Candidato: Akinola Akinjolagbe Samuel Folabi
E-mail: akinjolagbe.akinola@stud.unifi.it

Titolo: Explainable AI: studio ed implementazione di tecniche di machine learning interpretabili

Relatore: Andrea Ceccarelli
E-mail: andrea.ceccarelli@unifi.it

La tesi si concentra sull'argomento dell'Explainable Artificial Intelligence (XAI). Essendo l'Intelligenza Artificiale (IA) il principale punto di interesse della XAI si effettua una panoramica di entrambe. Si parte quindi esponendo le fondamenta teoriche dell'IA, in cui si analizzano in dettaglio anche i termini di Machine Learning, Reti Neurali e Deep Learning, per poi spiegare il perché della necessità di sistemi di Intelligenza Artificiale interpretabili. Proseguendo si presenta l'argomento dell'Explainable AI mostrandone caratteristiche, obbiettivi, contenuti e diverse tipologie.

Dopo una doverosa introduzione degli argomenti da trattare, si procede con la definizione di alcune delle tecniche interpretabili più utilizzate, distinguendo quelle che non hanno bisogno di ulteriori spiegazioni da quelle utilizzate per rendere i modelli di Deep Learning comprensibili. Proseguendo, si fanno dei cenni riguardo gli Adversarial Attacks in modo tale da poter rendere più chiari i loro utilizzi. Successivamente, si descrivono alcuni tra i più utilizzati strumenti per le implementazioni di XAI e Adversarial Attacks per poi evidenziare quelli che si utilizzeranno negli esperimenti, in specifico LIME, SHAP, tf-explain e ART.

In seguito, dopo una breve parentesi riguardo l'addestramento del modello ideale, la struttura e le motivazioni del capitolo, si arriva alla parte principale della tesi. A questo punto si effettuano vari esperimenti utilizzando cinque modelli per la classificazione di immagini, spiegandone i risultati utilizzando le tecniche scelte in precedenza e mostrando il loro funzionamento a seconda dei casi. Si terminano gli esperimenti effettuando degli Adversarial Attacks su un'immagine per poi analizzare e spiegare i risultati ottenuti.

Per concludere si discutono gli utilizzi attuali di queste tecniche analizzandone i lati positivi ed i limiti incontrati nella tesi, oltre a fornire una riflessione sui loro utilizzi futuri.