) :** Une autre métrique très populaire dans les problèmes de régression. Elle est calculée de la maniéré suivante : $\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$, où y_i représente les valeurs réelles et $\hat{y}_i$ représente les valeurs prédites.

  Les deux tableaux IV.5 et IV.6 illustrent les erreurs de MAE et RMSE des modèles (KPCA-RNN, UMAP-RNN, KPCA-GRU, UMAP-GRU, KPCA-LSTM, UMAP-LSTM) dans la prédiction des prix au comptant et à terme sur des fenêtres de taille 40, 60 et 80. Les résultats montrent que les modèles de réseaux de neurones combinés avec KPCA surpassent ceux combinés avec UMAP, à l'exception de la prédiction des prix au comptant où la MAE de UMAP-GRU est inférieure à celle de KPCA-GRU pour la fenêtre de taille 80, et dans la prédiction des prix à terme pour la fenêtre de taille 60 où UMAP-GRU surpasse KPCA-GRU.

Modèles	Fenêtre de taille 40		Fenêtre de taille 60		Fenêtre de taille 80	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
KPCA-RNN	**0.0280**	**0.0372**	**0.0284**	**0.0376**	**0.0276**	**0.0368**
UMAP-RNN	0.0304	0.0407	0.0320	0.0422	0.0303	0.0404
KPCA-GRU	**0.0272**	**0.0360**	**0.0288**	**0.0370**	0.0365	**0.0456**
UMAP-GRU	0.0312	0.0410	0.0377	0.0465	**0.0355**	0.0461
KPCA-LSTM	**0.0445**	**0.0532**	**0.0302**	**0.0394**	**0.0280**	**0.0368**
UMAP-LSTM	0.0637	0.0730	0.0381	0.0474	0.0306	0.0400

Tab. IV.5 : Erreurs des modèles proposés avec différentes tailles de fenêtre (prix au comptant)

Modèles	Fenêtre de taille 40		Fenêtre de taille 60		Fenêtre de taille 80	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
KPCA-RNN	**0.0325**	**0.0405**	**0.0282**	**0.0365**	**0.0290**	**0.0379**
UMAP-RNN	0.0383	0.0481	0.0422	0.0531	0.0389	0.0497
KPCA-GRU	**0.0371**	**0.0474**	0.0385	0.0480	**0.0263**	**0.0344**
UMAP-GRU	0.0393	0.0488	**0.0340**	**0.0435**	0.0289	0.0373
KPCA-LSTM	**0.0294**	**0.0384**	**0.0295**	**0.0380**	**0.0282**	**0.0368**
UMAP-LSTM	0.0362	0.0458	0.0335	0.0437	0.0350	0.0440

Tab. IV.6 : Erreurs des modèles proposés avec différentes tailles de fenêtre (prix à terme)

La figure IV.23 présente le graphique comparatif des prédictions des prix au comptant réalisées par les modèles KPCA-RNN, KPCA-GRU et KPCA-LSTM. Nous observons que le modèle KPCA-RNN se distingue par une excellente capacité de généralisation dans la prédiction, quelle que soit la taille de la fenêtre. Le modèle KPCA-GRU affiche également une excellente généralisation pour les fenêtres de taille 40 et 60, bien que sa performance diminue légèrement pour une fenêtre de taille 80. En ce qui concerne le modèle KPCA-LSTM, il démontre une très bonne généralisation pour les fenêtres de taille 60 et 80, mais une performance relativement moindre pour une fenêtre de taille 40.

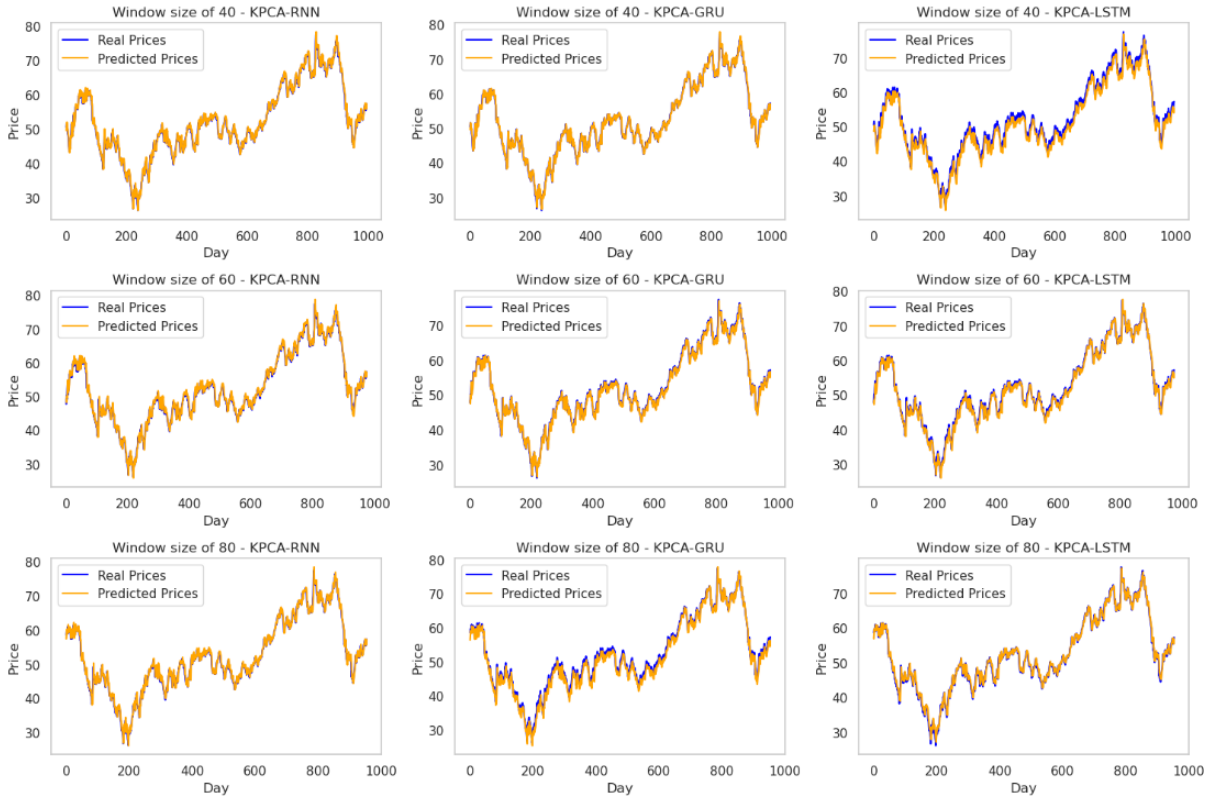

FIG. IV.23 : Prédiction des prix au comptant avec KPCA-RNN, KPCA-GRU et KPCA-LSTM

La figure IV.24 présente le graphique comparatif des prédictions des prix au comptant réalisées par les modèles UMAP-RNN, UMAP-GRU et UMAP-LSTM. Nous remarquons que le modèle UMAP-RNN se démarque également par son excellente capacité de généralisation dans la prédiction, indépendamment de la taille de la fenêtre. Le modèle UMAP-GRU affiche également une très bonne généralisation pour les fenêtres de taille 40 et 80, bien que sa performance diminue légèrement pour une fenêtre de taille 60. Quant au modèle UMAP-LSTM, il démontre une bonne généralisation pour les fenêtres de taille 60 et 80, mais une performance relativement moindre pour une fenêtre de taille 40.

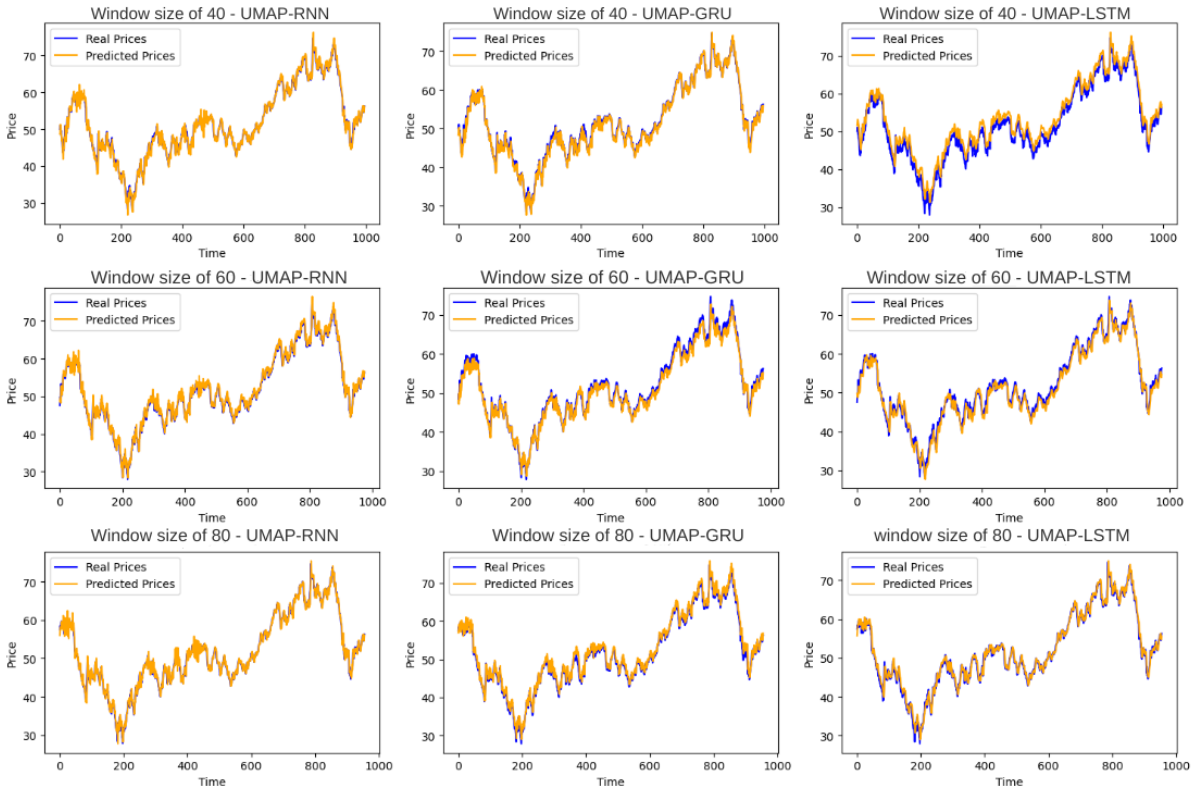

FIG. IV.24 : Prédiction des prix au comptant avec UMAP-RNN, UMAP-GRU et UMAP-LSTM

La figure IV.25 présente le graphique comparatif des prédictions des prix à terme réalisées par les modèles KPCA-RNN, KPCA-GRU et KPCA-LSTM. Nous remarquons que le modèle KPCA-RNN se distingue par son excellente capacité de généralisation pour toutes les tailles de fenêtre. Le modèle KPCA-GRU montre également une excellente généralisation pour la fenêtre de taille 80, mais une généralisation moins bonne pour les fenêtres de taille 40 et 60. En ce qui concerne le modèle KPCA-LSTM, il démontre une excellente généralisation pour toutes les fenêtres.

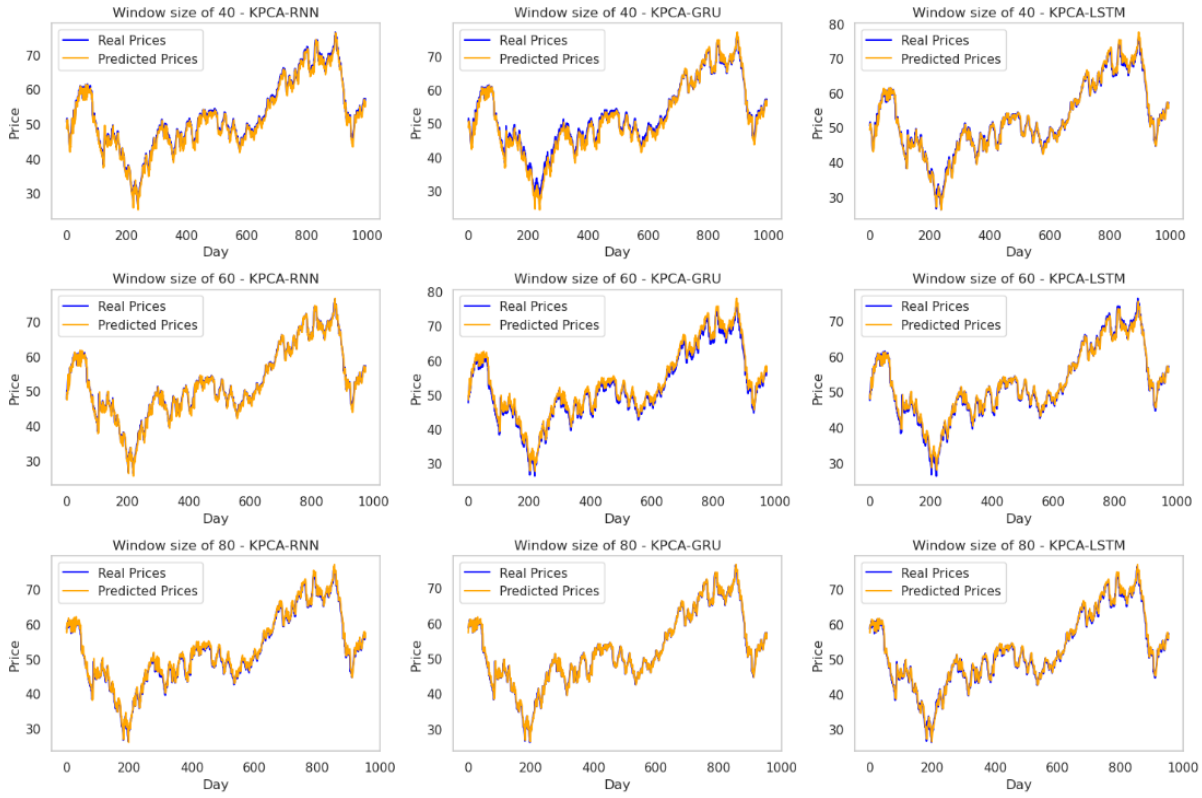

FIG. IV.25 : Prédiction des prix à terme avec KPCA-RNN, KPCA-GRU et KPCA-LSTM

La figure IV.26 présente le graphique comparatif des prédictions des prix à terme réalisées par les modèles UMAP-RNN, UMAP-GRU et UMAP-LSTM. Nous remarquons que le modèle UMAP-RNN offre une excellente généralisation, le modèle UMAP-GRU une très bonne généralisation et le modèle UMAP-LSTM une bonne généralisation, et ce, pour toutes les tailles de fenêtres temporelles.

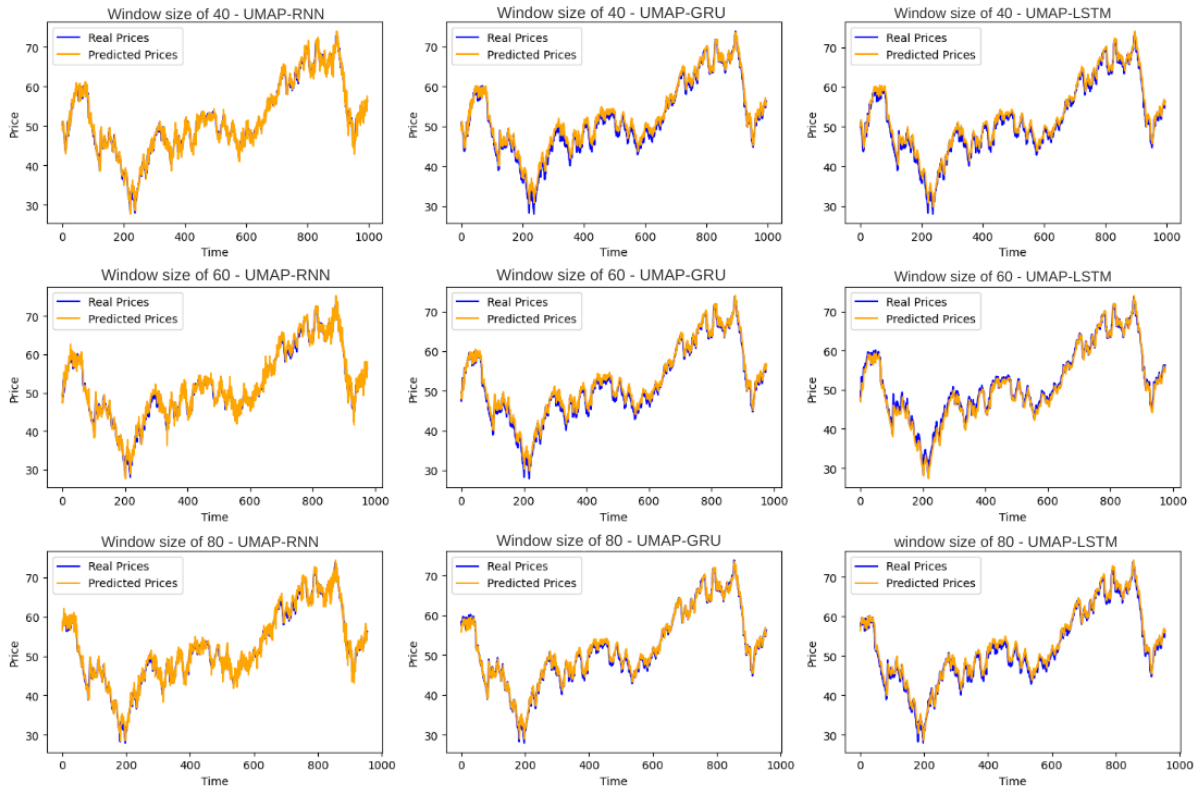

FIG. IV.26 : Prédiction des prix à terme avec UMAP-RNN, UMAP-GRU et UMAP-LSTM

IV.8.2 Comparaison des modèles proposés avec les modèles existants

Nous avons comparé nos modèles proposés aux modèles RNN, GRU et LSTM non réduits en dimensionnalité, ainsi qu'au modèle le plus performant de [47], le LLE-RNN. Comme le montrent les tableaux IV.7 et IV.8, nos modèles surpassent les modèles non combinés à des techniques de réduction de la dimensionnalité dans la prédiction des prix au comptant et à terme du pétrole.

Notre comparaison avec le modèle LLE-RNN de [47] révèle que plusieurs de nos modèles proposés montrent des performances supérieures dans la prédiction des prix au comptant et à terme du pétrole :

- **Dans la prédiction des prix au comptant**, les modèles KPCA-RNN et KPCA-GRU surpassent tous deux le modèle LLE-RNN pour toutes les fenêtres temporelles étudiées, à l'exception d'une seule où LLE-RNN présente une MAE inférieure à celle de KPCA-GRU. KPCA-LSTM, quant à lui, domine LLE-RNN pour les fenêtres de taille 60 et 80. UMAP-RNN affiche des performances supérieures à LLE-RNN dans tous les cas, excepté pour la RMSE de la fenêtre de taille 60. UMAP-GRU surpasse LLE-RNN dans toutes les situations sauf pour la fenêtre de taille 60, tandis que UMAP-LSTM est supérieur à ce dernier pour la fenêtre de taille 80.

- **Dans la prédiction des prix à terme**, le modèle KPCA-RNN surpasse LLE-RNN pour les fenêtres de taille 60 et 80, ainsi que dans la RMSE pour la fenêtre de

taille 40. De plus, UMAP-GRU est meilleur que LLE-RNN pour les fenêtres de taille 60 et 80. Pour la fenêtre de taille 80, KPCA-GRU affiche également de meilleures performances que LLE-RNN, tandis que KPCA-LSTM est systématiquement supérieur à LLE-RNN. Enfin, UMAP-LSTM dépasse LLE-RNN pour la fenêtre de taille 60 et dans la RMSE pour la fenêtre de taille 80.

Modèles	Fenêtre de taille 40		Fenêtre de taille 60		Fenêtre de taille 80	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
RNN	0.0396	0.0500	0.0596	0.0775	0.0584	0.0721
KPCA-RNN	**0.0280**	**0.0372**	**0.0284**	**0.0376**	**0.0276**	**0.0368**
UMAP-RNN	**0.0304**	**0.0407**	**0.0320**	0.0422	**0.0303**	**0.0404**
GRU	0.0685	0.0795	0.0836	0.0938	0.0465	0.0578
KPCA-GRU	**0.0272**	**0.0360**	**0.0288**	**0.0370**	0.0365	**0.0456**
UMAP-GRU	**0.0312**	**0.0410**	0.0377	0.0465	**0.0355**	**0.0461**
LSTM	0.0742	0.0855	0.0550	0.0668	0.0720	0.0849
KPCA-LSTM	0.0445	0.0532	**0.0302**	**0.0394**	**0.0280**	**0.0368**
UMAP-LSTM	0.0637	0.0730	0.0381	0.0474	**0.0306**	**0.0400**
LLE-RNN [47]	0.0314	0.0413	0.0323	0.0419	0.0362	0.0479

Tab. IV.7 : Comparaison des erreurs des modèles proposés avec ceux sans réduction de la dimensionnalité et avec le modèle LLE-RNN [47] (prix au comptant)

Modèles	Fenêtre de taille 40		Fenêtre de taille 60		Fenêtre de taille 80	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
RNN	0.0450	0.0562	0.0880	0.103	0.0420	0.0508
KPCA-RNN	0.0325	**0.0405**	**0.0282**	**0.0365**	**0.0290**	**0.0379**
UMAP-RNN	0.0383	0.0481	0.0422	0.0531	0.0389	0.0497
GRU	0.0840	0.1010	0.0790	0.0936	0.0577	0.0702
KPCA-GRU	0.0371	0.0474	0.0385	0.0480	**0.0263**	**0.0344**
UMAP-GRU	0.0393	0.0488	**0.0340**	**0.0435**	**0.0289**	**0.0373**
LSTM	0.0743	0.0872	0.0540	0.0649	0.0805	0.0948
KPCA-LSTM	**0.0294**	**0.0384**	**0.0295**	**0.0380**	**0.0282**	**0.0368**
UMAP-LSTM	0.0362	0.0458	**0.0335**	**0.0437**	0.0350	**0.0440**
LLE-RNN [47]	0.0319	0.0411	0.0344	0.0442	0.0337	0.0443

Tab. IV.8 : Comparaison des erreurs des modèles proposés avec ceux sans réduction de la dimensionnalité et avec le modèle LLE-RNN [47] (prix à terme)

IV.9 Conclusion

En conclusion de ce chapitre, nous avons présenté en détail chaque étape de notre approche pour la prédiction des prix du pétrole. Nous avons également comparé nos modèles à un modèle performant de la littérature, LLE-RNN [47], et avons démontré que nos modèles le surpassent pour plusieurs fenêtres temporelles.

Conclusion Générale et Perspectives

La prédiction des prix du pétrole est un enjeu crucial pour l'économie mondiale en raison de l'impact significatif que ces prix peuvent avoir sur divers secteurs économiques. Cependant, la complexité de cette tâche réside dans la nature non linéaire des nombreux facteurs qui influencent les prix du pétrole. Notre objectif dans ce mémoire a été d'élaborer une approche permettant d'améliorer la précision des prédictions en combinant des techniques non linéaires de réduction de la dimensionnalité avec des méthodes avancées de machine learning, notamment les réseaux de neurones récurrents (RNN) et leurs variantes.

Dans le premier chapitre, nous avons exploré et expliqué des concepts économiques relatifs au marché pétrolier. Cette base théorique est essentielle pour comprendre les mécanismes sous-jacents aux fluctuations des prix du pétrole, et constitue le fondement de notre travail de prédiction.

Le deuxième chapitre a été consacré aux séries temporelles, car les prix du pétrole évoluent dans le temps. Nous avons brièvement introduit ces concepts avant de détailler les réseaux de neurones récurrents et leurs variantes, ainsi que les techniques de réduction de la dimensionnalité que nous avons employées.

Ensuite, dans notre troisième chapitre, nous avons présenté les travaux antérieurs abordant les problématiques relatives à la prédiction des prix du pétrole. Cette revue de littérature nous a permis de comprendre les défis et les solutions existantes, et a guidé notre propre approche pour améliorer les prédictions.

Enfin, dans le quatrième chapitre, nous avons détaillé notre approche méthodologique. Nous avons analysé les données utilisées pour la prédiction, décrit les étapes de prétraitement, y compris la partition des jeux de données pour les séries temporelles et la préparation des séquences pour la prédiction. Nous avons appliqué les techniques de réduction de la dimensionnalité, à savoir KPCA et UMAP, et décrit les architectures des trois modèles de réseaux de neurones que nous avons utilisés. Nous avons ensuite évalué nos modèles et les avons comparés aux modèles RNN, LSTM et GRU sans réduction de la dimensionnalité, ainsi qu'au modèle combinant LLE et RNN de [47].

Nos résultats ont montré que l'utilisation de techniques non linéaires de réduction de la dimensionnalité, notamment la KPCA et l'UMAP, améliore significativement la performance des modèles de réseaux de neurones. Nos modèles proposés basés sur KPCA et UMAP ont surpassé le modèle le plus performant de [47], à savoir LLE-RNN, dans de nombreuses situations. Pour la prédiction des prix au comptant, les modèles KPCA-RNN, UMAP-RNN et KPCA-GRU ont généralement surpassé le modèle LLE-RNN pour toutes les fenêtres temporelles (40, 60 et 80), sauf dans deux cas spécifiques : l'UMAP-RNN a montré une erreur RMSE plus élevée que le LLE-RNN pour une fenêtre temporelle de 60,

tandis que le KPCA-GRU a affiché une erreur MAE supérieure pour une fenêtre de 80. De plus, le modèle UMAP-GRU a surpassé le LLE-RNN pour les fenêtres de taille 40 et 80, l'UMAP-LSTM pour la fenêtre de taille 80, et le KPCA-LSTM pour les fenêtres de taille 60 et 80. Concernant la prédiction des prix à terme, le modèle KPCA-LSTM surpasse le modèle LLE-RNN pour toutes les fenêtres temporelles (40, 60 et 80). Le KPCA-RNN a également surpassé le LLE-RNN pour les fenêtres de 60 et 80, ainsi qu'en termes de RMSE pour la fenêtre de taille 40. Le KPCA-GRU surpasse le LLE-RNN dans la fenêtre de taille 80, tandis que l'UMAP-GRU a présenté des performances supérieures dans les fenêtres de taille 60 et 80, et l'UMAP-LSTM surpasse le LLE-RNN pour la fenêtre de taille 60 et en termes de RMSE pour la fenêtre de taille 80.

Ces résultats prometteurs confirment que l'intégration de techniques de réduction de la dimensionnalité telles que KPCA et UMAP peut renforcer les capacités prédictives des modèles de réseaux de neurones récurrents dans le domaine complexe de la prédiction des prix du pétrole. Cette contribution a été acceptée pour présentation à la conférence internationale AMIE'2024 (Applied Mathematics in Economics 2024), qui se tiendra les 23-24 octobre 2024 à Bejaia.

Dans les perspectives à venir, nous envisageons d'appliquer la technique des encodeurs pour comparer les performances des méthodes mathématiques de réduction de la dimensionnalité à celles des méthodes de machine learning. Nous prévoyons également de tester les transformers et de les comparer aux modèles de réseaux de neurones récurrents ainsi qu'à leurs variantes. En parallèle, nous explorerons la possibilité de généraliser ces modèles pour la prédiction des prix d'autres matières premières.

Bibliographie

[1] Xuedong Liang et al. « Crude oil price prediction using deep reinforcement learning ». en. In : *Resources Policy* 81 (mars 2023), p. 103363. issn : 03014207. doi : 10.1016/j.resourpol.2023.103363. url : https://linkinghub.elsevier.com/retrieve/pii/S0301420723000715.

[2] M. Waqar Ahmed. *Understanding Mean Absolute Error (MAE) in Regression : A Practical Guide*. en. Août 2023. url : https://medium.com/@m.waqar.ahmed/understanding-mean-absolute-error-mae-in-regression-a-practical-guide-26e80ebb97df (visité le 20/03/2024).

[3] Thaís C Azevedo et al. « The behavior of West Texas Intermediate crude-oil and refined products prices volatility before and after the 2008 financial crisis : an approach through analysis of futures contracts ». en. In : *Ingeniare. Revista chilena de ingeniería* 23.3 (sept. 2015), p. 395-405. issn : 0718-3305. doi : 10.4067/S0718-33052015000300008. url : http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-33052015000300008&lng=en&nrm=iso&tlng=en.

[4] Yun Bai et al. « Crude oil price forecasting incorporating news text ». en. In : *International Journal of Forecasting* 38.1 (jan. 2022), p. 367-383. issn : 01692070. doi : 10.1016/j.ijforecast.2021.06.006. url : https://linkinghub.elsevier.com/retrieve/pii/S0169207021001060.

[5] Y. Bengio, P. Simard et P. Frasconi. « Learning long-term dependencies with gradient descent is difficult ». In : *IEEE Transactions on Neural Networks* 5.2 (mars 1994), p. 157-166. issn : 1045-9227, 1941-0093. doi : 10.1109/72.279181. url : https://ieeexplore.ieee.org/document/279181/.

[6] Léa Boluze. *Produits dérivés : définition et utilisation*. fr. Mars 2019. url : https://www.capital.fr/entreprises-marches/produits-derives-1331983 (visité le 15/04/2024).

[7] MERIEM BOUZEGHOUB et AMYRA TOUATI. « LES PRODUITS DERIVES ». fr. In : *La Revue des Sciences Commerciales* 16.4 (avr. 2017), p. 184-196. url : https://www.asjp.cerist.dz/en/article/70917.

[8] Nicolas Carnot et Caterine Hagège. « Le marché pétrolier : » in : *Économie & prévision* n o 166.5 (déc. 2004), p. 127-136. issn : 0249-4744. doi : 10.3917/ecop.166.0127. url : https://www.cairn.info/revue-economie-et-prevision-1-2004-5-page-127.htm?ref=doi.

[9] Kyunghyun Cho et al. *On the Properties of Neural Machine Translation : Encoder-Decoder Approaches*. arXiv :1409.1259 [cs, stat]. Oct. 2014. URL : http://arxiv.org/abs/1409.1259.

[10] *Cours Du Pétrole : Découvrez Notre Solution*. fr-FR. URL : https://www.defthedge.com/gestion-matieres-premieres/cours-du-petrole/ (visité le 12/06/2024).

[11] *Cours du pétrole : qu'est-ce que le Brent ?* fr-FR. Sept. 2023. URL : https://www.fioulmarket.fr/astuces-conseils/petrole-brent-quest-que-cest (visité le 12/06/2024).

[12] *CS 230 - Pense-bête de réseaux de neurones récurrents*. URL : https://stanford.edu/~shervine/l/fr/teaching/cs-230/pense-bete-reseaux-neurones-recurrents (visité le 19/02/2024).

[13] James Daniel. « Time Series Analysis of Brent Crude Oil Prices Per Barrel : A Box-Jenkins Approach ». In : (juin 2023).

[14] Chao Deng, Liang Ma et Taishan Zeng. « Crude Oil Price Forecast Based on Deep Transfer Learning : Shanghai Crude Oil as an Example ». en. In : *Sustainability* 13.24 (déc. 2021), p. 13770. ISSN : 2071-1050. DOI : 10.3390/su132413770. URL : https://www.mdpi.com/2071-1050/13/24/13770.

[15] Saul Dobilas. *UMAP Dimensionality Reduction — An Incredibly Robust Machine Learning Algorithm*. en. Fév. 2024. URL : https://towardsdatascience.com/umap-dimensionality-reduction-an-incredibly-robust-machine-learning-algorithm-b5acb01de568 (visité le 26/03/2024).

[16] Dunarea de Jos University of Galati, Romania et Maria Christina Enache. « Data Analysis with Pandas ». In : *Annals of Dunarea de Jos University of Galati. Fascicle I. Economics and Applied Informatics* 25.2 (juill. 2019), p. 69-74. ISSN : 15840409, 2344441X. DOI : 10.35219/eai1584040933. URL : http://www.eia.feaa.ugal.ro/images/eia/2019_2/Enache1.pdf.

[17] Kenneth Ezukwoke et Samaneh Zareian. « KERNEL METHODS FOR PRINCIPAL COMPONENT ANALYSIS (PCA) A comparative study of classical and kernel PCA ». en. In : (2019). DOI : 10.13140/RG.2.2.17763.09760. URL : http://rgdoi.net/10.13140/RG.2.2.17763.09760.

[18] Hongxiang Fan et al. « Comparison of Long Short Term Memory Networks and the Hydrological Model in Runoff Simulation ». en. In : *Water* 12.1 (jan. 2020), p. 175. ISSN : 2073-4441. DOI : 10.3390/w12010175. URL : https://www.mdpi.com/2073-4441/12/1/175.

[19] Abberrahmani Fares. *GUIDE PRATIQUE DES SERIES TEMPORELLES MACROECONOMIQUES ET FINANCIERES AVEC EVIEWS 9.5*. E - Learning Université de Béjaïa. 2017. URL : https://elearning.univ-bejaia.dz/course/view.php?id=5273 (visité le 08/05/2024).

[20] Florian. *Pourquoi le prix du pétrole varie ?* fr-FR. Juin 2021. URL : https://bpsuperfioul.fr/pourquoi-prix-petrole-varie/ (visité le 16/04/2024).

[21] Lili Guo et al. « Forecasting crude oil futures price using machine learning methods : Evidence from China ». en. In : *Energy Economics* 127 (nov. 2023), p. 107089. issn : 01409883. doi : 10.1016/j.eneco.2023.107089. url : https://linkinghub.elsevier.com/retrieve/pii/S014098832300587X.

[22] Sepp Hochreiter et Jürgen Schmidhuber. « Long Short-Term Memory ». en. In : *Neural Computation* 9.8 (nov. 1997), p. 1735-1780. issn : 0899-7667, 1530-888X. doi : 10.1162/neco.1997.9.8.1735. url : https://direct.mit.edu/neco/article/9/8/1735-1780/6109.

[23] John Hunt. « Introduction to Matplotlib ». en. In : *Advanced Guide to Python 3 Programming*. Cham : Springer International Publishing, 2019, p. 35-42. isbn : 9783030259426 9783030259433. doi : 10.1007/978-3-030-25943-3_5. url : http://link.springer.com/10.1007/978-3-030-25943-3_5.

[24] Hanan G. Al-Jasoor et Samaher Al-Janabi. « Oil Price Prediction Using Deep Neural Network Technique Gated Recurrent Unit (GRU) and Multivariate Analysis ». In : *2022 22nd International Conference on Computational Science and Its Applications (ICCSA)*. Malaga, Spain : IEEE, juill. 2022, p. 22-27. isbn : 9781665455848. doi : 10.1109/ICCSA57511.2022.00014. url : https://ieeexplore.ieee.org/document/10064400/.

[25] Ferdin Joe John Joseph, Sarayut Nonsiri et Annop Monsakul. « Keras and TensorFlow : A Hands-On Experience ». en. In : *Advanced Deep Learning for Engineers and Scientists*. Sous la dir. de Kolla Bhanu Prakash et al. Cham : Springer International Publishing, 2021, p. 85-111. isbn : 9783030665180 9783030665197. doi : 10.1007/978-3-030-66519-7_4. url : https://link.springer.com/10.1007/978-3-030-66519-7_4.

[26] Zohreh Karimi. « scikit-learn-Quick-Review ». en. In : (2021). doi : 10.13140/RG.2.2.14605.67043. url : http://rgdoi.net/10.13140/RG.2.2.14605.67043.

[27] Zohreh Karimi et Numpy. « NumPy Quick Review ». en. In : (2021). doi : 10.13140/RG.2.2.28097.58728. url : http://rgdoi.net/10.13140/RG.2.2.28097.58728 (visité le 03/06/2024).

[28] Raphael Kassel. *ARIMA : Modèle de prédiction des séries temporelles*. fr-FR. Fév. 2021. url : https://datascientest.com/arima-series-temporelles.

[29] Yassin Khalifa, Danilo Mandic et Ervin Sejdić. « A review of Hidden Markov models and Recurrent Neural Networks for event detection and localization in biomedical signals ». en. In : *Information Fusion* 69 (mai 2021), p. 52-72. issn : 15662535. doi : 10.1016/j.inffus.2020.11.008. url : https://linkinghub.elsevier.com/retrieve/pii/S1566253520304140.

[30] *Le pétrole : qu'est-ce que c'est ?* fr. url : https://www.xtb.com/fr/formation/petrole (visité le 16/04/2024).

[31] *Les recettes pétrolières de l'Irak ont atteint un niveau record en février, au plus haut depuis 8 ans | Connaissances des énergies*. fr. Mars 2022. url : https://www.connaissancedesenergies.org/afp/les-recettes-petrolieres-de-lirak-ont-atteint-un-niveau-record-en-fevrier-au-plus-haut-depuis-8-ans-220305 (visité le 10/06/2024).

[32] Louis. *Les séries temporelles avec Python (3/4) - Éléments théoriques et exemples*. fr. Juin 2021. URL : https://blog.statoscop.fr/timeseries-3.html (visité le 12/05/2024).

[33] *Marché du pétrole : qu'est-ce que le Dubaï Light ?* fr-FR. Mai 2017. URL : https://www.fioulmarket.fr/astuces-conseils/tout-savoir-sur-le-fioul/marche-du-petrole-qu-est-ce-que-le-dubai-light (visité le 16/04/2024).

[34] Nathalie Mayer. *Alerte rouge : la production de plastique pourrait tripler d'ici 2060 !* fr. URL : https://www.futura-sciences.com/planete/actualites/pollution-plastique-alerte-rouge-production-plastique-pourrait-tripler-ici-2060-112802/ (visité le 08/06/2024).

[35] Leland McInnes, John Healy et James Melville. *UMAP : Uniform Manifold Approximation and Projection for Dimension Reduction*. arXiv :1802.03426 [cs, stat]. Sept. 2020. URL : http://arxiv.org/abs/1802.03426.

[36] *Oil Market Report - November 2023 – Analysis*. en-GB. Nov. 2023. URL : https://www.iea.org/reports/oil-market-report-november-2023 (visité le 10/06/2024).

[37] Jean-Luc Parouty et al. « Formation Introduction au Deep Learning (FIDLE) ». In : 2022. URL : https://fidle.cnrs.fr.

[38] Razvan Pascanu, Tomas Mikolov et Yoshua Bengio. *On the difficulty of training Recurrent Neural Networks*. arXiv :1211.5063 [cs]. Fév. 2013. URL : http://arxiv.org/abs/1211.5063.

[39] *Pétrole*. fr-FR. URL : https://lelementarium.fr/focus/focus-4/ (visité le 08/06/2024).

[40] *Pic de la production d'électricité d'origine fossile : déjà une réalité pour de nombreux pays... | Connaissances des énergies*. fr. Oct. 2023. URL : https://www.connaissancedesenergies.org/pic-de-la-production-delectricite-dorigine-fossile-deja-une-realite-pour-de-nombreux-pays-240426 (visité le 10/06/2024).

[41] Damien Rolon-Mérette et al. « Introduction to Anaconda and Python : Installation and setup ». In : *The Quantitative Methods for Psychology* 16.5 (mai 2020), S3-S11. ISSN : 2292-1354. DOI : 10.20982/tqmp.16.5.S003. URL : http://www.tqmp.org/SpecialIssues/vol16-5/S003.

[42] Boumediène Souiki. « Les modèles à changement de régime : Application sur le marché du pétrole ». Thèse de doct. Déc. 2020.

[43] *Transport in Transition*. en. URL : https://www.dnv.com/publications/transport-in-transition-242808/ (visité le 08/06/2024).

[44] Xiaojia Wang et al. « LSTM-Based Broad Learning System for Remaining Useful Life Prediction ». en. In : *Mathematics* 10.12 (juin 2022), p. 2066. ISSN : 2227-7390. DOI : 10.3390/math10122066. URL : https://www.mdpi.com/2227-7390/10/12/2066.

[45] Michael Waskom. « seaborn : statistical data visualization ». In : *Journal of Open Source Software* 6.60 (avr. 2021), p. 3021. ISSN : 2475-9066. DOI : 10.21105/joss.03021. URL : https://joss.theoj.org/papers/10.21105/joss.03021.

[46] Cao XIAO et Jimeng SUN. « Recurrent Neural Networks (RNN) ». en. In : *Introduction to Deep Learning for Healthcare*. Cham : Springer International Publishing, 2021, p. 111-135. ISBN : 9783030821838 9783030821845. DOI : 10.1007/978-3-030-82184-5_7. URL : https://link.springer.com/10.1007/978-3-030-82184-5_7.

[47] Lei YAN, Yuting ZHU et Haiyan WANG. « Selection of Machine Learning Models for Oil Price Forecasting : Based on the Dual Attributes of Oil ». en. In : *Discrete Dynamics in Nature and Society* 2021 (oct. 2021). Sous la dir. de Daqing GONG, p. 1-16. ISSN : 1607-887X, 1026-0226. DOI : 10.1155/2021/1566093. URL : https://www.hindawi.com/journals/ddns/2021/1566093/.

[48] Chengkai ZHANG et al. « Real-time prediction of rate of penetration by combining attention-based gated recurrent unit network and fully connected neural networks ». en. In : *Journal of Petroleum Science and Engineering* 213 (juin 2022), p. 110396. ISSN : 09204105. DOI : 10.1016/j.petrol.2022.110396. URL : https://linkinghub.elsevier.com/retrieve/pii/S0920410522002820.

[49] Kexian ZHANG et Min HONG. « Forecasting crude oil price using LSTM neural networks ». In : *Data Science in Finance and Economics* 2.3 (2022), p. 163-180. ISSN : 2769-2140. DOI : 10.3934/DSFE.2022008. URL : https://www.aimspress.com/article/doi/10.3934/DSFE.2022008.

[50] Zhengling ZHAO et al. « A novel hybrid model with two-layer multivariate decomposition for crude oil price forecasting ». en. In : *Energy* 288 (fév. 2024), p. 129740. ISSN : 03605442. DOI : 10.1016/j.energy.2023.129740. URL : https://linkinghub.elsevier.com/retrieve/pii/S0360544223031341.

Résumé

Prédire avec précision les prix du pétrole est essentiel pour anticiper les fluctuations du marché, informer les stratégies d'investissement et garantir la stabilité économique mondiale. Plusieurs études ont été menées pour prévoir les prix du pétrole brut. Dans la plupart des cas où les facteurs influençant les prix du pétrole ont été pris en considération, les réseaux de neurones récurrents (RNN) et leurs variantes comme les mémoires à long terme et court terme (LSTM) ainsi que les unités récurrentes à portes (GRU) se sont révélés efficaces pour capturer les relations non linéaires entre ces facteurs. Cependant, ces modèles nécessitent encore des améliorations et des renforcements. Dans ce mémoire, nous proposons une approche basée sur des méthodes non linéaires de réduction de la dimensionnalité et de machine learning. Nous utilisons deux méthodes de réduction de la dimensionnalité non linéaire : l'Analyse en Composantes Principales à Noyau (KPCA) et l'Approximation Uniforme et Projection de Manifold (UMAP). Ensuite, nous construisons six modèles basés sur les RNN, LSTM et GRU. Nos modèles visent à prédire avec précision les prix au comptant et à terme du pétrole brut.
Après une étude comparative des résultats de prévision, nous avons constaté que les méthodes KPCA et UMAP renforcent et améliorent significativement les performances des modèles RNN, LSTM et GRU. Les modèles combinés avec KPCA surpassent généralement ceux combinés avec UMAP. Il est à noter que nos modèles présentent une meilleure robustesse en termes d'exactitude par rapport aux approches existantes utilisant le même jeu de données.

Mots clés : Méthode de réduction de la dimensionnalité, Machine Learning, Prix au comptant et à terme du pétrole, RNN, LSTM, GRU

Abstract

Accurately predicting oil prices is crucial for anticipating market fluctuations, informing investment strategies, and ensuring global economic stability. Several studies have been conducted to forecast crude oil prices. In most cases where factors influencing oil prices have been considered, Recurrent Neural Networks (RNN) and their Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) variants have proven effective in capturing the nonlinear relationships among these factors. However, these models still require further improvements and enhancements.
In this thesis, we propose an approach based on nonlinear methods of dimensionality reduction and machine learning. We employ two nonlinear dimensionality reduction methods : Kernel Principal Component Analysis (KPCA) and Uniform Manifold Approximation and Projection (UMAP). Subsequently, we construct six models based on RNN, LSTM, and GRU. Our models aim to accurately predict spot and futures prices of crude oil.
Following a comparative study of forecasting results, we found that KPCA and UMAP methods significantly enhance and improve the performance of RNN, LSTM, and GRU models. Models combined with KPCA generally outperform those combined with UMAP. It is noteworthy that our models exhibit greater robustness in terms of accuracy compared to existing approaches using the same dataset.

Keywords : Dimensionality Reduction methods , Machine Learning, Spot and Future oil Prices, RNN, LSTM, GRU

ملخص

توقع أسعار النفط بدقة أمر أساسي لتوقع التقلبات في السوق، وإعلام استراتيجيات الاستثمار و ضمان الاستقرار الاقتصادي العالمي. تم إجراء عدة دراسات لتوقع أسعار النفط الخام. في معظم الحالات التي تم فيها اعتبار العوامل التي تؤثر في أسعار النفط، ثبت أن الشبكات العصبية المتكررة (RNN) والمتغيرات المشتقة منها مثل الذاكرة طويلة قصيرة المدى (LSTM) و الوحدة المتكررة ذات البوابات (GRU) فعّالة في التقاط العلاقات غير الخطية بين هذه العوامل. ومع ذلك، فإن هذه النماذج تتطلب تحسينات وتعزيزات إضافية.

في هذه الأطروحة، نقترح نهجاً يستند إلى طرق غير خطية لتقليص البُعد والتعلم الآلي. نحن نستخدم اثنتين من طرق تقليص البُعد غير الخطية : تحليل المكونات الرئيسية بالنواة (KPCA) و تقريب وإسقاط البنية الموحدة (UMAP). بعد ذلك، نقوم ببناء ستة نماذج استناداً إلى RNN، LSTM و GRU. تهدف نماذجنا إلى توقع أسعار النفط الخام الفورية و العقود الآجلة بدقة.

بعد دراسة مقارنة لنتائج التنبؤ، تبين أن طرق KPCA و UMAP تعزز بشكل كبير أداء النماذج RNN، LSTM و GRU. النماذج المجتمعة مع KPCA تفوق عادة تلك المجتمعة مع UMAP. يجدر بالذكر أن نماذجنا تظهر متانة أكبر من حيث الدقة مقارنة بالنهج الحالية باستخدام نفس مجموعة البيانات.

الكلمات الرئيسية : أساليب تقليل الأبعاد، التعلم الآلي، أسعار النفط الفورية و العقود الآجلة، RNN، LSTM، GRU