Università degli Studi di Salerno

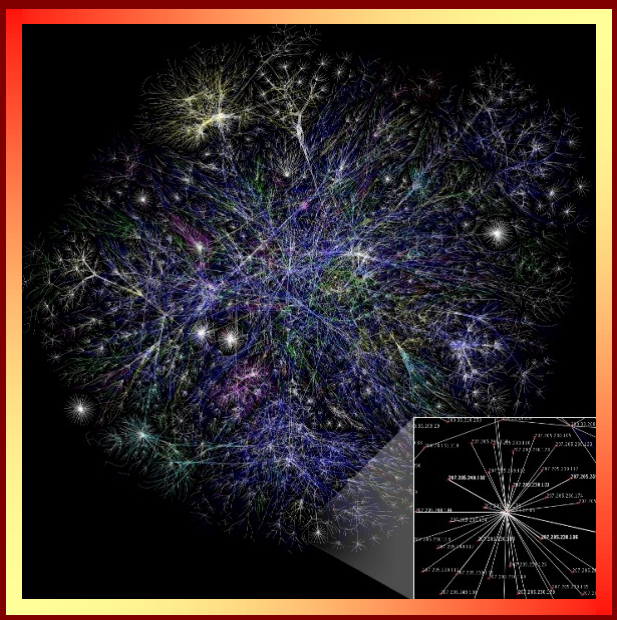
Dipartimento di Informatica

Dottorato di Ricerca in Informatica – XXXIV Ciclo

Tesi di Dottorato/Ph.D. Thesis

Application of machine learning techniques to biological big data

Alessia Auriemma Citarella



Supervisor: **Prof. Genoveffa Tortora**

Ph.D. Program Director: **Prof. Andrea De Lucia**

AA 2020/2021

Curriculum Computer Science and Information Technology

Università degli Studi di Salerno

Dipartimento di Informatica

Dottorato di Ricerca in Informatica
XXXIV Ciclo

TESI DI DOTTORATO / PH.D. THESIS

Application of machine learning techniques to biological big data

ALESSIA AURIEMMA CITARELLA

SUPERVISOR: **PROF. GENOVEFFA TORTORA**

PHD PROGRAM DIRECTOR: **PROF. ANDREA DE LUCIA**

A.A 2020/2021

ABSTRACT

To date, has been the primary driver of global innovation, competitiveness and cultural development. It is also a powerful engine for creating new job opportunities, expanding market segments and inspiring new horizons where new skills and specialties can compete. From this perspective, we are constantly pushed to investigate the ICT industry and its interconnections with other areas, such as new biomedical technology.

In the past twenty years, the development and increase of new diagnostic methods in the medical field has made available a huge amount of data capable of being stored and analyzed in order to extract important new knowledge.

In the biological field, the data produced by the sequencing techniques and the available databases provide a lot of information on multiple levels that can be integrated with each other. The ability to integrate and analyze data from multiple sources is vital in order to collect real benefits and speed up outputs thanks to the high computational possibilities of some tools.

Technology has the potential to dramatically change the conception of medicine and, at the same time, it plays a critical role in advanced diagnostics systems in making decisions intrinsic to patient care. Developing high-quality, accurate Artificial Intelligence (AI) resources improves work of clinicians by intervening on prevention, diagnosis and treatment of many pathologies. Some modern AI and computer science technologies, in general, encompass the power of clinical laboratory devices, allowing diagnostic activity to be carried out even outside of laboratories. Medical aids and new advanced diagnostic equipment are increasingly relying on qualified experts in the field to supplement medical evaluations and assist in diagnosis.

In this context, the focus of this work was on two main topics. First one, we explored additional Machine Learning and Deep Learning techniques that can guarantee a better classification of melanoma images even on clinical datasets with lower image quality. The goal is to improve melanoma early detection,

which is now a limiting factor for first-line therapies in this tumor pathology. Many of the research in the literature utilize similar strategies but use various approaches: some try to extract information directly from the image (such as color, plot and pixel density), while others try to extract functions based on guided lines of dermatologists (such as ABCDE and the Seven-Point Checklist). The majority of these researches are conducted using higher-resolution dermoscopic pictures. The purpose of the research is to identify novel features for melanoma classification that may be applied to less detailed images using advanced learning techniques.

The second contribution of this thesis is addressed to the classification of proteins. Researches focused on the possibility of exploring further molecular descriptors in addition to those already present in the literature to classify these proteins and to build new tools able to explore the complex interaction between proteins in a visual and intuitive way. In this line of research, the visualization of biological data was also taken into consideration. The work has mostly concentrated on the presentation tools of biological ontologies in order to develop user-friendly systems that allow end users to interact and extrapolate information more easily. This is useful for the complexity of the biological system that can be explored by the integration of omics disciplines. These sciences attempt to analyze the biological system holistically using biological Big Data, mainly proteomic, genomic, transcriptomic and metabolomic data. The latter are the most important groupings of organic compounds for the study of the functioning of living organisms.

ABSTRACT IN ITALIANO

Finora, la tecnologia è stata il motore principale dell'innovazione, della competitività e dello sviluppo culturale globali. È anche un potente motore per creare nuove opportunità di lavoro, espandere segmenti di mercato e ispirare nuovi orizzonti in cui nuove competenze e specialità possono competere. Da questo punto di vista, siamo costantemente spinti a indagare l'industria ICT e le sue interconnessioni con altre aree, come le nuove tecnologie

biomediche. Negli ultimi vent'anni, lo sviluppo e l'incremento di nuove metodiche diagnostiche in campo medico ha reso disponibile un'enorme quantità di dati in grado di essere archiviati e analizzati al fine di estrarre nuove importanti conoscenze. In campo biologico, i dati prodotti dalle tecniche di sequenziamento e le banche dati disponibili forniscono molte informazioni su più livelli che possono essere integrate tra loro. La capacità di integrare e analizzare dati provenienti da più fonti è fondamentale per raccogliere benefici reali e velocizzare gli output grazie alle elevate possibilità computazionali di alcuni strumenti. La tecnologia ha il potenziale per cambiare radicalmente la concezione della medicina e, allo stesso tempo, svolge un ruolo fondamentale nei sistemi diagnostici avanzati nel prendere decisioni intrinseche alla cura del paziente. Lo sviluppo di risorse di intelligenza artificiale (AI) accurate e di alta qualità migliora il lavoro dei medici intervenendo sulla prevenzione, la diagnosi e il trattamento di molte patologie. Alcune moderne tecnologie di intelligenza artificiale e informatica, in generale, racchiudono la potenza dei dispositivi di laboratorio clinico, consentendo di svolgere attività diagnostiche anche al di fuori dei laboratori. I presidi medici e le nuove apparecchiature diagnostiche avanzate si affidano sempre più a esperti qualificati del settore per integrare le valutazioni mediche e assistere nella diagnosi. In questo contesto, il focus di questo lavoro è stato su due temi principali. Innanzitutto, abbiamo esplorato ulteriori tecniche di Machine Learning e Deep Learning in grado di garantire una migliore classificazione delle immagini del melanoma anche su set di dati clinici con una qualità dell'immagine inferiore. L'obiettivo è migliorare la diagnosi precoce del melanoma, che ora è un fattore limitante per le terapie di prima linea in questa patologia tumorale. Molte delle ricerche in letteratura utilizzano strategie simili ma usano approcci diversi: alcune cercano di estrarre informazioni direttamente dall'immagine (come colore, trama e densità di pixel), mentre altre cercano di estrarre funzioni basate su linee guidate dai dermatologi (come l' ABCDE e la Seven-Point Checklist). La maggior parte di queste ricerche viene condotta utilizzando immagini dermoscopiche ad alta risoluzione. Lo scopo della ricerca è identificare nuove caratteristiche per la classificazione del melanoma che possono essere applicate a immagini meno dettagliate utilizzando tecniche di

apprendimento avanzate. Il secondo contributo di questa tesi è rivolto alla classificazione delle proteine. La ricerca si è concentrata sulla possibilità di esplorare ulteriori descrittori molecolari oltre a quelli già presenti in letteratura per classificare queste proteine e per costruire nuovi strumenti in grado di esplorare la complessa interazione tra proteine in modo visivo e intuitivo. In questo filone di ricerca è stata presa in considerazione anche la visualizzazione dei dati biologici. Il lavoro si è concentrato principalmente sugli strumenti di presentazione delle ontologie biologiche al fine di sviluppare sistemi user-friendly che consentano agli utenti finali di interagire ed estrapolare le informazioni più facilmente. Ciò si inserisce nell'ottica della complessità del sistema biologico che può essere esplorato dall'integrazione delle discipline omiche. Queste scienze tentano di analizzare il sistema biologico olisticamente utilizzando Big data biologici, principalmente dati proteomici, genomici, transcriptomici e metabolomici. Questi ultimi sono i raggruppamenti più importanti di composti organici per lo studio del funzionamento degli organismi viventi.