Abchiche-Mimouni, Nadia, Leila Amgoud, and Farida Zehraoui. 2023. “Explainable Ensemble Classification Model Based on Argumentation.” In *AAMAS*, 2367–69. https://doi.org/10.5555/3545946.3598936.

Amgoud, Leila. 2021a. “Explaining Black-Box Classification Models with Arguments.” In *ICTAI*, 791–95. https://doi.org/10.1109/ICTAI52525.2021.00126.

———. 2021b. “Non-Monotonic Explanation Functions.” In *ECSQARU*, 19–31. https://doi.org/10.1007/978-3-030-86772-0\2.

———. 2023. “Explaining Black-Box Classifiers: Properties and Functions.” *Int. J. Approx. Reason.* 155: 40–65. https://doi.org/10.1016/J.IJAR.2023.01.004.

Amgoud, Leila, and Jonathan Ben-Naim. 2022. “Axiomatic Foundations of Explainability.” In *IJCAI*, 636–42. https://doi.org/10.24963/IJCAI.2022/90.

Amgoud, Leila, Philippe Muller, and Henri Trenquier. 2023a. “Argument-Based Explanation Functions.” In *AAMAS*, 2373–75. https://doi.org/10.5555/3545946.3598938.

———. 2023b. “Leveraging Argumentation for Generating Robust Sample-Based Explanations.” In *IJCAI*, 3104–11. https://doi.org/10.24963/IJCAI.2023/346.

Arenas, Marcelo, Daniel Báez, Pablo Barceló, Jorge Pérez, and Bernardo Subercaseaux. 2021. “Foundations of Symbolic Languages for Model Interpretability.” In *NeurIPS*, 11690–701. https://proceedings.neurips.cc/paper/2021/hash/60cb558c40e4f18479664069d9642d5a-Abstract.html.

Arenas, Marcelo, Pablo Barceló, Leopoldo E. Bertossi, and Mikaël Monet. 2021. “The Tractability of SHAP-Score-Based Explanations for Classification over Deterministic and Decomposable Boolean Circuits.” In *AAAI*, 6670–78. https://doi.org/10.1609/AAAI.V35I8.16825.

———. 2023. “On the Complexity of SHAP-Score-Based Explanations: Tractability via Knowledge Compilation and Non-Approximability Results.” *J. Mach. Learn. Res.* 24: 63:1–58. http://jmlr.org/papers/v24/21-0389.html.

Arenas, Marcelo, Pablo Barceló, Miguel A. Romero Orth, and Bernardo Subercaseaux. 2022. “On Computing Probabilistic Explanations for Decision Trees.” In *NeurIPS*. http://papers.nips.cc/paper\files/paper/2022/hash/b8963f6a0a72e686dfa98ac3e7260f73-Abstract-Conference.html.

Audemard, Gilles, Steve Bellart, Louenas Bounia, Frédéric Koriche, Jean-Marie Lagniez, and Pierre Marquis. 2021. “On the Computational Intelligibility of Boolean Classifiers.” In *KR*, 74–86. https://doi.org/10.24963/KR.2021/8.

———. 2022a. “On Preferred Abductive Explanations for Decision Trees and Random Forests.” In *IJCAI*, 643–50. https://doi.org/10.24963/IJCAI.2022/91.

———. 2022b. “Trading Complexity for Sparsity in Random Forest Explanations.” In *AAAI*, 5461–69. https://doi.org/10.1609/AAAI.V36I5.20484.

Audemard, Gilles, Steve Bellart, Jean-Marie Lagniez, and Pierre Marquis. 2023. “Computing Abductive Explanations for Boosted Regression Trees.” In *IJCAI*, 3432–41. https://doi.org/10.24963/IJCAI.2023/382.

Audemard, Gilles, Frédéric Koriche, and Pierre Marquis. 2020. “On Tractable XAI Queries Based on Compiled Representations.” In *KR*, 838–49. https://doi.org/10.24963/KR.2020/86.

Audemard, Gilles, Jean-Marie Lagniez, Pierre Marquis, and Nicolas Szczepanski. 2023a. “Computing Abductive Explanations for Boosted Trees.” In *AISTATS*, 4699–4711. https://proceedings.mlr.press/v206/audemard23a.html.

———. 2023b. “On Contrastive Explanations for Tree-Based Classifiers.” In *ECAI*, 117–24. https://doi.org/10.3233/FAIA230261.

Bassan, Shahaf, and Guy Katz. 2023. “Towards Formal XAI: Formally Approximate Minimal Explanations of Neural Networks.” In *TACAS*, 187–207. https://doi.org/10.1007/978-3-031-30823-9\10.

Béjar, Ramón, António Morgado, Jordi Planes, and João Marques-Silva. 2023. “On Logic-Based Explainability with Partially Specified Inputs.” *CoRR* abs/2306.15803. https://doi.org/10.48550/ARXIV.2306.15803.

Biradar, Gagan, Yacine Izza, Elita Lobo, Vignesh Viswanathan, and Yair Zick. 2023. “Axiomatic Aggregations of Abductive Explanations.” *CoRR* abs/2310.03131. https://doi.org/10.48550/ARXIV.2310.03131.

Boumazouza, Ryma, Fahima Cheikh Alili, Bertrand Mazure, and Karim Tabia. 2020. “A Symbolic Approach for Counterfactual Explanations.” In *SUM*, 270–77. https://doi.org/10.1007/978-3-030-58449-8\21.

———. 2021. “ASTERYX: A Model-Agnostic SaT-basEd appRoach for sYmbolic and Score-Based eXplanations.” In *CIKM*, 120–29. https://doi.org/10.1145/3459637.3482321.

———. 2023. “Symbolic Explanations for Multi-Label Classification.” In *ICAART*, 342–49. https://doi.org/10.5220/0011668700003393.

Carbonnel, Clément, Martin C. Cooper, and João Marques-Silva. 2023. “Tractable Explaining of Multivariate Decision Trees.” In *KR*, 127–35. https://doi.org/10.24963/KR.2023/13.

Colnet, Alexis de, and Pierre Marquis. 2022. “On the Complexity of Enumerating Prime Implicants from Decision-DNNF Circuits.” In *IJCAI*, 2583–90. https://doi.org/10.24963/IJCAI.2022/358.

Cooper, Martin C., and João Marques-Silva. 2021. “On the Tractability of Explaining Decisions of Classifiers.” In *CP*, 21:1–18. https://doi.org/10.4230/LIPICS.CP.2021.21.

———. 2023. “Tractability of Explaining Classifier Decisions.” *Artif. Intell.* 316: 103841.

https://doi.org/10.1016/J.ARTINT.2022.103841.

Cooper, Martin, and Leila Amgoud. 2023. “Abductive Explanations of Classifiers Under Constraints: Complexity and Properties.” In *ECAI*, 469–76. https://doi.org/10.3233/FAIA230305.

Darwiche, Adnan. 2020. “Three Modern Roles for Logic in AI.” In *PODS*, 229–43. https://doi.org/10.1145/3375395.3389131.

———. 2023. “Logic for Explainable AI.” In *LICS*, 1–11. https://doi.org/10.1109/LICS56636.2023.10175757.

Darwiche, Adnan, and Auguste Hirth. 2020. “On the Reasons Behind Decisions.” In *ECAI*, 712–20. https://doi.org/10.3233/FAIA200158.

———. 2023. “On the (Complete) Reasons Behind Decisions.” *J. Log. Lang. Inf.* 32 (1): 63–88. https://doi.org/10.1007/S10849-022-09377-8.

Darwiche, Adnan, and Chunxi Ji. 2022. “On the Computation of Necessary and Sufficient Explanations.” In *AAAI*, 5582–91. https://doi.org/10.1609/AAAI.V36I5.20498.

Darwiche, Adnan, and Pierre Marquis. 2021. “On Quantifying Literals in Boolean Logic and Its Applications to Explainable AI.” *J. Artif. Intell. Res.* 72: 285–328. https://doi.org/10.1613/JAIR.1.12756.

———. 2022. “On Quantifying Literals in Boolean Logic and Its Applications to Explainable AI (Extended Abstract).” In *IJCAI*, 5718–21. https://doi.org/10.24963/IJCAI.2022/797.

Gorji, Niku, and Sasha Rubin. 2022. “Sufficient Reasons for Classifier Decisions in the Presence of Domain Constraints.” In *AAAI*, 5660–67. https://doi.org/10.1609/AAAI.V36I5.20507.

Huang, Xuanxiang, Martin C. Cooper, António Morgado, Jordi Planes, and João Marques-Silva. 2023. “Feature Necessity & Relevancy in ML Classifier Explanations.” In *TACAS*, 167–86. https://doi.org/10.1007/978-3-031-30823-9\9.

Huang, Xuanxiang, Yacine Izza, Alexey Ignatiev, Martin C. Cooper, Nicholas Asher, and João Marques-Silva. 2022. “Tractable Explanations for d-DNNF Classifiers.” In *AAAI*, 5719–28. https://doi.org/10.1609/AAAI.V36I5.20514.

Huang, Xuanxiang, Yacine Izza, Alexey Ignatiev, and João Marques-Silva. 2021. “On Efficiently Explaining Graph-Based Classifiers.” In *KR*, 356–67. https://doi.org/10.24963/KR.2021/34.

Huang, Xuanxiang, Yacine Izza, and João Marques-Silva. 2023. “Solving Explainability Queries with Quantification: The Case of Feature Relevancy.” In *AAAI*, 3996–4006. https://doi.org/10.1609/AAAI.V37I4.25514.

Huang, Xuanxiang, and João Marques-Silva. 2023a. “From Decision Trees to Explained Decision Sets.” In *ECAI*, 1100–1108. https://doi.org/10.3233/FAIA230384.

———. 2023b. “From Robustness to Explainability and Back Again.” *CoRR* abs/2306.03048. https://doi.org/10.48550/ARXIV.2306.03048.

Hurault, Aurélie, and João Marques-Silva. 2023. “Certified Logic-Based Explainable AI - the Case of Monotonic Classifiers.” In *TAP*, 51–67. https://doi.org/10.1007/978-3-031-38828-6\4.

Ignatiev, Alexey. 2020. “Towards Trustable Explainable AI.” In *IJCAI*, 5154–58. https://doi.org/10.24963/IJCAI.2020/726.

Ignatiev, Alexey, Martin C. Cooper, Mohamed Siala, Emmanuel Hebrard, and João Marques-Silva. 2020. “Towards Formal Fairness in Machine Learning.” In *CP*, 846–67. https://doi.org/10.1007/978-3-030-58475-7\49.

Ignatiev, Alexey, Yacine Izza, Peter J. Stuckey, and João Marques-Silva. 2022. “Using MaxSAT for Efficient Explanations of Tree Ensembles.” In *AAAI*, 3776–85. https://doi.org/10.1609/AAAI.V36I4.20292.

Ignatiev, Alexey, and João Marques-Silva. 2021. “SAT-Based Rigorous Explanations for Decision Lists.” In *SAT*, 251–69. https://doi.org/10.1007/978-3-030-80223-3\18.

Ignatiev, Alexey, Nina Narodytska, Nicholas Asher, and João Marques-Silva. 2020. “From Contrastive to Abductive Explanations and Back Again.” In *AIxIA*, 335–55. https://doi.org/10.1007/978-3-030-77091-4\21.

Ignatiev, Alexey, Nina Narodytska, and João Marques-Silva. 2019a. “Abduction-Based Explanations for Machine Learning Models.” In *AAAI*, 1511–19. https://doi.org/10.1609/AAAI.V33I01.33011511.

———. 2019b. “On Relating Explanations and Adversarial Examples.” In *NeurIPS*, 15857–67. https://proceedings.neurips.cc/paper/2019/hash/7392ea4ca76ad2fb4c9c3b6a5c6e31e3-Abstract.html.

Izza, Yacine, Xuanxiang Huang, Alexey Ignatiev, Nina Narodytska, Martin C. Cooper, and João Marques-Silva. 2023. “On Computing Probabilistic Abductive Explanations.” *Int. J. Approx. Reason.* 159: 108939. https://doi.org/10.1016/J.IJAR.2023.108939.

Izza, Yacine, Alexey Ignatiev, and João Marques-Silva. 2022. “On Tackling Explanation Redundancy in Decision Trees.” *J. Artif. Intell. Res.* 75: 261–321. https://doi.org/10.1613/JAIR.1.13575.

———. 2023. “On Tackling Explanation Redundancy in Decision Trees (Extended Abstract).” In *IJCAI*, 6900–6904. https://doi.org/10.24963/IJCAI.2023/779.

Izza, Yacine, Alexey Ignatiev, Peter J. Stuckey, and João Marques-Silva. 2023. “Delivering Inflated Explanations.” *CoRR* abs/2306.15272. https://doi.org/10.48550/ARXIV.2306.15272.

Izza, Yacine, and João Marques-Silva. 2021. “On Explaining Random Forests with SAT.” In *IJCAI*, 2584–91. https://doi.org/10.24963/IJCAI.2021/356.

Ji, Chunxi, and Adnan Darwiche. 2023. “A New Class of Explanations for Classifiers with Non-Binary Features.” In *JELIA*, 106–22. https://doi.org/10.1007/978-3-031-43619-2\8.

Liu, Xinghan, and Emiliano Lorini. 2021. “A Logic for Binary Classifiers and Their Explanation.” In *CLAR*, 302–21. https://doi.org/10.1007/978-3-030-89391-0\17.

———. 2022. “A Logic of "Black Box" Classifier Systems.” In *WoLLIC*, 158–74. https://doi.org/10.1007/978-3-031-15298-6\10.

———. 2023. “A Unified Logical Framework for Explanations in Classifier Systems.” *J. Log. Comput.* 33 (2): 485–515. https://doi.org/10.1093/LOGCOM/EXAC102.

Malfa, Emanuele La, Rhiannon Michelmore, Agnieszka M. Zbrzezny, Nicola Paoletti, and Marta Kwiatkowska. 2021. “On Guaranteed Optimal Robust Explanations for NLP Models.” In *IJCAI*, 2658–65. https://doi.org/10.24963/IJCAI.2021/366.

Marques-Silva, João. 2022. “Logic-Based Explainability in Machine Learning.” In *Reasoning Web*, 24–104. https://doi.org/10.1007/978-3-031-31414-8\2.

———. 2023. “Disproving XAI Myths with Formal Methods - Initial Results.” In *27th International Conference on Engineering of Complex Computer Systems, ICECCS 2023, Toulouse, France, June 14-16, 2023*, edited by Yamine Aït-Ameur, Ferhat Khendek, and Dominique Méry, 12–21. IEEE. https://doi.org/10.1109/ICECCS59891.2023.00012.

Marques-Silva, João, Thomas Gerspacher, Martin C. Cooper, Alexey Ignatiev, and Nina Narodytska. 2020. “Explaining Naive Bayes and Other Linear Classifiers with Polynomial Time and Delay.” In *NeurIPS*. https://proceedings.neurips.cc/paper/2020/hash/eccd2a86bae4728b38627162ba297828-Abstract.html.

———. 2021. “Explanations for Monotonic Classifiers.” In *ICML*, 7469–79. http://proceedings.mlr.press/v139/marques-silva21a.html.

Marques-Silva, João, and Alexey Ignatiev. 2022. “Delivering Trustworthy AI Through Formal XAI.” In *AAAI*, 12342–50. https://doi.org/10.1609/AAAI.V36I11.21499.

———. 2023. “No Silver Bullet: Interpretable ML Models Must Be Explained.” *Frontiers Artif. Intell.* 6. https://doi.org/10.3389/FRAI.2023.1128212.

Shi, Weijia, Andy Shih, Adnan Darwiche, and Arthur Choi. 2020. “On Tractable Representations of Binary Neural Networks.” In *KR*, 882–92. https://doi.org/10.24963/KR.2020/91.

Shih, Andy, Arthur Choi, and Adnan Darwiche. 2018. “A Symbolic Approach to Explaining Bayesian Network Classifiers.” In *IJCAI*, 5103–11. https://doi.org/10.24963/IJCAI.2018/708.

———. 2019. “Compiling Bayesian Network Classifiers into Decision Graphs.” In *AAAI*, 7966–74. https://doi.org/10.1609/AAAI.V33I01.33017966.

Van den Broeck, Guy, Anton Lykov, Maximilian Schleich, and Dan Suciu. 2021. “On the Tractability of SHAP Explanations.” In *AAAI*, 6505–13. https://doi.org/10.1609/AAAI.V35I7.16806.

———. 2022. “On the Tractability of SHAP Explanations.” *J. Artif. Intell. Res.* 74: 851–86. https://doi.org/10.1613/JAIR.1.13283.

Wäldchen, Stephan, Jan MacDonald, Sascha Hauch, and Gitta Kutyniok. 2021. “The Computational Complexity of Understanding Binary Classifier Decisions.” *J. Artif. Intell. Res.* 70: 351–87. https://doi.org/10.1613/JAIR.1.12359.

Wu, Min, Haoze Wu, and Clark W. Barrett. 2022. “VeriX: Towards Verified Explainability of Deep Neural Networks.” *CoRR* abs/2212.01051. https://doi.org/10.48550/ARXIV.2212.01051.

———. 2023. “VeriX: Towards Verified Explainability of Deep Neural Networks.” In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, edited by Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine. http://papers.nips.cc/paper\files/paper/2023/hash/46907c2ff9fafd618095161d76461842-Abstract-Conference.html.

Yu, Jinqiang, Alexey Ignatiev, and Peter J. Stuckey. 2023a. “From Formal Boosted Tree Explanations to Interpretable Rule Sets.” In *CP*, 38:1–21. https://doi.org/10.4230/LIPICS.CP.2023.38.

———. 2023b. “On Formal Feature Attribution and Its Approximation.” *CoRR* abs/2307.03380. https://doi.org/10.48550/ARXIV.2307.03380.

Yu, Jinqiang, Alexey Ignatiev, Peter J. Stuckey, Nina Narodytska, and João Marques-Silva. 2023. “Eliminating the Impossible, Whatever Remains Must Be True: On Extracting and Applying Background Knowledge in the Context of Formal Explanations.” In *AAAI*, 4123–31. https://doi.org/10.1609/AAAI.V37I4.25528.