1.	Record Nr.	UNISA996546839603316
	Autore	Fan Lixin
	Titolo	Digital Watermarking for Machine Learning Model [[electronic resource]] : Techniques, Protocols and Applications / / edited by Lixin Fan, Chee Seng Chan, Qiang Yang
	Pubbl/distr/stampa	Singapore : , : Springer Nature Singapore : , : Imprint : Springer, , 2023
	ISBN	981-19-7554-X
	Edizione	[1st ed. 2023.]
	Descrizione fisica	1 online resource (233 pages)
	Altri autori (Persone)	ChanChee Seng YangQiang
	Disciplina	005.82
	Soggetti	Machine learning Data protection Image processing—Digital techniques Computer vision Image processing Machine Learning Data and Information Security Computer Imaging, Vision, Pattern Recognition and Graphics Image Processing
	Lingua di pubblicazione	Inglese
	Formato	Materiale a stampa
	Livello bibliografico	Monografia
	Nota di contenuto	Part I. Preliminary -- Chapter 1. Introduction -- Chapter 2. Ownership Verification Protocols for Deep Neural Network Watermarks -- Part II Techniques -- Chapter 3. ModelWatermarking for Image Recovery DNNs -- Chapter 4. The Robust and Harmless ModelWatermarking -- Chapter 5. Protecting Intellectual Property of Machine Learning Models via Fingerprinting the Classification Boundary -- Chapter 6. Protecting Image Processing Networks via Model Water -- Chapter 7. Watermarks for Deep Reinforcement Learning -- Chapter 8. Ownership Protection for Image Captioning Models -- Chapter 9.Protecting Recurrent Neural Network by Embedding Key -- Part III Applications -- Chapter 10. FedIPR: Ownership Verification for Federated Deep Neural Network Models -- Chapter 11. Model Auditing For Data Intellectual Property .

Sommario/riassunto

Machine learning (ML) models, especially large pretrained deep learning (DL) models, are of high economic value and must be properly protected with regard to intellectual property rights (IPR). Model watermarking methods are proposed to embed watermarks into the target model, so that, in the event it is stolen, the model's owner can extract the pre-defined watermarks to assert ownership. Model watermarking methods adopt frequently used techniques like backdoor training, multi-task learning, decision boundary analysis etc. to generate secret conditions that constitute model watermarks or fingerprints only known to model owners. These methods have little or no effect on model performance, which makes them applicable to a wide variety of contexts. In terms of robustness, embedded watermarks must be robustly detectable against varying adversarial attacks that attempt to remove the watermarks. The efficacy of model watermarking methods is showcased in diverse applications including image classification, image generation, image captions, natural language processing and reinforcement learning. This book covers the motivations, fundamentals, techniques and protocols for protecting ML models using watermarking. Furthermore, it showcases cutting-edge work in e.g. model watermarking, signature and passport embedding and their use cases in distributed federated learning settings.