Amélioration des performances de l'Intelligence Artificielle

Moez Krichen

Amélioration des performances de l'Intelligence Artificielle

Moez Krichen

Laboratoire ReDCAD, Université de Sfax, Tunisie
moez.krichen@redcad.org

Résumé. L'avènement de l'apprentissage automatique (ML) et de l'intelligence artificielle (AI) a entraîné une transformation significative dans de multiples industries, car cela a facilité l'automatisation des emplois, l'extraction d'informations précieuses à partir de vastes ensembles de données, et la facilitation de processus de prise de décision sophistiqués. Néanmoins, optimiser l'efficacité est devenu un domaine de recherche critique en raison de la complexité croissante et des besoins en ressources des systèmes d'AI. Ce document fournit un examen approfondi de plusieurs techniques et méthodologies visant à améliorer l'efficacité du ML et de l'AI. Dans cette étude, nous enquêtons sur de nombreux domaines de recherche se rapportant à l'AI. Ces domaines incluent les améliorations algorithmiques, les techniques d'accélération matérielle, les méthodes de prétraitement des données, les approches de compression des modèles, les cadres de calcul distribué, les stratégies écoénergétiques, les concepts fondamentaux liés à l'AI, l'évaluation de l'efficacité de l'AI, et les méthodologies formelles. De plus, nous nous engageons dans un examen des obstacles et des avenues prospectives dans ce domaine particulier. Ce document offre une analyse complète de nombreux sujets, visant à équiper les chercheurs et les praticiens d'une compréhension exhaustive des stratégies pour améliorer l'efficacité au sein des systèmes de ML et d'AI.

1 Introduction

Le développement de l'apprentissage automatique (ML) et de l'intelligence artificielle (AI) a entraîné des changements significatifs dans plusieurs secteurs, menant à l'automatisation des emplois, à l'extraction de perspectives à partir de vastes ensembles de données et à la facilitation de processus de prise de décision avancés. L'utilisation des systèmes d'AI est de plus en plus répandue dans de nombreux domaines, englobant l'identification d'images, le traitement du langage naturel, les voitures autonomes et les recommandations personnalisées (198; 70). Néanmoins, la demande pour améliorer l'efficacité de ces systèmes d'AI a émergé comme un champ d'étude crucial, étant donné leur sophistication croissante et leurs besoins en ressources (131; 160).

Les différentes phases du ML et de l'AI comprennent :

- Collecte de données : Cela implique la collecte de données à partir de bases de données, d'APIs et de capteurs pour les algorithmes de ML ou d'AI. La qualité des données affecte grandement les résultats des algorithmes.
- Prétraitement des données : Nettoyage, organisation et formatage des données brutes de la phase précédente pour l'analyse. Cela comprend la suppression des nombres manquants, le traitement des valeurs aberrantes et la mise à l'échelle des données pour standardiser et assurer la cohérence.
- Sélection et entraînement du modèle : Cette étape sélectionne un modèle de ML ou d'AI pour le problème. Entraîner le modèle avec des données prétraitées pour apprendre des motifs et prédire. Pour minimiser l'inexactitude prédite-réelle, les paramètres du modèle sont ajustés pendant l'entraînement.
- Évaluation du modèle : Le modèle doit être testé pour sa performance après l'entraînement. Cela implique de faire fonctionner le modèle sur de nouvelles données et de comparer ses prédictions aux valeurs réelles. La performance du modèle peut être évaluée à l'aide de l'exactitude, la précision, le rappel et le score F1.
- Déploiement : Après validation, le modèle peut être utilisé pour accomplir la tâche. Cela implique d'ajouter le modèle à un système logiciel ou une application et de créer une interface utilisateur.
- Surveillance et maintenance : Après le déploiement du modèle, il doit être surveillé et mis à jour pour rester précis et efficace. Cela inclut la surveillance des entrées de données, du comportement du modèle et des erreurs ou anomalies. Maintenir et mettre à jour le modèle pour refléter les conditions changeantes et les exigences est également nécessaire.

Le rôle de l'efficacité est d'une importance capitale dans la création et la mise en œuvre des systèmes de ML et d'AI (29). Le concept incorpore divers éléments fondamentaux, tels que l'application prudente des ressources informatiques, la réduction de la complexité temporelle, la consommation efficace de l'énergie et l'amélioration globale des performances du système (48; 3). Améliorer l'efficacité ne résulte pas seulement en des solutions d'AI accélérées et économiquement efficaces, mais joue également un rôle dans la promotion de la durabilité et de la scalabilité. Cela, à son tour, facilite la mise en œuvre de systèmes d'AI sur des appareils à ressources limitées et dans des applications étendues.

L'objectif principal de cette étude est de fournir une analyse complète des différentes approches et tactiques qui peuvent être employées pour améliorer l'efficience dans les domaines du ML et de l'AI. Cette étude couvrira une large gamme de sujets, incluant les avancées dans les algorithmes, l'accélération matérielle, le prétraitement des données, la compression des modèles, le calcul distribué, les techniques écoénergétiques, les principes fondamentaux de l'AI, la mesure de l'efficacité de l'AI, et l'application des méthodes formelles. Notre étude vise à offrir aux chercheurs et aux praticiens une compréhension approfondie des différentes approches qui peuvent être employées pour améliorer l'efficacité des systèmes de ML et d'AI. Ceci sera réalisé en menant un examen détaillé de ces domaines.

L'objectif principal de cet examen approfondi est d'offrir des perspectives précieuses sur les dernières avancées, méthodologies et stratégies optimales qui peuvent être em-

ployées pour améliorer l'efficacité dans les domaines du ML et de l'AI (106; 170; 9; 164; 27; 127; 162; 4; 103; 110; 10). Discuter de ces sujets vise principalement à développer une corrélation entre la compréhension théorique et la mise en œuvre réelle (149; 105; 128; 129; 98; 205; 11; 5; 26; 138; 56; 13; 124; 139). Cette ressource fournira des insights précieux et des recommandations pour les chercheurs et les professionnels développant des systèmes d'AI qui montrent une performance supérieure tout en optimisant l'utilisation des ressources (157; 107; 130; 156; 8; 35; 33; 18; 81). De plus, les défis et les voies potentielles pour les avancements futurs dans ce domaine en rapide évolution seront examinés, permettant aux lecteurs d'anticiper et de naviguer dans les complexités associées à l'amélioration de l'efficacité du ML et de l'AI (153; 108; 109; 104; 102).

Les sections suivantes de ce document offrent un examen complet de nombreux aspects relatifs à l'amélioration de l'efficacité dans le ML et l'AI. La section 2 présente une exposition des concepts et principes essentiels de l'AI, équipant ainsi les lecteurs d'une compréhension complète de la discipline. Dans la section suivante, étiquetée comme la section 3, nous explorerons le sujet de la mesure de l'efficacité de l'AI. Cela impliquera un examen de diverses métriques utilisées à cette fin, ainsi qu'une analyse de la complexité computationnelle, de la complexité temporelle et des compromis existant entre l'exactitude et l'efficacité. La section 4 est dédiée à discuter des percées algorithmiques dans le domaine. Elle englobe plusieurs stratégies, incluant la sélection de caractéristiques, les algorithmes d'optimisation et les structures de modèle efficaces. L'emploi de matériel spécialisé, tel que les unités de traitement graphique (GPUs), les réseaux de portes programmables sur site (FPGAs) et les unités de traitement tensoriel (TPUs), pour optimiser les calculs d'AI est couvert dans la section 5, qui se concentre sur l'accélération matérielle. Dans la section 6, l'importance du prétraitement des données et de l'ingénierie des caractéristiques dans l'amélioration de l'efficacité est explorée. Cela inclut une analyse des approches pour traiter les données manquantes, la mise à l'échelle des caractéristiques et l'extraction des caractéristiques. Le document introduit de nombreuses méthodes pour la compression et l'optimisation des modèles dans la section 7. Ces stratégies incluent la réduction des paramètres, la quantification et la distillation des connaissances. Dans la section 8, un examen est mené sur les méthodologies de calcul distribué et de parallélisation pour améliorer la scalabilité des systèmes d'AI. Cette exploration englobe l'entraînement distribué, le parallélisme des modèles et l'apprentissage fédéré. La section 9 se plonge dans les solutions écoénergétiques, englobant une exploration des méthodologies visant à diminuer l'utilisation de l'énergie et à améliorer l'utilisation du matériel. La section 10 se concentre sur l'application des méthodes formelles pour augmenter l'efficacité, englobant la vérification formelle, les méthodologies de validation et la modélisation formelle des algorithmes d'AI. Dans la section 11, les défis et les limites liés à l'amélioration de l'efficacité sont discutés, se concentrant sur les considérations éthiques et les compromis entre l'efficacité et l'exactitude. La section suivante, la section 12, fournit un aperçu complet des principales découvertes et contributions décrites dans cet article académique. Elle souligne l'importance de l'efficacité comme force motrice pour faire avancer le ML et l'AI.

2 Préliminaires sur l'IA

L'objectif de cette partie est de présenter une introduction complète aux concepts et principes fondamentaux de l'IA afin d'établir une base solide pour comprendre les discussions ultérieures sur l'amélioration de l'efficacité. Le terme "IA" fait référence au domaine de l'informatique qui se concentre sur l'avancement des systèmes informatiques intelligents capables de répliquer les capacités cognitives humaines.

2.1 Comparaison entre les systèmes d'IA et les systèmes de programmation classique

L'IA et les systèmes de programmation classique sont deux méthodologies informatiques distinctes pour la résolution de problèmes. Les systèmes de programmation classique dépendent d'instructions claires et de règles préétablies pour exécuter des tâches. Les programmeurs écrivent directement le code spécifiant les instructions logiques et les procédures séquentielles nécessaires pour accomplir un objectif prédéterminé. Cette méthodologie offre un niveau élevé de précision et de prévisibilité, la rendant appropriée pour des emplois caractérisés par des règles clairement spécifiées et des corrélations entrée-sortie non ambiguës. En contraste, les systèmes d'IA emploient des algorithmes de ML et des modèles basés sur les données pour évaluer les motifs et prendre des décisions. Les systèmes d'IA peuvent acquérir des connaissances à partir des données et modifier leurs actions en conséquence, leur permettant de naviguer efficacement dans des circonstances complexes et en constante évolution. La programmation classique a une performance exceptionnelle dans des environnements caractérisés par le déterminisme, mais l'IA montre des capacités remarquables dans la gestion de l'incertitude et la réalisation de prédictions. Néanmoins, les systèmes d'IA peuvent présenter une déficience en termes de transparence et poser des difficultés d'interprétation en raison de leur dépendance vis-à-vis des algorithmes complexes. D'autre part, les systèmes de programmation classique offrent un niveau de transparence plus élevé en raison du codage explicite de la logique. Les deux approches possèdent des avantages et des inconvénients distincts, et leur pertinence dépend du domaine de problème particulier et des exigences.

2.2 Composants des systèmes d'IA

La première étape consiste à introduire les composants essentiels des systèmes d'IA, qui incluent la perception, la pensée et la prise de décision. La perception comprend la capacité cognitive d'acquérir et d'analyser des informations provenant de diverses sources, y compris des stimuli visuels et auditifs. La discipline englobe un éventail de tâches, ne se limitant pas à la reconnaissance d'images, la reconnaissance vocale et la compréhension du langage naturel. Le raisonnement est le processus cognitif par lequel les individus tirent des conclusions logiques ou font des inférences éclairées en utilisant les faits disponibles. L'entreprise actuelle implique l'application du raisonnement logique, du raisonnement probabiliste et de la pensée symbolique. Le processus de prise de décision implique l'évaluation et la sélection du cours d'action le plus avantageux parmi diverses options tout en considérant de nombreux critères et objectifs.

2.3 Algorithmes de ML

Le ML, qui est un sous-ensemble de l'IA, joue un rôle pivot dans de nombreux systèmes d'IA contemporains. Ce document présente un examen complet des algorithmes de ML, englobant les domaines de l'apprentissage supervisé, de l'apprentissage non supervisé et de l'apprentissage par renforcement (RL). L'apprentissage supervisé est une approche de ML qui implique la formation de modèles à l'aide de données annotées pour fournir des prédictions ou des classifications. Les algorithmes englobent la régression linéaire, les machines à vecteurs de support et les réseaux de neurones. En contraste, l'apprentissage non supervisé est une méthode qui identifie des motifs et des structures dans des données qui manquent de labels explicites. Les approches d'apprentissage non supervisé englobent le clustering, la réduction de dimensionnalité et les modèles génératifs. Le RL facilite l'acquisition d'un comportement optimal par les agents par essais et erreurs, en utilisant un mécanisme basé sur la récompense.

2.4 Données dans les systèmes d'IA

De plus, nous examinons l'importance des données au sein des systèmes d'IA et les obstacles liés à la collecte, au prétraitement et à l'annotation des données. Utiliser des ensembles de données de haute qualité et diversifiés est primordial dans la formation de modèles d'IA résilients. Dans cette étude, nous enquêtons sur la représentation des caractéristiques, qui concerne la conversion de données non traitées en un format approprié pour l'utilisation dans les méthodes de ML. L'utilisation d'approches d'ingénierie des caractéristiques, telles que la sélection et l'extraction de caractéristiques, a été introduite pour améliorer l'efficacité et l'efficacité des systèmes d'IA. En outre, nous abordons les implications éthiques et les biais potentiels qui émergent dans les systèmes d'IA en raison de la collecte de données et des processus de prise de décision algorithmique.

3 Mesure de l'efficacité de l'IA

La mesure de l'efficacité de l'IA revêt une importance significative pour comprendre les nombreux compromis associés aux ressources computationnelles, à la complexité temporelle, à l'exactitude, et à d'autres éléments pertinents.

3.1 Métriques de performance

Nous présentons les métriques de performance couramment utilisées pour évaluer les systèmes d'IA (75). Soit Y représentant les vraies étiquettes d'un ensemble de données, et soit Y' représentant les étiquettes prédites par un modèle d'IA. Pour les tâches de classification, l'exactitude est définie comme :

$$\text{Exactitude} = \frac{\text{Nombre d'instances correctement prédites}}{\text{Nombre total d'instances}}$$

La précision et le rappel sont définis comme :

$$\text{Précision} = \frac{\text{Vrais positifs}}{\text{Vrais positifs} + \text{Faux positifs}}$$

$$\text{Rappel} = \frac{\text{Vrais positifs}}{\text{Vrais positifs} + \text{Faux négatifs}}$$

Le score F1, qui équilibre la précision et le rappel, est défini comme :

$$\text{Score F1} = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

Pour les tâches de régression, des métriques telles que l'erreur quadratique moyenne (MSE), l'erreur absolue moyenne (MAE), et le coefficient de détermination (R-carré) peuvent être définies en fonction des valeurs prédites Y' et des valeurs réelles Y.

3.2 Complexité computationnelle

La complexité computationnelle des algorithmes d'IA fait référence à la mesure des ressources informatiques nécessaires pour exécuter un algorithme en fonction de la taille de l'entrée. L'analyse de la complexité temporelle utilise fréquemment la notation Big O, qui fournit une limite supérieure sur le taux de croissance du temps d'exécution d'un algorithme. Par exemple, une complexité temporelle de $O(n)$ indique que le temps d'exécution augmente linéairement avec la taille de l'entrée n. De même, la complexité spatiale mesure la quantité de mémoire nécessaire pour un algorithme.

3.3 Complexité temporelle

Le concept de complexité temporelle approfondit l'examen du temps de calcul requis par les systèmes d'IA pour résoudre un problème donné. Les complexités temporelles couramment rencontrées incluent :

- $O(1)$: Complexité constante, où le temps d'exécution ne dépend pas de la taille de l'entrée.
- $O(\log n)$: Complexité logarithmique, où le temps d'exécution croît logarithmiquement avec la taille de l'entrée.
- $O(n)$: Complexité linéaire, où le temps d'exécution croît linéairement avec la taille de l'entrée.
- $O(n^2)$: Complexité quadratique, où le temps d'exécution croît quadratiquement avec la taille de l'entrée.
- $O(2^n)$: Complexité exponentielle, où le temps d'exécution croît exponentiellement avec la taille de l'entrée.

Comprendre la complexité temporelle des algorithmes d'IA permet aux chercheurs et aux praticiens de faire des choix éclairés concernant la sélection d'algorithmes, en considérant les ressources disponibles et le niveau d'efficacité requis.

3.4 Compromis : Exactitude vs. Efficacité

L'optimisation de l'IA nécessite souvent une utilisation prudente des ressources computationnelles pour atteindre un équilibre entre l'exactitude et l'efficacité (76; 173). L'utilisation d'algorithmes approximatifs, d'approches d'échantillonnage et de simplifications algorithmiques est répandue pour atténuer les exigences de traitement tout en maintenant des niveaux de précision satisfaisants (61; 20). Ces méthodologies emploient l'introduction délibérée de fautes contrôlées ou la réduction de la complexité du problème pour améliorer l'efficacité. La mesure dans laquelle l'exactitude est compromise dépend de la méthodologie particulière utilisée et des exigences de l'application donnée.

3.5 Autres considérations

En plus des mesures de performance et de la complexité computationnelle, divers autres facteurs influencent l'efficacité des systèmes d'IA. L'interprétabilité du modèle concerne la capacité à comprendre et à expliquer les prédictions d'un modèle d'IA. La notion de robustesse aux attaques adverses se rapporte à la capacité d'un modèle à maintenir sa performance même lorsqu'il est exposé à des entrées méticuleusement conçues pour le tromper. Les problèmes de scalabilité englobent l'évaluation de la capacité d'un système d'IA à gérer efficacement de vastes ensembles de données ou des fonctions au sein de paramètres de calcul distribué.

.

4 Avancées Algorithmiques

Les avancées dans les algorithmes ont entraîné une transformation significative dans un large éventail d'applications de l'IA, conduisant à des niveaux améliorés d'efficacité, de précision, de sécurité et de robustesse. Cette étude examine les développements algorithmiques significatifs dans plusieurs secteurs de l'IA, soulignant leur influence et leurs conséquences potentielles (96; 78; 180; 186).

4.1 Apprentissage Profond

Le domaine de l'IA a vu l'émergence de l'apprentissage profond (Deep Learning, DL) comme un paradigme clé, permettant des progrès notables dans la vision par ordinateur, le traitement du langage naturel et la reconnaissance vocale. Les modèles de DL, tels que les réseaux neuronaux convolutionnels (CNNs) et les réseaux neuronaux récurrents (RNNs), ont considérablement transformé les tâches de reconnaissance d'image et de parole grâce à leur capacité à acquérir de manière autonome des représentations hiérarchiques à partir des données. La structure hiérarchique inhérente aux modèles de DL leur permet de capturer efficacement des motifs complexes et des interdépendances, résultant en une performance exceptionnelle dans divers domaines. L'utilisation du DL a facilité des avancées significatives dans le traitement du langage naturel, permettant la réalisation de nombreuses tâches, y compris la traduction automatique, l'analyse des sentiments et la réponse aux questions. Le succès des algorithmes de DL peut être

attribué à la présence de vastes ensembles de données étiquetées et aux capacités de traitement offertes par la technologie contemporaine.

4.2 Apprentissage par Renforcement

Des progrès significatifs ont été réalisés dans le développement des algorithmes d'apprentissage par renforcement (RL), qui permettent de former des agents à prendre des décisions optimales dans des environnements dynamiques. Les algorithmes de RL apprennent par interaction avec un environnement, recevant des retours sous forme de récompenses ou de pénalités pour leurs actions. Les algorithmes de RL notables incluent Q-learning, les gradients de politique et les approches acteur-critique. Ces algorithmes ont été appliqués avec succès à divers problèmes, y compris les jeux, la robotique et les systèmes autonomes. Le RL a démontré des réalisations remarquables, telles qu'AlphaGo, qui a battu les champions du monde au jeu de Go, et plus récemment, les modèles DALL-E et CLIP d'OpenAI, qui montrent le potentiel du RL dans la génération de contenu créatif et l'apprentissage multimodal.

4.3 Modèles Génératifs

L'objectif principal des modèles génératifs est de développer un modèle qui capture précisément la distribution de probabilité sous-jacente d'un ensemble de données donné, permettant la génération de nouveaux exemples inédits. L'attention accordée aux modèles génératifs ces dernières années est attribuée à leur capacité à produire des sorties réalistes et de haute qualité. Notamment, les réseaux antagonistes génératifs (GANs) ont émergé comme une catégorie puissante de modèles qui acquièrent la capacité de générer de nouveaux échantillons en formant un réseau générateur en opposition à un réseau discriminateur, favorisant une dynamique compétitive. Les GANs ont démontré des réalisations notables dans diverses disciplines, englobant la synthèse d'images, le transfert de style et l'augmentation de données. Les autoencodeurs variationnels (VAEs) constituent une catégorie remarquable de modèles génératifs qui peuvent acquérir une représentation latente des données et ensuite créer de nouveaux échantillons en échantillonnant à partir de l'espace latent acquis. Les modèles génératifs possèdent un potentiel significatif dans divers domaines, y compris la génération de contenu, la synthèse de données et la simulation.

4.4 Algorithmes Évolutifs

Les algorithmes évolutifs sont basés sur les principes fondamentaux de l'évolution biologique et sont utilisés pour aborder des problèmes d'optimisation complexes (64; 132). Ces algorithmes emploient une population de solutions potentielles et les raffinent progressivement à travers des opérations de sélection, de croisement et de mutation. Les algorithmes génétiques, la programmation génétique et les techniques évolutives sont reconnus comme des instances notables d'algorithmes évolutifs. Ces techniques ont été efficacement utilisées dans plusieurs problèmes, tels que l'optimisation, la robotique et la conception. Les algorithmes évolutifs possèdent des avantages inhérents dans des situations caractérisées par la complexité, la non-linéarité, ou une

compréhension limitée du domaine du problème. Ils ont la capacité d'explorer une gamme diversifiée d'alternatives et possèdent la capacité de découvrir des réponses innovantes et non traditionnelles qui peuvent échapper aux méthodes d'optimisation conventionnelles.

4.5 Approches Interdisciplinaires

La recherche en IA a progressivement adopté des méthodologies interdisciplinaires, intégrant des concepts et des méthodologies de diverses autres disciplines. Cette approche permet aux chercheurs de tirer parti des insights et des techniques d'autres domaines pour améliorer les algorithmes d'IA. Par exemple, les réseaux neuronaux graphiques sont apparus comme un outil puissant en incorporant des principes de la théorie des graphes, permettant une analyse et une prédiction efficaces dans les données structurées en graphes. Les algorithmes de jeu adversarial combinent des concepts de la théorie des jeux et de l'IA pour créer des agents capables de concurrencer des experts humains dans des jeux complexes. Les architectures cognitives s'inspirent de la science cognitive pour simuler des processus de pensée de type humain et permettre un comportement intelligent. De plus, l'apprentissage automatique informé par la physique combine des principes de la physique et de l'apprentissage automatique pour améliorer la modélisation et la prédiction des systèmes physiques. Les approches interdisciplinaires offrent des perspectives nouvelles et contribuent au développement d'algorithmes d'IA plus puissants et polyvalents capables de s'attaquer à des problèmes réels de plus en plus complexes.

Diverses technologies d'accélération matérielle, y compris les GPU, les TPU, les FPGA, les circuits intégrés spécifiques à une application (ASIC) et l'informatique quantique, ont été développées pour surmonter les contraintes des architectures traditionnelles et augmenter les capacités computationnelles et l'efficacité. Les méthodologies hybrides utilisent diverses techniques d'accélération matérielle pour améliorer à la fois la performance et l'efficacité. Les progrès dans l'accélération matérielle ont amélioré la formation, l'inférence et l'efficacité globale des modèles d'IA dans divers domaines.

5 Accélération Matérielle

Cette section explore l'importance de l'accélération matérielle dans l'accélération de l'avancement et du déploiement des modèles d'IA. La complexité croissante des algorithmes d'IA et leur dépendance aux données peuvent limiter les architectures informatiques conventionnelles, conduisant à des contraintes de puissance de traitement et d'efficacité. Les approches d'accélération matérielle sont employées pour relever ces défis en tirant parti de composants et d'architectures matériels spécialisés expressément conçus pour les tâches d'IA (45; 62; 17; 92; 143).

5.1 Unités de Traitement Graphique (GPUs)

L'incorporation de GPUs a été cruciale pour accélérer les calculs d'IA. Initialement conçus pour la génération visuelle, les GPUs ont démontré une compétence exception-

nelle dans le traitement parallèle, les rendant bien adaptés pour effectuer des tâches d'entraînement et d'inférence en apprentissage profond (91; 174). Les GPUs démontrent une compétence exceptionnelle à exécuter plusieurs calculs simultanément, capitalisant sur leur architecture hautement parallèle pour améliorer la vitesse des opérations matricielles et des calculs de réseaux neuronaux. Le parallélisme inhérent des GPUs facilite le traitement efficace de vastes ensembles de données et l'entraînement de modèles d'apprentissage profond complexes. Actuellement, les GPUs ont gagné une traction significative tant dans la recherche académique que dans les applications industrielles. Cette adoption généralisée a facilité l'entraînement et le déploiement de modèles d'IA à grande échelle par les chercheurs et les praticiens.

5.2 Unités de Traitement Tensoriel (TPUs)

Les TPUs sont un type d'accélérateur matériel spécialisé spécifiquement développé par Google. Ces accélérateurs sont destinés avec l'objectif principal d'améliorer la performance globale des activités de ML (93; 65). Les TPUs sont des dispositifs de calcul conçus pour effectuer efficacement des opérations tensorielles, qui sont fondamentales pour les processus de calcul impliqués dans les calculs d'apprentissage profond. Les TPUs sont capables d'accélérer considérablement les charges de travail d'IA. Les TPUs présentent une aptitude exceptionnelle pour des opérations d'entraînement et d'inférence étendues à grande échelle, grâce à leurs remarquables capacités de calcul et leur économie d'énergie. En utilisant des TPUs, les chercheurs et les entreprises peuvent atteindre des durées d'entraînement accélérées, réduire les dépenses et améliorer l'efficacité des modèles dans leurs efforts d'IA.

5.3 Field-Programmable Gate Arrays (FPGAs)

Les FPGAs offrent une approche flexible pour augmenter les capacités de calcul des applications d'IA via l'accélération matérielle (28; 94; 52). Contrairement aux GPUs et TPUs, les FPGAs possèdent l'attribut unique de reprogrammabilité, permettant l'implémentation d'architectures matérielles sur mesure optimisées pour certains workloads d'IA. Les FPGAs démontrent une performance exceptionnelle dans les situations nécessitant un faible délai temporel, et un traitement immédiat des données et ont été employés avec succès dans des domaines incluant la vision par ordinateur, le calcul en périphérie, et le streaming de données.

5.4 Circuits Intégrés Spécifiques à une Application (ASICs)

Les ASICs se réfèrent à des puces matérielles spécialisées méticuleusement conçues pour exécuter certaines tâches efficacement (55; 146). Les ASICs sont des circuits électroniques sur mesure spécifiquement adaptés pour optimiser la performance de calcul et l'efficacité énergétique des calculs d'apprentissage profond dans le domaine de l'IA (69; 89). Les ASICs ont la capacité d'atteindre des améliorations notables en vitesse et en économie d'énergie par rapport aux unités de traitement central (CPUs) conventionnelles et autres accélérateurs matériels en éliminant le besoin de calcul généraliste. Les ASICs sont importants dans les situations nécessitant une inférence d'IA efficace,

en particulier dans des domaines tels que les véhicules autonomes, la robotique, et les centres de données.

5.5 Informatique Quantique

L'utilisation des phénomènes quantiques dans l'informatique quantique a la capacité de transformer significativement le domaine de l'IA en permettant l'exécution de calculs à une échelle sans précédent (147; 158; 201; 159; 188). Les ordinateurs quantiques emploient des bits quantiques, communément appelés qubits, qui possèdent la capacité de représenter et de manipuler des informations dans plusieurs états simultanément. Le parallélisme intrinsèque observé dans l'informatique quantique peut améliorer significativement l'efficacité de calcul de certaines activités d'IA, notamment l'optimisation, l'apprentissage automatique, et la simulation, en tirant parti d'une accélération exponentielle. L'utilisation d'ordinateurs quantiques est actuellement explorée pour tirer parti des méthodes d'apprentissage automatique quantique, qui ont le potentiel de révéler des motifs et des corrélations précédemment inconnus au sein des ensembles de données. Néanmoins, le domaine de l'informatique quantique est actuellement dans sa phase naissante, nécessitant d'importants efforts de recherche et de développement pour surmonter les obstacles technologiques et construire des systèmes quantiques fonctionnels et extensibles adaptés aux applications d'IA.

5.6 Approches Hybrides

Les méthodologies hybrides utilisent diverses techniques d'accélération matérielle pour optimiser à la fois la performance et l'efficacité. Les systèmes hybrides peuvent atteindre des capacités de calcul améliorées et optimiser l'allocation des ressources en exploitant efficacement les avantages offerts par diverses architectures matérielles (32; 136; 2; 90). Un exemple illustratif d'une stratégie hybride pourrait impliquer l'utilisation de GPUs pour exécuter des tâches de traitement parallèle, de TPUs pour effectuer des opérations tensorielles, et de FPGAs pour faciliter l'accélération matérielle sur mesure. Les combinaisons mentionnées peuvent être personnalisées pour répondre aux critères spécifiques des algorithmes d'IA, permettant une exécution rationalisée et une performance globale améliorée. Les techniques hybrides offrent également un degré d'adaptabilité pour accommoder les workloads d'IA dynamiques et tirer parti des dernières avancées en technologie d'accélération matérielle. La progression de la technologie d'IA est anticipée pour être accompagnée par l'utilisation prédominante d'architectures hybrides, qui contribueront grandement à l'accélération de la recherche, du développement et de la mise en œuvre de l'IA.

6 Prétraitement des données et Ingénierie des caractéristiques

La préparation des données implique la conversion de données non traitées en un format raffiné et organisé. En contraste, l'ingénierie des caractéristiques génère des caractéristiques significatives et pertinentes à partir des données données (68; 151;

82; 202; 73; 1). Ces techniques sont d'une importance capitale pour affiner la qualité des données et optimiser la performance des modèles d'IA. Dans cette étude, nous examinons une gamme de méthodologies et de facteurs relatifs à la préparation des données et à l'ingénierie des caractéristiques pour améliorer l'efficacité des algorithmes d'IA.

6.1 Nettoyage des données

La première étape du prétraitement des données implique le nettoyage des données, un processus visant à identifier et à corriger toute imperfection présente dans l'ensemble de données qui pourrait entraver une analyse précise (86; 40; 111). La technique implique la gestion des valeurs manquantes, la correction des erreurs, et l'élimination des valeurs aberrantes. Diverses techniques statistiques peuvent gérer les données manquantes, y compris l'imputation par la moyenne ou la médiane, et des approches plus avancées, telles que la régression ou l'imputation multiple. Pour maintenir l'exactitude et la fiabilité des données, toute erreur, telle que des erreurs typographiques ou des incohérences, est corrigée. Dans l'analyse des données, il est coutumier de détecter et d'atténuer les valeurs aberrantes, qui se rapportent à des points de données manifestant des valeurs extraordinaires qui s'écartent significativement de la plupart de l'ensemble des données. Les valeurs aberrantes peuvent être identifiées par une gamme de techniques statistiques et traitées soit en les excluant de l'ensemble de données soit en mettant en œuvre des ajustements pour minimiser leur impact sur l'analyse.

6.2 Intégration des données

Le processus d'intégration des données implique la consolidation de données provenant de plusieurs sources pour créer un ensemble de données unifié adapté aux fins analytiques (41; 71; 120; 49). La technique susmentionnée est communément nécessaire lors de la manipulation de données provenant de bases de données diverses, de fichiers, ou de formats. Le processus d'intégration des données englobe l'identification et la résolution des incohérences, la consolidation des informations redondantes, et l'harmonisation des structures de données. Diverses techniques, y compris l'appariement des données, la liaison des enregistrements, et la fusion des données, sont utilisées pour déterminer et combiner les données pertinentes de différentes sources. L'utilisation d'un ensemble de données intégré permet une vue d'ensemble complète des données, améliorant l'exactitude de l'analyse et de la modélisation.

6.3 Transformation des données

Les techniques de transformation des données sont utilisées pour changer les données en un format approprié pour l'analyse et la modélisation. Les processus susmentionnés englobent la normalisation, la mise à l'échelle, et le codage (112; 15; 95). La normalisation est un processus visant à standardiser l'échelle des variables, atténuant ainsi le biais potentiel qui pourrait survenir d'une variable unique exerçant une influence disproportionnée sur l'étude en raison de sa magnitude relativement plus élevée. Pour ce faire, on peut employer des approches de mise à l'échelle telles que la standardisation

ou la mise à l'échelle min-max (166; 25; 171). Le codage des variables catégorielles est essentiel pour les exprimer sous un format numérique (34; 176). Diverses techniques, telles que le codage one-hot, le codage par étiquettes, et le codage ordinal, sont employées pour transformer les données catégorielles en un format susceptible d'être traité par les algorithmes d'IA.

6.4 Extraction de caractéristiques

L'extraction de caractéristiques est un processus qui implique la génération de nouvelles caractéristiques à partir d'un ensemble de données existant pour capturer des motifs et des informations pertinents (97; 72). L'objectif de cette procédure est de réduire la dimensionnalité des données tout en conservant des attributs significatifs. Diverses techniques, telles que l'analyse en composantes principales (PCA) (115; 100), l'analyse discriminante linéaire (LDA) (6; 200; 77), ou le hachage de caractéristiques (134; 115), peuvent être employées pour extraire des caractéristiques significatives. Le processus d'extraction de caractéristiques permet une représentation concise et informative des relations complexes et des motifs, améliorant l'efficacité de la modélisation.

6.5 Sélection de caractéristiques

L'objectif de la sélection de caractéristiques est de choisir le sous-ensemble de caractéristiques le plus pertinent et utile pour la modélisation (47; 36; 30; 53). Cette procédure aide à atténuer le surapprentissage, à améliorer l'interprétabilité du modèle, et à améliorer l'efficacité informatique. Diverses techniques, telles que les méthodes de filtrage (par exemple, l'analyse de corrélation) (39; 24), les méthodes d'enveloppe (par exemple, l'élimination récursive de caractéristiques) (31; 194), ou les méthodes d'incorporation (par exemple, la régularisation Lasso) (12; 172; 42), peuvent être employées pour la sélection de caractéristiques. Le processus de sélection de caractéristiques permet aux modèles de prioriser les caractéristiques les plus informatives, leur permettant de se concentrer sur les aspects les plus pertinents des données. En conséquence, cela conduit à une performance prédictive améliorée.

6.6 Traitement des données catégorielles et textuelles

Les données catégorielles et textuelles nécessitent des techniques de prétraitement spécifiques pour les rendre adaptées aux algorithmes d'IA (74). Les variables catégorielles peuvent être codées en one-hot, où chaque catégorie est représentée par une variable indicatrice binaire. Les données textuelles subissent souvent une tokenisation (152), où le texte est divisé en mots individuels ou en n-grammes, suivi de techniques telles que le stemming (167), la lemmatisation (148), ou la suppression des mots vides (114) pour réduire la dimensionnalité et améliorer la représentation du texte. De plus, les données textuelles peuvent être transformées en représentations numériques en utilisant des méthodes comme la fréquence du terme-inverse de la fréquence des documents (TF-IDF) (168) ou des plongements de mots comme Word2Vec ou GloVe (184; 87). Ces techniques de prétraitement permettent d'utiliser des données catégorielles et textuelles dans les modèles d'IA.

6.7 Équilibrage des données

Le processus d'équilibrage des données est d'une importance capitale dans le contexte des ensembles de données déséquilibrés, lorsqu'il existe une disparité substantielle dans le nombre d'instances à travers plusieurs classes (145). Les ensembles de données déséquilibrés ont le potentiel de conduire à des modèles qui présentent un biais envers la classe majoritaire. Des méthodes telles que le suréchantillonnage de la classe minoritaire, comme la Technique de Sur-échantillonnage de Minorité Synthétique (SMOTE) (175; 37), ou le sous-échantillonnage de la classe majoritaire (46), peuvent être employées pour atténuer cette préoccupation. De plus, il est possible d'utiliser des techniques algorithmiques telles que l'apprentissage sensible au coût (58) ou des méthodes d'ensemble (telles que le boosting) (203; 154) pour attribuer une importance ou un accent plus grands à la classe minoritaire lors de la formation du modèle. Le processus d'équilibrage des données est essentiel pour améliorer l'efficacité et l'impartialité du système d'IA en permettant au modèle d'apprendre correctement des deux classes, améliorant ainsi sa performance globale.

7 Compression et Optimisation des Modèles

Avec la croissance continue de la taille et de la complexité des modèles d'IA, il devient impératif de réduire leurs demandes computationnelles et de mémoire tout en maintenant les niveaux de performance (121). Les approches de compression et d'optimisation des modèles abordent cette difficulté en réduisant la taille, la complexité et les exigences de traitement des modèles d'IA tout en préservant leur précision et efficacité (150; 54). Ces stratégies facilitent la mise en œuvre de modèles d'IA sur des dispositifs aux ressources limitées, résultant en une amélioration de la vitesse d'inférence et une augmentation de la scalabilité (22).

7.1 Élagage

L'objectif de l'élagage des modèles est de réduire les dimensions des modèles d'IA en supprimant les caractéristiques superflues ou relativement moins significatives (191; 126). Cette méthodologie implique d'identifier et de supprimer les connexions, neurones ou couches entières qui contribuent peu à la performance globale du modèle. L'élagage peut être effectué soit pendant la phase d'entraînement, soit comme une action subséquente à la fin de l'entraînement. Pendant le processus d'entraînement, il est courant de remettre à zéro les petits poids ou d'éliminer les connexions basées sur certains critères, tels que l'élagage basé sur l'importance (117; 79). Les stratégies d'élagage post-entraînement (113; 116), telles que l'élagage des poids basé sur l'importance ou l'élagage itératif, sont employées pour évaluer le modèle appris et éliminer les paramètres redondants. L'élagage a pour double objectif de réduire la taille du modèle et d'améliorer la performance de calcul lors de l'inférence en réduisant le nombre de paramètres à évaluer.

7.2 Quantification

Les stratégies de quantification visent principalement à réduire le niveau de précision des paramètres du modèle, passant souvent de représentations en virgule flottante à des représentations à point fixe (63). La quantification est cruciale pour réduire l'utilisation de la mémoire et les demandes de traitement des modèles d'IA en codant les valeurs numériques avec moins de bits. Des méthodes telles que la quantification uniforme (197; 19) ou la quantification basée sur le regroupement (193; 195) sont employées pour regrouper des valeurs comparables et les représenter à l'aide d'un livre de codes commun. Cette approche vise à minimiser la perte d'informations. Les techniques de formation consciente de la quantification impliquent l'intégration de contraintes de quantification tout au long de la procédure d'entraînement, garantissant la résilience du modèle à la réduction de précision. Récemment, il y a eu des progrès notables dans le domaine, notamment en quantification uniquement entière et en quantification de précision mixte. Ces développements ont atteint un équilibre harmonieux entre la réduction de la précision et l'amélioration de la performance du modèle. En conséquence, ils ont facilité le déploiement efficace des modèles sur du matériel aux capacités de calcul limitées.

7.3 Distillation des Connaissances

Le processus de distillation des connaissances implique le transfert de connaissances d'un modèle enseignant, caractérisé par sa structure grande et complexe, à un modèle étudiant, conçu pour être plus petit et plus efficace (125; 67). Le modèle enseignant sert de mécanisme directeur en offrant des cibles douces ou des distributions de probabilité au lieu de cibles rigides pendant le processus d'entraînement (38). Le modèle étudiant s'efforce de répliquer les actions du modèle enseignant en réduisant l'écart entre leurs prévisions respectives (182). Le processus de distillation des connaissances permet au modèle étudiant d'acquérir les informations et capacités de généralisation du modèle enseignant tout en affichant une forme plus condensée et efficace sur le plan computationnel (141). Cette technique s'avère particulièrement avantageuse dans les scénarios où les modèles d'IA doivent être déployés sur des dispositifs aux capacités de calcul limitées ou lorsqu'il est nécessaire d'atteindre une performance semblable à celle d'un ensemble avec un seul modèle (83).

7.4 Approximation de Faible Rang

L'objectif des approches d'approximation de faible rang est de réduire la complexité et les demandes de calcul des modèles d'IA en approximant leurs matrices de poids avec des matrices de rang inférieur (140). La méthodologie actuelle tire parti de la reconnaissance que les matrices de poids présentent souvent une redondance et peuvent être efficacement approximées en utilisant un ensemble réduit de vecteurs de base. Des méthodes telles que la décomposition en valeurs singulières (SVD) (43; 60), la décomposition tensorielle (165), et la factorisation matricielle (183; 177) sont employées pour décomposer les matrices de poids en composants de rang inférieur. En utilisant des représentations de rang inférieur pour approximer les matrices de poids originales,

l'utilisation de la mémoire et les demandes de calcul du modèle sont considérablement réduites tout en maintenant ses propriétés fondamentales et sa performance.

7.5 Conception d'Architecture

Les stratégies de conception d'architecture mettent principalement l'accent sur le développement de modèles d'IA à la fois plus efficaces et plus légers, dès les premières étapes du processus de conception. Ces méthodologies englobent la réévaluation du cadre du modèle, la mise en œuvre de modifications architecturales, ou l'adoption de concepts de conception innovants pour améliorer l'efficacité. Diverses techniques, telles que l'élagage du réseau pendant la recherche d'architecture (196), la construction de blocs de réseau compacts, l'utilisation de connexions de saut ou de connexions résiduelles (133; 169; 192), et la mise en œuvre de convolutions séparables en profondeur (80; 181; 99), ont le potentiel de résulter en des modèles plus efficaces en termes de paramètres et de ressources de traitement. Le processus de conception d'architecture prend en compte la performance du modèle ainsi que les limitations imposées par l'environnement de déploiement prévu. Cela permet le développement de modèles personnalisés pour répondre aux exigences d'applications spécifiques ou de plateformes matérielles.

7.6 Distillation de Modèle

La distillation de modèle fait référence à la formation d'un modèle à plus petite échelle pour répliquer la fonctionnalité démontrée par un modèle plus grand et plus complexe (23; 189). Cette approche se distingue de la distillation des connaissances puisqu'elle met l'accent sur la formation d'un modèle autonome plutôt que sur le transfert de connaissances à partir d'un modèle existant. Tout au long de la procédure de formation, le modèle plus grand adopte la position d'un enseignant, tandis que le modèle plus petit s'efforce de répliquer ses prédictions ou représentations internes. Le processus de distillation de modèle suit une approche similaire à celle de l'apprentissage supervisé traditionnel mais avec les sorties du modèle enseignant utilisées comme valeurs cibles. Cette méthodologie facilite la création de modèles simplifiés qui démontrent une efficacité computationnelle tout en maintenant une performance comparable à des modèles plus complexes.

7.7 AutoML et Recherche d'Architecture Neuronale

Les méthodologies AutoML et de recherche d'architecture neuronale (NAS) facilitent l'automatisation des procédures de compression et d'optimisation des modèles à travers l'utilisation d'algorithmes qui recherchent des architectures idéales ou effectuent une construction automatisée de modèles (85; 137; 57; 88). Ces méthodologies visent à explorer une vaste gamme de configurations de modèles potentielles, d'hyperparamètres et d'algorithmes d'optimisation pour découvrir les configurations qui sont à la fois hautement efficaces et performantes (179; 185; 155; 122). AutoML et NAS emploient diverses stratégies telles que l'apprentissage par renforcement, les algorithmes évolutifs et l'optimisation basée sur le gradient pour faciliter le processus de recherche.

En utilisant des procédures automatisées pour la construction et l'optimisation des modèles, ces techniques facilitent le développement de modèles personnalisés qui adhèrent aux limitations de performance et de ressources prédéterminées. Ceci résulte en des économies substantielles de temps et de travail pour les praticiens en IA.

8 Calcul Distribué et Parallélisation

La section suivante explore les principes, les méthodologies et les avantages du calcul distribué et de la parallélisation dans le cadre des systèmes d'IA (101). Avec la croissance continue du volume de données et l'augmentation de la complexité des modèles d'IA, l'utilisation du calcul distribué et de la parallélisation est apparue comme une approche viable pour relever les défis de scalabilité (178). Ces techniques permettent le traitement et l'analyse efficaces de grands ensembles de données et l'accélération des processus d'entraînement et d'inférence de l'IA. Grâce à l'utilisation de diverses ressources informatiques opérant en collaboration, ces méthodologies facilitent des calculs d'IA plus rapides et plus efficaces. Cette section se penche sur de multiples facettes du calcul distribué et de la parallélisation, incluant les architectures systèmes, le parallélisme des données, le parallélisme des modèles, les mécanismes de synchronisation, les systèmes de gestion de cluster, le parallélisme des GPU et des accélérateurs, et leurs applications dans l'entraînement et l'inférence distribués (199).

8.1 Architectures Systèmes

L'importance des conceptions de systèmes est cruciale pour permettre le calcul distribué et la parallélisation. Ces entités sont le cadre principal pour organiser et coordonner les ressources et les tâches distribuées. Diverses méthodologies architecturales, telles que client-serveur, maître-esclave, pair-à-pair et modèles hybrides, peuvent être utilisées dans la conceptualisation des systèmes distribués (190; 21; 14). Les cadres architecturaux sont responsables de l'établissement des protocoles et des méthodes qui fournissent la communication, la collaboration et l'échange de données et de tâches de calcul entre les nœuds informatiques. La détermination de l'architecture système repose sur plusieurs facteurs, y compris la scalabilité, la tolérance aux pannes, les schémas de communication et les besoins spéciaux inhérents à l'application d'IA.

8.2 Parallélisme des Données

Le parallélisme des données est une méthodologie computationnelle qui implique la réplication d'un modèle unique sur plusieurs nœuds informatiques (118; 163). Chaque nœud traite alors indépendamment un sous-ensemble unique des données disponibles en parallèle. Dans le calcul distribué, des nœuds distincts se voient attribuer des sous-ensembles spécialisés de données à des fins d'entraînement ou d'inférence. Ensuite, ces nœuds traitent individuellement les données qui leur sont attribuées, effectuant des calculs sur leurs sous-ensembles respectifs (16; 51). Après cela, les nœuds procèdent à l'échange de gradients ou de résultats pour permettre un traitement collaboratif. Le parallélisme des données est une stratégie hautement efficace, particulièrement lorsque

l'ensemble de données peut être divisé en lots ou sous-ensembles plus petits pouvant être traités individuellement en parallèle sans interdépendances. Cette technologie permet un entraînement et une inférence distribués efficaces en partageant la charge de calcul parmi plusieurs nœuds, réduisant ainsi le temps global de traitement des données.

8.3 Parallélisme des Modèles

Le parallélisme des modèles est une technique qui implique la partition d'un modèle d'IA substantiel en composants plus petits et la distribution de ces composants sur plusieurs nœuds de traitement (135; 142). Chaque nœud individuel est chargé d'effectuer un certain sous-ensemble des tâches de calcul ou des couches dans le modèle. Le parallélisme des modèles est utilisé lorsque un nœud informatique unique manque de la mémoire ou des capacités de traitement nécessaires pour accueillir l'ensemble du modèle (204; 204). En partitionnant le modèle en composants plus petits et en les dispersant parmi des nœuds distincts, les demandes de calcul et de mémoire sont réparties, facilitant ainsi l'accommodation du modèle dans les ressources existantes. La mise en œuvre du parallélisme des modèles nécessite une coordination minutieuse et une communication efficace entre les nœuds pour garantir l'échange précis des résultats intermédiaires et la synchronisation du processus global.

8.4 Mécanismes de Synchronisation

Les méthodes de synchronisation sont cruciales pour coordonner le calcul et la communication parmi les nœuds dispersés dans les systèmes de calcul parallèle (187; 161). Ces procédures assurent la synchronisation appropriée des nœuds tout au long des processus d'entraînement et d'inférence, maintenant ainsi la cohérence et la précision. Des points de synchronisation sont mis en place pour faciliter l'alignement des calculs et permettre le flux d'informations entre les nœuds, incluant mais sans s'y limiter, les gradients, les paramètres du modèle et les résultats intermédiaires. Les méthodes couramment employées pour synchroniser les nœuds distribués et assurer des résultats cohérents comprennent la synchronisation par barrière, la moyenne des paramètres et l'agrégation des gradients. La mise en œuvre d'algorithmes de synchronisation efficaces est d'une importance capitale pour atteindre une performance optimale dans les systèmes de calcul parallèle tout en atténuant les risques associés aux situations de course et aux incohérences de données.

8.5 Systèmes de Gestion de Cluster

Les systèmes de gestion de cluster offrent une infrastructure complète et des outils qui simplifient le déploiement et la gestion des systèmes d'IA distribués (44; 144; 123). Ces systèmes gèrent l'allocation des ressources, planifient les travaux, assurent la tolérance aux pannes et équilibrent la charge de travail dans un environnement distribué. Les systèmes de gestion de cluster ont pour but d'abstraire l'infrastructure matérielle sous-jacente, offrant ainsi une interface consolidée pour gérer efficacement les ressources distribuées. Cette abstraction permet aux utilisateurs de se concentrer sur le développement et l'exécution des applications d'IA. Des exemples notables de systèmes de gestion

de cluster incluent Apache Hadoop, Apache Spark, Kubernetes et Apache Mesos. Ces systèmes peuvent être mis à l'échelle, tolérer les pannes et montrer de la flexibilité, les rendant bien adaptés pour mener de vastes calculs d'IA distribués.

8.6 Parallélisme des GPU et des Accélérateurs

Les approches de parallélisation englobent plus que le calcul distribué ; elles comprennent également l'utilisation des GPU et d'autres accélérateurs pour exploiter les capacités de traitement (119). Les GPU et les accélérateurs matériels spécialisés, tels que les TPUs, possèdent d'importantes capacités de calcul parallèle qui ont le potentiel d'améliorer grandement la vitesse des opérations d'IA. En déléguant des opérations informatiquement exigeantes, telles que les multiplications matricielles ou les convolutions, aux GPU ou aux accélérateurs spécialisés, les modèles d'IA peuvent atteindre des améliorations significatives de performance. Des méthodes telles que le parallélisme des GPU et le calcul de précision mixte facilitent l'utilisation optimale des ressources matérielles, améliorant ainsi la vitesse des processus d'entraînement et d'inférence.

8.7 Entraînement et Inférence Distribués

Les techniques de calcul distribué et de parallélisation peuvent être efficacement employées dans les activités d'entraînement et d'inférence des modèles d'IA (84). L'entraînement distribué implique la répartition de la charge de calcul sur de nombreux nœuds, résultant en une convergence accélérée et une réduction de la durée d'entraînement. L'entraînement distribué s'avère avantageux dans les scénarios impliquant d'importantes ressources informatiques, telles que de grands ensembles de données ou des modèles complexes. Dans l'inférence distribuée, la répartition de la charge de travail parmi de nombreux nœuds permet un traitement accéléré et amélioré des demandes d'inférence. L'inférence distribuée joue un rôle crucial dans les applications d'IA nécessitant des capacités en temps réel ou à haut débit, privilégiant la faible latence et la scalabilité comme facteurs clés. L'utilisation du calcul distribué et de la parallélisation permet l'accélération des activités d'entraînement et d'inférence, facilitant ainsi la mise en œuvre des systèmes d'IA à grande échelle.

Le sujet englobe plusieurs domaines, incluant les Architectures Systèmes, le Parallélisme des Données, le Parallélisme des Modèles, les Mécanismes de Synchronisation, les Systèmes de Gestion de Cluster, le Parallélisme des GPU et des Accélérateurs, et l'Entraînement et l'Inférence Distribués. Ces techniques facilitent le traitement efficace, la scalabilité et l'accélération des systèmes d'IA en concevant des systèmes distribués, en répliquant des modèles, en partitionnant des tâches, en mettant en œuvre la synchronisation, en gérant des clusters, et en utilisant des GPU et des accélérateurs.

9 Stratégies Écoénergétiques

L'intensification des calculs et la popularité croissante des charges de travail d'IA ont conduit à une focalisation accrue sur l'optimisation de la consommation d'énergie. L'optimisation de l'utilisation de l'énergie résulte non seulement en une diminution des

dépenses opérationnelles mais améliore également la longévité des batteries des dispositifs et atténue les conséquences écologiques associées aux systèmes d'IA. Cette section explore de nombreuses méthodologies et considérations pour atteindre des systèmes d'IA écoénergétiques. En utilisant ces tactiques, les entreprises et les individus peuvent atteindre un équilibre harmonieux entre capacités computationnelles et utilisation de l'énergie, favorisant ainsi l'établissement d'implémentations d'IA durables.

9.1 Optimisation des Modèles

L'objectif principal des stratégies d'optimisation des modèles est de minimiser les demandes computationnelles et de mémoire des modèles d'IA, entraînant une conservation de l'énergie. Les techniques mentionnées incluent la compression de modèle, la quantification et l'élagage. L'objectif de la compression de modèle est de réduire la taille du modèle en supprimant les composants redondants ou moins influents, menant à une consommation réduite de mémoire et à des exigences de traitement moindres pendant les processus d'entraînement et d'inférence. Le processus de quantification résulte en une diminution du niveau de précision des paramètres d'un modèle, facilitant ainsi une utilisation plus efficace des ressources matérielles et menant à une réduction de la consommation d'énergie. Le processus d'élagage implique l'élimination des connexions ou poids insignifiants au sein du modèle, menant à une structure plus condensée et écoénergétique. En mettant en œuvre ces stratégies d'optimisation, les modèles d'IA peuvent atteindre des niveaux de performance similaires tout en utilisant moins de ressources, entraînant une conservation substantielle de l'énergie.

9.2 Accélération Matérielle

L'accélération matérielle améliore l'efficacité énergétique en déléguant des activités computationnelles des processeurs polyvalents à des composants matériels dédiés. Les GPU, TPU et FPGA sont des accélérateurs matériels fréquemment employés dans les systèmes d'IA. Le but de ces accélérateurs est d'optimiser l'exécution de calculs ciblés, facilitant ainsi des processus d'IA accélérés et écoénergétiques. En utilisant l'accélération matérielle, les tâches d'IA peuvent être exécutées efficacement avec un débit accru et une consommation d'énergie réduite par rapport à une dépendance exclusive sur des processeurs conventionnels. L'intégration d'accélérateurs matériels spécialisés dans les systèmes d'IA facilite l'achèvement efficace d'opérations computationnellement exigeantes, résultant en une conservation de l'énergie à plus grande échelle.

9.3 Optimisation à l'Exécution

L'objectif principal des stratégies d'optimisation à l'exécution est de minimiser la consommation d'énergie lors de l'exécution des charges de travail d'IA. Les stratégies mentionnées incluent la planification des tâches, l'allocation des ressources et la gestion de l'équilibre des charges de travail. Grâce à une allocation stratégique des ressources, l'optimisation à l'exécution permet l'exécution des charges de travail d'IA d'une manière qui maximise l'efficacité énergétique. Des techniques telles que l'ajustement dynamique de la tension et de la fréquence (DVFS) sont employées pour moduler

la fréquence de fonctionnement et la tension des processeurs en réponse aux exigences de la charge de travail. Cette approche vise à optimiser l'utilisation de l'énergie tout en maintenant les niveaux de performance. De plus, des méthodes d'équilibrage des charges de travail sont employées pour diviser la charge computationnelle parmi plusieurs ressources, atténuant la sous-utilisation des ressources et facilitant une exécution écoénergétique.

9.4 Conception Basse Consommation

L'objectif principal des principes de conception basse consommation est de réduire la consommation d'énergie au niveau matériel. Cela implique le développement de CPU, systèmes de mémoire et interfaces d'entrée/sortie écoénergétiques. Des techniques telles que la mise hors tension et l'arrêt d'horloge sont mises en œuvre pour aborder l'inefficacité énergétique. Ces techniques incluent la désactivation sélective ou la limitation de l'alimentation électrique aux composants qui sont inactifs ou sous-utilisés. De plus, l'intégration de composants basse consommation et écoénergétiques, tels que les CPU basse puissance et les technologies de mémoire, est cruciale pour atteindre la conservation de l'énergie dans les systèmes d'IA. L'utilisation d'approches de conception basse consommation a le potentiel d'améliorer significativement l'efficacité énergétique du matériel d'IA, menant à une réduction de la consommation d'énergie pendant les opérations d'IA.

9.5 Formation Sensible à l'Énergie

La formation sensible à l'énergie fait référence à la méthodologie de formation de modèles d'IA en se concentrant sur l'optimisation de l'utilisation de l'énergie. L'incorporation de la consommation d'énergie comme métrique tout au long du processus de formation du modèle est un élément crucial de cette entreprise. L'objectif principal de la formation sensible à l'énergie est d'atteindre un état d'équilibre équilibré entre la performance d'un modèle et son efficacité énergétique. Ceci est accompli par l'évaluation et l'optimisation de la consommation d'énergie associée aux algorithmes de formation et aux hyperparamètres. La méthodologie comprend une gamme de techniques, dont l'une implique l'incorporation de la régularisation sensible à l'énergie. Cette technique impose une pénalité sur la complexité du modèle, encourageant ainsi la création de représentations écoénergétiques. Intégrer des considérations énergétiques dans la méthodologie de formation rend possible le développement de modèles d'IA qui démontrent des exigences de traitement réduites et une meilleure efficacité énergétique.

9.6 Optimisation au Niveau Système

Les techniques d'optimisation au niveau du système impliquent l'examen et l'amélioration complets de la structure globale d'un système d'IA, y compris ses composants, pour atteindre une efficacité énergétique maximale. Cela englobe l'avancement de protocoles de communication efficaces, l'amélioration des modèles d'accès à la mémoire et la minimisation des transferts de données entre différents composants. En employant une conception précise de l'architecture système, il devient possible de réduire

la consommation d'énergie par la réduction des transferts de données et l'optimisation de l'utilisation des ressources. De plus, le processus d'optimisation au niveau du système englobe la sélection méticuleuse de composants matériels appropriés, tels que les CPU et systèmes de mémoire écoénergétiques, dans le but de construire des systèmes d'IA qui présentent une efficacité énergétique. En considérant l'ensemble du système, il est possible d'optimiser l'efficacité énergétique à travers tous les composants, menant à des diminutions significatives de l'utilisation de l'énergie.

9.7 Surveillance et Analyse de l'Énergie

Les méthodologies de surveillance et d'analyse de l'énergie offrent des aperçus précieux sur les modèles de consommation d'énergie et facilitent l'identification des domaines potentiels d'amélioration. En surveillant la consommation d'énergie dans divers composants du système pendant les activités d'IA, les entreprises peuvent découvrir des processus ou des goulets d'étranglement qui consomment des quantités significatives d'énergie. Ces données peuvent être utilisées pour améliorer l'utilisation de l'énergie en se concentrant sur des domaines spécifiques pour l'amélioration, tels que des algorithmes inefficaces, des déséquilibres dans l'allocation des ressources ou des configurations matérielles. L'utilisation de la surveillance et de l'analyse de l'énergie facilite la mise en œuvre stratégique de mesures d'économie d'énergie et favorise un processus continu d'amélioration de l'efficacité énergétique. En comprenant les modèles d'utilisation de l'énergie, les organisations peuvent prendre des décisions éclairées visant à optimiser la consommation d'énergie au sein de leurs systèmes d'IA.

Le sujet englobe plusieurs domaines, incluant l'Optimisation des Modèles, l'Accélération Matérielle, l'Optimisation à l'Exécution, la Conception Basse Consommation, la Formation Sensible à l'Énergie, l'Optimisation au Niveau Système, et la Surveillance et l'Analyse de l'Énergie. Les stratégies mentionnées mettent principalement l'accent sur la réduction des exigences computationnelles et de mémoire, l'utilisation de l'accélération matérielle, l'optimisation des opérations à l'exécution, la mise en œuvre de principes de conception basse consommation, l'intégration de considérations énergétiques dans les processus de formation, l'optimisation de la structure globale du système, et la surveillance continue de la consommation d'énergie pour améliorer l'efficacité énergétique.

10 Utilisation des Méthodes Formelles pour Améliorer l'Efficacité de l'IA

Cette section explore l'utilisation d'approches formelles pour améliorer l'efficacité des systèmes d'IA (66; 50). Les approches formelles offrent un cadre mathématique robuste pour l'examen, la vérification et la validation des attributs des systèmes, garantissant ainsi précision, fiabilité et efficacité. Grâce à l'utilisation d'approches formelles, les entreprises ont la capacité d'optimiser plusieurs facettes des systèmes d'IA, conduisant à une amélioration substantielle de leur efficacité globale (7; 59; 110). Cette section se penche sur l'utilisation des approches formelles pour améliorer l'efficacité des systèmes d'IA, facilitant la mise en œuvre de déploiements fiables.

10.1 Spécification et Modélisation Formelles

Le processus de définition et de modélisation formelles inclut l'utilisation de langages mathématiques et de formalismes pour articuler précisément le comportement et les caractéristiques des systèmes d'IA. Les méthodologies mentionnées facilitent une représentation claire et sans équivoque des besoins du système, garantissant une compréhension complète de la performance et du comportement du système. Grâce à la spécification rigoureuse des exigences des systèmes d'IA, les développeurs peuvent détecter des inefficacités potentielles ou des défauts de conception tôt dans le processus de développement. L'utilisation de la modélisation formelle facilite l'examen du comportement du système dans de nombreuses circonstances, permettant ainsi l'amélioration des composants cruciaux et l'élimination des goulots d'étranglement problématiques. En employant des techniques de spécification et de modélisation formelles, il est possible de construire des systèmes d'IA qui privilégient l'efficacité, résultant en une performance améliorée et une utilisation réduite des ressources.

10.2 Vérification Formelle

Les techniques de vérification formelle utilisent le raisonnement mathématique pour examiner et établir rigoureusement les attributs d'un système. Grâce à l'utilisation de techniques de vérification formelle, les entreprises ont la capacité de s'assurer que les systèmes d'IA sont conformes aux exigences et limitations prédéfinies. La vérification formelle permet une analyse complète du comportement du système, facilitant la vérification de la précision et l'identification de fautes ou vulnérabilités potentielles. De plus, l'utilisation de techniques de vérification formelle peut améliorer les conceptions de système en identifiant les calculs redondants ou superflus, résultant ainsi dans le développement d'algorithmes plus efficaces. La vérification formelle permet aux entreprises d'optimiser l'efficacité des systèmes d'IA tout en garantissant leur fiabilité et conformité aux normes prédéfinies.

10.3 Optimisation par Synthèse Formelle

Les approches de synthèse formelle sont conçues pour développer automatiquement des conceptions de système efficaces, en utilisant des spécifications formelles et des objectifs d'optimisation. Grâce à l'utilisation de la synthèse formelle, les entreprises ont la capacité d'explorer de nombreuses alternatives de conception et de développer automatiquement des configurations idéales pour les systèmes d'IA. Le processus de synthèse formelle prend en compte les objectifs de performance, les limitations imposées par les ressources disponibles et d'autres critères de conception, facilitant ainsi le développement de systèmes d'IA qui présentent une haute efficacité. Cette méthodologie permet l'examen des compromis de conception et la détermination des configurations de système optimales, résultant en une utilisation améliorée des ressources et une efficacité énergétique. Grâce à l'utilisation des approches de synthèse formelle, les entreprises ont la capacité d'optimiser le processus de conception et de développer des systèmes d'IA qui présentent des niveaux élevés d'efficacité et d'efficacité.

10.4 Analyse de Performance Formelle

Les méthodologies d'analyse de performance formelle se concentrent principalement sur la quantification et l'optimisation des attributs de performance des systèmes. Ces techniques emploient des modèles mathématiques et des analyses pour évaluer les paramètres de performance d'un système, y compris le débit, le temps de réponse et la consommation de ressources. Grâce à l'utilisation de l'analyse de performance formelle, les entreprises ont la capacité de discerner les goulots d'étranglement et les inefficacités qui peuvent exister au sein des systèmes d'IA. Les données mentionnées peuvent ensuite être utilisées pour améliorer les éléments cruciaux, optimiser l'allocation des ressources et affiner les paramètres du système pour atteindre des niveaux idéaux de performance et d'efficacité. L'utilisation de l'analyse de performance formelle permet aux entreprises de prendre des décisions éclairées concernant la conception et la configuration du système, résultant ainsi dans le développement de systèmes d'IA hautement efficaces.

10.5 Raffinement et Décomposition Formels

L'application de techniques de raffinement et de décomposition formels implique la décomposition systématique de systèmes d'IA complexes en modules discrets et gérables, tout en garantissant précision et optimisation de la performance. Ces méthodologies facilitent la progression incrémentale des systèmes d'IA, commençant par des exigences globales et les raffinant progressivement en composants prêts pour l'exécution. Grâce à l'utilisation de techniques de raffinement et de décomposition formels, les entreprises ont la capacité de découvrir et de réduire efficacement les inefficacités dès le début du processus de développement. Cette méthodologie permet l'amélioration des éléments individuels du système, garantissant le fonctionnement efficace de chaque module et son intégration harmonieuse dans le système plus large. L'utilisation des approches de raffinement et de décomposition formels joue un rôle significatif dans l'avancement des systèmes d'IA caractérisés par leur organisation et leur efficacité.

10.6 Tests et Évaluation Formels

Le but des méthodologies de tests et d'évaluation formels est d'évaluer méthodiquement le comportement et la performance des systèmes d'IA par rapport aux spécifications et exigences formelles. Ces techniques utilisent des méthodologies formelles pour développer et mettre en œuvre des cas de test qui évaluent de manière exhaustive la fonctionnalité et la performance d'un système. Grâce à l'utilisation de méthodologies de tests et d'évaluation formels, les entreprises ont la capacité de découvrir et de résoudre des préoccupations potentielles, garantissant ainsi que les systèmes d'IA fonctionnent comme prévu et répondent aux exigences de performance. Les méthodologies de test formelles facilitent la mesure de l'efficacité du système, l'identification des domaines à améliorer et l'optimisation des composants importants pour atteindre des niveaux élevés d'efficacité. L'intégration de méthodologies de test et d'évaluation formels dans le processus de développement peut améliorer significativement l'efficacité et la fiabilité des systèmes d'IA pour les entreprises.

10.7 Intégration avec les Outils et Cadres des Méthodes Formelles

Les outils et cadres des méthodes formelles offrent aux développeurs l'assistance et l'infrastructure nécessaires pour employer efficacement les procédures formelles. Ces outils fournissent une gamme de fonctionnalités, incluant des langages de spécification formels, des vérificateurs de modèles, des démonstrateurs de théorèmes et des outils d'analyse de performance. En intégrant des méthodes formelles, des outils et des cadres dans le développement de l'IA, les entreprises peuvent faciliter la mise en œuvre d'approches formelles, simplifier les tâches d'analyse et de vérification, et garantir l'utilisation efficace des méthodes formelles tout au long du cycle de vie du développement du système. L'utilisation de cette connexion permet aux développeurs d'exploiter efficacement les capacités des approches formelles, résultant ainsi dans des systèmes d'IA plus efficaces et fiables.

Grâce aux approches formelles, les systèmes d'IA peuvent être développés, évalués et améliorés avec un haut degré de rigueur et de précision. L'amélioration de l'efficacité des systèmes d'IA peut être réalisée à travers diverses approches,

11 Défis et Orientations Futures

Le domaine de l'IA est dans un état constant de développement, offrant à la fois des perspectives prometteuses et des obstacles notables. Cette section se concentre sur l'élucidation de plusieurs problèmes significatifs rencontrés lors du développement et de la mise en œuvre des systèmes d'IA et explore les avenues potentielles pour des progrès futurs dans la résolution de ces défis.

11.1 IA Éthique et Responsable

L'augmentation de la prévalence des technologies d'IA nécessite une forte emphase sur le développement et le déploiement éthiques et responsables de l'IA. Il est impératif pour les organisations de reconnaître et de traiter de manière proactive les problèmes liés au biais, à l'équité, à la transparence et à la responsabilité au sein des systèmes d'IA. Les directions futures comprennent l'impératif d'établir des cadres robustes et des règles complètes pour la mise en œuvre éthique de l'IA. De plus, il est crucial de plaider pour la formation d'équipes inclusives et diversifiées engagées dans le développement de l'IA tout en intégrant les considérations éthiques à travers l'ensemble du cycle de vie de l'IA.

11.2 Confidentialité et Sécurité des Données

La dépendance croissante à la collecte et au traitement étendus des données suscite des appréhensions concernant la protection et la confidentialité des données. Préserver les données confidentielles tout en obtenant des aperçus significatifs à partir des ensembles de données pose un obstacle considérable. Les domaines potentiels pour l'exploration future englobent l'avancement des méthodologies d'IA qui priorisent la

préservation de la vie privée, l'intégration de systèmes sécurisés pour l'échange et le stockage des données, et l'établissement de cadres de gouvernance des données robustes pour soutenir les droits des individus à la vie privée.

11.3 Interprétabilité et Explicabilité

Les difficultés relatives à l'interprétabilité et à l'explicabilité des modèles d'IA continuent d'être d'une grande importance. Obtenir une compréhension complète des processus décisionnels employés par les systèmes d'IA est crucial pour maintenir les principes de transparence et favoriser la confiance. La trajectoire future de ce domaine implique la poursuite de modèles d'IA interprétables et explicables, l'avancement de la recherche dans les méthodologies d'interprétabilité indépendantes du modèle, et l'établissement de normes d'interprétabilité et d'explicabilité dans des secteurs cruciaux comme la santé et la finance.

11.4 Adaptabilité et Généralisation

Les systèmes d'IA rencontrent fréquemment des défis lorsqu'il s'agit de s'ajuster efficacement à des circonstances nouvelles ou inconnues et d'extrapoler les connaissances à travers des disciplines disparates. Améliorer les capacités d'adaptabilité et de généralisation de l'IA est primordial lorsqu'on envisage son application dans des scénarios réels. Les avenues potentielles futures d'exploration englobent des investigations dans les méthodologies d'apprentissage par transfert, d'apprentissage par méta et d'apprentissage tout au long de la vie, qui visent à faciliter l'acquisition et l'application efficaces d'informations par les systèmes d'IA à travers diverses situations.

11.5 Impact Social et Économique

La mise en œuvre étendue de l'IA porte d'importantes ramifications sociales et économiques. La considération des ramifications potentielles de l'IA sur l'emploi, l'inégalité et les dynamiques sociales est d'une importance capitale. Les directions futures englobent la poursuite de recherches interdisciplinaires visant à comprendre les ramifications sociétales de l'IA. De plus, des efforts seront faits pour avancer le déploiement responsable de l'IA en mettant en œuvre une législation et une réglementation. En outre, l'établissement de collaborations entre l'industrie, le milieu universitaire et les décideurs politiques sera priorisé pour garantir que les avantages de l'IA soient distribués de manière inclusive.

11.6 Durabilité et Efficacité Énergétique

Les problèmes croissants en IA sont dus à la consommation d'énergie des systèmes d'IA et à leur impact environnemental subséquent. Les perspectives futures clés incluent le développement d'algorithmes d'IA écoénergétiques, l'optimisation des architectures matérielles et la promotion de pratiques informatiques durables. De plus, l'investigation des sources d'énergie renouvelables et l'avancement des systèmes d'IA qui contribuent

activement aux objectifs de durabilité sont des mesures impératives pour atténuer l'impact environnemental de la technologie d'IA.

11.7 Collaboration et Régulation

Le développement et le déploiement de l'IA nécessitent les efforts collaboratifs de multiples parties prenantes, englobant les chercheurs, les décideurs politiques, les représentants de l'industrie et le grand public. Les directions futures englobent l'établissement de plateformes et de programmes collaboratifs visant à faciliter le partage des connaissances. De plus, ces directions impliquent de s'attaquer aux préoccupations juridiques et éthiques par la coopération internationale, ainsi que le développement de cadres réglementaires qui équilibrent efficacement l'innovation et la responsabilité.

11.8 Avancement de la Recherche et de l'Éducation en IA

L'avancement ultérieur de l'IA dépend du développement continu de la recherche et de l'éducation en IA. Les domaines potentiels pour l'exploration future englobent la culture de partenariats de recherche multidisciplinaires, la fourniture de soutien pour les efforts de recherche en IA en accès ouvert, et l'avancement des alternatives pour l'apprentissage tout au long de la vie pour habiliter les individus avec l'aptitude à comprendre, créer et interagir avec la technologie d'IA de manière compétente.

11.9 Intégration avec les Outils et Cadres des Méthodes Formelles

Les directions futures dans le domaine englobent l'amélioration de l'intégration des méthodes formelles, des outils et des cadres. Ceci implique de simplifier le processus d'incorporation des méthodes formelles dans le développement de l'IA, ainsi que l'avancement de la convivialité et de la disponibilité des outils formels. De plus, des efforts sont faits pour faciliter l'amalgame des techniques de vérification formelle, de synthèse et de test dans les flux de travail de développement d'IA conventionnels. L'incorporation de techniques formelles fournira aux développeurs la capacité d'utiliser efficacement les méthodes formelles, améliorant ainsi l'efficacité, la fiabilité et la confiance dans les systèmes d'IA.

En reconnaissant et en confrontant ces obstacles, le domaine de l'IA peut progresser de manière consciencieuse et durable. Les directions futures englobent l'impératif de collaboration interdisciplinaire, les dimensions éthiques à prendre en compte, l'établissement de cadres réglementaires et la poursuite continue de l'innovation. Ces mesures sont nécessaires pour garantir que les technologies d'IA soient exploitées au profit de l'humanité, tout en abordant et en minimisant les risques et obstacles potentiels.

12 Conclusion

L'IA s'est imposée comme une technologie transformatrice avec le potentiel de révolutionner divers aspects de l'existence humaine. Cette étude explore une gamme de

sujets importants liés à l'IA et offre un examen approfondi des avancements, des applications pratiques, des défis et des orientations futures potentielles de l'IA.

La recherche a commencé par analyser les principes fondamentaux de l'IA et les concepts de base qui forment sa base, y compris l'apprentissage automatique (ML), l'apprentissage profond (DL) et les réseaux de neurones. Au cours de notre discours, nous avons analysé les progrès significatifs réalisés dans l'investigation de l'IA à côté de la large gamme d'implémentations observées dans diverses industries telles que la santé, la banque, les transports et le divertissement.

En outre, un examen a été mené sur les avancements réalisés dans les méthodologies et cadres de l'IA, y compris les CNN utilisés pour la reconnaissance d'image, les RNN employés pour la modélisation de séquence, et les modèles génératifs utilisés pour le développement de contenu. Cette étude examine l'importance des ensembles de données annotées, l'utilisation de l'apprentissage par transfert et l'incorporation des techniques d'apprentissage par renforcement dans la formation de systèmes d'IA robustes et compétents.

Les implications éthiques entourant l'IA ont également été délibérées, soulignant l'impératif d'un progrès conscient de l'IA, d'équité, d'ouverture et de responsabilité. Les problèmes présentés par les biais, les préoccupations de confidentialité, l'interprétabilité et les implications socioéconomiques possibles de l'IA ont été soigneusement abordés. Aborder ces difficultés nécessite une collaboration interdisciplinaire, l'établissement de cadres législatifs et l'incorporation de concepts éthiques dans les algorithmes d'IA et les processus de prise de décision.

De plus, nous avons souligné l'importance de la scalabilité, de l'efficacité et de la durabilité dans les systèmes d'IA. L'optimisation de l'utilisation de l'énergie et le développement d'algorithmes et d'architectures matérielles écoénergétiques seront cruciaux pour atténuer l'impact environnemental et réduire les coûts opérationnels à mesure que les charges de travail de l'IA continuent d'augmenter en intensité computationnelle.

Bien que des avancements considérables aient été réalisés dans le domaine de l'IA, il existe une multitude de difficultés qui nécessitent une attention et une résolution. Les problèmes incluent les considérations éthiques, la confidentialité et la sécurité des données, l'interprétabilité, l'adaptabilité, les effets sociaux et économiques, la durabilité, la collaboration et l'avancement de la recherche et de l'éducation en IA. Aborder ces difficultés nécessite l'établissement d'efforts collaboratifs parmi les chercheurs, les décideurs politiques et les acteurs de l'industrie afin de garantir la mise en œuvre responsable et avantageuse de la technologie d'IA.

En résumé, l'IA possède une capacité significative à apporter des changements transformateurs dans notre société. Cependant, elle pose également de nombreux obstacles qui nécessitent une considération réfléchie et une résolution. À travers la culture d'un développement responsable, de normes éthiques et d'innovations continues, le potentiel de l'IA peut être utilisé pour améliorer plusieurs aspects de l'existence humaine. Cela inclut l'amélioration de la qualité de vie, le raffinement des processus de prise de décision, la facilitation des avancées scientifiques et la lutte efficace contre des problèmes sociétaux complexes. En adoptant une méthodologie appropriée, l'IA peut potentiellement émerger comme un instrument puissant pour faciliter un changement constructif,

produisant ainsi des avantages pour les individus, les communautés et la communauté mondiale.

Références

[1] Mohamed S Abdalzaher, Sayed SR Moustafa, HE Abdel Hafiez, and Walid Farid Ahmed. An optimized learning model augment analyst decisions for seismic source discrimination. *IEEE Transactions on Geoscience and Remote Sensing*, 60 :1–12, 2022.

[2] Mohamed S Abdalzaher, M Sami Soliman, and Sherif M El-Hady. Seismic intensity estimation for earthquake early warning using optimized machine learning model. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[3] Abdiansah Abdiansah and Retantyo Wardoyo. Time complexity analysis of support vector machines (svm) in libsvm. *Int. J. Comput. Appl*, 128(3) :28–34, 2015.

[4] Q Abu Al-Haija and M Krichen. A lightweight in-vehicle alcohol detection using smart sensing and supervised learning. computers 2022, 11, 121, 2022.

[5] Qasem Abu Al-Haija, Moez Krichen, and Wejdan Abu Elhaija. Machine-learning-based darknet traffic detection system for iot applications. *Electronics*, 11(4) :556, 2022.

[6] Marion Olubunmi Adebiyi, Micheal Olaolu Arowolo, Moses Damilola Mshelia, and Oludayo O Olugbara. A linear discriminant analysis and classification model for breast cancer diagnosis. *Applied Sciences*, 12(22) :11455, 2022.

[7] Faouzi Adjed, Mallek Mziou-Sallami, Frédéric Pelliccia, Mehdi Rezzoug, Lucas Schott, Christophe Bohn, and Yesmina Jaafra. Coupling algebraic topology theory, formal methods and safety requirements toward a new coverage metric for artificial intelligence models. *Neural Computing and Applications*, 34(19) :17129–17144, 2022.

[8] Qasem Abu Al-Haija and Moez Krichen. Analyzing malware from api call sequences using support vector machines. In *International Conference on Cybersecurity, Cybercrimes, and Smart Emerging Technologies*, pages 27–39. Springer International Publishing Cham, 2022.

[9] Omar Azib Alkhudaydi, Moez Krichen, and Ans D Alghamdi. A deep learning methodology for predicting cybersecurity attacks on the internet of things. *Information*, 14(10) :550, 2023.

[10] Hamoud Alshammari, Karim Gasmi, Ibtihel Ben Ltaifa, Moez Krichen, Lassaad Ben Ammar, and Mahmood A Mahmood. Olive disease classification based on vision transformer and cnn models. *Computational Intelligence and Neuroscience*, 2022, 2022.

[11] Hamoud Alshammari, Karim Gasmi, Moez Krichen, Lassaad Ben Ammar, Mohamed Osman Abdelhadi, Ammar Boukrara, and Mahmood A Mahmood. Optimal deep learning model for olive disease diagnosis based on an adaptive genetic algorithm. *Wireless Communications and Mobile Computing*, 2022 :1–13, 2022.

[12] Ali Alwehaibi, Marwan Bikdash, Mohammad Albogmi, and Kaushik Roy. A study of the performance of embedding methods for arabic short-text sentiment analysis using deep learning approaches. *Journal of King Saud University-Computer and Information Sciences*, 34(8) :6140–6149, 2022.

[13] Hashem Alyami, Wael Alosaimi, Moez Krichen, and Roobaea Alroobaea. Monitoring social distancing using artificial intelligence for fighting covid-19 virus spread. *International Journal of Open Source Software and Processes (IJOSSP)*, 12(3) :48–63, 2021.

[14] Sarina Aminizadeh, Arash Heidari, Shiva Toumaj, Mehdi Darbandi, Nima Jafari Navimipour, Mahsa Rezaei, Samira Talebi, Poupak Azad, and Mehmet Unal. The applications of machine learning techniques in medical data processing based on distributed computing and the internet of things. *Computer Methods and Programs in Biomedicine*, page 107745, 2023.

[15] Nikos Andriopoulos, Aristeidis Magklaras, Alexios Birbas, Alex Papalexopoulos, Christos Valouxis, Sophia Daskalaki, Michael Birbas, Efthymios Housos, and George P Papaioannou. Short term electric load forecasting based on data transformation and statistical machine learning. *Applied Sciences*, 11(1) :158, 2020.

[16] Ignacio Arnaldo, Kalyan Veeramachaneni, Andrew Song, and Una-May O'Reilly. Bring your own learner : A cloud-based, data-parallel commons for machine learning. *IEEE Computational Intelligence Magazine*, 10(1) :20–32, 2015.

[17] Adam Auten, Matthew Tomei, and Rakesh Kumar. Hardware acceleration of graph neural networks. In *2020 57th ACM/IEEE Design Automation Conference (DAC)*, pages 1–6. IEEE, 2020.

[18] Rubby Aworka, Lontsi Saadio Cedric, Wilfried Yves Hamilton Adoni, Jérémie Thouakesseh Zoueu, Franck Kalala Mutombo, Charles Lebon Mberi Kimpolo, Tarik Nahhal, and Moez Krichen. Agricultural decision system based on advanced machine learning models for yield prediction : Case of east african countries. *Smart Agricultural Technology*, 2 :100048, 2022.

[19] Manijeh Bashar, Kanapathippillai Cumanan, Alister G Burr, Hien Quoc Ngo, Erik G Larsson, and Pei Xiao. Energy efficiency of the cell-free massive mimo uplink with optimal uniform quantization. *IEEE Transactions on Green Communications and Networking*, 3(4) :971–987, 2019.

[20] Christian Beck, Weinan E, and Arnulf Jentzen. Machine learning approximation algorithms for high-dimensional fully nonlinear partial differential equations and second-order backward stochastic differential equations. *Journal of Nonlinear Science*, 29 :1563–1619, 2019.

[21] Ron Bekkerman, Mikhail Bilenko, and John Langford. *Scaling up machine learning : Parallel and distributed approaches*. Cambridge University Press, 2011.

[22] Anthony Berthelier, Thierry Chateau, Stefan Duffner, Christophe Garcia, and Christophe Blanc. Deep model compression and architecture optimization for embedded systems : A survey. *Journal of Signal Processing Systems*, 93 :863–878, 2021.

[23] Max Biggs, Wei Sun, and Markus Ettl. Model distillation for revenue optimiza-

tion : Interpretable personalized pricing. In *International Conference on Machine Learning*, pages 946–956. PMLR, 2021.

[24] Andrea Bommert, Xudong Sun, Bernd Bischl, Jörg Rahnenführer, and Michel Lang. Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis*, 143 :106839, 2020.

[25] Dmitrii Borkin, Andrea Némethová, German Michal'čonok, and Konstantin Maiorov. Impact of data normalization on classification model accuracy. *Research Papers Faculty of Materials Science and Technology Slovak University of Technology*, 27(45) :79–84, 2019.

[26] Wadii Boulila, Maha Driss, Eman Alshanqiti, Mohamed Al-Sarem, Faisal Saeed, and Moez Krichen. Weight initialization techniques for deep learning algorithms in remote sensing : Recent trends and future perspectives. *Advances on Smart and Soft Computing : Proceedings of ICACIn 2021*, pages 477–484, 2022.

[27] Zakaria Boulouard, Mariyam Ouaissa, Mariya Ouaissa, Farhan Siddiqui, Mutiq Almutiq, and Moez Krichen. An integrated artificial intelligence of things environment for river flood prevention. *Sensors*, 22(23) :9485, 2022.

[28] Sérgio Branco, André G Ferreira, and Jorge Cabral. Machine learning in resource-scarce embedded systems, fpgas, and end-devices : A survey. *Electronics*, 8(11) :1289, 2019.

[29] Erik Brynjolfsson and ANDREW Mcafee. Artificial intelligence, for real. *Harvard business review*, 1 :1–31, 2017.

[30] Jie Cai, Jiawei Luo, Shulin Wang, and Sheng Yang. Feature selection in machine learning : A new perspective. *Neurocomputing*, 300 :70–79, 2018.

[31] Murat Canayaz. Classification of diabetic retinopathy with feature selection over deep features using nature-inspired wrapper methods. *Applied Soft Computing*, 128 :109462, 2022.

[32] Maurizio Capra, Beatrice Bussolino, Alberto Marchisio, Muhammad Shafique, Guido Masera, and Maurizio Martina. An updated survey of efficient hardware architectures for accelerating deep convolutional neural networks. *Future Internet*, 12(7) :113, 2020.

[33] Lontsi Saadio Cedric, Wilfried Yves Hamilton Adoni, Rubby Aworka, Jérémie Thouakesseh Zoueu, Franck Kalala Mutombo, Moez Krichen, and Charles Lebon Mberi Kimpolo. Crops yield prediction based on machine learning models : Case of west african countries. *Smart Agricultural Technology*, 2 :100049, 2022.

[34] Patricio Cerda, Gaël Varoquaux, and Balázs Kégl. Similarity encoding for learning with dirty categorical variables. *Machine Learning*, 107(8-10) :1477–1494, 2018.

[35] Oumaima Chakir, Abdeslam Rehaimi, Yassine Sadqi, Moez Krichen, Gurjot Singh Gaba, Andrei Gurtov, et al. An empirical assessment of ensemble methods and traditional machine learning techniques for web-based attack detection in industry 5.0. *Journal of King Saud University-Computer and Information*

Sciences, 35(3) :103–119, 2023.

[36] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1) :16–28, 2014.

[37] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote : synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16 :321–357, 2002.

[38] Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11933–11942, 2022.

[39] Marianne Cherrington, Fadi Thabtah, Joan Lu, and Qiang Xu. Feature selection : filter methods performance challenges. In *2019 International Conference on Computer and Information Sciences (ICCIS)*, pages 1–4. IEEE, 2019.

[40] Xu Chu, Ihab F Ilyas, Sanjay Krishnan, and Jiannan Wang. Data cleaning : Overview and emerging challenges. In *Proceedings of the 2016 international conference on management of data*, pages 2201–2206, 2016.

[41] Andrew Collins and Yin Yao. Machine learning approaches : Data integration for disease prediction and prognosis. *Applied computational genomics*, pages 137–141, 2018.

[42] Donatello Conte, Jean-Yves Ramel, Nicolas Sidere, Muhammad Muzzamil Luqman, Benoit Gaüzere, Jaume Gibert, Luc Brun, and Mario Vento. A comparison of explicit and implicit graph embedding methods for pattern recognition. In *Graph-Based Representations in Pattern Recognition : 9th IAPR-TC-15 International Workshop, GbRPR 2013, Vienna, Austria, May 15-17, 2013. Proceedings 9*, pages 81–90. Springer, 2013.

[43] Julio Cesar Stacchini de Souza, Tatiana Mariano Lessa Assis, and Bikash Chandra Pal. Data compression in smart distribution systems via singular value decomposition. *IEEE transactions on smart grid*, 8(1) :275–284, 2015.

[44] Christina Delimitrou and Christos Kozyrakis. Quasar : Resource-efficient and qos-aware cluster management. *ACM SIGPLAN Notices*, 49(4) :127–144, 2014.

[45] Lei Deng, Guoqi Li, Song Han, Luping Shi, and Yuan Xie. Model compression and hardware acceleration for neural networks : A comprehensive survey. *Proceedings of the IEEE*, 108(4) :485–532, 2020.

[46] Debashree Devi, Saroj K Biswas, and Biswajit Purkayastha. A review on solution to class imbalance problem : Undersampling approaches. In *2020 international conference on computational performance evaluation (ComPE)*, pages 626–631. IEEE, 2020.

[47] Pradip Dhal and Chandrashekhar Azad. A comprehensive survey on feature selection in the various fields of machine learning. *Applied Intelligence*, pages 1–39, 2022.

[48] Payal Dhar. The carbon impact of artificial intelligence. *Nat. Mach. Intell.*, 2(8) :423–425, 2020.

[49] Xin Luna Dong and Theodoros Rekatsinas. Data integration and machine learning : A natural synergy. In *Proceedings of the 2018 international conference on management of data*, pages 1645–1650, 2018.

[50] Tommaso Dreossi, Daniel J Fremont, Shromona Ghosh, Edward Kim, Hadi Ravanbakhsh, Marcell Vazquez-Chanlatte, and Sanjit A Seshia. Verifai : A toolkit for the formal design and analysis of artificial intelligence-based systems. In *International Conference on Computer Aided Verification*, pages 432–442. Springer, 2019.

[51] Nikoli Dryden, Tim Moon, Sam Ade Jacobs, and Brian Van Essen. Communication quantization for data-parallel training of deep neural networks. In *2016 2nd Workshop on Machine Learning in HPC Environments (MLHPC)*, pages 1–8. IEEE, 2016.

[52] Javier Duarte, Philip Harris, Scott Hauck, Burt Holzman, Shih-Chieh Hsu, Sergo Jindariani, Suffian Khan, Benjamin Kreis, Brian Lee, Mia Liu, et al. Fpga-accelerated machine learning inference as a service for particle physics computing. *Computing and Software for Big Science*, 3 :1–15, 2019.

[53] Ritik Dutta, Varun Gohil, and Atishay Jain. Effect of feature hashing on fair classification. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, pages 365–366. 2020.

[54] Nahid Eddermoug, Abdeljebar Mansour, Mohamed Sadik, Essaid Sabir, and Mohamed Azmi. klm-ppsa v. 1.1 : machine learning-augmented profiling and preventing security attacks in cloud environments. *Annals of Telecommunications*, pages 1–27, 2023.

[55] Norman Einspruch. *Application specific integrated circuit (ASIC) technology*, volume 23. Academic Press, 2012.

[56] Mourad Ellouze, Seifeddine Mechti, Moez Krichen, Vinayakumar Ravi, and Lamia Hadrich Belguith. A deep learning approach for detecting the behaviour of people having personality disorders towards covid-19 from twitter. *International Journal of Computational Science and Engineering*, 25(4) :353–366, 2022.

[57] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search : A survey. *The Journal of Machine Learning Research*, 20(1) :1997–2017, 2019.

[58] Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C Prati, Bartosz Krawczyk, Francisco Herrera, Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C Prati, et al. Cost-sensitive learning. *Learning from imbalanced data sets*, pages 63–78, 2018.

[59] Nathan Fulton and André Platzer. Safe reinforcement learning via formal methods : Toward safe control through proof and learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[60] George W Furnas, Scott Deerwester, Susan T Durnais, Thomas K Landauer, Richard A Harshman, Lynn A Streeter, and Karen E Lochbaum. Information retrieval using a singular value decomposition model of latent semantic structure. In *ACM SIGIR Forum*, volume 51, pages 90–105. ACM New York, NY, USA,

2017.

[61] Zoubin Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553) :452–459, 2015.

[62] Deepak Ghimire, Dayoung Kil, and Seong-heum Kim. A survey on efficient convolutional neural networks and hardware acceleration. *Electronics*, 11(6) :945, 2022.

[63] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference. In *Low-Power Computer Vision*, pages 291–326. Chapman and Hall/CRC, 2022.

[64] Sacha Gobeyn, Ans M Mouton, Anna F Cord, Andrea Kaim, Martin Volk, and Peter LM Goethals. Evolutionary algorithms for species distribution modelling : A review in the context of machine learning. *Ecological Modelling*, 392 :179–195, 2019.

[65] Ashish Gondimalla, Noah Chesnut, Mithuna Thottethodi, and TN Vijaykumar. Sparten : A sparse tensor accelerator for convolutional neural networks. In *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*, pages 151–165, 2019.

[66] Frederik Gossen, Tiziana Margaria, and Bernhard Steffen. Towards explainability in machine learning : The formal methods way. *IT Professional*, 22(4) :8–12, 2020.

[67] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation : A survey. *International Journal of Computer Vision*, 129 :1789–1819, 2021.

[68] Stefan Grafberger, Julia Stoyanovich, and Sebastian Schelter. Lightweight inspection of data preprocessing in native machine learning pipelines. In *Conference on Innovative Data Systems Research (CIDR)*, 2021.

[69] Neha Gupta. Introduction to hardware accelerator systems for artificial intelligence and machine learning. In *Advances in Computers*, volume 122, pages 1–21. Elsevier, 2021.

[70] Michael Haenlein and Andreas Kaplan. A brief history of artificial intelligence : On the past, present, and future of artificial intelligence. *California management review*, 61(4) :5–14, 2019.

[71] Rihan Hai, Christos Koutras, Andra Ionescu, Ziyu Li, Wenbo Sun, Jessie Van Schijndel, Yan Kang, and Asterios Katsifodimos. Amalur : Data integration meets machine learning. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 3729–3739. IEEE, 2023.

[72] Saqib Hakak, Mamoun Alazab, Suleman Khan, Thippa Reddy Gadekallu, Praveen Kumar Reddy Maddikunta, and Wazir Zada Khan. An ensemble machine learning approach through effective feature extraction to classify fake news. *Future Generation Computer Systems*, 117 :47–58, 2021.

[73] Omar Hamdy, Hanan Gaber, Mohamed S Abdalzaher, and Mahmoud Elhadidy. Identifying exposure of urban area to certain seismic hazard using machine lear-

ning and gis : A case study of greater cairo. *Sustainability*, 14(17) :10722, 2022.

[74] John T Hancock and Taghi M Khoshgoftaar. Survey on categorical data for neural networks. *Journal of Big Data*, 7(1) :1–41, 2020.

[75] Guy S Handelman, Hong Kuan Kok, Ronil V Chandra, Amir H Razavi, Shiwei Huang, Mark Brooks, Michael J Lee, and Hamed Asadi. Peering into the black box of artificial intelligence : evaluation metrics of machine learning methods. *American Journal of Roentgenology*, 212(1) :38–43, 2019.

[76] Meng Hao, Hongwei Li, Xizhao Luo, Guowen Xu, Haomiao Yang, and Sen Liu. Efficient and privacy-enhanced federated learning for industrial artificial intelligence. *IEEE Transactions on Industrial Informatics*, 16(10) :6532–6542, 2019.

[77] Tyler L Hayes and Christopher Kanan. Lifelong machine learning with deep streaming linear discriminant analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 220–221, 2020.

[78] Gabriel Paes Herrera, Michel Constantino, Jen-Je Su, and Athula Naranpanawa. The use of icts and income distribution in brazil : A machine learning explanation using shap values. *Telecommunications Policy*, 47(8) :102598, 2023.

[79] Wenjing Hong, Peng Yang, Yiwen Wang, and Ke Tang. Multi-objective magnitude-based pruning for latency-aware deep neural network compression. In *International Conference on Parallel Problem Solving from Nature*, pages 470–483. Springer, 2020.

[80] Rashidul Hasan Hridoy, Tarek Habib, Ismail Jabiullah, Riazur Rahman, and Farruk Ahmed. Early recognition of betel leaf disease using deep learning with depth-wise separable convolutions. In *2021 IEEE region 10 symposium (TENSYMP)*, pages 1–7. IEEE, 2021.

[81] Olfa Hrizi, Karim Gasmi, Ibtihel Ben Ltaifa, Hamoud Alshammari, Hanen Karamti, Moez Krichen, Lassaad Ben Ammar, and Mahmood A Mahmood. Tuberculosis disease diagnosis based on an optimized machine learning model. *Journal of Healthcare Engineering*, 2022, 2022.

[82] Jianglin Huang, Yan-Fu Li, and Min Xie. An empirical analysis of data preprocessing for machine learning-based software cost estimation. *Information and software Technology*, 67 :108–127, 2015.

[83] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. *Advances in Neural Information Processing Systems*, 35 :33716–33727, 2022.

[84] Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V Le, Yonghui Wu, et al. Gpipe : Efficient training of giant neural networks using pipeline parallelism. *Advances in neural information processing systems*, 32, 2019.

[85] Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. *Automated machine learning : methods, systems, challenges*. Springer Nature, 2019.

[86] Ihab F Ilyas and Xu Chu. *Data cleaning*. Morgan & Claypool, 2019.

[87] Francesca Incitti, Federico Urli, and Lauro Snidaro. Beyond word embeddings : A survey. *Information Fusion*, 89 :418–436, 2023.

[88] Yesmina Jaafra, Jean Luc Laurent, Aline Deruyver, and Mohamed Saber Naceur. Reinforcement learning for neural architecture search : A review. *Image and Vision Computing*, 89 :57–66, 2019.

[89] Aabha Jain and Neha Sharma. Accelerated ai inference at cnn-based machine vision in asics : A design approach. *ECS Transactions*, 107(1) :5165, 2022.

[90] Conrad D James, James B Aimone, Nadine E Miner, Craig M Vineyard, Fredrick H Rothganger, Kristofor D Carlson, Samuel A Mulder, Timothy J Draelos, Aleksandra Faust, Matthew J Marinella, et al. A historical survey of algorithms and hardware architectures for neural-inspired and neuromorphic computing applications. *Biologically Inspired Cognitive Architectures*, 19 :49–64, 2017.

[91] Abhinav Jangda, Sandeep Polisetty, Arjun Guha, and Marco Serafini. Accelerating graph sampling for graph machine learning using gpus. In *Proceedings of the Sixteenth European Conference on Computer Systems*, pages 311–326, 2021.

[92] Pooja Jawandhiya. Hardware design for machine learning. *Int. J. Artif. Intell. Appl*, 9(1) :63–84, 2018.

[93] Norman Jouppi, Cliff Young, Nishant Patil, and David Patterson. Motivation for and evaluation of the first tensor processing unit. *ieee Micro*, 38(3) :10–19, 2018.

[94] Kaan Kara, Dan Alistarh, Gustavo Alonso, Onur Mutlu, and Ce Zhang. Fpga-accelerated dense linear machine learning : A precision-convergence trade-off. In *2017 IEEE 25th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, pages 160–167. IEEE, 2017.

[95] Kaan Kara, Ken Eguro, Ce Zhang, and Gustavo Alonso. Columnml : Column-store machine learning with on-the-fly data transformation. *Proceedings of the VLDB Endowment*, 12(4) :348–361, 2018.

[96] Athanasios Karapantelakis, Pegah Alizadeh, Abdulrahman Alabassi, Kaushik Dey, and Alexandros Nikou. Generative ai in mobile networks : a survey. *Annals of Telecommunications*, pages 1–19, 2023.

[97] Samina Khalid, Tehmina Khalil, and Shamila Nasreen. A survey of feature selection and feature extraction techniques in machine learning. In *2014 science and information conference*, pages 372–378. IEEE, 2014.

[98] Hajra Khan, Imran Fareed Nizami, Saeed Mian Qaisar, Asad Waqar, Moez Krichen, and Abdulaziz Turki Almaktoom. Analyzing optimal battery sizing in microgrids based on the feature selection and machine learning approaches. *Energies*, 15(21) :7865, 2022.

[99] Zahid Younas Khan and Zhendong Niu. Cnn with depthwise separable convolutions and combined kernels for rating prediction. *Expert Systems with Applications*, 170 :114528, 2021.

[100] Ferath Kherif and Adeliya Latypova. Principal component analysis. In *Machine Learning*, pages 209–225. Elsevier, 2020.

[101] Jin Kyu Kim, Qirong Ho, Seunghak Lee, Xun Zheng, Wei Dai, Garth A Gib-

son, and Eric P Xing. Strads : A distributed framework for scheduled model parallel machine learning. In *Proceedings of the Eleventh European Conference on Computer Systems*, pages 1–16, 2016.

[102] Moez Krichen. How artificial intelligence can revolutionize software testing techniques. In *International Conference on Innovations in Bio-Inspired Computing and Applications*, pages 189–198. Springer Nature Switzerland Cham, 2022.

[103] Moez Krichen. Les méthodes formelles sont-elles applicables à l'apprentissage automatique et à l'intelligence artificielle. 2022.

[104] Moez Krichen. Comment l'intelligence artificielle peut révolutionner les techniques de test de logiciels. 2023.

[105] Moez Krichen. Convolutional neural networks : A survey. *Computers*, 12(8) :151, 2023.

[106] Moez Krichen. Deep reinforcement learning. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–7. IEEE, 2023.

[107] Moez Krichen. Generative adversarial networks. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–7. IEEE, 2023.

[108] Moez Krichen. Renforcer la sécurité des contrats intelligents grâce à la puissance de l'intelligence artificielle. 2023.

[109] Moez Krichen. Strengthening the security of smart contracts through the power of artificial intelligence. *Computers*, 12(5) :107, 2023.

[110] Moez Krichen, Alaeddine Mihoub, Mohammed Y Alzahrani, Wilfried Yves Hamilton Adoni, and Tarik Nahhal. Are formal methods applicable to machine learning and artificial intelligence ? In *2022 2nd International Conference of Smart Systems and Emerging Technologies (SMARTTECH)*, pages 48–53. IEEE, 2022.

[111] Sanjay Krishnan, Michael J Franklin, Ken Goldberg, Jiannan Wang, and Eugene Wu. Activeclean : An interactive data cleaning framework for modern machine learning. In *Proceedings of the 2016 International Conference on Management of Data*, pages 2117–2120, 2016.

[112] Jacek Krupski, Waldemar Graniszewski, and Marcin Iwanowski. Data transformation schemes for cnn-based network traffic analysis : A survey. *Electronics*, 10(16) :2042, 2021.

[113] Woosuk Kwon, Sehoon Kim, Michael W Mahoney, Joseph Hassoun, Kurt Keutzer, and Amir Gholami. A fast post-training pruning framework for transformers. *Advances in Neural Information Processing Systems*, 35 :24101–24116, 2022.

[114] Dhara J Ladani and Nikita P Desai. Stopword identification and removal techniques on tc and ir applications : A survey. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 466–472. IEEE, 2020.

[115] Ahmed Lasisi and Nii Attoh-Okine. Principal components analysis and track

quality index : A machine learning approach. *Transportation Research Part C : Emerging Technologies*, 91 :230–248, 2018.

[116] Ivan Lazarevich, Alexander Kozlov, and Nikita Malinin. Post-training deep neural network pruning via layer-wise calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 798–805, 2021.

[117] Guiying Li, Peng Yang, Chao Qian, Richang Hong, and Ke Tang. Stage-wise magnitude-based pruning for recurrent neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[118] Hao Li, Asim Kadav, Erik Kruus, and Cristian Ungureanu. Malt : distributed data-parallelism for existing ml applications. In *Proceedings of the tenth european conference on computer systems*, pages 1–16, 2015.

[119] Peilong Li, Yan Luo, Ning Zhang, and Yu Cao. Heterospark : A heterogeneous cpu/gpu spark platform for machine learning algorithms. In *2015 IEEE International Conference on Networking, Architecture and Storage (NAS)*, pages 347–348. IEEE, 2015.

[120] Yifeng Li and Alioune Ngom. Data integration in machine learning. In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1665–1671. IEEE, 2015.

[121] Qi Liu, Tao Liu, Zihao Liu, Yanzhi Wang, Yier Jin, and Wujie Wen. Security analysis and enhancement of model compressed deep learning systems under adversarial attacks. In *2018 23rd Asia and South Pacific Design Automation Conference (ASP-DAC)*, pages 721–726. IEEE, 2018.

[122] Yuqiao Liu, Yanan Sun, Bing Xue, Mengjie Zhang, Gary G Yen, and Kay Chen Tan. A survey on evolutionary neural architecture search. *IEEE transactions on neural networks and learning systems*, 2021.

[123] Darius Malysiak and Uwe Handmann. An efficient framework for distributed computing in heterogeneous beowulf clusters and cluster-management. In *2014 IEEE 15th International Symposium on Computational Intelligence and Informatics (CINTI)*, pages 169–178. IEEE, 2014.

[124] Seifeddine Mechti, Moez Krichen, Dhouha Ben Noureddine, and Lamia H Belguith. A decision system for computational authors profiling : From machine learning to deep learning. *Concurrency and Computation : Practice and Experience*, 34(7) :e5985, 2022.

[125] Hefeng Meng, Zhiqiang Lin, Fan Yang, Yonghui Xu, and Lizhen Cui. Knowledge distillation in medical data mining : a survey. In *5th International Conference on Crowd Science and Engineering*, pages 175–182, 2021.

[126] Gaurav Menghani. Efficient deep learning : A survey on making deep learning models smaller, faster, and better. *ACM Computing Surveys*, 55(12) :1–37, 2023.

[127] Saeed Mian Qaisar, Nehal Alyamani, Asad Waqar, and Moez Krichen. Machine learning with adaptive rate processing for power quality disturbances identification. *SN Computer Science*, 3 :1–6, 2022.

[128] Saeed Mian Qaisar, Dalila Say, Salah Zidi, and Krichen Moez. Automated cate-

gorization of multiclass welding defects using the x-ray image augmentation and convolutional neural network. 2023.

[129] Alaeddine Mihoub, Moez Krichen, Mohannad Alswailim, Sami Mahfoudhi, and Riadh Bel Hadj Salah. Road scanner : A road state scanning approach based on machine learning techniques. *Applied Sciences*, 13(2) :683, 2023.

[130] Alaeddine Mihoub, Hosni Snoun, Moez Krichen, Riadh Bel Hadj Salah, and Montassar Kahia. Predicting covid-19 spread level using socio-economic indicators and machine learning techniques. In *2020 first international conference of smart systems and emerging technologies (SMARTTECH)*, pages 128–133. IEEE, 2020.

[131] Madison Milne-Ives, Caroline de Cock, Ernest Lim, Melissa Harper Shehadeh, Nick de Pennington, Guy Mole, Eduardo Normando, and Edward Meinert. The effectiveness of artificial intelligence conversational agents in health care : systematic review. *Journal of medical Internet research*, 22(10) :e20346, 2020.

[132] Seyedali Mirjalili. Evolutionary algorithms and neural networks. In *Studies in computational intelligence*, volume 780. Springer, 2019.

[133] Ricardo Pio Monti, Sina Tootoonian, and Robin Cao. Avoiding degradation in deep feed-forward networks by phasing out skip-connections. In *Artificial Neural Networks and Machine Learning–ICANN 2018 : 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*, pages 447–456. Springer, 2018.

[134] Damin Moon, JaeKoo Lee, and MyungKeun Yoon. Compact feature hashing for machine learning based malware detection. *ICT Express*, 8(1) :124–129, 2022.

[135] Sergio Moreno-Alvarez, Juan M Haut, Mercedes E Paoletti, and Juan A Rico-Gallego. Heterogeneous model parallelism for deep neural networks. *Neurocomputing*, 441 :1–12, 2021.

[136] Sayed SR Moustafa, Mohamed S Abdalzaher, Farhan Khan, Mohamed Metwaly, Eslam A Elawadi, and Nassir S Al-Arifi. A quantitative site-specific classification approach based on affinity propagation clustering. *IEEE Access*, 9 :155297–155313, 2021.

[137] Akram Mustafa and Mostafa Rahimi Azghadi. Automated machine learning for healthcare and clinical notes analysis. *Computers*, 10(2) :24, 2021.

[138] Pierre Stanislas Birame Ndong, Wilfried Yves Hamilton Adoni, Tarik Nahhal, Charles Kimpolo, Moez Krichen, Abdeltif EL Byed, Ismail Assayad, and Franck Kalala Mutombo. A face-mask detection system based on deep learning convolutional neural networks. In *Advances on Smart and Soft Computing : Proceedings of ICACIn 2021*, pages 273–283. Springer Singapore Singapore, 2021.

[139] Dhouha Ben Noureddine, Moez Krichen, Seifeddine Mechti, Tarik Nahhal, and Wilfried Yves Hamilton Adoni. An agent-based architecture using deep reinforcement learning for the intelligent internet of things applications. In *Advances on Smart and Soft Computing : Proceedings of ICACIn 2020*, pages 273–283. Springer Singapore, 2021.

[140] Kazuki Osawa, Akira Sekiya, Hiroki Naganuma, and Rio Yokota. Accelerating

matrix multiplication in deep learning by using low-rank approximation. In *2017 International Conference on High Performance Computing & Simulation (HPCS)*, pages 186–192. IEEE, 2017.

[141] Dae Young Park, Moon-Hyun Cha, Daesin Kim, Bohyung Han, et al. Learning student-friendly teacher networks for knowledge distillation. *Advances in neural information processing systems*, 34 :13292–13303, 2021.

[142] Jay H Park, Gyeongchan Yun, M Yi Chang, Nguyen T Nguyen, Seungmin Lee, Jaesik Choi, Sam H Noh, and Young-ri Choi. {HetPipe} : Enabling large {DNN} training on (whimpy) heterogeneous {GPU} clusters through integration of pipelined model parallelism and data parallelism. In *2020 USENIX Annual Technical Conference (USENIX ATC 20)*, pages 307–321, 2020.

[143] Jongse Park, Hardik Sharma, Divya Mahajan, Joon Kyung Kim, Preston Olds, and Hadi Esmaeilzadeh. Scale-out acceleration for machine learning. In *Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture*, pages 367–381, 2017.

[144] Ju-Won Park and Jaegyoon Hahm. Container-based cluster management platform for distributed computing. In *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA)*, page 34. The Steering Committee of The World Congress in Computer Science, Computer ..., 2015.

[145] Gaurav Parmar, Rimi Gupta, Tejas Bhatt, GJ Sahani, Brijeshkumar Y Panchal, and Hiren Patel. A review on data balancing techniques and machine learning methods. In *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pages 1004–1008. IEEE, 2023.

[146] Loukas Petrou, Kypros M Kossifos, Marco A Antoniades, and Julius Georgiou. The first family of application-specific integrated circuits for programmable and reconfigurable metasurfaces. *Scientific reports*, 12(1) :5826, 2022.

[147] Frank Phillipson. Quantum machine learning : Benefits and practical examples. In *QANSWER*, pages 51–56, 2020.

[148] Rio Pramana, Jonathan Jansen Subroto, Alexander Agung Santoso Gunawan, et al. Systematic literature review of stemming and lemmatization performance for sentence similarity. In *2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA)*, pages 1–6. IEEE, 2022.

[149] Saeed Mian Qaisar, Alaeddine Mihoub, Moez Krichen, and Humaira Nisar. Multirate processing with selective subbands and machine learning for efficient arrhythmia classification. *Sensors*, 21(4) :1511, 2021.

[150] Qing Qin, Jie Ren, Jialong Yu, Hai Wang, Ling Gao, Jie Zheng, Yansong Feng, Jianbin Fang, and Zheng Wang. To compress, or not to compress : Characterizing deep learning model compression for embedded inference. In *2018 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom)*, pages 729–736. IEEE, 2018.

[151] Azizur Rahman. Statistics-based data preprocessing methods and machine learning algorithms for big data analysis. *International Journal of Artificial Intelligence*, 17(2) :44–65, 2019.

[152] Abigail Rai and Samarjeet Borah. Study of various methods for tokenization. In *Applications of Internet of Things : Proceedings of ICCCIOT 2020*, pages 193–200. Springer, 2021.

[153] Shalli Rani, Ali Kashif Bashir, Moez Krichen, Abdulaziz Alshammari, et al. A low-rank learning based multi-label security solution for industry 5.0 consumers using machine learning classifiers. *IEEE Transactions on Consumer Electronics*, 2023.

[154] Elaheh Rashedi and Abdolreza Mirzaei. A hierarchical clusterer ensemble method based on boosting theory. *Knowledge-Based Systems*, 45 :83–93, 2013.

[155] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. A comprehensive survey of neural architecture search : Challenges and solutions. *ACM Computing Surveys (CSUR)*, 54(4) :1–34, 2021.

[156] Shashidhar Rudregowda, Sudarshan Patil Kulkarni, Gururaj HL, Vinayakumar Ravi, and Moez Krichen. Visual speech recognition for kannada language using vgg16 convolutional neural network. In *Acoustics*, volume 5, pages 343–353. MDPI, 2023.

[157] Dalila Say, Salah Zidi, Saeed Mian Qaisar, and Moez Krichen. Automated categorization of multiclass welding defects using the x-ray image augmentation and convolutional neural network. *Sensors*, 23(14) :6422, 2023.

[158] Maria Schuld and Francesco Petruccione. *Machine learning with quantum computers*. Springer, 2021.

[159] Maria Schuld, Ilya Sinayskiy, and Francesco Petruccione. An introduction to quantum machine learning. *Contemporary Physics*, 56(2) :172–185, 2015.

[160] Terrence J Sejnowski. The unreasonable effectiveness of deep learning in artificial intelligence. *Proceedings of the National Academy of Sciences*, 117(48) :30033–30038, 2020.

[161] Amir Sepehr, Oriol Gomis-Bellmunt, and Edris Pouresmaeil. Employing machine learning for enhancing transient stability of power synchronization control during fault conditions in weak grids. *IEEE Transactions on Smart Grid*, 13(3) :2121–2131, 2022.

[162] Souhir Sghaier, Moez Krichen, Abir Othman Elfaki, and Qasem Abu Al-Haija. Efficient machine-learning based 3d face identification system under large pose variation. In *International Conference on Computational Collective Intelligence*, pages 273–285. Springer International Publishing Cham, 2022.

[163] Christopher J Shallue, Jaehoon Lee, Joseph Antognini, Jascha Sohl-Dickstein, Roy Frostig, and George E Dahl. Measuring the effects of data parallelism on neural network training. *arXiv preprint arXiv :1811.03600*, 2018.

[164] R Shashidhar, S Patilkulkarni, Vinayakumar Ravi, HL Gururaj, and Moez Krichen. Audiovisual speech recognition based on a deep convolutional neural net-

work. *Data Science and Management*, 2023.

[165] Nicholas D Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E Papalexakis, and Christos Faloutsos. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on signal processing*, 65(13) :3551–3582, 2017.

[166] Dalwinder Singh and Birmohan Singh. Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, 97 :105524, 2020.

[167] Jasmeet Singh and Vishal Gupta. Text stemming : Approaches, applications, and challenges. *ACM Computing Surveys (CSUR)*, 49(3) :1–46, 2016.

[168] Nilam Nur Amir Sjarif, Nurulhuda Firdaus Mohd Azmi, Suriayati Chuprat, Haslina Md Sarkan, Yazriwati Yahya, and Suriani Mohd Sam. Sms spam message detection using term frequency-inverse document frequency and random forest algorithm. *Procedia Computer Science*, 161 :509–515, 2019.

[169] Hyeongseok Son and Seungyong Lee. Fast non-blind deconvolution via regularized residual networks with long/short skip-connections. In *2017 IEEE International Conference on Computational Photography (ICCP)*, pages 1–10. IEEE, 2017.

[170] Wiem Souai, Alaeddine Mihoub, Mounira Tarhouni, Salah Zidi, Moez Krichen, and Sami Mahfoudhi. Predicting at-risk students using the deep learning blstm approach. In *2022 2nd International Conference of Smart Systems and Emerging Technologies (SMARTTECH)*, pages 32–37. IEEE, 2022.

[171] Valery V Starovoitov and Yu I Golub. Data normalization in machine learning. In *Informatics*, volume 18, pages 83–96, 2021.

[172] Andrew Stolman, Caleb Levy, C Seshadhri, and Aneesh Sharma. Classic graph structural features outperform factorization-based graph embedding methods on community labeling. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pages 388–396. SIAM, 2022.

[173] Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel S Emer. Efficient processing of deep neural networks : A tutorial and survey. *Proceedings of the IEEE*, 105(12) :2295–2329, 2017.

[174] Sijun Tan, Brian Knott, Yuan Tian, and David J Wu. Cryptgpu : Fast privacy-preserving machine learning on the gpu. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 1021–1038. IEEE, 2021.

[175] Ahmad S Tarawneh, Ahmad BA Hassanat, Khalid Almohammadi, Dmitry Chetverikov, and Colin Bellinger. Smotefuna : Synthetic minority over-sampling technique based on furthest neighbour algorithm. *IEEE Access*, 8 :59069–59082, 2020.

[176] Cu Thi Thu Thuy, Kim Anh Tran, Cu Nguyen Giap, et al. Optimize the combination of categorical variable encoding and deep learning technique for the problem of prediction of vietnamese student academic performance. *International Journal of Advanced Computer Science and Applications*, 11(11), 2020.

[177] George Trigeorgis, Konstantinos Bousmalis, Stefanos Zafeiriou, and Björn W Schuller. A deep matrix factorization method for learning attribute repre-

sentations. *IEEE transactions on pattern analysis and machine intelligence*, 39(3) :417–429, 2016.

[178] Sujatha R Upadhyaya. Parallel approaches to machine learning—a comprehensive survey. *Journal of Parallel and Distributed Computing*, 73(3) :284–292, 2013.

[179] Lorenzo Vaccaro, Giuseppe Sansonetti, and Alessandro Micarelli. An empirical review of automated machine learning. *Computers*, 10(1) :11, 2021.

[180] João Vitorino, Isabel Praça, and Eva Maia. Towards adversarial realism and robust learning for iot intrusion detection and classification. *Annals of Telecommunications*, pages 1–12, 2023.

[181] Chandra Sekhar Vorugunti, Viswanath Pulabaigari, Rama Krishna Sai Subrahmanyam Gorthi, and Prerana Mukherjee. Osvfusenet : online signature verification by feature fusion and depth-wise separable convolution based deep learning. *Neurocomputing*, 409 :157–172, 2020.

[182] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence : A review and new outlooks. *IEEE transactions on pattern analysis and machine intelligence*, 44(6) :3048–3068, 2021.

[183] Shiping Wang, Witold Pedrycz, Qingxin Zhu, and William Zhu. Subspace learning for unsupervised feature selection via matrix factorization. *Pattern Recognition*, 48(1) :10–19, 2015.

[184] Shirui Wang, Wenan Zhou, and Chao Jiang. A survey of word embeddings based on deep learning. *Computing*, 102 :717–740, 2020.

[185] Jonathan Waring, Charlotta Lindvall, and Renato Umeton. Automated machine learning : Review of the state-of-the-art and opportunities for healthcare. *Artificial intelligence in medicine*, 104 :101822, 2020.

[186] Ahmad Samer Wazan and Frédéric Cuppens. Cybersecurity in networking : adaptations, investigation, attacks, and countermeasures. *Annals of Telecommunications*, 78(3-4) :133–134, 2023.

[187] Tongfeng Weng, Xiaolu Chen, Zhuoming Ren, Huijie Yang, Jie Zhang, and Michael Small. Synchronization of machine learning oscillators in complex networks. *Information Sciences*, 630 :74–81, 2023.

[188] Peter Wittek. *Quantum machine learning : what quantum computing means to data mining*. Academic Press, 2014.

[189] Zach Wood-Doughty, Isabel Cachola, and Mark Dredze. Model distillation for faithful explanations of medical code predictions. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 412–425, 2022.

[190] Hang Xu, Chen-Yu Ho, Ahmed M Abdelmoniem, Aritra Dutta, El Houcine Bergou, Konstantinos Karatsenidis, Marco Canini, and Panos Kalnis. Grace : A compressed communication framework for distributed machine learning. In *2021 IEEE 41st international conference on distributed computing systems (ICDCS)*, pages 561–572. IEEE, 2021.

[191] Sheng Xu, Anran Huang, Lei Chen, and Baochang Zhang. Convolutional neural network pruning : A survey. In *2020 39th Chinese Control Conference (CCC)*,

pages 7458–7463. IEEE, 2020.

[192] Jin Yamanaka, Shigesumi Kuwashima, and Takio Kurita. Fast and accurate image super resolution by deep cnn with skip connection and network in network. In *Neural Information Processing : 24th International Conference, ICONIP 2017, Guangzhou, China, November 14-18, 2017, Proceedings, Part II 24*, pages 217–225. Springer, 2017.

[193] Jianquan Yang, Yulan Zhang, Guopu Zhu, and Sam Kwong. A clustering-based framework for improving the performance of jpeg quantization step estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(4) :1661–1672, 2020.

[194] Pengyi Yang, Wei Liu, Bing B Zhou, Sanjay Chawla, and Albert Y Zomaya. Ensemble-based wrapper methods for feature selection and class imbalance learning. In *Advances in Knowledge Discovery and Data Mining : 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part I 17*, pages 544–555. Springer, 2013.

[195] Shuyuan Yang, RuiXia Wu, Min Wang, and Licheng Jiao. Evolutionary clustering based vector quantization and spiht coding for image compression. *Pattern Recognition Letters*, 31(13) :1773–1780, 2010.

[196] Seul-Ki Yeom, Philipp Seegerer, Sebastian Lapuschkin, Alexander Binder, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Pruning by explaining : A novel criterion for deep neural network pruning. *Pattern Recognition*, 115 :107899, 2021.

[197] Zhihang Yuan, Chenhao Xue, Yiqi Chen, Qiang Wu, and Guangyu Sun. Ptq4vit : Post-training quantization for vision transformers with twin uniform quantization. In *European Conference on Computer Vision*, pages 191–207. Springer, 2022.

[198] Caiming Zhang and Yang Lu. Study on artificial intelligence : The state of the art and future prospects. *Journal of Industrial Information Integration*, 23 :100224, 2021.

[199] Jilin Zhang, Hangdi Tu, Yongjian Ren, Jian Wan, Li Zhou, Mingwei Li, and Jue Wang. An adaptive synchronous parallel strategy for distributed machine learning. *IEEE Access*, 6 :19222–19230, 2018.

[200] Xinfeng Zhang, Dianning He, Yue Zheng, Huaibi Huo, Simiao Li, Ruimei Chai, and Ting Liu. Deep learning based analysis of breast cancer using advanced ensemble classifier and linear discriminant analysis. *IEEE access*, 8 :120208–120217, 2020.

[201] Yao Zhang and Qiang Ni. Recent advances in quantum machine learning. *Quantum Engineering*, 2(1) :e34, 2020.

[202] Alice Zheng and Amanda Casari. *Feature engineering for machine learning : principles and techniques for data scientists.* " O'Reilly Media, Inc.", 2018.

[203] Zhi-Hua Zhou. *Ensemble methods : foundations and algorithms.* CRC press, 2012.

[204] Yonghao Zhuang, Lianmin Zheng, Zhuohan Li, Eric Xing, Qirong Ho, Joseph Gonzalez, Ion Stoica, Hao Zhang, and Hexu Zhao. On optimizing the communication of model parallelism. *Proceedings of Machine Learning and Systems*, 5, 2023.

[205] Salah Zidi, Alaeddine Mihoub, Saeed Mian Qaisar, Moez Krichen, and Qasem Abu Al-Haija. Theft detection dataset for benchmarking and machine learning based classification in a smart grid environment. *Journal of King Saud University-Computer and Information Sciences*, 35(1) :13–25, 2023.