Sabrina Ermshaus

112714

Passau, 08th January 2024

Master Thesis – Prof. Dr. Jelena Mitrovic

A mixed-method evaluation of opportunities of AI in credit card fraud detection in banking

Bibliography

3D Secure. 2019. ‘Why Was 3-D Secure 1.0 Not Successful in Some Countries?’ 2019. https://3dsecure2.com/blog/why-was-3-d-secure-1.0-not-successful-in-some-countries/.

A. Bhowmik, M. Sannigrahi, D. Chowdhury, A. D. Dwivedi, and R. Rao Mukkamala. 2022. ‘DBNex: Deep Belief Network and Explainable AI Based Financial Fraud Detection’. In *2022 IEEE International Conference on Big Data (Big Data)*, 3033–42. https://doi.org/10.1109/BigData55660.2022.10020494.

Achituve, I., S. Kraus, and J. Goldberger. 2019. ‘Interpretable Online Banking Fraud Detection Based On Hierarchical Attention Mechanism’. *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, October, 1–6. https://doi.org/10.1109/MLSP.2019.8918896.

Ada Lovelace Institute. 2022. ‘Expert Explainer: The EU AI Act’. 2022. https://www.adalovelaceinstitute.org/wp-content/uploads/2022/04/Expert-explainer-The-EU-AI-Act-11-April-2022.pdf.

AL-Dosari, Khalifa, Noora Fetais, and Murat Kucukvar. 2022. ‘Artificial Intelligence and Cyber Defense System for Banking Industry: A Qualitative Study of AI Applications and Challenges’. *Cybernetics and Systems*, 1–29.

Alfaiz, N.S., and S.M. Fati. 2022. ‘Enhanced Credit Card Fraud Detection Model Using Machine Learning’. *Electronics (Switzerland)* 11 (4). https://doi.org/10.3390/electronics11040662.

Alvarez-Melis, David, and Tommi S Jaakkola. 2018. ‘On the Robustness of Interpretability Methods’. *arXiv Preprint arXiv:1806.08049*.

Aschi, Massimiliano, Susanna Bonura, Nicola Masi, Domenico Messina, and Davide Profeta. 2022. ‘Cybersecurity and Fraud Detection in Financial Transactions’. In *Big Data and Artificial Intelligence in Digital Finance: Increasing Personalization and Trust in Digital Finance Using Big Data and AI*, edited by John Soldatos and Dimosthenis Kyriazis, 269–78. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-94590-9_15.

Ashenden, Stephanie Kay, Aleksandra Bartosik, Paul-Michael Agapow, and Elizaveta Semenova. 2021. ‘Chapter 2 - Introduction to Artificial Intelligence and Machine Learning’. In *The Era of Artificial Intelligence, Machine Learning, and Data Science in the Pharmaceutical Industry*, edited by Stephanie Kay Ashenden, 15–26. Academic Press. https://doi.org/10.1016/B978-0-12-820045-2.00003-9.

Bahnsen, Alejandro Correa, Djamila Aouada, Aleksandar Stojanovic, and Björn Ottersten. 2016. ‘Feature Engineering Strategies for Credit Card Fraud Detection’. *Expert Systems with Applications* 51: 134–42.

Bao, Yang, Gilles Hilary, and Bin Ke. 2022. ‘Artificial Intelligence and Fraud Detection’. In *Innovative Technology at the Interface of Finance and Operations: Volume I*, edited by Volodymyr Babich, John R. Birge, and Gilles Hilary, 223–47. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-75729-8_8.

Bhat, M. Q., S. A. Alex, S. Nanda, and S. Goutham. 2022. ‘Qualitative Analysis of Anomaly Detection in Time Series’. *2022 4th International Conference on Circuits, Control, Communication and Computing (I4C)*, December, 250–53. https://doi.org/10.1109/I4C57141.2022.10057732.

Bhat, P. Ishwara. 2020. ‘Doctrinal Legal Research as a Means of Synthesizing Facts, Thoughts, and Legal Principles’. In *Idea and Methods of Legal Research*, edited by P. Ishwara Bhat, 143–68. Oxford University Press. https://doi.org/10.1093/oso/9780199493098.003.0005.

Biswas, J., A. Ahmed Mridha, M. Sakib Hossain, A. Subhra Trisha, M. Sabbir Ahmed, and M. Iqbal Hossain. 2023. ‘Interpretable Credit Card Fraud Detection Using Machine Learning Leveraging

SHAP'. *2023 IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT)*, July, 1206–11. https://doi.org/10.1109/ICEICT57916.2023.10245439.

Bockel-Rickermann, Christopher, Tim Verdonck, and Wouter Verbeke. 2023. 'Fraud Analytics::A Decade of Research'.

Bomhard, David, and Marieke Merkle. 2021. 'Der Aktuelle Kommissionsentwurf Und Praktische Auswirkungen'. *RDi* 2021: 276–83.

Boulieris, P., J. Pavlopoulos, A. Xenos, and V. Vassalos. 2023a. 'Fraud Detection with Natural Language Processing'. *Machine Learning*. https://doi.org/10.1007/s10994-023-06354-5.

———. 2023b. 'Fraud Detection with Natural Language Processing'. *Machine Learning*. https://doi.org/10.1007/s10994-023-06354-5.

Bowden, Mark G., Chitralakshmi K. Balasubramanian, Andrea L. Behrman, and Steven A. Kautz. 2008. 'Validation of a Speed-Based Classification System Using Quantitative Measures of Walking Performance Post-Stroke'. *Neurorehabilitation and Neural Repair* 22 (6): 672–75. https://doi.org/10.1177/1545968308318837.

Brownlee, Jason. 2020. 'Random Oversampling and Undersampling for Imbalanced Classification'. MachineLearningMastery.Com. 14 January 2020. https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/.

Bruxvoort, X. van, and M. van Keulen. 2021. 'Framework for Assessing Ethical Aspects of Algorithms and Their Encompassing Socio-Technical System'. *Applied Sciences (Switzerland)* 11 (23). https://doi.org/10.3390/app112311187.

Chen, Joy Iong-Zong, and Kong-Long Lai. 2021. 'Deep Convolution Neural Network Model for Credit-Card Fraud Detection and Alert'. *Journal of Artificial Intelligence and Capsule Networks* 3 (2): 101–12. https://doi.org/10.36548/jaicn.2021.2.003.

Cherkaoui, Rabab, and El Mokhtar En-Naimi. 2023. 'A Comparison of Machine Learning Algorithms for Credit Card Fraud Detection', NISS '23, . https://doi.org/10.1145/3607720.3607759.

Cirqueira, D., M. Helfert, and M. Bezbradica. 2021a. 'Towards Design Principles for User-Centric Explainable AI in Fraud Detection'. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12797: 21–40. https://doi.org/10.1007/978-3-030-77772-2_2.

———. 2021b. 'Towards Design Principles for User-Centric Explainable AI in Fraud Detection'. In , 12797 LNAI:21–40. https://doi.org/10.1007/978-3-030-77772-2_2.

Computerbild. 2023. 'Datenleck bei Postbank und Deutsche Bank'. computerbild.de. 10 July 2023. https://www.computerbild.de/artikel/cb-News-Sicherheit-Datenleck-bei-Postbank-und-Deutsche-Bank-36131331.html.

Dal Pozzolo, Andrea, Giacomo Boracchi, Olivier Caelen, Cesare Alippi, and Gianluca Bontempi. 2015. 'Credit Card Fraud Detection and Concept-Drift Adaptation with Delayed Supervised Information'. In , 1–8. IEEE.

Das, Shibsankar. 2022. 'Best Practices for Dealing With Concept Drift'. Neptune.Ai. 21 July 2022. https://neptune.ai/blog/concept-drift-best-practices.

Datacamp. 2023. 'What Is Online Machine Learning?' 2023. https://www.datacamp.com/blog/what-is-online-machine-learning.

Datatilsynet. 2023. 'How to Succeed with Transparency'. Datatilsynet. 2023. https://www.datatilsynet.no/en/regulations-and-tools/sandbox-for-artificial-intelligence/reports/how-to-succeed-with-transparency/.

Dwork, Cynthia. 2011. ‘A Firm Foundation for Private Data Analysis’. *Communications of the ACM* 54 (1): 86–95. https://doi.org/10.1145/1866739.1866758.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. ‘Calibrating Noise to Sensitivity in Private Data Analysis’. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, 265–84. Springer.

EDPB. 2019. ‘Leitlinien 2/2019 Für Die Verarbeitung Personenbezogener Daten Gemäß Artikel 6 Abs. 1 Lit. b DSGVO Im Zusammenhang Mit Der Erbringung von Online-Diensten Für Betroffene Personen’. 2019. https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines-art_6-1-b-adopted_after_public_consultation_de_0.pdf.

Eurobarometer. 2020. ‘Europeans’ Attitudes towards Cyber Security (Cybercrime) - Factsheets Germany’. 2020. https://europa.eu/eurobarometer/api/deliverable/download/file?deliverableId=71881.

European Central Bank. 2015. ‘Fourth Report on Card Fraud’. https://www.ecb.europa.eu/pub/pdf/other/4th_card_fraud_report.en.pdf.

European Payments Council. n.d. ‘Payments Threats and Fraud Trends Report 2023’. Accessed 29 December 2023. https://www.europeanpaymentscouncil.eu/sites/default/files/kb/file/2023-12/EPC181-23%20v1.0%202023%20Payments%20Threats%20and%20Fraud%20Trends%20Report_1.pdf.

Feng, Jiaming, Zheng Huang, Jie Guo, and Weidong Qiu. 2021. ‘UNSUPERVISED ANOMALY DETECTION FOR TIME SERIES WITH OUTLIER EXPOSURE’, SSDBM ’21, , 1–12. https://doi.org/10.1145/3468791.3468793.

FIS Global. 2019. ‘What Is Credit Card Processing? - Insights | Worldpay from FIS’. 2019. https://www.fisglobal.com/en/insights/merchant-solutions-worldpay/article/what-is-credit-card-processing.

Franck. 2022. ‘DS-GVO Art. 13 Informationspflicht Bei Erhebung von Personenbezogenen Daten Bei Der Betroffenen Person’. In *Datenschutz-Grundverordnung - Bundesdatenschutzgesetz*, edited by Gola and Heckmann, 3rd ed. Munich: Beck.

Fraunhofer-Institut. 2021. ‘Effiziente Betrugserkennung durch Maschinelles Lernen’. https://www.iais.fraunhofer.de/content/dam/iais/gf/bda/Downloads/Fraunhofer_IAIS_Whitepaper_Fraud.pdf.

Fritz-Morgenthal, S., B. Hein, and J. Papenbrock. 2022. ‘Financial Risk Management and Explainable, Trustworthy, Responsible AI’. *Frontiers in Artificial Intelligence* 5. https://doi.org/10.3389/frai.2022.779799.

Gabudeanu, L., I. Brici, C. Mare, I.C. Mihai, and M.C. Scheau. 2021. ‘Privacy Intrusiveness in Financial-Banking Fraud Detection’. *Risks* 9 (6). https://doi.org/10.3390/risks9060104.

Gaub, Daniela, and Mathias Kadler. 2022. ‘Der Einsatz von Künstlicher Intelligenz Im Forderungsmanagement’. *Rethinking Law* 02.2022: 57–63.

Gelder, Koen van. 2023. ‘Leading Reasons for Abandonment during Checkout in the U.S. 2023’. Statista. 9 October 2023. https://www.statista.com/statistics/1228452/reasons-for-abandonments-during-checkout-united-states/.

Gianini, Gabriele, Leopold Ghemmogne Fossi, Corrado Mio, Olivier Caelen, Lionel Brunie, and Ernesto Damiani. 2020. ‘Managing a Pool of Rules for Credit Card Fraud Detection by a Game Theory Based Approach’. *Future Generation Computer Systems* 102 (January): 549–61. https://doi.org/10.1016/j.future.2019.08.028.

Gianotti, Enrico, and Eduardo Damião da Silva. 2021. ‘Strategic Management of Credit Card Fraud: Stakeholder Mapping of a Card Issuer’. *Journal of Financial Crime* 28 (1): 156–69.

Hanae, A., B. Abdellah, E. Saida, and G. Youssef. 2023. ‘End-to-End Real-Time Architecture for Fraud Detection in Online Digital Transactions’. *International Journal of Advanced Computer Science and Applications* 14 (6): 749–57. https://doi.org/10.14569/IJACSA.2023.0140680.

Hine, Emmie, Claudio Novelli, Mariarosaria Taddeo, and Luciano Floridi. 2023. ‘Supporting Trustworthy AI Through Machine Unlearning’. *Available at SSRN*.

Hoppe, F, J Hohmann, M Knoll, C Kubik, and P Groche. 2019. ‘Feature-Based Supervision of Shear Cutting Processes on the Basis of Force Measurements: Evaluation of Feature Engineering and Feature Extraction’. *Procedia Manufacturing* 34: 847–56.

Hsin, Y.-Y., T.-S. Dai, Y.-W. Ti, M.-C. Huang, T.-H. Chiang, and L.-C. Liu. 2022. ‘Feature Engineering and Resampling Strategies for Fund Transfer Fraud with Limited Transaction Data and a Time-Inhomogeneous Modi Operandi’. *IEEE Access* 10: 86101–16. https://doi.org/10.1109/ACCESS.2022.3199425.

Huang, Deshan, Yu Lin, Zhaoxing Weng, and Jiajie Xiong. 2021. ‘Decision Analysis and Prediction Based on Credit Card Fraud Data’, ESCC ’21, , 20–26. https://doi.org/10.1145/3478301.3478305.

I. Achituve, S. Kraus, and J. Goldberger. 2019. ‘Interpretable Online Banking Fraud Detection Based On Hierarchical Attention Mechanism’. In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6. https://doi.org/10.1109/MLSP.2019.8918896.

Ji, Yingchao. 2021. ‘Explainable AI Methods for Credit Card Fraud Detection: Evaluation of LIME and SHAP through a User Study’.

Jing, Xiao-Yuan, Xinyu Zhang, Xiaoke Zhu, Fei Wu, Xinge You, Yang Gao, Shiguang Shan, and Jing-Yu Yang. 2019. ‘Multiset Feature Learning for Highly Imbalanced Data Classification’. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43 (1): 139–56.

Jurgovsky, Johannes, Michael Granitzer, Konstantin Ziegler, Sylvie Calabretto, Pierre-Edouard Portier, Liyun He-Guelton, and Olivier Caelen. 2018. ‘Sequence Classification for Credit-Card Fraud Detection’. *Expert Systems with Applications* 100: 234–45.

Kaupa, Clemens. 2016. ‘Methods and Approaches to Research Legal Questions in Your Thesis’. Legal Methodology. 14 December 2016. https://medium.com/legal-methodology/6-methods-and-approaches-to-research-legal-questions-in-your-thesis-8caada918bb5.

Leevy, J.L., J. Hancock, T.M. Khoshgoftaar, and A. Abdollah Zadeh. 2023. ‘Investigating the Effectiveness of One-Class and Binary Classification for Fraud Detection’. *Journal of Big Data* 10 (1). https://doi.org/10.1186/s40537-023-00825-1.

LexisNexis. 2022. ‘True Cost of Financial Crime Compliance Global Study 2022’. LexisNexis Risk Solutions. 2022. https://risk.lexisnexis.com/about-us/press-room/press-release/20221116-study-finds-fraud-costs.

———. 2023. ‘True Cost of Financial Crime Compliance Global Study 2023’. https://risk.lexisnexis.com/global/en/insights-resources/research/true-cost-of-financial-crime-compliance-study-global-report.

Liang, D., J. Wang, X. Gao, J. Wang, X. Zhao, and L. Wang. 2022. ‘Self-Supervised Pretraining Isolated Forest for Outlier Detection’. *2022 International Conference on Big Data, Information and Computer Network (BDICN)*, January, 306–10. https://doi.org/10.1109/BDICN55575.2022.00065.

Lin, T.-H., and J.-R. Jiang. 2021. ‘Credit Card Fraud Detection with Autoencoder and Probabilistic Random Forest’. *Mathematics* 9 (21). https://doi.org/10.3390/math9212683.

Linhart, Andrea. 2023. *Information Aus Der Blackbox. Zum Verhältnis von Transparenz Und Geheimnisschutz Am Beispiel Künstlicher Neuronaler Netze*. Schriften Zum 13.

Liu, Xiyang, Weihao Kong, Prateek Jain, and Sewoong Oh. 2022. 'DP-PCA: Statistically Optimal and Differentially Private PCA'. *Advances in Neural Information Processing Systems* 35: 29929–43.

Lucas, Yvan, and Johannes Jurgovsky. 2020. *Credit Card Fraud Detection Using Machine Learning: A Survey*.

Makki, Sara, Zainab Assaghir, Yehia Taher, Rafiqul Haque, Mohand-Said Hacid, and Hassan Zeineddine. 2019. 'An Experimental Study with Imbalanced Classification Approaches for Credit Card Fraud Detection'. *IEEE Access* 7: 93010–22.

Megdad, Mosa MM, Samy S Abu-Naser, and Bassem S Abu-Nasser. 2022. 'Fraudulent Financial Transactions Detection Using Machine Learning'. *International Journal of Academic Information Systems Research* 6 (3): 30–39.

Merriam Webster. 2023. 'INTERACTION'. 12 December 2023. https://www.merriam-webster.com/dictionary/interaction.

Miao, Jiaju, and Wei Zhu. 2022. 'Precision–Recall Curve (PRC) Classification Trees'. *Evolutionary Intelligence* 15 (3): 1545–69. https://doi.org/10.1007/s12065-021-00565-2.

Mienye, I.D., and Y. Sun. 2023a. 'A Deep Learning Ensemble With Data Resampling for Credit Card Fraud Detection'. *IEEE Access* 11: 30628–38. https://doi.org/10.1109/ACCESS.2023.3262020.

———. 2023b. 'A Machine Learning Method with Hybrid Feature Selection for Improved Credit Card Fraud Detection'. *Applied Sciences (Switzerland)* 13 (12). https://doi.org/10.3390/app13127254.

Mill, E., W. Garn, N. Ryman-Tubb, and C. Turner. 2023. 'Opportunities in Real Time Fraud Detection: An Explainable Artificial Intelligence (XAI) Research Agenda'. *International Journal of Advanced Computer Science and Applications* 14 (5): 1172–86. https://doi.org/10.14569/IJACSA.2023.01405121.

Minku, Leandro L, Allan P White, and Xin Yao. 2009. 'The Impact of Diversity on Online Ensemble Learning in the Presence of Concept Drift'. *IEEE Transactions on Knowledge and Data Engineering* 22 (5): 730–42.

Mohseni, Sina, Niloofar Zarei, and Eric D Ragan. 2021. 'A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems'. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 11 (3–4): 1–45.

Molnar, Christoph. 2020. *Model-Agnostic Interpretable Machine Learning*. Munich: LeanPub.

Mozaffari, Sadaf. 2023. 'The Importance of Feature Selection and Feature Importance in Machine Learning | LinkedIn'. 2023. https://www.linkedin.com/pulse/importance-feature-selection-machine-learning-sadaf-mozaffari/.

Mueller.legal. 2023. 'Kreditkartenbetrug'. 2023. https://mueller.legal/de/anwalt-kreditkartenbetrug.

N. Boutaher, A. Elomri, N. Abghour, K. Moussaid, and M. Rida. 2020. 'A Review of Credit Card Fraud Detection Using Machine Learning Techniques'. In *2020 5th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech)*, 1–5. https://doi.org/10.1109/CloudTech49835.2020.9365916.

Narayan, V., and S. Ganapathisamy. 2022. 'Hybrid Sampling and Similarity Attention Layer in Bidirectional Long Short Term Memory in Credit Card Fraud Detection'. *International Journal of Intelligent Engineering and Systems* 15 (6): 35–44. https://doi.org/10.22266/ijies2022.1231.04.

Nesvijevskaia, A., S. Ouillade, P. Guilmin, and J.-D. Zucker. 2021. 'The Accuracy versus Interpretability Trade-off in Fraud Detection Model'. *Data and Policy* 3 (7). https://doi.org/10.1017/dap.2021.3.

Nguyen, Thanh Tam, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. 2022. 'A Survey of Machine Unlearning'. *arXiv Preprint arXiv:2209.02299*.

Patel, Harshil. 2022. ‘ML from Research to Production - Challenges, Best Practices and Tools [Guide]’. Neptune.Ai. 22 July 2022. https://neptune.ai/blog/ml-from-research-to-production.

Pedregosa, F, and E Triantafillou. 2023. ‘Announcing the First Machine Unlearning Challenge’. 29 June 2023. https://blog.research.google/2023/06/announcing-first-machine-unlearning.html.

Pratt, Kevin B, and Gleb Tschapek. 2003. ‘Visualizing Concept Drift’. In , 735–40.

PromptCloud. 2016. ‘How to Use Web Crawling to Detect Fraud’. 14 November 2016. https://www.promptcloud.com/blog/web-crawling-to-detect-fraud/.

Psychoula, I., A. Gutmann, P. Mainali, S.H. Lee, P. Dunphy, and F. Petitcolas. 2021. ‘Explainable Machine Learning for Fraud Detection’. *Computer* 54 (10): 49–59. https://doi.org/10.1109/MC.2021.3081249.

Raj, A. T., J. Shobana, V. K. Nassa, S. Painuly, M. Savaram, and M. Sridevi. 2023. ‘Enhancing Security for Online Transactions through Supervised Machine Learning and Block Chain Technology in Credit Card Fraud Detection’. *2023 7th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, October, 241–48. https://doi.org/10.1109/I-SMAC58438.2023.10290462.

Ranjbaran, G., D. Reforgiato Recupero, G. Lombardo, and S. Consoli. 2023. ‘Leveraging Augmentation Techniques for Tasks with Unbalancedness within the Financial Domain: A Two-Level Ensemble Approach’. *EPJ Data Science* 12 (1). https://doi.org/10.1140/epjds/s13688-023-00402-9.

Raval, J., P. Bhattacharya, N.K. Jadav, S. Tanwar, G. Sharma, P.N. Bokoro, M. Elmorsy, A. Tolba, and M.S. Raboaca. 2023. ‘RaKShA: A Trusted Explainable LSTM Model to Classify Fraud Patterns on Credit Card Transactions’. *Mathematics* 11 (8). https://doi.org/10.3390/math11081901.

Ryman-Tubb, Nick F, Paul Krause, and Wolfgang Garn. 2018. ‘How Artificial Intelligence and Machine Learning Research Impacts Payment Card Fraud Detection: A Survey and Industry Benchmark’. *Engineering Applications of Artificial Intelligence* 76: 130–57.

Sai, Chaithanya Vamshi, Debashish Das, Nouh Elmitwally, Ogerta Elezaj, and Md Baharul Islam. n.d. ‘Explainable Ai-Driven Financial Transaction Fraud Detection Using Machine Learning and Deep Neural Networks’. *Available at SSRN 4439980*.

SudoPurge. 2021. ‘4 Reasons Your Machine Learning Model Is Underperforming’. Medium. 2021. https://towardsdatascience.com/3-reasons-why-your-machine-learning-model-is-garbage-d643e6f0661.

Thennakoon, Anuruddha, Chee Bhagyani, Sasitha Premadasa, Shalitha Mihiranga, and Nuwan Kuruwitaarachchi. 2019. ‘Real-Time Credit Card Fraud Detection Using Machine Learning’. In , 488–93. IEEE.

Ti, Y.-W., Y.-Y. Hsin, T.-S. Dai, M.-C. Huang, and L.-C. Liu. 2022. ‘Feature Generation and Contribution Comparison for Electronic Fraud Detection’. *Scientific Reports* 12 (1). https://doi.org/10.1038/s41598-022-22130-2.

Tschantz, Michael Carl. 2022. ‘What Is Proxy Discrimination?’ In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1993–2003. Seoul Republic of Korea: ACM. https://doi.org/10.1145/3531146.3533242.

Wang, C., S. Chai, H. Zhu, and C. Jiang. 2023. ‘CAeSaR: An Online Payment Anti-Fraud Integration System With Decision Explainability’. *IEEE Transactions on Dependable and Secure Computing* 20 (3): 2565–77. https://doi.org/10.1109/TDSC.2022.3186733.

WP29. 2018. ‘Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679’. https://ec.europa.eu/newsroom/article29/redirection/document/51025.

Wu, B., X. Lv, A. Alghamdi, H. Abosaq, and M. Alrizq. 2023. ‘Advancement of Management Information System for Discovering Fraud in Master Card Based Intelligent Supervised Machine Learning and Deep Learning during SARS-CoV2’. *Information Processing and Management* 60 (2). https://doi.org/10.1016/j.ipm.2022.103231.

Wu, T.-Y., and Y.-T. Wang. 2021. ‘Locally Interpretable One-Class Anomaly Detection for Credit Card Fraud Detection’. *Proceedings - 2021 International Conference on Technologies and Applications of Artificial Intelligence, TAAI 2021*, 25–30. https://doi.org/10.1109/TAAI54685.2021.00014.

Wulf, Alexander J., and Ognyan Seizov. 2022. ‘“Please Understand We Cannot Provide Further Information”: Evaluating Content and Transparency of GDPR-Mandated AI Disclosures’. *AI & SOCIETY*, May. https://doi.org/10.1007/s00146-022-01424-z.

Xiuguo, W, and D Shengyong. 2022. ‘An Analysis on Financial Statement Fraud Detection for Chinese Listed Companies Using Deep Learning’. *IEEE Access* 10: 22516–32. https://doi.org/10.1109/ACCESS.2022.3153478.

Yale Ledger. n.d. ‘What Are the Risks of a Data Leak?’ Accessed 5 January 2024. https://campuspress.yale.edu/ledger/what-are-the-risks-of-a-data-leak/.

Yang, Yue, Chenyuan Liu, and Ningning Liu. 2020. ‘Credit Card Fraud Detection Based on CSat-Related AdaBoost’, ICCPR ’19, , 420–25. https://doi.org/10.1145/3373509.3373548.

Zou, Junyi, Jinliang Zhang, and Ping Jiang. 2019. ‘Credit Card Fraud Detection Using Autoencoder Neural Network’. *arXiv Preprint arXiv:1908.11553*.

Abbreviations

Abbreviation used in text	Full term
AI	Artificial intelligence
CCF	Credit card fraud
CCFD	Credit card fraud detection
FN	False negative
FP	False positive
LSTM	Long short-term memory
ML	Machine Learning
PD	Personal data
SAL	Similarity Attention Layer
TN	True negative
TP	True positive
XAI	Explainable AI

Outline

§ 1 Introduction

Fraud is defined as all acts of deceit to obtain a (financial) advantage in bad faith (Nesvijevskaia et al. 2021, 2). It is an issue that causes a significant economic damage every year. Current statistics estimate the losses to $85 billion for EMEA in 2022[1] (LexisNexis 2023, 9). The dark figure for the losses and fraud instances is presumed to be much higher (Fraunhofer-Institut 2021, 6). Digital payments, cryptocurrencies and AI technologies have been major factors in this increase (LexisNexis 2023, 11).

However, the numbers are increasingly steadily, therefore, the current efforts evidently are inept to tackle the issue adequately. This thesis aims to look at the challenges of fraud detection in general and researches whether current artificial intelligence (AI) can be a good tool to detect fraud attempts early and prevent it before damages are incurred.

A. Related works

Traditional fraud detection is heavily based on rules and therefore cannot adapt to evolving fraud schemes without human intervention (Boulieris et al. 2023a). Machine learning (ML), however, can detect fraudulent patterns itself and then flag these to humans.

Credit card fraud (CCF) is a significant problem that has far-reaching implications for individuals, businesses, and financial institutions. The use of advanced technology, such as AI-powered fraud detection systems, underscores the critical importance of enhancing cybersecurity measures, particularly in the realm of fraud detection and prevention. With the increasing frequency and sophistication of cyber threats targeting the financial sector, there is a pressing need to fortify the resilience of banking networks and information systems against such fraudulent activities. Apart from financial losses for both banks and their customers, credit card fraud can erode trust and confidence in the financial system.

I. Different types of fraud

Financial fraud is present in all different spaces, e.g. through financial statements, however this work focuses on consumers, specifically their use of credit cards.

[1] It has to be noted that this number is not limited to CCF only.

The primary method of defrauding customers involves social engineering. In the German market, a common approach is reaching out to a seller on Kleinanzeigen and falsely claiming to have made payments through the platform's channels. The seller is then prompted to create an account with the payment provider, mistakenly entering their payment details into a fraudulent form designed to mimic the Kleinanzeigen platform. With access to credit card data, fraudsters can initiate payments instantly or request authentication and acquire TAN from the user which they use for fraudulent transactions. (Mueller.legal 2023) This technique has been identified by the European Payments Council as one of the most widespread and perilous methods.[2] These scams persist due to counterfeit websites' high quality; fake merchants also utilize similar tactics such as fictitious import duty payments schemes. Card details can also be obtained from data leaks or digital skimmers. As evident from the above example, one general fraud idea (using Kleinanzeigen to obtain credit card data) can lead to different types of fraud that each have to be detected differently. The approaches change constantly.[3]

Mill et al. attempted to classify the observed patterns into different profiles, such as geo-location and spending profiling, that aim to aide in the establishing of appropriate AI systems (Mill et al. 2023, 1177). The author has observed in private practice that prior adding of a new device, cash payments from supermarkets or the sudden use of certain payment services, such as REVOLUT, were precursors for fraudulent activity.

Apart from the listed fraud approaches, the researcher has worked for clients who seem to have been hacked. According to client statements, they have not observed any suspicious messages or other activities, nor have they authorised transactions or new devices for the accounts. Banks were not able to make any statements as to how the authorisations took place or whether they had been hacked. In the case of one German major bank, most frauds took place in a time frame were the bank moved their IT to new platforms and all cards and authentication methods were temporarily disabled (Computerbild 2023).

II. Key challenges

Credit card fraud detection (CCFD), with or without employing AI, faces different challenges. The challenges must be considered when drafting solutions.

---

[2] Also referred to as Authorized Transactions Fraud *(European Payments Council, n.d., 59)*.

[3] This phenomenon is called ‚concept drift' in AI.

1. Real-time analysis

Today's business is fast paced. Amazon can only deliver when they have money. Thus, speed is an essential factor – also in bank transactions.

Traditional CCFD is reactive and can rarely proactively block transactions due to fraud suspicion. If the goal is not only to detect fraud after the fact, but rather to prevent unwanted transactions from being processed, CCFD must happen in the split seconds between initiating a transaction and completing authorisation. Processing a transaction routinely takes less than two seconds (FIS Global 2019). Prior research shows that 18 % of the checkout abandonments in the USA were due to prolonged checkout times (van Gelder 2023). Findings like these have been known to delay the implementation of security features, such as 3D Secure (3D Secure 2019). Hence speed is a key factor in the success of CCFD.

2. Prediction performance

No fraud management system is bulletproof. Human classification will have errors, as does every AI-generated. The results of those false predictions can vary and will be discussed in more detail[4].

All parties potentially suffer negative consequences when their predictions are faulty, i.e. a fraudulent transaction is not detected and therefore not blocked (false negative, FN) or a legitimate transaction is blocked (false positive, FP). All true predictions (processed legitimate transaction (TP) or blocked fraudulent transaction (TN)) are desired by banks and will increase customer satisfaction.

Depending on the setup of the fraud management, i.e. whether a negative prediction automatically blocks a transaction or flags for manual review, this can lead to delays and therefore missed sales opportunities. It is therefore vital for the success of the CCFD system to identify the reasons for faulty predictions. (Mill et al. 2023, 1176)

Most papers use accuracy, precision, recall, f1 and ROC-AUC to measure performance. Cost loss functions are increasingly being mentioned to give way to the fact that not all predictions incur the same negative outcomes to the parties involved. (Bockel-Rickermann, Verdonck, and Verbeke 2023, 16)

[4] Below §1.A.II.4, §3.A.I

3. Evolution of fraud schemes

It has been observed that fraudsters have gotten much more sophisticated in their approach. Through reporting on fraud schemes, they can observe current mitigation efforts and then adapt their scheme. (Fraunhofer-Institut 2021)

To detect new fraud schemes, the rules employed by banks to find them have to be developed accordingly. Early detection of new schemes can prevent high losses. As fraud management is rule based, the rules have to constantly adapt to the fraud schemes. Aside from evolving fraud schemes, fraud detection efforts also must adapt to new technology or payment developments. How we pay has evolved rapidly in the last decade. This change of relationship between the *independent variable*, the features indicating CCF, and the *target variable*, the fraud label, is coined concept drift in ML (Das 2022).

4. Customer satisfaction and costs

The focus in CCFD is primarily on financial aspect. However, customer satisfaction plays a crucial role as well. A rejected transaction can become cumbersome if a customer is travelling abroad and only carries one card and no cash; or if they need to pay for essential goods or products. Lack of trust with credit card information lead to cart abandoning in one of five shopping experiences (van Gelder 2023).

Customer satisfaction is one potential cost of CCF. The cost for customer service dealing with (alleged) fraud, the defrauded amount and incurred interest as well as legal fees incurred by banks, their payers and payees also have to be considered. (Nesvijevskaia et al. 2021, 8) Surveys show that the losses per fraudulent dollar have steadily increased to $4.23 in the USA.[5]

5. Limited data available

The lack in data for establishing CCF is a hurdle in various aspects: (1) little data is shared about CCFD approaches due to confidentiality reasons; (2) fraudulent transactions occur much less frequently than legitimate ones, therefore it is more difficult to learn from them; and (3) due to those factors, there is little reliable research that has been tested on large real-world datasets. All those reasons have hindered high-quality research and thus the advancement of CCFD AI. (N. Boutaher et al. 2020)

[5] There was a 16.2% increase since 2020 *(LexisNexis 2022)*.

III. Opportunities of AI in fraud detection

The following gives a brief overview of how AI tackles each of the five challenges flagged above.

CCFD has been a semi-automated, mostly manual process thus far. A credit card fraud detection system is typically composed of a set of five layers of control, including the data-driven model, which helps investigators by raising alerts on the most suspicious transactions. The primary goal of a data-driven model is to return precise alerts, as investigators might only check a few alerts per day.

Traditionally CCFD has been a semi-automated, mostly manual process thus far. Based on pre-determined rules, machines will scan transactions and accounts for suspicious transactions. The respective instances will be flagged for further inspection. The flags will be reviewed by humans upon availability. This process is labour-intensive, delayed and lacking dynamic. Therefore, it is not adaptive to the current challenges of CCFD and cannot be considered effective. (Aschi et al. 2022, 270)

The European Payments Council recommend real-time solutions that detect patterns and anomalies early on (European Payments Council, n.d., 6). AI is a tool that can (1) detect patterns and anomalies (2) across a large amount of data (3) in real-time.

In the current literature, the most used machine learning (ML) techniques are supervised techniques like Logistic Regression that always has high accuracy and works well with linear data; SVM that reduce the time of detection; Neural Networks which improve the classification rates. Supervised learning has been named as the method that addresses all the above challenges. (Bockel-Rickermann, Verdonck, and Verbeke 2023, 18)

The introduction of AI has brought with it a new problem: a lack of understandability. Most AI is a black box, i.e. its decisions are not transparent and cannot be retraced why it deemed a transaction fraudulent. This new research field is called explainable AI (XAI). XAI can be applied after training (*post hoc*) or be an *intrinsically interpretable* models, such as logistic regression or decision trees. Logistic regression is an intrinsic explainer, which uses the logistic function, also called *sigmoid function*, for calculating probabilities. Each weight indicates the size and direction of the impact that a certain input feature has on the prediction. One example for post-hoc explanations are *model-agnostic* systems that can be applied to any AI model. (Molnar 2020, 140)

Sabrina Ermshaus, A mixed-method evaluation of opportunities of AI in credit card fraud detection in banking

Nevskaja et al have researched the conflict of explainability and performance in different models. Their results are shown in Figure 1 and reflect that no

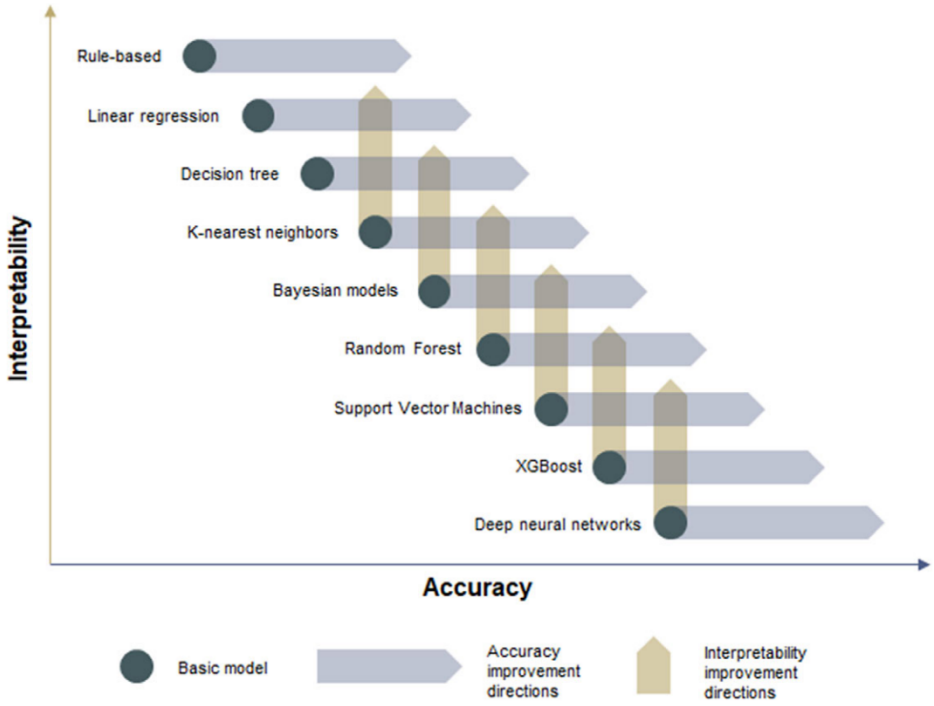

Figure 1 Explainability v performance, source: Nesvijevskaia et al. 2021

good compromise has been identified so far because experts only try to improve one metric, rarely both at the same time. The performance of intrinsic models is often criticised and therefore not applied for CCFD which needs accurate results. (Nesvijevskaia et al. 2021) Banks who have implemented solid CCFD have reported a significantly lower fraud rate and 14% lower fraud costs (LexisNexis 2022).

B. Research statement

As shown in the review, current research has not taken a holistic approach that considers both technological and legal perspectives in using AI while also looking at the specificities of CCFD. Thus, the developed solutions may not be fit for use in the field. This thesis aims to close that gap for CCFD in banks.

The research was designed from the perspective of a bank's General Counsel tasked with assessing whether AI can assist in combating credit card fraud. Based on that and the analysis of prior research, the following research questions have been formulated:

1) Can existing AI systems perform well enough to detect CCF?
    a) Which parameters are relevant in assessing good performance?
    b) Which regulations must banks comply with?

   c) Which AI systems perform well in CCFD?

2) Which technical aspects are relevant for robust and compliant CCFD?
3) Based on the results of 1) and 2), how, if at all, can banks implement AI-based CCFD systems?

To address the above questions, we will first examine pertinent regulations (qualitative research) and then conduct a literature review of quantitative research on current AI systems developed for identifying CCF. The methodology for the quantitative and qualitative research will be explained in § 2. Subsequently, requirements for the AI from both legal and technical perspectives will be outlined using qualitative methods. § 4 will assess the effectiveness of existing AI models, while § 5 will consolidate the findings of preceding sections to answer the aforementioned research questions.

§ 2 Methodology

The research followed a mixed method approach.

A. Step 1: Qualitative research

The qualitative part of the research was divided into two parts.

The author is a German bar-licensed attorney with experience in CCF litigation. As part of that, she has obtained relevant experience in the sector of banking and privacy law and is familiar with relevant legislation. This experience supplemented with further legal research guides the research of the legal framework by screening relevant laws using doctrinal research (P. I. Bhat 2020, 155 ff; Kaupa 2016).

First, we identified the relevant primary legal materials. These are further interpreted using secondary legal materials. Where applicable, we consulted relevant CCFD literature. Based on the literature for the legal framework, we employed snowballing to further explore the arising themes.

While cybersecurity and appropriate technological and operational measures are mandated by both the GDPR and the AI Act, due to time restraints these will not be included in our research.

B. Step 2: Quantitative research

Based on requirements identified from challenges (§1.A.II,III) and qualitative research (§3.A,B) we have collected and analysed relevant studies.

I. Data Collection

To identify relevant papers, a search with the prominent scientific databases SCOPUS, IEEE and ACM have been conducted with terms[6] led by research questions 1c and 2 (cf. Table 1). They were adapted to the functionality of each database and produced a total of 701 results.

[6] Boolean query: AI OR “artificial intelligence” OR DL OR “deep learning” OR ML OR “machine learning”) AND (“Credit card fraud” OR “card fraud” OR “card-fraud” OR “credit-fraud” OR “card cyber fraud” OR “transaction fraud” OR “payment fraud” OR “fraud detec*” OR “bank* fraud” OR “financ* fraud”))) AND ((explainab* OR transparency* OR interpretab*))

Table 1 Records obtained based on database

Database	Returned results
IEEE	68
SCOPUS	564
ACM	83

The whole selection was led by the research questions and the inclusion and exclusion criteria (Table 2). The process can be reviewed in the PRISMA (Figure 2).

Table 2 Inclusion and exclusion criteria

Included	Excluded
▪ English language ▪ Published in or after 2019 ▪ Journal articles ▪ Conference papers	▪ Crypto/bitcoin ▪ Non-financial products ▪ Financial aspects other than credit card fraud, e.g. money laundering, financial statement fraud or credit scoring ▪ Fraud systems for non-banks, e.g. e-commerce websites

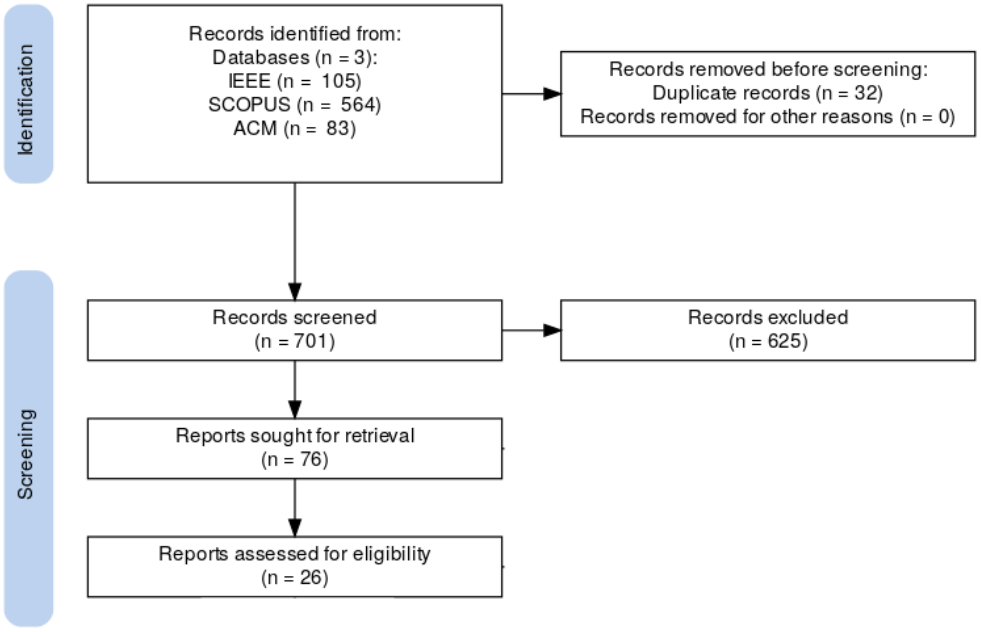

Figure 2 PRISMA diagram

II. Data Analysis

Due to time restraints, when a paper researched several AI systems, only the best performing models were included in this study.

Sabrina Ermshaus, A mixed-method evaluation of opportunities of AI in credit card fraud detection in banking

The remaining 26 papers were screened and evaluated according to the (1) pre-selected performance metrics, (2) methods used and (3) analytical challenges faced.

Where measures were missing, we computed them, when possible.

To determine the most conclusive metric, project goal and business objectives must be weighed against the data distribution and characteristics of each model. Accuracy computes how many times a classification was correct. This can be misleading when the dataset is imbalanced, thus installing false confidence in the prediction ability of the mode, as is the case for this project. (Makki et al. 2019, 93016)

Instead, the model must limit misclassifications of fraud[7] in an imbalanced dataset. For this purpose, recall is an efficient matter that can be calculated by the formula depicted in Figure 4. When the classification of both negatives is equally important, the f1-score is the preferred metric. The f1-Score is a weighted average of precision and recall, thus giving insight on false classifications. It is calculated as shown in Figure 5. (Miao and Zhu 2022, 1547)

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

Figure 3 Formula for F1 score

$$Recall = \frac{TP}{TP+FN}$$

Figure 4 Formula for recall

$$Precision = \frac{TP}{TP+FP}$$

Figure 5 Formula for precision

C. Step 3: Synthesis

The findings of the quantitative research are measured against the identified requirements from step 1 and are synthesised with industry insights to be able to answer research questions.

[7] Both FP and FN also referred to as minority class, must be limited.

§ 3 Developing a framework

As a first step, the legal and technical possibilities and boundaries for implementing AI in CCFD have to be explored. By establishing the legal framework, we want to see if regulation adds more requirements to what the AI has to be capable of. The next step is to look at the capabilities of current AI models and how they can assist in meeting the requirements and challenges of CCFD.

A. Legal framework

Apart from the general requirement to be reliable and be able to prove predictions due to general civil law baselines (Fraunhofer-Institut 2021, 8), there are further legal requirements a CCFD AI needs to pass.

I. Banking Law

Banks are regulated by national and EU laws. The statutes relevant to CCFD are found in the German Civil Code and its underlying EU regulations. We have looked at current legislation and further developments of EU law to establish a bank's (1) duty to process transaction requests, (2) liability for unprocessed and (3) fraudulent transactions, (4) duty to protect customer interests and (5) a duty to implement AI for CCFD efforts. Finally, we have researched how (6) customer satisfaction is a relevant business factor.

The cited BGB sections are implementations of the PSR-2. The EU is currently working on new payments legislation – the PSD-3 and PSR.[8] Any PSR norms cited below are future laws that are considered to ensure future-proof research.

1. Duty to process transaction requests and liability

Banks are required to process payment orders within a legally binding timeframe (pursuant to § 675s(1) BGB). If the bank has cause to suspect fraudulent activity and decides not to process a payment order, it can be made liable.

§ 675o(1) BGB requires immediate notification of unprocessed transactions, including the grounds for refusal and what to remedy in order to execute the transaction.

[8] We are referring to the proposal put forth by the Commission on 28/06/2023.

Pursuant to § 675y(1),(2),(6) BGB, a bank must reimburse an unprocessed payment in due time and reimburse all fees incurred by the payment recipient due to the. Therefore, an unprocessed FP leads to damage and customer dissatisfaction. If contractually agreed upon, a bank can block a credit card due to suspicions of fraud (Art. 51(2) PSR). Neither the BGB nor the PSR stipulate that banks reimburse any incurred costs, e.g. legal or additional fees.[9]

2. Liability for fraudulent transactions

Purs. § 675u(1) BGB) an unauthorised charge has to be reimbursed within two business days. A charge is unauthorised if no PIN was entered in card-present transaction or Strong Customer Authentication (§ 55(1)(2) ZAG) for online transactions was not obtained. The burden of proof for the authorisation is on the bank (§ 675w(1),(4) BGB). If the fraud was carried out using social engineering for instance, customers provide authorisation because of a wrongly assumed situation. In these cases, banks are not legally required to reimburse. Under the PSR, banks will be liable for impersonation fraud: In case someone impersonates a bank and defrauds a customer, the bank will have to reimburse the full amount (Art. 59(1) PSR).

3. Duty of care

Every company has a responsibility to care for the interests of their business partners (§ 241(2) BGB). Banks are also required to fulfil additional obligations: According to § 675m(1)(1) BGB, a payment service provider is obligated to protect payment instruments from unauthorised access. However, this duty specifically pertains to payment instruments such as credit cards and does not extend to entire bank accounts. The German Banking Act's[10] safeguards do not extend to consumer usage of credit cards, thus, rendering them irrelevant for this research.

4. Obligation to implement AI for CCFD

We researched relevant regulations as to whether banks are obligated to employ AI in their efforts to detect CCF.

Art. 83(1)(c) PSR requires banks to establish transaction monitoring mechanisms for preventing and detecting fraudulent transactions. These mechanisms are limited in terms of permissible data usage: user information;

[9] For example, a merchant could charge a customer late fees.

[10] German: Kreditwesengesetz (KWG)

account details including transaction history; specifics about transactions such as amount and recipient; and session data. The permission granted is restricted to the duration of the contract. Moreover, these mechanisms shall consider factors like compromised or stolen authentication devices along with known fraud schemes and malware. There is no explicit allowance or mandate for using AI technology in monitoring according to this regulation.

The *PCI DSS*[11] is not a national law, but rather an obligation stipulated by major credit card companies. Payment service providers, including banks, are contractually required to adhere to the 12 compliance principles. These principles aim to safeguard cardholder data and impose significant cybersecurity and data security responsibilities. The standard does not explicitly mandate the use of artificial intelligence or require fraud detection implementation.

Concludingly, no laws or contracts require AI to be used in CCFD, but rather infer that in can help with the monitoring obligations.

5. Customer satisfaction

While ensuring customer satisfaction is not a legal requirement, it is vital to the fiscal responsibilities that apply to banks. When legitimate payments are not processed, or a customer cannot pay while travelling internationally that causes major inconvenience. Customer satisfaction is potentially impacted by both false negatives and false positives. Customers might react positively when fraud is detected and prevented by their banks without their own involvement. Therefore, banks have an interest to achieve a high rate of true positives, but also to prevent both false positives and negatives. (Gabudeanu et al. 2021, 15)

Surveys show that customers are dissatisfied by calls asking them about the legitimacy of a charge (Ryman-Tubb, Krause, and Garn 2018). Not implementing CCFD can cause severe reputation damage (Fraunhofer-Institut 2021, 6).

Once a credit card transaction is processed, it cannot be revoked, even when reported immediately (§ 675k BGB). This leads to frustration with customers who have been victims of fraud. It can therefore be advisable to block transactions.

[11] The fourth version of the standard can be found at https://docs-prv.pcisecuritystandards.org/PCI%20DSS/Standard/PCI-DSS-v4_0.pdf.

II. Privacy

The GDPR governs all processing of personal data (PD). That roughly includes any activity with customer information. The GDPR has special safeguards for automated decision-making like that of AI.

PD is any data associated with or traceable to a natural person. In the banking arena, PD is an IBAN or a name. The natural person that is associated with a data point is considered a data subject under GDPR and is afforded specific rights. The IBAN of a business is not connected to a natural person and therefore not safeguarded by the GDPR. When data of a recipient or sender that is not the account holder of the monitored account is also included in AI, a third party becomes data subject.

Financial data is not deemed sensitive data under Art. 9 GDPR, therefore no specific safeguards need to be taken.

1. Grounds to process

The design of the GDPR generally prohibits all processing of PD, unless employed for at least one of six pre-defined purposes (Art. 6(1) GDPR).

CCFD could be considered a contractual duty that constitutes an exception under Art. 6(1)(b) GDPR, however, using AI is not essential (EDPB 2019, 16). Scholars agree that FD does, however, constitute a legitimate interest of the bank Art. 6(1)(f) GDPR. The term ‘necessary’ has to be interpreted narrowly, hence simplifying the process is not enough. (Franck 2022, para. 35)This also applies to processing of third party PD.

While there is grounds for PD processing, banks have to be careful that processing is kept to a minimum (Art 5(1)(c) GDPR). Data minimisation means to only process data in a matter that is adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed.

Whether data processing is kept to a minimum requires to ask: If the objective can be reached a different way, that is the minimum and the intended processing is prohibited. Banks will therefore have to determine how little PD they need by developing different AI. This could be achieved by using technologies such as feature engineering.[12]

[12] Below §3.B.I.5.

2. Transparency

Transparency is one of the guiding principles of GDPR. The duty to inform in a privacy policy and respond to data subject requests are reflections of this principle (Art. 5(1)(a) GDPR, Artt. 13-15 GDPR). Before implementing AI, the explainability of a model needs to be checked and maintained as part of the Privacy by Design principle. If it cannot be provided, the use of AI is prohibited (Franck 2022, para. 29).

In the privacy policy, it is necessary to include information about automated decision-making, its logic, scope and consequences as per Art. 13(2)(f) GDPR. This requirement ensures that individuals can exercise their data subject rights. However, the right to information may conflict with safeguarding trade secrets. While confidentiality may limit the scope of disclosure, it cannot eliminate it entirely (Rec. 63 GDPR). The transparency requirements are constrained by ensuring clarity for the individual. As a result, algorithms do not need to be revealed since they may not provide relevant insights for an average person; instead, providing details on the types of PD processed and their influence on predictions will suffice (Linhart 2023, 146).

For third parties, it will not be possible to inform them about their rights before initiating processing activity. This does not, however, violate Art. 13,14 GDPR though. Rec. 62 GDPR provides an exception for such disproportionate efforts.

3. Right to be forgotten

The right to be forgotten, or in this case the right to object (Art. 17 and Art. 21 GDPR resp.), is a significant data subject right, requiring the processor to cease all data processing once lodged. There are limitations if other parties' rights or legal duties require further processing. According to Art. 19(3) GDPR, deletion requests can be rejected when data is necessary for processing current claims. This is applicable to pending fraud investigations.

When developing the algorithm, it is important to determine how long PD is retained for training to comply with GDPR regulation. Retention timeframes have to be implemented once the processing is no longer necessary for the legitimate interests, Art. 6(1)(f) GDPR. Then processing must cease immediately and all PD has to be expunged. Extracting embedded PD may pose challenges in AI systems.

4. Automated decision-making, Art. 22 GDPR

Automated decision-making is generally prohibited purs. to Art. 22(1) GDPR. When an exception applies, a data subject must be given suitable information about the process and be afforded the right to a human decision.

The ECJ has recently decided on a request regarding German credit scoring provider SCHUFA[13]. In this landmark decision, they clarified the scope of Art. 22 GDPR as the cumulative condition of a (1) decision being made (2) based solely on automated processing that has (3) legal effects or similarly significantly affects subject (para. 42 ff.). The Court defines decision as an act which may affect data subject in many ways, citing Rec. 71 GDPR.

The proposed CCFD process delivers a decision, namely whether a requested transaction is fraudulent or not. This decision is based on automated processing using AI – routinely – without a human in the loop. The big question is whether the decision meets the significance threshold of Art. 22 GDPR.

The fraud classification does not lead to direct legal consequences. A rejected transaction may have negative consequences, but they are not of a legal nature. Hence, only other significant effects may be applicable.

WP29 defines significant effects as '*sufficiently great or important to be worthy of attention*', i.e. a substantial impact on the circumstances, actions, or decisions of the data subject (WP29 2018, 10). An account holder may be frustrated by a rejected transaction, but that does not impact their decision-making moving forward.

Therefore, Art. 22 GDPR does not apply to CCFD and the use of AI is permissible without further safeguards.

III. AI Act

Apart from the GDPR, the EU is currently working on agreeing on an AI Act. The latest available draft[14] is used as a basis for this work.

1. Overview of AI Act

The AI Act imposes different levels of security on AI system depending on their associated risks: unacceptable risk, high, low or minimal. All systems posing high risks to the fundamental rights of persons are regulated

[13] ECJ, C-634/21, judgement dated 07/12/2023.

[14] We refer to the draft adopted by the European parliament 14 June 2023.

specifically. Low risk systems must comply with transparency requirements and those systems posing no risk sanctioned by the act, may implement internally binding rules without a legal duty.

Figure 6 shows the different risk levels, some common examples for each level and the restrictions imposed by the AI Act.

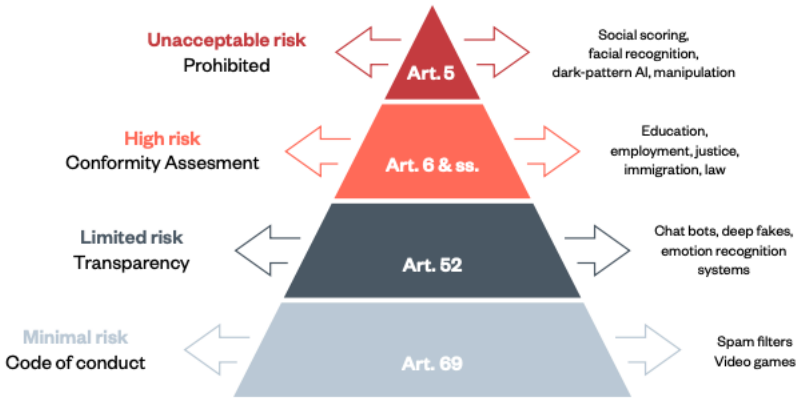

Figure 6 Risk structure of AI Act, source: (Ada Lovelace Institute 2022, 9)

The AI Act addresses users and providers[15] of AI systems. For the purposes of this paper, it is assumed that the banks employing the AI (user, purs. Art. 3(4) AI Act) to aide in their fraud prevention efforts will also be the provider (Art. 3 (2) AI Act) of the system. We will therefore not differentiate between the different operator structures and simply state the responsibilities applicable to both types of operators.

2. Assigned risk level

The different legal texts have differing opinions on what can be considered high risk. The most recent proposal has made clear that fraud prevention systems as envisioned in this paper will not be considered high risk systems. To determine which regulation applies to the investigated situation, fraud detection and prevention, the attributed risk level needs to be assessed.

Some argue, that due to Annex III 6 and Rec. 38 AI Act, this research system is classified as high-risk AI. This is not convincing, however, since the Act clearly stipulates that a system is only classified as high risk when used by law enforcement authorities. But the amended proposal also states in consideration 37: ‘However, AI systems foreseen by Union law for the purpose of detecting fraud in the offering of financial services should not be considered as high-risk under this Regulation’ and adds that in Annex III para. 5 (b)

[15] Sometimes also referred to as *operator*, cf. Art. 3 (8) AI Act.

accordingly. Therefore, we stipulate that the associated risk will be low and the further restrictions of Title III of the Act do not apply.

3. Requirements for low-risk AI

For any low-risk AI, the general principles of AI systems (Art. 4a AI Act) and Art. 69 AI Act are relevant. The general principles can be achieved by complying with transparency (Art. 52 AI Act) which is not obligatory purs. Art. 4a(2) AI Act. The duty to conduct a high-risk conformity assessment (Art. 43 AI Act) and provide technical documentation does not apply to this research project (Art. 11 AI Act).

a. Art. 4a AI Act

Art. 4a AI Act sets six ground principles, similar to those of the GDPR, that are applicable to all systems regardless of their attributed risks. According to the provision, AI systems should prioritise (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, and (6) social and environmental well-being.

Technical robustness and safety are achieved by minimising unforeseen harm through malfunctions or malicious actors. Systems should be developed and used in compliance with existing privacy and data protection rules. AI-generated decisions need to be traceable and understandable and made transparent to affected persons. Furthermore, it is important to prevent discriminatory impacts and unfair biases by incorporating diversity and non-discrimination principles in the development. Lastly, the AI systems should be operated sustainably, benefiting all human beings and monitoring their long-term impacts on individuals, society, and democracy.

b. Art. 52 AI Act

Under Art. 52 AI Act, a service user must be notified whether they are interacting with an AI system. No further information disclosure is warranted.

The scope of interaction is not clearly defined by the Act. A broad interpretation would go by the letter of the law and define interaction as any mutual or reciprocal action or influence between two or more parties. *(Merriam Webster 2023)*; 17 USC § 114(j)(7)

It is unclear whether Art. 52 applies to this specific tool because there is no direct human interaction between human and AI (Bomhard and Merkle 2021, 35).

While credit card fraud is a crime, the researched system will be employed by banks and not by law enforcement agencies hence why the crime investigation privilege is not applicable.

When comparing Art. 52 AI Act to Art. 22 GDPR, it is important to note that both regulations address the issue of transparency in AI systems, albeit with some differences in scope. Art. 52 AI Act requires service users to be notified when interacting with an AI system, but it does not mandate further information disclosure. On the other hand, Art. 22 GDPR provides individuals with the right to be informed about the functionality of automated decision-making processes, encouraging transparency and trust-building around these systems. (Linhart 2023, 69–71)

The provision in the AI Act might be more limited in scope compared to the transparency requirements of Art. 22 GDPR. While Art. 52 AI Act only requires notifying users of their interaction with an AI system, Art. 22 GDPR emphasises the right to be informed about the system's functionality in automated decision-making processes. It encourages creators of algorithmic decision-making systems to build trust and increase transparency around these systems, aligning with the principles of fairness and accountability outlined in both legislations.

c. Art. 69 AI Act: Code of Conduct

Under Art. 69 AI Act, entities providing AI systems of any risk level are encouraged to draw up codes of conduct inspired by the requirements set out in Title III Chapter 2 AI Act, i.e. voluntarily comply with further assessment and documentation duties. This is not of further interest for this paper.

B. Technological framework

Considering the challenges specific to CCFD and the legal requirements, an AI model must adopt certain requirements. CCFD entails three problem-imminent issues: unbalanced data, poor data quality, concept shift, and privacy. Additionally, there are legal obligations for the AI system to avoid bias against customers, minimize PD processing, enable data deletion, and provide explanations for its classifications. AI has solutions to mitigate these issues and at the same time brings with it its own challenges that have to be dealt with. AI can address these concerns while simultaneously presenting its own concerns.

I. AI means to combat problems

First, we have researched how AI might mitigate class imbalance, concept shift and poor data quality.

CCFD challenges	Possible AI solutions
Class imbalance	Sampling
	Feature selection
	Cost-sensitivity
	Ensemble learning
Concept drift	Online learning
	Machine unlearning
	Periodic model retraining
Poor label quality	Unsupervised learning
	Human intervention
Bias	No PD
	Principal component analysis (PCA)
	Feature engineering
Privacy	Machine unlearning
	Principal component analysis (PCA)
	Homomorphic encryption

Table 3 List of problems and possible AI solutions

Table 3 shows the CCFD issues and solutions AI technology can present. We will present some of the solutions below. The solutions not presented here, will be elaborated on below.

1. Cost-sensitivity

Cost-sensitive learning refers to a form of machine learning that considers the expenses linked with misclassification. In contrast to conventional learning methods, cost-sensitive learning aims to reduce the overall cost of misclassification by explicitly acknowledging the costs connected with various types of errors. (Jing et al. 2019, 143)

2. Sampling

The training data can be synthetically modified by adding minority data (*oversampling*) or removing majority data (*undersampling*). Oversampling increases the number of instances in the minority class, preventing information loss and capturing patterns more effectively. The duplication can lead to

overfitting and model bias (Brownlee 2020). Undersampling reduces the number of instances in the majority class, reducing training time and preventing overfitting. It may also lead to information loss and incomplete representation of the majority class (Megdad, Abu-Naser, and Abu-Nasser 2022, 36).

Combining oversampling and undersampling techniques is more effective in handling class imbalance than using them individually. A prominent combination is SMOTE-ENN which is often used in CCFD because of its good results. (Zou, Zhang, and Jiang 2019, 3)

3. Online learning

Online, or incremental, learning will mitigate concept drift by constantly feeding the model new data to retrain and learn new patterns. The technique is very sensitive and can easily decrease its accuracy when fed with poor data. It can also be difficult to explain due to its complexity. (Datacamp 2023)

4. Machine unlearning

Machine unlearning seeks to reverse the effects of its targeted datapoints, classes and features. This is done by readjusting weights or retraining a model (Nguyen et al. 2022, 3). Machine unlearning is a relatively new research field, but has already been endorsed by scholars for its qualities, most prominently in privacy preservation (Hine et al. 2023, 2). Google has recently launched a challenge to explore machine unlearning's potential further (Pedregosa and Triantafillou 2023).

5. Feature engineering

Feature engineering is used to leverage raw data to generate or optimise variables. The goal is to simplify and accelerate the machine learning process and enhance its performance. A bad feature can have a direct impact on model output. Inputting high-value and easily-processable data can also improve resource needs. (Patel 2022)

Categorical features need to be encoded into a binary format to be machine readable. This can be done using techniques such as one-hot encoding. (Ashenden et al. 2021, 17)

a. Feature selection

Feature selection refers to identifying and reducing the size of the important features. This facilitates computational efficiency in machine learning and

data analysis algorithms. It also enhances the overall output quality derived from these algorithms. By removing irrelevant features, feature selection can help reduce overfitting and improve the generalization of the model, leading to better performance on unseen data. (Mozaffari 2023)

b. Feature extraction

Feature extraction describes decreasing the size of a dataset by combining existing features into new ones. This results in a reduced number of features, making the data more manageable for algorithms to handle, while preserving original relationships and relevant information. (Hoppe et al. 2019, 849)

Principal Component Analysis (PCA) is a frequently-employed extraction technique for reducing dimensionality and adding privacy in ML (Liu et al. 2022). Nevertheless, when utilised with data gathered from multiple individuals, PCA has the potential to release PD. Differential privacy is a recognised adaptation of PCA that is also applied to share U.S. Census data. (Dwork et al. 2006)

II. Issues arising from AI

Second, we analysed which new problems the use of AI introduces to CCFD.

1. Bias

Bias can be inherent to the training data. Bias can also be rooted in the structure of the methods used. This bias can also develop in the course of training.

To limit bias, training and test data has to be ‘cleansed’. While it cannot be completely eradicated and classification will never be entirely objective, regular testing for bias can keep it to a minimum (Fritz-Morgenthal, Hein, and Papenbrock 2022, 5).

Bias is confounded by the imbalance of data and concept drift. When the bias is based on specific types of PD, it can discriminate against a specific group of customers. Geographical and occupational data as well as the websites they interact with, are attributes in credit card data that can cause a bias (Tschantz 2022, 2001).

2. Explainability

ML models are often called black boxes because it is not retraceable how and why a classification was made. As explored earlier, current regulation

needs a bank to be transparent as to how they derive at their CCF predictions. This transparency is generally referred to as *explainability* in the AI industry.

Local explainability is the explanation of a model decision about a single transaction. This would be needed for an Art. 15 GDPR access request. A global explanation gives insight on the model's overall operation which is required to comply with Art. 13,14 GDPR. (Sai et al., n.d., 8)

Explainability can be achieved in two ways: *Intrinsically interpretable* systems, such as decision trees, are often used for problems that are highly regulated and therefore need to produce understandable predictions. While they easily achieve explainability, they often lack the accuracy needed for reliable results. Other algorithms need *post-hoc explanation* that are applied after training. Most post-hoc XAI is model-agnostic and can therefore be applied to any given algorithm. SHAP and LIME are prominent examples. (Gianini et al. 2020, 551)

LIME models the predictions of the underlying black-box model through the use of local surrogate models in order to provide explanations for individual predictions. Essentially, LIME adjusts feature values in a single data sample within a simpler local model and observes how it impacts the output. The SHAP approach explains instance predictions by calculating each feature's contribution using Shapley values. In simple terms, SHAP assigns importance to each feature by analysing how every possible combination of features affects the output. It takes a long time to compute. (Sai et al., n.d.)

There is no consensus as to the scope and labelling of explainability. The focus for this research is to provide understandability – for the customer and for the bank's fraud manager to be able to understand why a transaction was classified and to provide the GDPR-mandated transparency. As explained above[16], the relevant information lies in how and why a prediction was made.

3. Speed

The speed of algorithm execution is impacted by the type of classifier, the size of data and number of analysed features. High-value computational resources can also accelerate prediction. Cost-sensitive learning is another positive factor. Decision trees are known to be fast classifiers. (Bowden et al. 2008; SudoPurge 2021)

[16] Cf. §3.A.II.2.

III. Algorithm types

ML algorithms can broadly be typified into three categories: supervised, unsupervised and semi-supervised algorithms. Unsupervised algorithms can classify unlabelled data, while supervised algorithms use labelled data. The semi-supervised type leverages the benefits of both by incorporating unlabelled data based on the labelled ones. (AL-Dosari, Fetais, and Kucukvar 2022, 3)

In a highly-imbalanced dataset such as fraud supervised learners focus on the majority class and have an increased error rate for the minority class. The performance of semi-supervised systems can be improved. Data imbalance can be mitigated by addressing the balance in pre-processing or weight assignment. In supervised learning, the objective is to predict results for new data with known outcomes. In contrast, unsupervised learning aims to derive insights from extensive datasets without predefined expectations, allowing the machine learning process to identify distinct patterns within the dataset. (Feng et al. 2021) In CCFD, due to its poor label quality, both strengths need to be leveraged. Hence, others suggest combining different algorithms to improve detection accuracy by identifying different types of frauds and other real-life scenarios. This approach is termed *ensemble learning*. (Jurgovsky et al. 2018, 245)

C. Required features of AI model

Based on the findings from the legal and technical frameworks in this chapter and the identified challenges from the previous one, we have identified the following requirements for CCFD AI:

(1) Prediction performance
(2) Adaptable to concept drift
(3) High speed
(4) Customer satisfaction
(5) Explainability
(6) Privacy compliance
(7) No bias or discrimination

The requirements are ranked according to their importance for successful CCFD and the findings of the literature review in §1.

§ 4 Model evaluation & selection

We have identified 26 research papers that employ various approaches to CCFD through the use of AI. The specifications and results of each system were extracted, and they were subsequently ranked based on their f1 value. We conducted a more detailed examination of the top seven models to identify any common factors contributing to their success. After, a comprehensive analysis of the full dataset was performed in order to gain insights into the pre-defined requirements. When different models and specifications were tested within a single paper, we only present the overall best performing model according to our selected metrics.

It should be noted that there is considerable scarcity in training data for AI, as evidenced by our data analysis: out of the screened papers, 13 used a dataset from a European bank dating back to 2013 that contained fewer than 300,000 transactions. The utilised data underwent reduction using principal component analysis (PCA), resulting in it being devoid of personal information. Additionally, [15,20] cited slightly different sample sizes.

The table in the annex shows the results of all reviewed 26 studies.

A. Prediction performance

Performance is measured using the f1 score. If this score was not provided, we tried computing it using other reported results, or we used the AUC. In a second round we look at recall.

The top AI model [5] is an AdaBoost classifier with Random Forest as base estimator. The model achieved a perfect f1 value of 1, indicating no faulty predictions. While these results are impressive, it is important to consider that there is limited information about the dataset, which contains 1,000,000 synthetic samples and lacks detail about the synthesiser used by the researcher. As a result, its validity cannot be verified and there is uncertainty regarding how well it would perform on real-world data. Additionally, this research has not undergone peer-review or been cited by other researchers yet; therefore, we have excluded it from our further review to ensure reliable results.

For prediction performance, we will first present the architecture of the seven top-performing models and then show emerging patterns. The prediction results and an overview of the AI models, any employed feature engineering and sampling method can be viewed in Table 4.

I. Architecture overview of top-performing models

Table 4 Prediction of performance of top 7 models

	f1-score	Recall	AUC	AI system researched	Feature Engineering	Sampling Method	European Dataset
15	0.998	1.000	1.000	LSTM, GRU	Feature Extraction, Genetic Algorithm, Information Gain	SMOTE ENN	Yes
14	0.995	0.997	0.990	Information Gain, Genetic Algorithm		SMOTE ENN	Yes
16	0.992	0.992		Bi-LSTM	Similarity Attention Layer (SAL)	SMOTE ENN	Yes
19	0.990	0.983		Random Forest	Variational Autoencoders	SMOTE	Yes
26	0.987	0.999	0.970	AdaCost	Feature Extraction 'TPC'	SMOTE	Yes
20	0.985	0.980	0.970	LSTM (enhanced, named by authors 'RaKShA')	Exploratory Analysis Normalisation		Yes
22	0.976			Platform's method, Logistic Regression, Random Forest, DNN, XGBoost (named by authors 'CAeSaR')	Random forest: k fields		

[15] is a stacking ensemble of LSTM and GRU as level-0-learners and MLP as level-1-learner. The level-0-learners follow two different sequential approaches (LSTM and GRU) ensuring a diverse prediction range, yet both known for their robustness and good performance. Their research is reproduced using a second smaller dataset that underperforms slightly.

[14] wraps information gain and genetic algorithm. Information gain is a widely-used feature selection tool that weighs the most important features more heavily. Thus, it enables a classifier to perform better. The most significant attributes are identified by building a statistical threshold. GA is based on generations that are also computed using statistics. GA is inspired by natural selection's 'survival of the fittest'. The GA evaluates fitness statistically. Special attention has to be paid to the fitness value: When statistically evaluating fitness the minority class has to be included [14:9]. The performance of the model was confirmed using additional datasets [14:12].

The research conducted in [16] introduced a variation of the LSTM learner used in [15] called Bi-LSTM[17]. The model further leverages similarity attention layer (SAL), a new feature engineering technique. It takes the context of vector collection and generates new hidden vector that is assigned new weight.

[17] Short for *BiDirectional LSTM*. Reported f1(BiLSTM)=.6625; f1(LSTM)=.5677.

Sabrina Ermshaus, A mixed-method evaluation of opportunities of AI in credit card fraud detection in banking

The research strategy in [19] was compromised of testing different single algorithms on the EU dataset and pairing them with different sample methods. The best combination was a Random Forest algorithm using Variational Autoencoder. None of these methods otherwise ranked among the top seven models. The system ranks in the middle of the top 7, suggesting that they could boost results in a diverse ensemble. An advantage of that Autoencoder is the generation of an extra feature, thus introducing less bias.

The authors in [26] had a stronger industry and cost focus. They use AdaCost, an AdaBoost-based ensemble learning with an added function computing costs for each classification type as described in Figure 7. **Error! Reference source not found.**The result is coined 'TPC' for *total profit of classification*, to measure the performance of the classifier. The metric considers (1) transaction amount, (2) customer satisfaction, (3) profit, (4) potential profit and (5) added admin cost. They report a TPC of $20,513,872 when using CSat-Ada; $20,239,205 for AdaBoost. This is a difference of $300,000 for a transaction volume of roughly 285,000. The study reports 18 FP and 30 FN classifications. The number of FN is higher compared to that of the AdaBoost classifier. This result is not explained by the authors.

	Actual Legitimate $y_{true} = 0$	Actual Fraud $y_{true} = 1$
Predict Legitimate $y_{pred} = 0$	*True Positive* *Satisfaction* *Loan Profit*	*False Positive* *Transaction* *Satisfaction*
Predict Fraud $y_{pred} = 1$	*False Negative* *Admin* *Satisfaction* *Potential Profit*	*True Negative* *Admin* *Satisfaction*

Figure 7 Cost attributed to classification type; red for added costs and green for saved costs Source: [26]

The RaKShA researched in [20] detects CCF using X-LSTM[18] and stores the results in smart contract to make traceable and more trustworthy. The results can be reviewed in the blockchain. They compute how a genuine transaction would look for a customer (feature reduction) and then (1) save that into the blockchain and smart contract and (2) input that pattern into LSTM. The RaKShA uses one-hot encoding and normalisation to improve data quality. The XAI further prioritises the data before sending it into the classifier. The research uses a second dataset to train on credit approval to enhance classification power. Therefore, this might skew the results and potentially introduce bias if the data will suggest that creditworthiness is lower due to formerly declined credit applications.

[18] Short for an ensemble of LSTM and AdaBoost with SHAP as XAI.

Sabrina Ermshaus, A mixed-method evaluation of opportunities of AI in credit card fraud detection in banking

[22] uses an ensemble of three different classifiers to approach the data from diverse angles: Association evaluation, anomaly detection in spending behaviour; Subsequent analysis, repeat transaction over span of short time; Risk review e.g. specific merchants. Function results fed into credit card where a XGBoost will select the final decision strategy. The final decisions made in the Center Control are output to the 3 function modules at regular intervals to retrain to account for new schemes. Data is pre-processed by Random Forest feature selection. The algorithms rely on different PD, such as frequent IP and registered home address. This has the potential for bias and the processing bears significant risks. The full workflow can be reviewed in Figure 8.

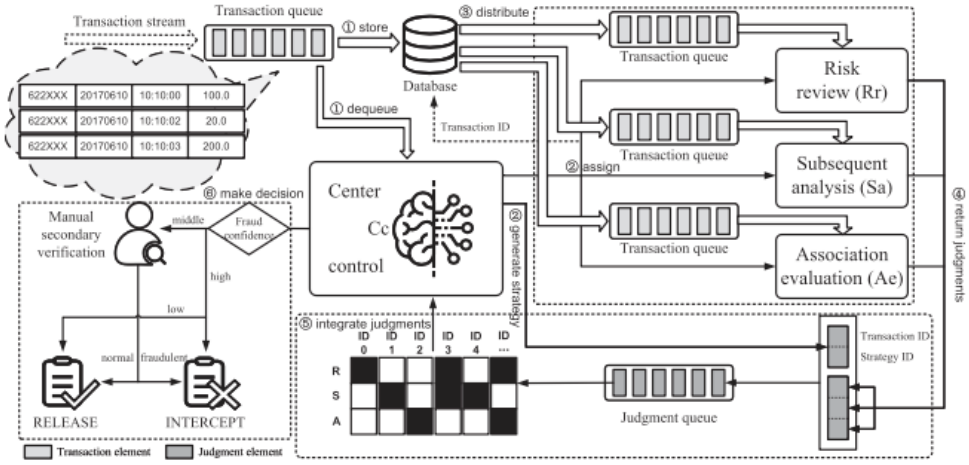

Figure 8 Workflow of CAeSaR Source: [22]

II. Observed patterns among top classifiers

71.4 % of the top models use SMOTE as (one of two) sampling technique to account for the class imbalance. The average reported approvement by Sampling is 0.159 in the f1-score. [19] Models that do use SMOTE-ENN rank among the top three.

None of the datasets used by the top-performing models contain PD. The first model that does ranks 10th. All but [22] use the comparatively small European dataset. [22] ranks 7th and has about 10 times more transactions to train on and classify.

B. Adaptability to adopt to concept drift

One way to measure adaptability to concept drift is to look at the performance over a course of time. There were some difficulties in the assessment of this

[19] The following f1-values without prior sampling were reported: [15]* 0.935; [16]* 0.665; [19] 0.863 [23] 0.690; [26] 0.830. Results marked '*' use SOMTE-ENN.

requirement[20], hence we had to pivot and investigate the performance of the AI systems in a more generalist way.

The dataset used in [22] with a sample size of more than 2m transactions that have been assigned their fraud labels through careful investigation. The reliability of the data for concept drift therefore is much higher. The authors describe that the Center control is retrained at regular intervals based on Reinforcement Learning [22:2570].

C. Speed

Again, only [22] includes data on the speed of a classification for one single transaction. It reports the CAeSAr was the fastest of all tested models at 19.7 ms. They attribute that to their novel Center control architecture which, in contrast to the other methods, does not require all function models to be run. [22:2573]

D. Explainability

Not all papers who reported using XAI[21] produce explainable output, e.g. visuals or feature tables. We will only present the seven papers with actual explainability output.

Explainability in [1] is achieved by outputting a table explaining the highest weighted transactions and features across a sequence of transactions for TP, FN, FP. The output for TP and TN is cohesive and comprehensive, giving a good description of the decision process. For the FP, the description is cohesive, the highest weighted feature (‘transaction status’), however, is not explained in the paper, nor can it be made sense of.

[4] provides examples in a beeswarm plot and a heatmap for two unidentifiable classifications. None deliver specific explanations.

The authors of [5,17,20] present global explainability in the form of a SHAP diagram (Figure 9), with pink colour indicating a high relevance and blue a low one. On the X-axis they report the SHAP value. The Y-axis indicates the feature names in descending order of importance. Pink means high feature value, blue low. Each dot represents a transaction. For instance, ratio to median purchase price has a significant impact on the prediction. Depending on its value, it can either forecast fraud or legitimacy.

---

[20] See below §3.**Error! Reference source not found.**.

[21] N(XAI) = 12.

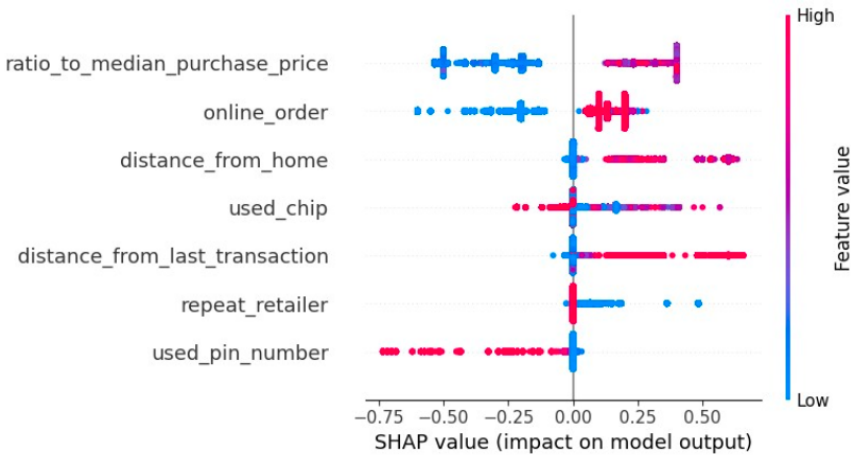

Figure 9 SHAP diagram Source: [5]

[20] includes SHAP diagrams that do not have any qualitative significance as to what drives a decision due to PCA performed on dataset. Time is 10th variable on impact on mode output.

Figure 10 shows the local explanation for [22], with a black label indicating a fraudulent transaction. Comparing the records, the current transaction (record 5) matches or closely resembles the features of the fraudulent records 2-4.

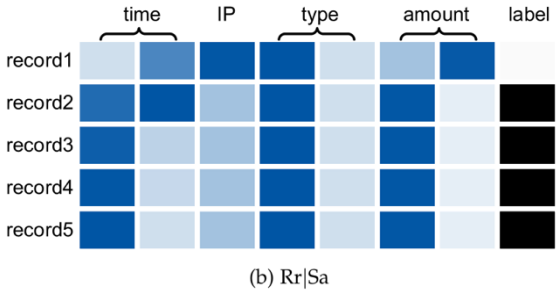

Figure 10 Explanation visualisation in [22], with record 1-4 being historic transactions and record 5 the predicted transaction.

[24] provides diagrams for the AutoEncoder, the C explainer and a general explainer. The explanations were not comprehensible to the author.

E. Privacy compliance

Only four of the datasets are cited to contain raw PD. Only two [3,18] ranks in the ten best-performing models.

[3] pseudonymised payer and payee names by assigning them IDs. [22] Their study uses anomaly detection and tracks if an ID is matched with several instances of fraud or where the payee lives. [3:252] reports that those three features are the most significant in prediction. [18] uses the geographic location in which the transaction was initiated. [8] handles raw account numbers, from both payer and payee. In [17], most PD is used: from names, behaviour data, to location data, most of the 24 features contain PD. Neither [17] nor [8] report employing any privacy-preserving measures.

F. Limitations

We have to assume that the fraud labels contain noise, i.e. it is possible that false positives are actually true positives because they were incorrectly labelled non-fraudulent by banks (Nesvijevskaia et al. 2021, 9). It is not clear how the bank has established their labelling decision. Keeping this in mind, some models may have performed better than reported due to an increased number of false predictions.

None of the top-performing algorithms are unsupervised. This is unexpected compared to the reported results in literature.

[22] The dataset is synthetic, however, the model would process personal data so for the purposes of evaluating privacy it will be considered as containing personal data.

§ 5 Discussion

The above evaluation has uncovered whether the defined requirements can be met by the 25 papers we have analysed. The findings must be matched with current industry insights resulting in a proposal for banks.

A. Prediction performance

The performance of the reviewed papers was already good with the top performer scoring 0.998 at f1, 1 at recall and 0.997 at precision. This indicates no FP and few FN, i.e. no wrongly blocked transactions and few fraudulent transactions being processed. These results are good considering the involved costs and expected customer satisfaction.

I. Factors

Comparing the AI models in [14-16] is feasible as they use the same dataset, and similar pre-processing, presenting an equivalent setting for comparison.

Poor label quality needs to be improved before introducing data for training. This especially true for supervised and semi-supervised models which are dependent on the accuracy of the label.

1. Ensemble learning

Ten of the papers have used ensemble learning already, others have mentioned that they see a combination of their tested systems as an opportunity to improve results. This suggestion can be affirmed by the good results ensemble learners have shown [14,15,20,22,25]. This is underlined by the excellent comparability of [14,15] who use the same dataset and pre-processing.

Supervised models may overfit training data, resulting in poor generalisation when introduced to atypical new data. Unsupervised learning is highly reliant on the quality of the data and may necessitate human verification of the results. In imbalanced datasets like fraud detection, supervised learners tend to focus on the majority class which can lead to increased error rates for minority classes. Therefore, it would be advised to combine a supervised and unsupervised algorithm.

Based on the multiple good results, we suggest including LSTM in the ensemble. LSTMs always need data to work on and therefore will need at least five prior transactions to reach ample performance *(Jurgovsky et al. 2018, 239)*. At the same time, they memorise history and recognise irregularities,

e.g. unusual withdrawals or spending behaviour (Xiuguo and Shengyong 2022, 22518). Due to its advantages and reported results in literature, the ensemble should also include a supervised learner. Further, the reinforcement method AdaCost [26] should be considered to account for cost reduction.

2. Sampling with SMOTE-ENN

71.4 % of the top models use SMOTE as (one of two) sampling technique to account for the class imbalance. Models that do use SMOTE-ENN rank among the top three. The models using SMOTE are using the same European dataset. Thus, it is possible that SMOTE only works well for this specific dataset. Unfortunately, the only other research using SMOTE without using that specific dataset [18] does not report any data without using SMOTE. However, the effects of SMOTE in combatting imbalanced data has been shown in CCFD for over a decade *(Jurgovsky et al. 2018, 235)*.

According to the findings of [26] using SMOTE could have saved banks over 2 million euros.[23] They criticise SMOTE not being efficient enough in cost-sensitive problems (Yang, Liu, and Liu 2020, 421). Based on the results in [14-16] adding ENN should lead to more savings.

3. Feature engineering

Bahnsen et. al propose keeping a transaction history by aggregative past transaction and matching that against every future transaction to ascertain its legitimacy in a new feature (2016). The algorithm can be set to match for any given timeframe and element, such as merchant category. Feature aggregation is more effective using online transactions than card present purchases (Jurgovsky et al. 2018, 241). As discussed above, the cost-sensitivity measure has been a relevant contributor [26]. It will be discussed below in §5.C.I.

4. Blockchain

Authors of [20] are not convinced using blockchain is feasible for CCFD due to privacy concerns and the needed time and other resources. They do not address the security issues that might come with using the blockchain and how public it is. The advantages are not convincing enough to consider blockchain fur further research.

[23] Calculated based on the results of the model with and without using SMOTE.

5. Webscraping

Web scraping is employed in various fields, such as identifying copyright violations (PromptCloud 2016). It also has potential in identifying new fraud schemes or detecting leaked PD. The exposure of credit card data is a risk factor for CCF (Nesvijevskaia et al. 2021, 7) and employing this method could enhance predictive quality and improve customer satisfaction by providing personalised attention.

B. Adaptability to concept drift

We were unable to make reliable findings on the adaptability to concept drift because of (1) the short time span covered by the datasets, (2) inadequate results output, and (3) poor labelling.

50% of the models used a 2013 dataset from a European bank. The transactions all took place in one month. We were unable to retain insights on the trends and concept drift at that time consulting the European Central Bank's Fraud Report (2015) for that year. While it cannot be ruled out with absolute certainty that some concept drift did occur, it is unlikely that it made a significant difference over such a short period of time. Apart from that, payment behaviour and fraud schemes have evolved in the past eleven years. Thus, it is uncertain whether the tested models will be able to detect current fraud schemes.

None of the papers included graphs on the prediction for performance over time. If a time graph shows significant performance decrease after a time that can indicate concept drift (Pratt and Tschapek 2003). Hence, such visuals could have aided in evaluating the concept drift adaptability.

Poor labelling, i.e. fraudulent transaction prematurely marked as legitimate, can contribute to concept drift. Most datasets follow this premature approach and therefore contain wrong labels (Dal Pozzolo et al. 2015, 1). When no description is provided, it must be assumed that at least a small fraction of labels is wrong. The labelling issue can be solved by better CCFD and using unsupervised ML.

Ensemble learning with model weighting can also combat concept drift. It will leverage the differing detection methods and then create an average that will decide whether a new pattern has emerged. (Minku, White, and Yao 2009)

The only way to really account for prediction performance under concept drift is to monitor the model over a longer period. From literature review and with the use of a diversified ensemble learner, ideally including Random Forest or a Genetic Algorithm, we are confident that concept drift can be detected and be accounted for in the predictions. The adaptability to new fraud

schemes will become even more important in the feature with fraudsters increasingly employing more sophisticated schemes. (Lucas and Jurgovsky 2020, 14–15)

C. Customer satisfaction

LexisNexis has interviewed 1,181 crime compliance strategists at global financial institutions in 2023 and compiled their findings in a report: 82% were focussed on compliance with regulations and 83% on strengthening their fraud management by using data. The highest priority, however, was improving customer experience while improving fraud management. (LexisNexis 2023, 6) 79% of Germans believe they are increasingly at risk of becoming victims of cybercrime. About half of the Germans also believe they are unable to protect themselves or are protected by websites. (Eurobarometer 2020)

Because of these statistics, it can be assumed that customers will welcome automatic rejections of suspicious transactions. Once a credit card transaction is processed, it cannot be revoked, even when reported immediately. According to the author's experience, this leads to frustration with customers and drives them to contact customer service. It is likely that the transaction will not be reimbursed. To increase customer satisfaction, it is therefore advisable to block transactions, especially for larger sums.

I. Cost sensitivity

Achituve et al. report that banks often set a probability threshold that determines when transactions need to be blocked. The reason for this is user experience. [1:4] The chosen thresholds are not publicised and therefore could not be used for this research.

As reported above, [26] has developed a novel feature for including costs in CCFD. This aggregate feature must be combined with other AI to improve overall performance. [1] supports that notion.

II. Speed

The majority of banks agree that speed is a critical factor in successful CCFD. (LexisNexis 2023, 7) As cited in §1.A.II.1, speed is essential in keeping bounce rates low.

The speed of classifying single transactions can depend on various factors such as the computational resources, the complexity of the model, and the size of the dataset. In real-time CCFD systems, the classification of single

transactions is typically expected to be performed within milliseconds to ensure timely identification of fraudulent activities (Chen and Lai 2021, 107).

The efficiency and speed of the classification process is crucial for real-time CCFD. The reported speed of 19 ms [22] is within a reasonable timeframe for a fraud detection mechanism.

To ensure high speed, online learning needs to be implemented. In order for online learners to operate at high speed, the data cannot be too complex which highlights the advantages of feature engineering. The ML should be connected to a bank API that feeds it with aggregated features of a pending transaction. (Thennakoon et al. 2019, 492)[24]

D. Explainability

Being able to explain why a transaction was blocked is important for compliance reasons. However, explainability may conflict with two objectives: prediction performance and confidentiality. Furthermore, explainability has to be provided for two different groups.

Other literature suggests that not all XAI is trustworthy. A major issue in explainability is the reliability of the output: not all techniques are equally robust. LIME is volatile, i.e. it will output different results for the same model. It is therefore not recommended to be implemented for legally-mandated explanations. (Alvarez-Melis and Jaakkola 2018, 69; Molnar 2020, 217) The SHAP value is a robust metric. It does take more computation power, however, explanations are not generated under time pressure. Thus, speed is not a relevant factor. (Psychoula et al. 2021, 52)

I. Explainability vs. prediction

Out of the top 7, only [22] produce relevant explanations. The other top performers did not consider explainability for undisclosed reasons. From these results we can still not reliably conclude that explainability and good prediction are mutually exclusive.

II. Explainability vs. confidentiality

Fraud detection systems are often kept confidential to prevent fraudsters from adjusting their techniques to evade detection. Disclose of AI details could potentially enable fraudsters to adapt their approach to bypass the

[24] Find a real-life example of this below §5.E.III.

prevention mechanisms. The objective of the processing can be compromised by issuing too much information. This is a limit that is recognised by the GDPR and will therefore have to applied to this scenario as well. Why-explanations will suffice for individual decisions, while privacy policies will need how-explanations. (Mohseni, Zarei, and Ragan 2021, 10; Molnar 2020, 18)

III. Explainability for customers

Customers must be made aware as to how their PD is used for making predictions. The explanations must be provided in a user-centric way, considering their lack of knowledge on AI (Cirqueira, Helfert, and Bezbradica 2021a, 2). People want short and easily-understandable explanations, even if that does not reproduce the complexity of the system (Molnar 2020, 37). This can be ensured using visualisations or even interactive diagrams (Mill et al. 2023, 11). We have found the table provided in [1] to be understood easily for local results. It provides the information required by the GDPR in simple laymen's terms. For global explanations, the SHAP value can be used in the form of a diagram, as shown in [5] (cf. Figure 9).

Customers are concerned their data is being mishandled. Thus, it is essential to lower suspicion and be truthful and transparent in these explanations. From a business perspective and to increase the acceptance of AI, it is recommended to reassure customers that their account is safe and what the next steps are. (Wulf and Seizov 2022, 5)

IV. Explainability for professionals

Albeit a small sample size of 59, it is interesting to see that a majority of surveyed fraud experts found SHAP to be a trustworthy and helpful explainer (Ji 2021, 30). They need to be able to ascertain whether a flagged transaction is in fact fraudulent and why. [22:2575] Semantic explanations can assist fraud teams in understanding trends and will significantly reduce their workloads (Cirqueira, Helfert, and Bezbradica 2021b, 32 f). Understanding how a prediction was made will also shed light onto emerging fraud schemes (Datatilsynet 2023). These insights can aide in developing customer education and prevention programmes (Gianotti and Damião da Silva 2021, 160).

E. Privacy compliance

Our research shows that PD is not needed for CCFD.

I. Performance of models using PD

The top eight models either use pseudonymised data or no PD. It was impossible to ascertain whether the European dataset used pseudonymised or anonymised PD.

It could not be seen from the collective results whether the PD in the European datasets was pseudonymised or anonymised. Anonymity would be achieved if no transaction could be retraced by using a different feature in the set.

The use of PD in [3] raises several privacy concerns: First, their pseudonymisation technique of simply assigning IDs is not sufficient to protect. Second, it is possible that with the matching of IDs negative consequences, such as investigations, arise. This is especially bad since not just the bank's customer as either a payer or payee is affected, but also a third party sending or receiving a transaction deemed fraudulent. Others propose that differential privacy is a promising solution direction where data are more aggregated (van Bruxvoort and van Keulen 2021, 7; Dwork 2011, 88)

Even if PD were useful in increasing prediction performance, research shows that customers would be willing to sacrifice some of their privacy rights to adequately protect themselves from fraud: 70% prefer a suspicious transaction be blocked immediately, rather than being notified at the end of the day without blocking. (Gabudeanu et al. 2021, 15)

The GDPR mandates that it is ascertained whether PD is necessary for the robust CCFD. Therefore, the banks must evaluate whether performance requires PD for high-quality predictions. If it does, feature engineering and pseudonymisation must be used. When PD processing must cease, a machine unlearning method must be applied.

II. No bias/discrimination

Both the GDPR and the AI Act place a strong emphasis on a zero tolerance for bias and discrimination. The risk of bias and discrimination is significantly reduced when no PD included.

Starting out, the data introduced for training has to be of good quality, i.e. be sampled to achieve balance and exclude any factors that have a high potential for prejudice, e.g. name, gender, residence location. This is easily achieved if there is no (raw) PD in the data. (Bao, Hilary, and Ke 2022, 230)

III. Industry example

We can learn from German debt collector PAIR leverages AI to find out how to receive a response from debtors and recommend a promising payment plan. They employ a reinforcement learner called *QLearning*. The relevant data is stored in two different databases: the AI is fed inferred and aggregated data stored in a separate database. The two are connected via API. An example for this inferred data: one datapoint could be 'name contains first and last name' rather than the raw email address. The PAIR approach has three significant advantages as the separate databases and inferred/aggregated data foster privacy: (1) The AI does not handle raw PD. (2) PD can be deleted when needed by removing it from the database.[25] (3) The AI gets a lot less complex and therefore computation time can potentially be reduced. This preserves privacy without sacrificing prediction power. (Gaub and Kadler 2022, 62) Due to the lack of raw PD, discrimination can also be limited.

F. Limitations

Throughout this paper the lack of industry and customer focus has been criticised. During the evaluation of this thesis, the author also hit limitations due to a lack of contact with banks and their customers. With her private practice experience she can estimate the customer needs, but not the limitations of banks. To really be able to answer the research question, more industry data is needed.

[25] Authors warn that data needs to be pseudonymised further so they cannot be traced back to an individual.

§ 6 Conclusion & Outlook

Based on our findings, CCFD AI has to comply with GDPR and BGB laws. Additionally, it will be regulated by the new PSD3 and PSR as well as the AI Act. According to the selected metrics, we found that ensemble learners with LSTM with SMOTE-ENN sampling and different feature engineering methods are algorithms that can predict CCFD reliably. These predictions can be made with limited to no PD, also reducing the risk of bias. While useful explanations can already be generated, their faithfulness has to be improved We can therefore recommend adopting automatic AI in CCFD efforts. This will improve customer satisfaction by more proactively addressing fraud.

The TCP feature developed by Yang et. al represents those industry needs in one feature: By learning from false predictions, it can improve prediction performance. That is combined with the overall consideration of costs and customer satisfaction. This feature should be re-evaluated with other observations that promise improvements in the other domains. The findings need to be discussed with industry experts and learn from other sectors to be even more reliable.

Several EU regulations were passed recently that will change the dynamics of the researched issue. The new Data Governance Act facilitates data sharing among banks. This promises further research on this topic and wider availability of datasets. The learnings can further be explored in regulatory sandboxes introduced by the novel AI Act. Additionally, the new PSR will improve security measure and customer education which might decrease, albeit miniscule, the number of fraud.

When looking at privacy concerns, prior research has rarely included legislation, customer perspectives and technical concerns in one paper. This thesis is a starting point for a combined approach of legal and technological viewpoints. A supplementation with industry insights promises even more reliable results.

Appendix

Appendix 1 Overview of studies

	Authors	Title	Data-set	Sample Year	Sample Size
1	**Achituve et. al.** (2019)	Interpretable Online Banking Fraud Detection Based On Hierarchical Attention Mechanism	Real	2017	26,100,000
2	**Alfaiz & Fati** (2022)	Enhanced Credit Card Fraud Detection Model Using Machine Learning	Real	2013	284,807
3	**Bhat et. al.** (2022)	Qualitative Analysis of Anomaly Detection in Time Series	Synth	-	82,864
4	**Bhowmik et. al.** (2022)	DBNex: Deep Belief Network and Explainable AI based Financial Fraud Detection	Real	2013	284,807
5	**Biswas et. al.** (2023)	Interpretable Credit Card Fraud Detection Using Machine Learning Leveraging SHAP	Synth	-	1,000,000
6	**Boulieris et. al.** (2023)	Fraud detection with natural language processing	Real	2020	105,303
7	**Cherkaoui & En-Naimi** (2023)	A comparison of machine learning algorithms for credit card fraud detection	Real	2013	284,808
8	**Hanae et. al.** (2023)	End-to-End Real-time Architecture for Fraud Detection in Online Digital Transactions	Synth	-	1,000,000
9	**Hsin et. al.** (2022)	Feature Engineering and Resampling Strategies for Fund Transfer Fraud with Limited Transaction Data and a Time-Inhomogeneous Modi Operandi	Real	2018	-
10	**Huang et. al.** (2021)	Decision Analysis and Prediction Based on Credit Card Fraud Data	Real	2013	284,807
11	**Leevy et. al.** (2023)	Investigating the effectiveness of one-class and binary classification for fraud detection	Real	2013	284,807
12	**Liang et. al.** (2022)	Self-supervised Pretraining Isolated Forest for Outlier Detection	Real	2013	284,807
13	**Lin & Jiang** (2021)	Credit card fraud detection with autoencoder and probabilistic random forest	Real	2013	284,807
14	**Mienye & Sun** (2023)	A Machine Learning Method with Hybrid Feature Selection for Improved Credit Card Fraud Detection	Real	2013	284,807
15	**Mienye & Sun** (2023)	A Deep Learning Ensemble With Data Resampling for Credit Card Fraud Detection	Real	2013	283,807
16	**Narayan & Ganapathisamy** (2022)	Hybrid Sampling and Similarity Attention Layer in Bidirectional Long Short Term Memory in Credit Card Fraud Detection	Real	2013	284,807
17	**Psychoula et. al.** (2021)	Explainable Machine Learning for Fraud Detection	Real	2017	2,856,000
18	**Raj et. al.** (2023)	Enhancing Security for Online Transactions through Supervised Machine Learning and Block Chain Technology in Credit Card Fraud Detection	Real	-	500,000
19	**Ranjbaran et. al.** (2023)	Leveraging augmentation techniques for tasks with unbalancedness within the financial domain: a two-level ensemble approach	Real	-	284,315
20	**Raval et. al.** (2023)	RaKShA: A Trusted Explainable LSTM Model to Classify Fraud Patterns on Credit Card Transactions	Real	2013	284,808
21	**Ti et. al.** (2022)	Feature generation and contribution comparison for electronic fraud detection	Real	-	-
22	**Wang et. al.** (2023)	CAeSaR: An Online Payment Anti-Fraud Integration System With Decision Explainability	Real	2017	2,856,000
23	**Wu et. al.** (2023)	Advancement of management information system for discovering fraud in master card based intelligent supervised machine learning and deep learning during SARS-CoV2	Real	-	-
24	**Wu & Wang** (2021)	Locally Interpretable One-Class Anomaly Detection for Credit Card Fraud Detection	Real	-	-
25	**Xu et. al.** (2023)	Efficient fraud detection using deep boosting decision trees	Real	2013	284,807
26	**Yang et. al.** (2020)	Credit Card Fraud Detection Based on CSat-Related AdaBoost	Real	2013	284,807

Note: Real = Real World; Synth = Synthetic; - = Not specified