Affinement des Données pour l'Excellence en IA: Techniques Avancées de Prétraitement et d'Ingénierie des Caractéristiques

Moez Krichen

Affinement des Données pour l'Excellence en IA: Techniques Avancées de Prétraitement et d'Ingénierie des Caractéristiques

Moez Krichen

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

Résumé. La qualité des données et l'ingéniosité de l'ingénierie des caractéristiques sont cruciales pour le succès des modèles d'intelligence artificielle (IA). Ce document explore des méthodologies sophistiquées de prétraitement des données et d'ingénierie des caractéristiques pour optimiser la performance des algorithmes d'IA. À travers le nettoyage des données, l'intégration, la transformation, l'extraction et la sélection de caractéristiques, ainsi que le traitement spécialisé des données catégorielles et textuelles et l'équilibrage des données, nous présentons des stratégies visant à améliorer la qualité des données. Ces processus sont essentiels pour construire des modèles d'IA robustes, capables de généraliser à partir d'ensembles de données complexes et diversifiés, soulignant l'importance d'une préparation méticuleuse des données dans le cycle de vie du développement de l'IA.

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 (94; 35). 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 (68; 81).

Dans le domaine en évolution rapide de l'intelligence artificielle (IA), la capacité des modèles à interpréter, apprendre et prendre des décisions de manière efficace est intrinsèquement liée à la qualité des données sur lesquelles ils sont entraînés (53; 87; 6; 83; 16; 64; 82; 2; 50; 57; 7). La préparation des données, qui englobe des processus critiques tels que le prétraitement des données et l'ingénierie des caractéristiques, joue un rôle déterminant dans le succès des applications d'IA (74; 52; 65; 66; 47; 98; 8; 3; 15; 70; 32; 10; 63; 71). Ce n'est pas simplement une étape préliminaire, mais un

élément fondamental qui peut définir la trajectoire de la performance d'un modèle (80; 54; 67; 79; 5; 21; 19; 12; 41). Alors que les données brutes sont souvent imparfaites, incomplètes, et dispersées à travers diverses sources, leur transformation en un ensemble structuré et significatif est essentielle pour débloquer le potentiel des algorithmes d'IA (77; 55; 56; 51; 49).

La préparation des données est un domaine complexe qui nécessite une attention méticuleuse à de multiples facettes, allant du nettoyage des données et de l'intégration à des techniques plus sophistiquées telles que la transformation, l'extraction, et la sélection de caractéristiques. Chacune de ces étapes contribue à affiner et à optimiser l'ensemble de données, assurant que les modèles d'IA disposent d'un substrat riche et pertinent à partir duquel apprendre. De plus, le traitement spécialisé des données catégorielles et textuelles ainsi que l'équilibrage des données jouent un rôle crucial dans la réduction des biais et l'amélioration de la capacité des modèles à interpréter correctement la variété et la complexité des informations.

Ce document explore l'impact profond que ces techniques de préparation des données peuvent avoir sur la performance des modèles d'IA. Nous mettons en lumière l'importance d'une approche systématique et réfléchie dans la préparation des données, illustrant comment des stratégies bien conçues de prétraitement et d'ingénierie des caractéristiques peuvent significativement améliorer la robustesse, l'efficacité, et l'exactitude des systèmes d'IA. En démontrant l'interconnexion entre la qualité des données et la performance des modèles, nous soulignons l'importance cruciale d'accorder une attention dédiée à la préparation des données dans le cycle de vie du développement de l'IA, mettant en avant les défis et les opportunités inhérents à cette phase fondamentale.

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 (34; 75; 42; 96; 38; 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.

2 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 (43; 25; 58). 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.

3 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 (26; 36; 62; 30). 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.

4 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 (59; 11; 45). 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 (84; 14; 88). Le codage des variables catégorielles est essentiel pour les exprimer sous un format numérique (20; 91). 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.

5 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 (46; 37). 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) (61; 48), l'analyse discriminante linéaire (LDA) (4; 95; 40), ou le hachage de caractéristiques (69; 61), 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 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 (29; 22; 17; 31). 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) (24; 13), les méthodes d'enveloppe (par exemple, l'élimination récursive de caractéristiques) (18; 93), ou les méthodes d'incorporation (par exemple, la régularisation Lasso) (9; 89; 27), 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.

7 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 (39). 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 (76), où le texte est divisé en mots individuels ou en n-grammes, suivi de techniques telles que le stemming (85), la lemmatisation (73), ou la suppression des mots vides (60) 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) (86) ou des plongements de mots comme Word2Vec ou GloVe (92; 44). Ces techniques de prétraitement permettent d'utiliser des données catégorielles et textuelles dans les modèles d'IA.

8 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 (72). 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) (90; 23), ou le sous-échantillonnage de la classe majoritaire (28), peuvent être employées pour atténuer cette préoccupation. De plus, il est possible d'utiliser des techniques algorithmiques telles que l'apprentissage sensible au coût (33) ou des méthodes d'ensemble (telles que le boosting) (97; 78) 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.

9 Conclusion

Les techniques de prétraitement des données et d'ingénierie des caractéristiques occupent une place centrale dans l'optimisation des modèles d'intelligence artificielle. Comme démontré dans ce document, le processus méticuleux de nettoyage, d'intégration, de transformation, d'extraction et de sélection de caractéristiques, ainsi que le traitement spécialisé des données catégorielles et textuelles et l'équilibrage des données, sont essentiels pour préparer les ensembles de données à une analyse efficace. Ces étapes préliminaires permettent non seulement d'améliorer la performance et la précision des modèles d'IA mais aussi de réduire le risque de surapprentissage et d'augmenter leur capacité à généraliser à partir de données nouvelles ou non vues. En conséquence, l'accent mis sur une préparation rigoureuse des données peut significativement propulser les capacités des systèmes d'IA, soulignant l'importance d'une fondation solide de données pour l'excellence en intelligence artificielle.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

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

[19] 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.

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

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.

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

[27] 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.

[28] 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.

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] Omar Hamdy, Hanan Gaber, Mohamed S Abdalzaher, and Mahmoud Elhadidy. Identifying exposure of urban area to certain seismic hazard using machine learning and gis : A case study of greater cairo. *Sustainability*, 14(17) :10722, 2022.

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

[40] 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.

[41] 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.

[42] 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.

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

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

[45] 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.

[46] Samina Khalid, Tehmina Khalil, and Shamila Nasreen. A survey of feature selec-

tion and feature extraction techniques in machine learning. In *2014 science and information conference*, pages 372–378. IEEE, 2014.

[47] 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.

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

[49] 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.

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

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

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

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

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

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

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

[57] 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.

[58] 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.

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

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] Saeed Mian Qaisar, Dalila Say, Salah Zidi, and Krichen Moez. Automated categorization of multiclass welding defects using the x-ray image augmentation and convolutional neural network. 2023.

[66] 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.

[67] 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.

[68] 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.

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

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

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

[79] 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.

[80] 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.

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

[82] 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.

[83] R Shashidhar, S Patilkulkarni, Vinayakumar Ravi, HL Gururaj, and Moez Krichen. Audiovisual speech recognition based on a deep convolutional neural network. *Data Science and Management*, 2023.

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

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

[86] 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.

[87] 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.

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

[89] 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.

[90] 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.

[91] 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.

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

[93] 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.

[94] 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.

[95] 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.

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

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

[98] 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.