

Machine learning & semantic web technologies for cancer care

Citation for published version (APA):

Mahadevaiah, G. (2020). *Machine learning & semantic web technologies for cancer care*. Maastricht University. <https://doi.org/10.26481/dis.20200409gm>

Document status and date:

Published: 01/01/2020

DOI:

[10.26481/dis.20200409gm](https://doi.org/10.26481/dis.20200409gm)

Document Version:

Publisher's PDF, also known as Version of record

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy

If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

Summary

Geetha Mahadevaiah

Industrialization and automation of healthcare are crucial to transform the healthcare industry towards cost effectiveness, better outcomes, improved patient and staff satisfaction. With the high availability of cloud infrastructure, network bandwidth, cost efficient hardware and software technologies such as Semantic Web, AI, deep learning, this transformation is rapidly progressing. In this research, the role of Semantic Web technologies and machine learning in improving the clinical decision making for cancer care was studied. Chapter 2 elaborates on a method of using semantic technologies and machine learning to automatically detect cancer cases and classify cancer types, from unstructured medical reports.

The requirement of data privacy and security for healthcare data limits the sharing and accessing of data for research and clinical purposes. In Chapter 3, state of the art techniques for de-identification of patient notes in electronic health records were analyzed for their performance and a machine learning technique NeuroNER was studied for applicability of to Indian PHI data. This technique was further improved for reliability and performance, by the inclusion of a new model trained on Indian PHI data. Various methods addressing the security aspects of medical data stored on the Semantic Web was studied. This led to the design of an innovative framework, described in Chapter 4, to manage the data access with the appropriate rights and required patient consent for the data stored in Semantic Web format (RDF).

In addition to the large amounts of textual data stored in medical records and reports, there is valuable and vast data stored in medical images, such as CT, MR and Radiotherapy. DICOM format is hierarchical and is not easy to mine. Chapter 5 describes a Semantic technology based architecture to extract clinically relevant information from the Radiotherapy DICOM data and make it easily findable using natural language like queries.

As a subsequent next step, Radiomics techniques were explored to increase the clinical findings from DICOM data. Radiomics is an evolving medical field for discovering deeper information from Radiology images to aid diagnosis and treatment. Chapter 6 explains a method for automating the classification of non-small lung tumors based on histopathology using Radiomics. Chapter 7 shows that Radiomics techniques can be extended further to gain deeper clinical insights by computing the Fractal dimension of a tumor to classify non small lung tumors. Fractals can be an important Radiomics feature, providing information not only about the Gross Tumor Volume (GTV) structure, but also help in characterization of the tumor.

Finally, system level needs and techniques to facilitate effective deployment and usage of clinical decision support in hospitals was studied. In Chapter 8, a system level architecture of a Cloud solution for image analytics, that can be deployed in a hospital, is described. Chapter 9 provides an overview of the selection, acceptance, commissioning and quality assurance processes required in hospitals for effective deployment of machine learning based clinical decision support systems.

In conclusion, Chapter 10 provides a synopsis of the thesis and discusses the future advancements in the field of automation for healthcare.

Samenvatting

Industrialisering en automatisering zijn essentieel om de gezondheidszorg kosteneffectiever te maken, de resultaten te verbeteren en de tevredenheid van patiënten en personeel te vergroten. Dankzij de grote beschikbaarheid van technologische middelen, zoals de cloud-infrastructuur, netwerkbandbreedte, kostenefficiënte hardware en software zoals het semantische web, AI en deep learning, vindt deze transformatie in hoog tempo plaats. In dit onderzoek werd gekeken hoe semantische-webtechnologieën en automatisch leren kunnen worden benut om de klinische besluitvorming met betrekking tot de zorg bij kanker te verbeteren. In hoofdstuk 2 wordt uitleg gegeven over een methode waarbij semantische technologieën en automatisch leren worden gebruikt voor de automatisch detectie van gevallen van kanker en de classificering van vormen van kanker op basis van niet-gestructureerde medische dossiers.

Als gevolg van de wet- en regelgeving met betrekking tot gegevensbescherming en de beveiliging van zorggegevens kunnen deze gegevens slechts beperkt worden gedeeld en gebruikt voor onderzoek en klinische doeleinden. In hoofdstuk 3 wordt geanalyseerd hoe goed geavanceerde technieken voor de anonimisering van patiëntgegevens in elektronische dossiers werken en wordt gekeken naar de toepassing van de NeuroNER-techniek voor automatisch leren op privacygevoelige gegevens uit India. De betrouwbaarheid en prestaties van deze techniek werden verder verbeterd door een nieuw model toe te voegen dat werd ontwikkeld aan de hand van de gegevens uit India. Er werd gekeken naar verschillende methoden om om te gaan met de beveiligingsaspecten van medische gegevens die op het semantische web zijn opgeslagen. Op basis hiervan werd een innovatief raamwerk ontwikkeld waarmee de toegang tot gegevens opgeslagen in RDF (het formaat om gegevens te delen via het semantische web) kan worden beheerd, rekening houdend met de benodigde rechten en toestemming van de patiënt.

Er zijn niet alleen enorm veel gegevens in tekstvorm opgeslagen in medische dossiers en rapporten, maar ook in medische beelden, zoals CT-, MR- en röntgenbeelden is gigantisch veel waardevolle informatie vastgelegd. Maar de DICOM-indeling is hiërarchisch en kan niet eenvoudig doorzocht worden. In hoofdstuk 5 wordt een architectuur op basis van semantische technologie beschreven waarmee klinisch relevante informatie kan worden geëxtraheerd uit radiotherapeutische DICOM-gegevens en eenvoudig gezocht kan worden met bijvoorbeeld zoekopdrachten.

Als volgende stap werd gekeken of Radiomics technieken konden worden toegepast om meer klinische informatie uit DICOM-gegevens te verkrijgen. Radiomics is een wetenschapsgebied in ontwikkeling dat erop gericht is om meer informatie te extraheren uit radiologische beelden ter ondersteuning van de diagnose en behandeling. In hoofdstuk 6 wordt een methode uitgelegd voor het automatiseren van de classificatie van niet-kleincellige longkanker (NSCLC) op basis van histopathologie met behulp van Radiomics technieken. In hoofdstuk 7 wordt aangetoond dat de toepassing van Radiomics technieken verder kan worden uitgebreid om diepergaande klinische inzichten te verwerven door de fractale dimensie van een tumor te berekenen voor de classificering van NSCLC. Fractals kunnen een belangrijke functie hebben bij Radiomics technieken; niet alleen geven ze informatie over het 'gross tumor volume' (GTV), maar daarnaast kunnen ze helpen bij het bepalen van de kenmerken van de tumor.

Tot slot werd er gekeken naar behoeften en technieken op ziekenhuisniveau om de effectieve toepassing van ondersteuning bij klinische besluitvorming in ziekenhuizen te vereenvoudigen. In

hoofdstuk 8 wordt een architectuur op systeemniveau beschreven, gebaseerd op een cloudoplossing voor beeldanalyse die in ziekenhuizen kan worden toegepast. In hoofdstuk 9 wordt een overzicht gegeven van de processen voor selectie, aanvaarding, inbedrijfstelling en kwaliteitsborging die nodig zijn in ziekenhuizen voor een effectieve toepassing van automatisch lerende systemen ter ondersteuning van de klinische besluitvorming.

Afsluitend wordt er in hoofdstuk 10 een samenvatting gegeven van het proefschrift en worden de toekomstige ontwikkelingen op het gebied van automatisering in de gezondheidszorg besproken.

