Medical Physics Unit and Department of Physics,
McGill University, Montréal, Québec, Canada

DOCTORAL THESIS

The use of radiomics and natural language processing to detect pain in the simulation-CT images of patients undergoing radiotherapy for bone metastasis

Hossein Naseri

April 15, 2023

A thesis submitted to McGill University in partial fulfillment of
the requirements of the degree of *Doctor of Philosophy*

Abstract

Cancer develops when cells lose their ability to control division and form a tumor. Malignant tumor cells can invade nearby tissues or spread (metastasize) to other parts of the body. Bone is one of the most common sites for cancer to metastasize to. Bone Metastases (BM) can result in inflammation, structural damage, and severe pain. 70 to 90% of patients with BM suffer from severe pain. Therefore, detecting and controlling BM-associated pain has the potential to improve the quality of life of BM patients.

This thesis project aimed to develop and evaluate an Artificial Intelligence (AI) pipeline for detecting pain in cancer patients with BM by combining information from clinical texts and radiographic images. The project fits within an ultimate research goal of enabling early prediction and management of BM pain before it becomes distressing. It addressed three specific objectives in three studies: 1) Construction of a Natural Language Processing (NLP) pipeline to extract pain scores from consultation notes, 2) Construction of a radiomics pipeline to extract BM lesion features from radiographic images, and 3) Development of a radiomics-based machine-learning model of pain in patients with BM by combining NLP-quantified pain scores with radiomic features.

In the first study, we trained and tested an NLP pipeline using publicly-available hospital discharge notes and achieved a precision and recall of 0.86 and 0.83 in detecting sentence level pain scores. The pipeline was then used to automatically extract and classify note-level pain from clinical notes at our institution with 0.925 F1 score.

In the second study, a radiomics model was generated based on a novel lesion-centerpoint-based geometric Regions Of Interest (ROIs). The geometric ROIs were automatically delineated around lesion centerpoints that were manually pinpointed by radiation oncologists on CT images. This allowed us to greatly simplify the data preparation process. We demonstrated that, the introduced pipeline was successful in differentiating BM from healthy bones.

In the third study, a Machine Learning (ML) pipeline was developed to detect pain in cancer patients with thoracic spinal BM. The study used data from 176 patients treated at Cedar Cancer Center between January 2016 and September 2019. Our NLP pipeline was used to extract pain scores from radiation oncology consultation notes. Radiomics features were extracted from each ROI, and various ML classifiers were evaluated using precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (ROC-AUC). The results showed that the pipeline was successful in differentiating between painful and painless BM lesions with an accuracy, specificity, and ROC-AUC of 0.82, 0.85, and 0.83, respectively.

Overall in this thesis, we developed a robust radiomics pipeline to identify painful BM lesions in CT images. Our pipeline is fast and scalable as it is trained using NLP-extracted pain scores from clinical notes and it requires just centerpoints to identify BM lesions in CT images. This work represents the first step in building a clinically-practical pain detection pipeline and is consistent with the ultimate goal of better managing pain in patients with BM.

Abrégé

Le cancer se développe lorsque les cellules perdent leur capacité à contrôler la division et forment une tumeur. Les cellules de tumeur maligne peuvent envahir les tissus voisins ou se propager (métastaser) à d'autres parties du corps. Les os sont l'un des sites les plus courants pour les métastases cancéreuses. Les Métastases Osseuses (MO) peuvent entraîner une inflammation, des dommages structurels et une douleur intense. De 70 à 90% des patients atteints de MO souffrent de douleur intense. Par conséquent, détecter et contrôler la douleur associée à la MO a le potentiel d'améliorer la qualité de vie des patients atteints de MO.

Ce projet de thèse visait à développer et à évaluer un pipeline IA pour détecter la douleur chez les patients atteints de MO en combinant des informations provenant de notes cliniques et d'images radiographiques. Le projet s'inscrit dans un objectif de recherche qui permettra ultimement la prédiction et la gestion précoce de la douleur MO avant qu'elle ne devienne angoissante. Il comporte trois objectifs spécifiques dans trois études: 1) Construction d'un pipeline du traitement du langage naturel (NLP) pour extraire les scores de douleur à partir des notes de consultation, 2) Construction d'un pipeline radiomique pour extraire les caractéristiques de lésion MO à partir d'images radiographiques, et 3) Développement d'un modèle d'apprentissage machine basé sur la radiomique pour prédire la douleur chez les patients atteints de MO en combinant les scores de douleur quantifiés par NLP et les caractéristiques radiomiques.

Dans la première étude, nous avons formé et testé un pipeline NLP à l'aide de notes de congé d'hôpital disponibles publiquement. Globalement, nous avons obtenu une précision et un rappel de 0.86 est 0.83 dans la détection de la douleur. Le pipeline a ensuite été utilisé pour extraire et classer automatiquement la douleur à partir des notes cliniques de notre institution.

Dans la seconde étude, un modèle radiomique a été généré sur la base de régions d'intérêt (ROIs) géométriques originales, basées sur le centre de la lésion. Les ROIs géométriques ont été délimitées automatiquement autour des centres des lésions ayant été repérées manuellement par des radiothérapeutes à partir des images CT. Cela nous a permis de simplifier considérablement le processus de préparation des données. Nous avons démontré que le pipeline a pu différencier le MO des os sains avec succès.

Dans la troisième étude, un pipeline d'apprentissage automatique a été développé pour détecter la douleur chez les patients atteints de cancer avec MO de la colonne thoracique. L'étude a utilisé les données de 176 patients traités au centre de cancer des Cèdres entre janvier 2016 et septembre 2019. Notre pipeline NLP a été utilisé pour extraire les scores de douleur des notes de consultation en radiothérapie. Des caractéristiques radiomiques ont été extraites de chaque ROI, et utilisées par divers classificateurs d'apprentissage automatique qui ont été évalués en utilisant la précision, le rappel, le score F1 et l'aire sous la courbe (AUC) des courbes d'efficacité du récepteur. Les résultats ont montré que le pipeline a réussi à différencier les lésions BM douloureuses et indolores avec une précision, une spécificité et une AUC de 0.82, 0.85 et 0.83, respectivement.

Dans l'ensemble, nous avons développé dans cette thèse un pipeline radiomique robuste pour identifier les lésions MO douloureuses à partir d'images CT. Notre pipeline est rapide et évolutif, car il est construit en utilisant des scores de douleur extraits du NLP des notes cliniques et qu'il nécessite simplement des points centraux pour identifier les lésions MO dans les images CT. Ce travail représente la première étape dans la construction d'un pipeline de détection de la douleur pratique en clinique et est cohérent avec le but ultime qui est de mieux gérer la douleur chez les patients avec MO.

Acknowledgements

First and foremost, I want to express my sincere gratitude to my supervisor Dr. John Kildea who has supported me throughout the entire Ph.D. with his advice, patience, and knowledge. I have had the privilege to work for such a considerate, dedicated, and motivating supervisor and mentor. I deeply appreciate his support and work in helping me achieve my Ph.D.; without him, this thesis would not have been completed or written. One simply could not wish for a better or friendlier supervisor.

In addition, I would like to express my gratitude to Dr. Kameran Kafi, Dr. Marc David, and Dr. Peter Savadjiev who have served as members of my committee and provided me with vital medical and technical guidance when I first began this project. Also, I would like to thank Dr. Sonia Skamene, Marwan Tolba, Mame Daro Faye, Paul Ramia, and Julia Khriguian for their assistance in labeling imaging data and medical notes. I consider myself extremely fortunate to have had their guidance and collaboration all through my doctoral studies.

I would want to say a big “thank you” to the NICE-ROKS group at Kildea Lab for all their help and friendship. Thank you Haley Patrick for helping me prepare my data, for reviewing my writings, and for simply being an awesome friend. Thank you Aixa X. Andrade, Esteban Sepulveda, Jonathan Yeo, and Thierry Lefebvre for helping me adjust to life in Montreal. Thank you Logan Montgomery for patiently assisting me in preparing for my preliminary exam. Thank you Kayla O’Sullivan, Felix Mathew, and James Manalad for your helpful feedback on my writings and for staying in touch with me despite our geographical separation during the COVID-19 pandemic. Thank you Luc Galarneau for your assistance with the French translation of my abstract. It was a pleasure working with you and everyone else who has come and gone from NICE-ROKS group during my time here, and I appreciate everything you’ve taught me.

I would like to express my gratitude to Dr. Jan Seuntjens, the former director of McGill University's Medical Physics Unit (MPU), Dr. Shirin Enger, and Dr. Ives Levesque, the current members of the MPU's co-directorship, and administrative coordinators Margery Knewstubb and Tatjana Nisic. I am incredibly appreciative of the work you have put in to making the MPU a welcoming, diverse, and enjoyable place to work in the scientific community. I acknowledge the space and computing resources provided by the MPU at the MUHC.

Tiohtiá:ke, Montreal, has a long history of serving as a gathering site for several First Nations; as a result, I understand and value the historic connections and continued presence of Indigenous people on these lands and waters.

I also thank the Natural Sciences and Engineering Research Council of Canada (NSERC), Fonds de recherche du Québec - Sante (FRQS), McGill Graduate and Postdoctoral Studies, Department of Physics and Faculty of Medicine of McGill University, The Research Institute of the McGill University Health Centre (RI-MUHC), and Mitacs Accelerate and NSERC-CREATE Responsible Health and Healthcare Data Science (SDRDS) programs for funding and support throughout my Ph.D. studies.

Finally, I would especially like to thank my wife, who has always been both my best friend and my best colleague. Ruby, you were there for me at every turn of this long journey with your unwavering love, support, and encouragement. Yasha :*

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence.
BM	Bone Metastases.
DT	Decision Tree.
EBRT	External Beam Radiation Therapy.
EHRs	Electronic Health Records.
FPR	False positive Rate.
FS	Feature Selection.
GPR	Gaussian Process Regression.
HU	Hounsfield Units.
IBSI	Image Biomarker Standardization Initiative.
ICA	Independent Component Analysis.
IMRT	Intensity-Modulated Radiation Therapy.
kNN	k-Nearest Neighbors.
LASSO	Least Absolute Shrinkage and Selection Operator logistic regression algorithm.
ML	Machine Learning.
MO	Métastases Osseuses.
NB	Gaussian Naive Bayes.
NLP	Natural Language Processing.
NN	Neural Network.
NNet	Neural Networks.
NSAIDs	Non-Steroidal Anti-Inflammatory Drugs.
PCA	Principal Component Analysis.

QDA	Quadratic Discriminant Analysis.
RBF	Radial Basis Function.
ReLU	Rectified Linear Unit.
RF	Random Forest.
RFECV	Recursive Feature Elimination with Cross-Validation.
ROC	Receiver Operating Characteristic.
ROC-AUC	Area Under the Receiver Operating Characteristic Curve.
ROI	Region Of Interest.
ROIs	Regions Of Interest.
RT	Radiation Therapy.
SMOTE	Synthetic Minority Oversampling Technique.
SVM	Support Vector Machine.
TPR	True Positive Rate.
TREE	Decision-Tree-Based.
UMLS	Unified Medical Language System.
VDP	Verbally-Declared Pain.
VT	Variance Threshold.

Preface and Contribution of Authors

The idea for this project was initiated by my supervisor, John Kildea, in 2015, when he was doing the medical physics chart check of a patient receiving palliative radiotherapy for bone metastases. While checking the CT simulation image of the patient, who an obviously painful metastatic bone lesion, it occurred to him that maybe we can quantify and predict pain by just looking at patients' radiographic images. We sat down together and devised a project to use artificial intelligence to analyze radiographic images of patients with bone metastasis and quantify their pain scores.

This project has resulted in two published and one submitted original manuscripts (Chapters 3, 4 & 5) that are integrated to this thesis. To the best of the authors' knowledge, we were the first group who: 1) developed a lesion-centerpoint-based radiomics pipeline to separate metastatic and healthy bone lesions; 2) developed a generalizable Natural Language Processing (NLP) pipeline to detect pain from consultation notes of cancer patients; and 3) combined NLP extracted pain scores with lesion-centerpoint-based radiomics features to build a predictive model of cancer pain for patients with spinal bone metastasis. The contributions of each author to the manuscripts are detailed below.

Journal publications

1. Chapter 3: *Naseri H, Kafi K, Skamene S, Tolba M, Faye MD, Ramia P, Khriguian J, Kildea J.* "Development of a generalizable natural language processing pipeline to extract physician-reported pain from clinical reports: Generated using publicly-available datasets and tested on institutional clinical reports for cancer patients with bone metastases". J Biomed Inform. 120:103864 (2021)

   I designed this study based on the idea presented by Dr. John Kildea. I gathered information from publicly-available sources and our institution and developed an application for labeling “ground truth” pain scores. I created the natural language processing approach, performed the tests, and wrote the manuscript. Throughout the study, Dr. Kamran Kafi,Dr. Sonia Skamene, Marwan Tolba, Mame Daro Faye, Paul Ramia, and Julia Khriguian supplied expert-identified pain scores as well as expert knowledge and counseling. Dr. John Kildea served as the study’s director and contributed both financial and computational resources to the project. He also edited the manuscript. All authors reviewed the paper and approved the final manuscript.

2. Chapter 4: *Naseri H, Skamene S, Tolba M, Faye MD, Ramia P, Khriguian J, Patrick H, Andrade Hernandez AX, David M, Kildea J* “Radiomics-based machine learning models to distinguish between metastatic and healthy bone using lesion-center-based geometric regions of interest”. Scientific Reports. 12: 9866 (2022)

   I designed this study in consultation with Dr. John Kildea. Under an approved Research Ethics Board protocol, I gathered CT images of patients from our institution and developed an application for labeling ground truth bone lesion-centers in images. I created the Radiomics and machine learning pipelines, conducted the analysis, and wrote the manuscript. Dr. Marc David participated in project conceptualization and methodology planning. Dr. Sonia Skamene, Marwan Tolba, Mame Daro Faye, Paul Ramia, and Julia Khriguian supplied expert-identified metastatic bone lesion-center points as well as expert knowledge throughout this study. Haley Patrick and Aixa Andrade Hernandez participated in data collection and identified the center points of healthy bones in CT images. Dr. John Kildea contributed to the conceptualization, investigation, supervision, funding acquisition, and editing of the original draft. The final manuscript was reviewed and approved by all of the authors.

3. Chapter 5: *Naseri H, Skamene S, Tolba M, Faye MD, Ramia P, Khriguian J, David M, Kildea J* “A scalable radiomics- and NLP- based machine learning pipeline to distinguish between painful and painless thoracic spinal bone metastases: Algorithm Development and Validation” (Accepted for publication in JMIR-AI - in press)

   Dr. John Kildea and I collaborated on the design of this study. I gathered CT scans and patient consultation records from our institution’s database. I developed the pipelines for machine learning, radiomics, and natural language processing, carried out the analysis, and wrote the manuscript. Dr. Marc David was involved in the conceptualization of the idea. Dr. Sonia Skamene, Marwan Tolba, Mame Daro Faye, Paul Ramia, and Julia Khriguian contributed their expertise to this study by providing expert-identified metastatic bone lesion centers and expert-extracted pain scores. Dr. John Kildea helped with the draft's conception, investigation, supervision, funding procurement, and editing. All of the writers reviewed and approved the final manuscript.

Open source code contribution

During my Ph.D. research I developed three open-source software applications.

1. *Naseri, H.*, "texTRACTOR; physician-reported pain scoring tool" (2020)

   texTRACTOR (https://github.com/hn617/texTRACTOR) is a Python Flask-based web application that allows users to annotate physician-reported pain scores in selected patients' consultation notes. This tool was utilized in the studies described in Chapters 3 and 5.

1. *Naseri, H.*, "diCOMBINE; 3D DICOM visualization and annotation tool" (2021)

   diCOMBINE (https://github.com/hn617/diCOMBINE) is a web application for 3D DICOM visualization and lesion segmentation that allows users to review 3D DICOM images of selected patients and label their lesion centerpoints. It also enables users to cross-validate lesions that have been annotated by other users. This tool was used in the research described in Chapters 4 and 5.

3. *Naseri, H.*, "paINDICATOR; a pain detection pipeline" (2021)

   paINDICATOR (https://github.com/hn617/paINDICATOR) is a python-based radiomics pipeline for identifying pain in cancer patients using consultation notes and radiographic images. Chapters 4 and 5 explain the pipeline.

Chapter 1

Introduction

1.1 Cancer

If any of the cells in our body lose their ability to control cell division, they may disrupt homeostasis and can form a tumor (also called a neoplasm). A tumor is called benign (non-cancerous) if it grows in a single location and does not have the ability to spread into nearby tissues. Cancerous or malignant tumors, on the other hand, can invade nearby tissues or spread to other parts of the body [1]. As a malignant tumor grows, it requires more oxygen and nutrients. Therefore, it creates new blood vessels. These blood vessels allow cancer cells to enter the bloodstream or lymph system and spread to other parts of the body [2]. Cancer that has spread to another part of the body is referred to as metastatic cancer, whereas the original cancer is referred to as the primary tumor.

The pathological examination of a tumor's histology, along with a physical exam, and imaging, are used to determine the type of the cancer as well as its potential spread. This process is called cancer staging [3]. Staging is an essential step in deciding which anti-cancer treatment to pursue and in providing a preliminary evaluation of the patient's prognosis (such as their chance of survival). The "TNM staging" system, which classifies the size of the primary tumor (T), the absence or presence of regional lymph node invasion (N), and the absence or presence of distant metastases (M), is the most widely used system in cancer evaluation [3].

Almost all kinds of cancer have the potential to metastasize anywhere in the body. The lungs, liver, bones, and brain are the most common sites for cancer metastasis [2, 4, 5]. Figure 1.1 illustrates some of the common cancers and their most common metastasis sites. Cancer that spreads to the bones is referred to as Bone Metastases (BM) or metastatic bone disease.

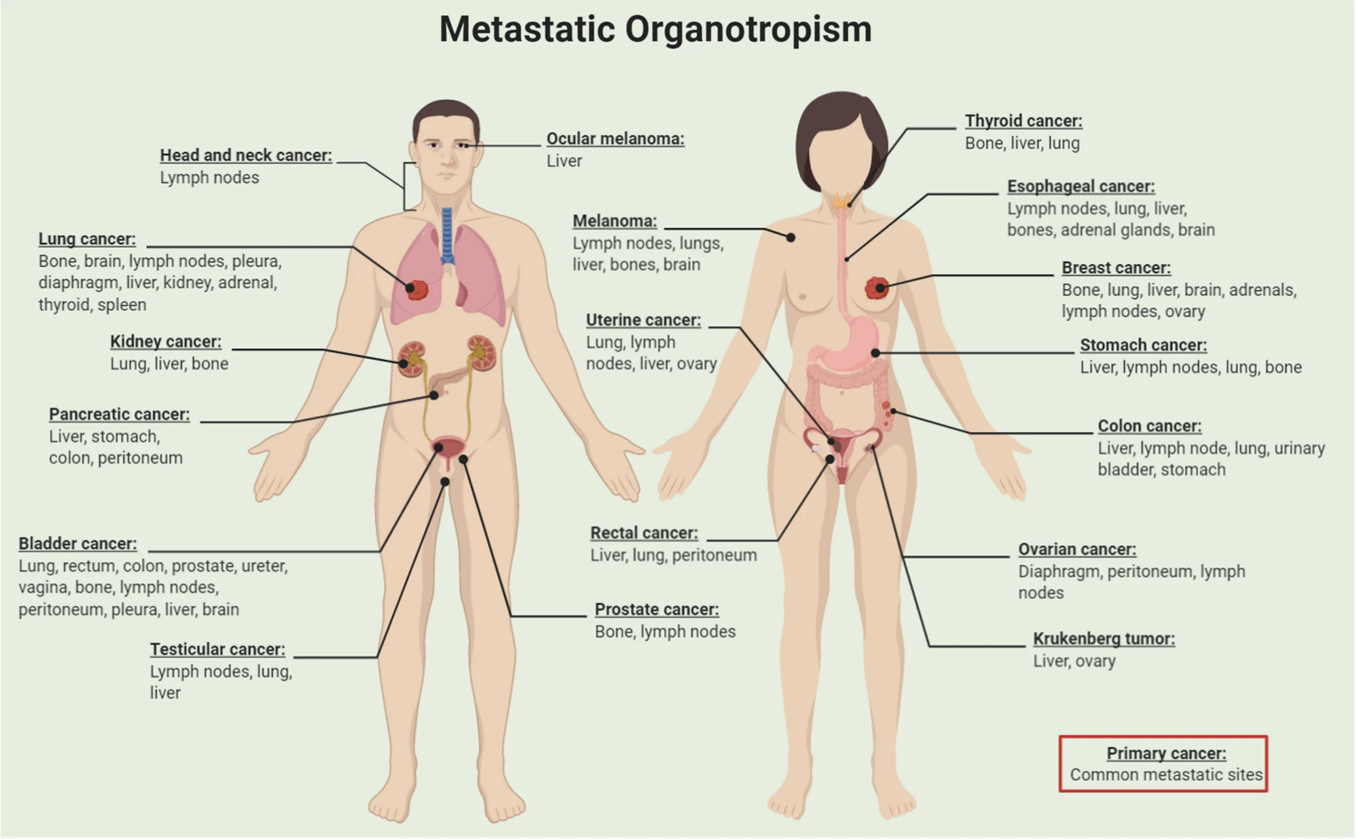

Metastatic Organotropism
Head and neck cancer: Lymph nodes
Ocular melanoma: Liver
Lung cancer: Bone, brain, lymph nodes, pleura, diaphragm, liver, kidney, adrenal, thyroid, spleen
Melanoma: Lymph nodes, lungs, liver, bones, brain
Kidney cancer: Lung, liver, bone
Uterine cancer: Lung, lymph nodes, liver, ovary
Pancreatic cancer: Liver, stomach, colon, peritoneum
Bladder cancer: Lung, rectum, colon, prostate, ureter, vagina, bone, lymph nodes, peritoneum, pleura, liver, brain
Rectal cancer: Liver, lung, peritoneum
Prostate cancer: Bone, lymph nodes
Testicular cancer: Lymph nodes, lung, liver
Thyroid cancer: Bone, liver, lung
Esophageal cancer: Lymph nodes, lung, liver, bones, adrenal glands, brain
Breast cancer: Bone, lung, liver, brain, adrenals, lymph nodes, ovary
Stomach cancer: Liver, lymph nodes, lung, bone
Colon cancer: Liver, lymph node, lung, urinary bladder, stomach
Ovarian cancer: Diaphragm, peritoneum, lymph nodes
Krukenberg tumor: Liver, ovary
Primary cancer: Common metastatic sites

Figure 1.1: Major cancers and their most common metastasis sites. Figure reprinted from [6] under a Creative Commons license.

1.2 Bone metastases

The Canadian Cancer Society predicts that one in every two Canadians will be diagnosed with cancer during their lifetime [7]. Unfortunately, more than 50% of cancers are diagnosed at the stage when they have already metastasized to another part of the body [8]. It is estimated that up to 70% of cancers metastasize to the bone in their metastatic stage [9].

Breast and prostate cancers are the most common types of cancer that spread to the bones, followed by lung and kidney cancers [10, 11]. BM can result in inflammation, structural damage, and more frequently, excruciating pain.

1.3 Cancer pain

Pain is the most common symptom of metastatic cancers. Various studies have shown that 70-90% of patients with metastatic cancer suffer from severe pain [12–14]. Cancer pain is multifaceted, with biological, psychological, social, and cultural factors all playing a role in how it is experienced [15].

Tumors induce pain in a number of different ways. They can press on or invade healthy innervated tissue, create inflammation or infection, or release chemicals that amplify the sensation of pain in response to stimuli that would otherwise be painless [16]. In metastatic cancers, pain can be caused by a more than one mechanism. A patient with metastatic breast cancer, for instance, may suffer from neuropathic pain from chemotherapy in addition to abdominal pain from liver metastases and back pain from spinal metastases. In such a circumstance, pain management would be insufficient unless each potential source of pain is carefully explored and evaluated

[17]. In Section 1.7 I will go into more detail about pain assessment.

1.4 Pain management in bone metastases

The clinical treatment of cancer pain is typically based on one or more of the following three modalities: Non-Steroidal Anti-Inflammatory Drugs (NSAIDs), opioid therapy, and Radiation Therapy (RT) [18]. NSAIDs are the primary method of pain control in many diseases, including cancer. They provide significant pain relief for patients with mild to moderate pain. In addition, the anti-inflammatory effect of NSAIDs make them the ideal treatment for bone pain caused by inflammation during extensive tissue invasion and destruction. However, NSAIDs are ineffective for treating severe BM pain due to the lack of long-lasting effects and potential side effects of their prolonged use. Opioids are the second-most frequently prescribed medication for treating cancer pain. Since opioids provide longer-lasting pain relief than NSAIDs, more than 80% of BM patients use them to manage moderate to severe BM pain [13]. Significant adverse effects of opioids include physiological dependence and addiction. Another significant issue is sensitization, which reduces opioids' effectiveness when used for extended periods of time. RT is considered the most effective method of pain management for metastatic cancer patients. Several studies have reported that RT of painful BM can provide 60-90% of patients with either partial or complete pain relief [19–22].

1.5 Radiation therapy

RT (also known as radiotherapy) is regarded as one of the most effective cancer treatments, alongside chemotherapy and surgery [21]. It can be used to cure cancer (curative-intent RT) and/or control symptoms by slowing down tumor growth (palliative-intent RT). Curative-intent RT is a cancer treatment that uses high doses of radiation to destroy cancer cells while causing minimal harm to healthy cells and surrounding tissues. Over 50% of cancer patients undergo RT in specialized cancer centers over the course of their treatments [23].

RT can be delivered from within the body by positioning a radioactive source in or near the tumor (brachytherapy) or from outside the body by employing a device to direct the beam at the tumor (External Beam Radiation Therapy, or EBRT). High-energy photon EBRT is by far the most common form of RT. High-energy X-ray photons (MeV range) are generated in a linear accelerator, or linac, by accelerating electrons and then colliding them with a tungsten target to produce bremsstrahlung X-ray photons. The X-ray beam is then focused on the area of the tumor, where it can cause DNA damage and ultimately kill the rapidly-dividing cancer cells.

To minimize the likelihood of causing long-term collateral damage to normal cells, curative RT is typically administered as multiple daily fractions of small doses. Moreover, several beam delivery techniques have been developed to match the shape of the external beam to the tumor and minimize the radiation dose to normal tissue. The most commonly used technique is called three-dimensional conformal RT (3D-CRT) which uses radiation beams that are collimated and combined from different directions to match the shape of the tumor. Another commonly-used technique is Intensity-Modulated Radiation Therapy (IMRT), which modifies both the intensities and the directions of multiple beams to closely match the shape of the tumor. IMRT may be delivered with multiple static beams or as one or more arcs of radiation delivered continuously as the beam delivery device (head of the linac) rotates around the patient [24].

1.6 Radiotherapy for palliative care

The goal of palliative RT, unlike curative RT, is not to cure the cancer per se, but rather to alleviate or eliminate their symptoms and improve the patient's quality of life [25]. Therefore, a single large dose of radiation can be utilized safely without causing significant side effects. It has been demonstrated that palliative RT is an effective, time-efficient, well-tolerated, and cost-effective method for managing cancer symptoms [26]. About 35%-40% of all palliative RTs are intended for palliation of painful BM [20, 27] and over 60% of patients with BM report significant pain alleviation after palliative RT [28].

Despite the fact that RT is an effective treatment option for palliation of BM pain, some studies suggest that RT is frequently administered too late in the progression of the disease [29, 30]. A study by Rosen et al. (2020) [31], showed that up to 60% of patients treated for painful BM had evidence of their lesions being visible on images taken within the four months prior to their RT. Multiple studies have shown that providing BM patients with early palliative care improves both their overall survival and quality of life [31–33]. One study showed that the likelihood of experiencing severe pain was reduced by at least 30% in patients who received early palliative care [34]. An early palliation approach begins with the identification and prioritization of patients who would benefit from such care.

1.7 Pain assessment in bone metastases

A number of measures have been developed to assess pain in cancer patients [35, 36]. The 11-point numeric rating scale, a verbal rating scale (mild, moderate, or severe), or a visual analogue scale [37] that uses drawings of faces are examples of uni-dimensional pain intensity scales. They can be used to validate the presence of pain, learn some basic details about it, track its progression over time, and assess how well pain management is working [38]. Multidimensional pain assessment tools, such as the Brief Pain Inventory and the McGill Pain Questionnaire Haefeli2006PainAssessment, can be used to measure the location and intensity of pain in cancer patients as well as any disability or associated symptoms brought on by it. These tools can help in the identification of symptom clusters as well as the systematic assessment of the psychological and physiological aspects of pain [37].

Due to the subjective nature of these assessment methodologies, there is considerable variation in the objective measure of pain recorded by healthcare practitioners [39]. Moreover, there has been a documented lack of sufficient recording of pain evaluation using these methods, which leads to inappropriate pain monitoring and pain treatment [40]. Several clinician-driven tools and recommendations have been proposed to resolve these deficiencies [17]. Adapting standardized assessment methods, more frequent and systematic collection of patient-reported pain outcomes, and improved communication between patients and healthcare providers (doctors, nurses, and other healthcare workers) are among the proposed solutions [41–43].

While standard pain assessment techniques can capture current and historical pain patterns in cancer patients, they have very limited and inconsistent predictive potential for future pain [44–46]. As a result, they are ineffective in identifying and prioritizing patients who would benefit from early palliative care. To alleviate this burden, numerous tools and guidelines for early palliative care have been proposed in recent years. [47–54]. However, recent studies have drawn attention to the limitations of these tools and proposed data-driven methods to find indicators predictive of cancer outcomes that can then be used to identify patients who would benefit from early palliative care [55, 56]. The abundance of patient-centric cancer data stored in routinely-recorded Electronic Health Records (EHRs) in conjunction with the availability of modern Artificial Intelligence (AI) technologies make it conceivably possible to build data-driven tools to predict cancer outcomes such as survival, pain, and distress.

1.8 Artificial intelligence for outcome prediction

The vast majority of the cancer-related data contained within EHRs, such as oncology consultation notes and radiographic images, are, unfortunately, stored in an unstructured formats. However, by utilizing AI techniques, it is possible to process these unstructured data and extract structured information from them. In addition to this, AI models are able to “learn” from the data in order to recognize statistical trends, which in turn can be used to draw certain conclusions, predict outcomes, and improve treatment procedures. For instance as is the subject of this thesis, Natural Language Processing (NLP) techniques can extract quantifiable information from clinical texts, radiomics technologies can extract tumor phenotype information from radiographic images, and machine-learning techniques can learn from both text and radiomic data and predict future events such as tumor growth, survival, and distress. The next chapter of this thesis provides a more in-depth explanation of these strategies. Although various authors have developed NLP and radiomics strategies and have demonstrated their potential in predicting patient outcomes, as yet no group has reported a strategy to combine these AI technologies in order to detect pain in cancer patients with BM. This thesis reports the development and implementation of such a combined NLP and radiomics machine-learning pipeline.

1.9 Thesis objectives

The hypothesis of this thesis research project is that by combining cancer symptoms (extracted from clinical texts) with tumor imaging phenotypes (extracted from radiographic images), it is possible to radiographically detect pain in cancer patients with BM using AI. The ultimate goal of this line of research is to enable the early prediction and management of BM pain in patients before it becomes distressing. This project had three main objectives (each with respective sub-objectives) that built towards the overarching objective of developing, implementing, and evaluating an AI pipeline to detect pain in the simulation-CT images of cancer patients with BM.

Objective 1: Construct an NLP pipeline to extract pain scores from the consultation notes of patients: a) Use a publically-available de-identified clinical note database to develop a generalizable NLP pipeline to process unstructured clinical notes and quantify physician-reported pain scores. b) Verify the generalizability of the developed NLP pipeline by processing retrospectively-collected radiation oncology consultation notes of cancer patients with BM and extracting physician-reported pain scores from them.

Objective 2: Construct a radiomics pipeline to extract BM lesion features from radiographic images of patients: a) Build a novel tool for fast labeling of BM lesion-centerpoints in radiographic images (CT scans) and generate lesion-center-based geometric Regions Of Interest (ROIs). b) Develop a methodology to analyzesimulation CT images and extract radiomics imaging phenotypes using lesion-center-based geometric ROIs

(Chapter 4). c) Verify the feasibility of building a radiomics-based machine learning pipeline to distinguish between healthy and metastatic bone lesions using the methodology developed in objective 2.a.

Objective 3: Combine the NLP-quantified pain scores extracted in objective 1 with the radiomic features extracted in objective 2, to develop and evaluate a radiomics-based machine-learning model of pain in patients with BM.

1.10 Thesis organization

This thesis consists of six chapters and one appendix. Because it is a manuscript-based thesis, each chapter is self-contained, and there is some overlap of concepts and references between chapters.

Chapter 2 is broken into four sections that describe the methods employed to accomplish the objectives of this thesis: NLP, radiomics, and machine learning, and performance evaluation. Each section discusses the relevant mathematical and computational foundations as well as actual and potential clinical applications, challenges, and best practices.

Chapter 3 describes the first original manuscript based on objective 1. Two independent publicly-available de-identified hospital discharge summary corpora were used to build a generalizable NLP pipeline for pain extraction from clinical notes. Following that, the NLP pipeline was validated by analyzing and scoring pain in radiation oncology consultation notes of cancer patients treated at our institution.

Chapter 4 describes the second original manuscript, which is based on objective 2. In this study, a novel lesion-centerpoint labeling application was introduced, and then the feasibility of using it for radiomics-based studies was investigated. Using the introduced lesion-centerpoint-based geometric ROIs, a radiomics-based machine-learning pipeline was developed to separate metastatic bone lesions from healthy bone.

Chapter 5 describes the third original manuscript. In this study, methods from chapters 3 and 4 were combined to build an NLP-radiomics pipeline to detect pain in the simulation-CT images of patients with spinal BM treated using RT at our institution. The pipeline took into account the normalization of images as well as the imbalance in the proportion of patients who experience different levels of pain.

Chapter 6 summarizes the scientific originality of each manuscript and how they were integrated to meet the thesis's overarching objective. In addition, it discusses the significance of our findings for future prediction and management of BM pain, as well as some of the clinical limitations and potential research opportunities that are associated with them.

Bibliography

[1] A. Patel, Benign vs Malignant Tumors, JAMA Oncology 6 (9) (2020) 1488–1488. doi:10.1001/JAMAONCOL.2020.2592.
URL https://jamanetwork.com/journals/jamaoncology/fullarticle/2768634

[2] National Cancer Institute, Metastatic Cancer: When Cancer Spreads (11 2020).
URL https://www.cancer.gov/types/metastatic-cancer

[3] D. Hartl, S. Leboulleux, J. Hadoux, A. Berdelou, I. Breuskin, J. Guerlain, M. Schlumberger, TNM Classification, Surgery of the Thyroid and Parathyroid Glands (2022) 440–446doi:10.1016/B978-0-323-66127-0.00047-8.
URL https://www.ncbi.nlm.nih.gov/books/NBK553187/

[4] P. K. Newton, J. Mason, N. Venkatappa, M. S. Jochelson, B. Hurt, J. Nieva, E. Comen, L. Norton, P. Kuhn, Spatiotemporal progression of metastatic breast cancer: a Markov chain model highlighting the role of early metastatic sites, npj Breast Cancer 2015 1:1 1 (1) (2015) 1–9. doi:10.1038/npjbcancer.2015.18.
URL https://www.nature.com/articles/npjbcancer201518

[5] M. Riihimäki, H. Thomsen, K. Sundquist, J. Sundquist, K. Hemminki, Clinical landscape of cancer metastases, Cancer Medicine 7 (11) (2018) 5534. doi:10.1002/CAM4.1697.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6246954/

[6] J. Fares, M. Y. Fares, H. H. Khachfe, H. A. Salhab, Y. Fares, Molecular principles of metastasis: a hallmark of cancer revisited, Signal Transduction and Targeted Therapy 2020 5:1 5 (1) (2020) 1–17. doi:10.1038/s41392-020-0134-x.
URL https://www.nature.com/articles/s41392-020-0134-x

[7] L. Smith, S. Bryan, P. De, R. Rahal, A. Shaw, D. Turner, C. Manitoba, M. K. Hannah Weir, R. Woods, M. Dixon, Canadian Cancer Statistics Advisory Committee. Canadian Cancer Statistics, Canadian Cancer Society (2018).

[8] Martine Bomb, Sara Hiom, Harpal Kumar, Jodie Moffat, Nick Ormiston-Smith, Liz Woolf, Sarah Woolnough, Saving lives, averting costs An analysis of the financial implications of achieving earlier diagnosis of colorectal, lung and ovarian cancer , Tech. rep., Cancer Research UK (9 2014).
URL moz-extension://c0403e0d-3ea5-1849-973d-037c1c57bc1e/enhanced-reader.html?openApp&pdf=https%3A%2F%2Fwww.cancerresearchuk.org%2Fsites%2Fdefault%2Ffiles%2Fsaving_lives_averting_costs.pdf%3F_gl%3D1*huc8ce*_ga*MTc2Nzg0NzY3Ni4xNjYzMzQwMzIw*_ga_58736Z2GNN*MTY2MzUzODA3Ni40LjEuMTY2MzUzODEzOS42MC4wLjA.%26_ga%3D2.232784670.1280083254.1663538076-1767847676.1663340320

[9] R. E. Coleman, Roodman, Smith, Body, Suva, Vessella, Clinical Features of Metastatic Bone Disease and Risk of Skeletal Morbidity, Clinical Cancer Research 12 (20) (2006) 6243s–6249s. doi:10.1158/1078-0432.CCR-06-0931.
URL https://aacrjournals.org/clincancerres/article/12/20/6243s/191442/Clinical-Features-of-Metastatic-Bone-Disease-and

[10] F. Macedo, K. Ladeira, F. Pinho, N. Saraiva, N. Bonito, L. Pinto, F. Gonçalves, Bone Metastases: An Overview, Oncology reviews 11 (1) (2017). doi:10.4081/ONCOL.2017.321.
URL https://pubmed.ncbi.nlm.nih.gov/28584570/

[11] A. Jayarangaiah, A. K. Kemp, P. T. Kariyanna, Bone Metastasis, Functional Imaging in Oncology: Clinical Applications - Volume 2 (2022) 1389–1410doi:10.1007/978-3-642-40582-2{_}34.
URL https://www.ncbi.nlm.nih.gov/books/NBK507911/

[12] A. Turabi, A. R. Plunkett, The application of genomic and molecular data in the treatment of chronic cancer pain, Journal of surgical oncology 105 (5) (2012) 494–501. doi:10.1002/JSO.21707.
URL https://pubmed.ncbi.nlm.nih.gov/22441902/

[13] I. Ahmad, M. M. Ahmed, M. F. Ahsraf, A. Naeem, A. Tasleem, M. Ahmed, M. S. Farooqi, Pain Management in Metastatic Bone Disease: A Literature Review, Cureus 10 (9) (9 2018). doi:10.7759/CUREUS.3286.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6235631/

[14] S. Mercadante, Malignant bone pain: pathophysiology and treatment, Pain 69 (1-2) (1997) 1–18. doi:10.1016/S0304-3959(96)03267-8.
URL https://pubmed.ncbi.nlm.nih.gov/9060007/

[15] R. K. Portenoy, Treatment of cancer pain, Lancet (London, England) 377 (9784) (2011) 2236–2247. doi:10.1016/S0140-6736(11)60236-5.
URL https://pubmed.ncbi.nlm.nih.gov/21704873/

[16] C. F. Von Gunten, Pathophysiology of pain in cancer, Journal of Pediatric Hematology/Oncology 33 (SUPPL. 1) (4 2011). doi:10.1097/MPH.0B013E31821218A7.
URL https://journals.lww.com/jpho-online/Fulltext/2011/04001/Pathophysiology_of_Pain_in_Cancer.3.aspx

[17] J. Logan, C. Cluxton, The Challenge of Cancer Pain Assessment, The Ulster Medical Journal 88 (1) (2019) 43.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6342038/

[18] H. Nersesyan, K. V. Slavin, Current aproach to cancer pain management: Availability and implications of different treatment options, Therapeutics and Clinical Risk Management 3 (3) (2007) 381.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2386360/

[19] T. Nomiya, K. Teruyama, H. Wada, K. Nemoto, Time course of pain relief in patients treated with radiotherapy for cancer pain: A prospective study, Clinical Journal of Pain 26 (1) (2010) 38–42. doi:10.1097/AJP.0B013E3181B0C82C.
URL https://journals.lww.com/clinicalpain/Fulltext/2010/01000/Time_Course_of_Pain_Relief_in_Patients_Treated.6.aspx

[20] W. M. Sze, M. Shelley, I. Held, M. Mason, Palliation of metastatic bone pain: single fraction versus multifraction radiotherapy - a systematic review of the randomised

trials, The Cochrane database of systematic reviews 2002 (2) (1 2004). doi:10.1002/14651858.CD004721.
URL https://pubmed.ncbi.nlm.nih.gov/15106258/

[21] F. De Felice, A. Piccioli, D. Musio, V. Tombolini, The role of radiation therapy in bone metastases management, Oncotarget 8 (15) (2017) 25691. doi:10.18632/ONCOTARGET.14823.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5421962/

[22] R. McDonald, K. Ding, M. Brundage, R. M. Meyer, A. Nabid, P. Chabot, G. Coulombe, S. Ahmed, J. Kuk, A. R. Dar, A. Mahmud, A. Fairchild, C. F. Wilson, J. S. Wu, K. Dennis, C. DeAngelis, R. K. Wong, L. Zhu, S. Chan, E. Chow, Effect of Radiotherapy on Painful Bone Metastases: A Secondary Analysis of the NCIC Clinical Trials Group Symptom Control Trial SC.23, JAMA Oncology 3 (7) (2017) 953–959. doi:10.1001/JAMAONCOL.2016.6770.
URL https://jamanetwork.com/journals/jamaoncology/fullarticle/2601221

[23] Y. Lievens, M. Gospodarowicz, S. Grover, D. Jaffray, D. Rodin, J. Torode, M. L. Yap, E. Zubizarreta, Global impact of radiotherapy in oncology: Saving one million lives by 2035, Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology 125 (2) (2017) 175–177. doi:10.1016/J.RADONC.2017.10.027.
URL https://pubmed.ncbi.nlm.nih.gov/29173397/

[24] A. Taylor, M. E. Powell, Intensity-modulated radiotherapy—what is it?, Cancer Imaging 4 (2) (2004) 68. doi:10.1102/1470-7330.2004.0003.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1434586/

[25] K. Spencer, R. Parrish, R. Barton, A. Henry, Palliative radiotherapy, BMJ 360 (3 2018). doi:10.1136/BMJ.K821.
URL https://www.bmj.com/content/360/bmj.k821

[26] S. T. Lutz, J. Jones, E. Chow, Role of radiation therapy in palliative care of the patient with cancer, Journal of clinical oncology : official journal of the American Society of

Clinical Oncology 32 (26) (2014) 2913–2919. doi:10.1200/JCO.2014.55.1143.
URL https://pubmed.ncbi.nlm.nih.gov/25113773/

[27] K. Spencer, E. Morris, E. Dugdale, A. Newsham, D. Sebag-Montefiore, R. Turner, G. Hall, A. Crellin, 30 day mortality in adult palliative radiotherapy–A retrospective population based study of 14,972 treatment episodes, Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology 115 (2) (2015) 264–271. doi:10.1016/J.RADONC.2015.03.023.
URL https://pubmed.ncbi.nlm.nih.gov/25861831/

[28] E. Chow, K. Harris, G. Fan, M. Tsao, W. M. Sze, Palliative radiotherapy trials for bone metastases: a systematic review, Journal of clinical oncology : official journal of the American Society of Clinical Oncology 25 (11) (2007) 1423–1436. doi:10.1200/JCO.2006.09.5281.
URL https://pubmed.ncbi.nlm.nih.gov/17416863/

[29] T. Morita, T. Akechi, M. Ikenaga, Y. Kizawa, H. Kohara, T. Mukaiyama, T. Nakaho, N. Nakashima, Y. Shima, T. Matsubara, Y. Uchitomi, Late referrals to specialized palliative care service in Japan, Journal of clinical oncology : official journal of the American Society of Clinical Oncology 23 (12) (2005) 2637–2644. doi:10.1200/JCO.2005.12.107.
URL https://pubmed.ncbi.nlm.nih.gov/15728219/

[30] C. Zimmermann, N. Swami, M. Krzyzanowska, B. Hannon, N. Leighl, A. Oza, M. Moore, A. Rydall, G. Rodin, I. Tannock, A. Donner, C. Lo, Early palliative care for patients with advanced cancer: A cluster-randomised controlled trial, The Lancet 383 (9930) (2014) 1721–1730. doi:10.1016/S0140-6736(13)62416-2.
URL http://www.thelancet.com/article/S0140673613624162/fulltext

[31] D. B. Rosen, C. D. Benjamin, J. C. Yang, C. Doyle, Z. Zhang, C. A. Barker, M. Vaynrub, T. J. Yang, E. F. Gillespie, Early palliative radiation versus observation for high-risk asymptomatic or minimally symptomatic bone metastases: study protocol for a randomized controlled trial, BMC Cancer 20 (1) (2020) 1–12. doi:10.1186/S12885-020-07591-W/TABLES/3.

URL https://bmccancer.biomedcentral.com/articles/10.1186/s12885-020-07591-w

[32] C. Zimmermann, R. Riechelmann, M. Krzyzanowska, G. Rodin, I. Tannock, Effectiveness of specialized palliative care: a systematic review, JAMA 299 (14) (2008) 1698–1709. doi:10.1001/JAMA.299.14.1698.
URL https://pubmed.ncbi.nlm.nih.gov/18398082/

[33] J. S. Temel, J. A. Greer, A. Muzikansky, E. R. Gallagher, S. Admane, V. A. Jackson, C. M. Dahlin, C. D. Blinderman, J. Jacobsen, W. F. Pirl, J. A. Billings, T. J. Lynch, Early Palliative Care for Patients with Metastatic Non–Small-Cell Lung Cancer, New England Journal of Medicine 363 (8) (2010) 733–742. doi:10.1056/NEJMOA1000678/SUPPL{_}FILE/NEJMOA1000678{_}DISCLOSURES.PDF.
URL https://www.nejm.org/doi/full/10.1056/nejmoa1000678

[34] E. Bandieri, D. Sichetti, M. Romero, C. Fanizza, M. Belfiglio, L. Buonaccorso, F. Artioli, F. Campione, G. Tognoni, M. Luppi, Impact of early access to a palliative/supportive care intervention on pain management in patients with cancer, Annals of oncology : official journal of the European Society for Medical Oncology 23 (8) (2012) 2016–2020. doi:10.1093/ANNONC/MDS103.
URL https://pubmed.ncbi.nlm.nih.gov/22565123/

[35] T. J. Keay, The mind-set of pain assessment (1 2005). doi:10.1016/j.jamda.2004.12.011.

[36] K. O. Anderson, Assessment tools for the evaluation of pain in the oncology patient, Current pain and headache reports 11 (4) (2007) 259–264. doi:10.1007/S11916-007-0201-9.
URL https://pubmed.ncbi.nlm.nih.gov/17686388/

[37] M. Haefeli, A. Elfering, Pain assessment, European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society 15 Suppl 1 (Suppl 1) (1 2006). doi:10.1007/S00586-005-1044-X.
URL https://pubmed.ncbi.nlm.nih.gov/16320034/

[38] J. Raphael, J. Hester, S. Ahmedzai, J. Barrie, P. Farqhuar-Smith, J. Williams, C. Urch, M. I. Bennett, K. Robb, B. Simpson, M. Pittler, B. Wider, C. Ewer-Smith, J. DeCourcy, A. Young, C. Liossi, R. McCullough, D. Rajapakse, M. Johnson, R. Duarte, E. Sparkes, Cancer Pain: Part 2: Physical, Interventional and Complimentary Therapies; Management in the Community; Acute, Treatment-Related and Complex Cancer Pain: A Perspective from the British Pain Society Endorsed by the UK Association of Palliative Medicine and the Royal College of General Practitioners, Pain Medicine 11 (6) (2010) 872–896. doi:10.1111/J.1526-4637.2010.00841.X.
URL https://academic.oup.com/painmedicine/article/11/6/872/1853290

[39] S. Deandrea, M. Montanari, L. Moja, G. Apolone, Prevalence of undertreatment in cancer pain. A review of published literature, Annals of oncology : official journal of the European Society for Medical Oncology 19 (12) (2008) 1985–1991. doi:10.1093/ANNONC/MDN419.
URL https://pubmed.ncbi.nlm.nih.gov/18632721/

[40] S. M. Weinstein, D. Romanus, E. M. Lepisto, C. Reyes-Gibby, C. Cleeland, R. Greene, C. Muir, J. Niland, Documentation of pain in comprehensive cancer centers in the United States: a preliminary analysis, Journal of the National Comprehensive Cancer Network : JNCCN 2 (2) (2004) 173–180. doi:10.6004/JNCCN.2004.0015.
URL https://pubmed.ncbi.nlm.nih.gov/19777706/

[41] R. Adam, P. Murchie, Why are we not controlling cancer pain adequately in the community?, The British journal of general practice : the journal of the Royal College of General Practitioners 64 (626) (2014) 438–439. doi:10.3399/BJGP14X681229.
URL https://pubmed.ncbi.nlm.nih.gov/25179046/

[42] K. Nuseir, M. Kassab, B. Almomani, Healthcare Providers' Knowledge and Current Practice of Pain Assessment and Management: How Much Progress Have We Made?, Pain research & management 2016 (2016). doi:10.1155/2016/8432973.
URL https://pubmed.ncbi.nlm.nih.gov/27965524/

[43] M. J. Hjermstad, J. Gibbins, D. F. Haugen, A. Caraceni, J. H. Loge, S. Kaasa, Pain assessment tools in palliative care: an urgent need for consensus, Palliative medicine

22 (8) (2008) 895–903. doi:10.1177/0269216308095701.
URL https://pubmed.ncbi.nlm.nih.gov/18799513/

[44] E. Veirman, D. M. Van Ryckeghem, A. De Paepe, O. J. Kirtley, G. Crombez, Multidimensional screening for predicting pain problems in adults: A systematic review of screening tools and validation studies, Pain Reports 4 (5) (9 2019). doi:10.1097/PR9.0000000000000775.
URL https://journals.lww.com/painrpts/Fulltext/2019/10000/Multidimensional_screening_for_predicting_pain.7.aspx

[45] S. Falk, A. H. Dickenson, Pain and nociception: mechanisms of cancer-induced bone pain, Journal of clinical oncology : official journal of the American Society of Clinical Oncology 32 (16) (2014) 1647–1654. doi:10.1200/JCO.2013.51.7219.
URL https://pubmed.ncbi.nlm.nih.gov/24799469/

[46] E. Bruera, T. Schoeller, R. Wenk, T. Maceachern, S. Marcelino, J. Hanson, M. Suarez-Almazor, A Prospective Multicenter Assessment of the Edmonton Staging System for Cancer Pain, Journal of Pain and STmptom Ma˜t 10 (1995) 348–355.

[47] R. I. Walsh, G. Mitchell, L. Francis, M. L. Van Driel, What Diagnostic Tools Exist for the Early Identification of Palliative Care Patients in General Practice? A systematic review, Journal of palliative care 31 (2) (2015) 118–123. doi:10.1177/082585971503100208.
URL https://pubmed.ncbi.nlm.nih.gov/26201214/

[48] A. Health Services, Integrating an Early Palliative Approach into Advanced Cancer Care, Cancer Care Alberta (2021).
URL http://www.ahs.ca/guru

[49] J. Downar, R. Goldman, R. Pinto, M. Englesakis, N. K. Adhikari, The ”surprise question” for predicting death in seriously ill patients: a systematic review and meta-analysis, CMAJ : Canadian Medical Association journal = journal de l’Association medicale canadienne 189 (13) (2017) E484–E493. doi:10.1503/CMAJ.160775.
URL https://pubmed.ncbi.nlm.nih.gov/28385893/

[50] B. O'Neill, A. Rodway, ABC of palliative care: Care in the community, BMJ 316 (7128) (1998) 373–377. doi:10.1136/BMJ.316.7128.373.
URL https://www.bmj.com/content/316/7128/373

[51] C. Clifford, K. Thomas, J. Armstrong-Wilson, Going for Gold: the Gold Standards Framework programme and accreditation in primary care, End of Life Journal 6 (e000028) (2016). doi:10.1136/eoljnl.
URL http://eolj.bmj.com/

[52] Y.-F. Yen, Y.-L. Lee, H.-Y. Hu, W.-J. Sun, M.-C. Ko, C.-C. Chen, W. K. Wong, D. E. Morisky, S.-J. Huang, D. Chu, Early palliative care: the surprise question and the palliative care screening tool-better together, BMJ Supportive & Palliative Care 12 (2022) 211–217. doi:10.1136/bmjspcare-2019-002116.
URL http://spcare.bmj.com/

[53] K. L. Schreiber, N. Zinboonyahgoon, K. M. Flowers, V. Hruschak, K. G. Fields, M. E. Patton, E. Schwartz, D. Azizoddin, M. Soens, T. King, A. Partridge, A. Pusic, M. Golshan, R. R. Edwards, Prediction of Persistent Pain Severity and Impact 12 Months After Breast Surgery Using Comprehensive Preoperative Assessment of Biopsychosocial Pain Modulators, Annals of surgical oncology 28 (9) (2021) 5015. doi:10.1245/S10434-020-09479-2.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8280248/

[54] S. S. Hwang, V. T. Chang, D. L. Fairclough, B. Kasimis, Development of a cancer pain prognostic scale, Journal of Pain and Symptom Management 24 (4) (2003) 366–378. doi:10.1016/S0885-3924(02)00488-8.
URL https://pubmed.ncbi.nlm.nih.gov/12505205/

[55] J. Downar, P. Wegier, P. Tanuseputro, Early Identification of People Who Would Benefit From a Palliative Approach—Moving From Surprise to Routine, JAMA Network Open 2 (9) (2019) e1911146–e1911146. doi:10.1001/JAMANETWORKOPEN.2019.11146.
URL https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2749774

[56] Y. ElMokhallalati, S. H. Bradley, E. Chapman, L. Ziegler, F. E. Murtagh, M. J. Johnson, M. I. Bennett, Identification of patients with potential palliative care needs: A systematic review of screening tools in primary care, Palliative Medicine 34 (8) (2020) 989–1005. doi:10.1177/0269216320929552.
URL https://journals.sagepub.com/doi/10.1177/0269216320929552

Chapter 2

An introduction to artificial intelligence tools in medicine

2.1 Natural Language Processing

NLP is a computational strategy that synthesizes large databases of free-text data and extracts meaningful quantitative data from them. This chapter focuses on the workflow and application of NLP algorithms to medical text documents for precision medicine.

2.1.1 Electronic health records

Electronic Health Records (EHRs) are digitized medical records of patients that are used in daily healthcare administration, delivery, and research. These records include tabular data (such as administrative and billing data, patient demographics, and lab and test results), unstructured data (such as clinician-written medical histories, consultation notes, and discharge summary notes), and imaging data (like radiology images).

More than half of all EHRs and up to 70% of patient data utilized by doctors and outcome researchers are unstructured data (primarily in the form of free-text) [1]. Because they have little to no content, organization, or quality standardization, these unstructured notes are rarely used to improve clinical decision support at the point of care, rather than before or after it. Thus, to this end, it would be extremely beneficial if such abundance of free-text data could be converted to quantitative data.

NLP combined with Machine Learning (ML) provides a computational means for synthesizing massive databases of free-text data and extracting quantitative structured data from them. The NLP algorithms have the potential to be used to improve clinical decision support by drawing conclusions from EHRs at the point of care when tabular or patient-reported structured data are not available [2].

2.1.2 Natural language processing for clinical decision support

For over 20 years, NLP algorithms have been utilized on clinical notes for a variety of clinical point of care decision support applications including outcome prediction, keyword searches, diagnosis categorization, and cancer phenotyping extraction [3–8].

For information extraction, there are two basic NLP approaches: rule-based, and ML-based. Each method's success is determined by the nature of the problem. In a rule-based approach, predefined sets of rules are used to process text and extract information. This method works well when the desired goal can be attained with a small set of rules. A rule-based strategy, for example, is the best choice for extracting a given symptom or outcome from a particular sort of semi-organized clinical note [9]. However, when the target information appears in a wide variety of contexts within the free text and cannot be extracted with a small number of rules, the rule-based approach fails [10]. In this thesis, we will discuss several strategies for addressing this challenge to construct generalizable rule-based NLP algorithms for symptom extraction.

In ML-based methods, supervised computer algorithms are trained to discover patterns from a labeled set of free text notes. Models based on ML are often trained on a large dataset of labeled data. In general, ML-based approaches perform worse than rule-based approaches despite their better generalizability [9]. As a result, these models are better suited to projects with a lot of data variance or where the NLP rules are not easily applied. For example, ML-based natural language processing has been used to extract tumor features and event distribution [7, 11].

2.1.3 Thesaurus of biomedical vocabulary

To improve the effectiveness of medical information retrieval, NLP pipelines are normally combined with standard thesauri of biomedical vocabulary. Such thesauri are designed to categorize clinical terminologies into standardized tables with a unique code for each medical concept. Several studies have proved that the use of a medical thesaurus can significantly improve the retrieval efficiency of an NLP pipeline for data mining, keyword search, abbreviation mapping, and extracting hierarchical relationships between medical terminologies [12]. The Unified Medical Language System (UMLS) Metathesaurus® [13], maintained by the US National Library of Medicine (NLM), is the largest and the most commonly used thesaurus in the biomedical domain.

Metathesaurus aggregates more than 200 biomedical databases containing over two million terms, 900,000 biomedical concepts, and 12 million relationships. Metathesaurus includes vocabulary from various databases, including ICD [14], SNOMED CT [15], MeSH [16], and Gene Ontology [17]. The majority of them are updated on a weekly basis. Figure 2.1 illustrates some of the databases that are included in Metathesaurus. It provides medical definitions for each concept, together with its lexical variants, synonyms, hierarchical relationships with other concepts, and semantic tags. Semantic tags, assigned by Metathesaurus editors, are used to categorize each concept to one of the 135 high-level categories [https://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html]. Examples of semantic tags are "Pharmacologic Substance", "Clinical Drug", "Sign or Symptom", and "Disease or Syndrome". Metathesaurus is a freely accessible database (under the UMLS user agreement and license https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html) that may be downloaded to local repositories or queried using SQL, XML, and APIs.

2.1.4 MetaMap NLP tool

In addition to the UMLS Metathesaurus, the NLM offers the MetaMap NLP tool [18, 19], which extracts biomedical terms from clinical text and maps them to the UMLS concepts. MetaMap is widely used in medical NLP applications [20]. The MetaMap's processing workflow is represented in Figure 2.2. First, MetaMap breaks the text into tokens using "space delimiters" and maps user defined acronyms. Then, it uses the SPECIALIST Lexicon [21] algorithm to find and parse the sentences, and extract all the phrases. Next, MetaMap finds all the possible variants of each phrase and maps them to the UMLS Metathesaurus and

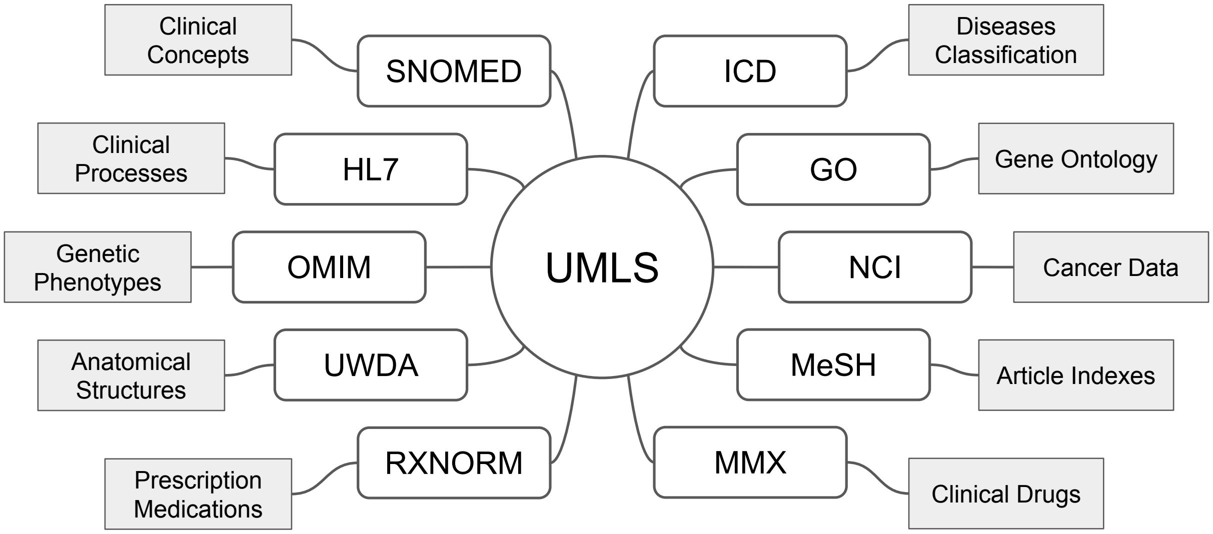


Figure 2.1: Some of the UMLS Metathesaurus vocabulary databases. Full list of the databases is available at (https://www.nlm.nih.gov/research/umls/sourcereleasedocs/index.html).

extract biomedical concepts. Finally, as explained in [22, 23], MetaMap assesses the concept's relevance to the original text and assigns a similarity score for all matching Metathesaurus concepts. MetaMap has a number of add-on packages that can be enabled or disabled. For example, Word Sense Disambiguation (WSD) is a tool that evaluates the semantic consistency of each mapped concept with its neighbors and removes disambiguation [24], and the NegEx, negation detection, algorithm is a tool for determining whether or not medical phrases in the corpus have been negated.

MetaMap is highly configurable in its processing and output options. The default human readable output includes parsed phrases as well as all Metathesaurus concepts that map some or all of the phrases, preferred names for concepts, negation indexes, similarity scores, and UMLS semantic tags (see Figure 2.3). Since MetaMap is a phrase-by-phrase processor and generates many potential mappings for each phrase, it is relatively slow. For example,

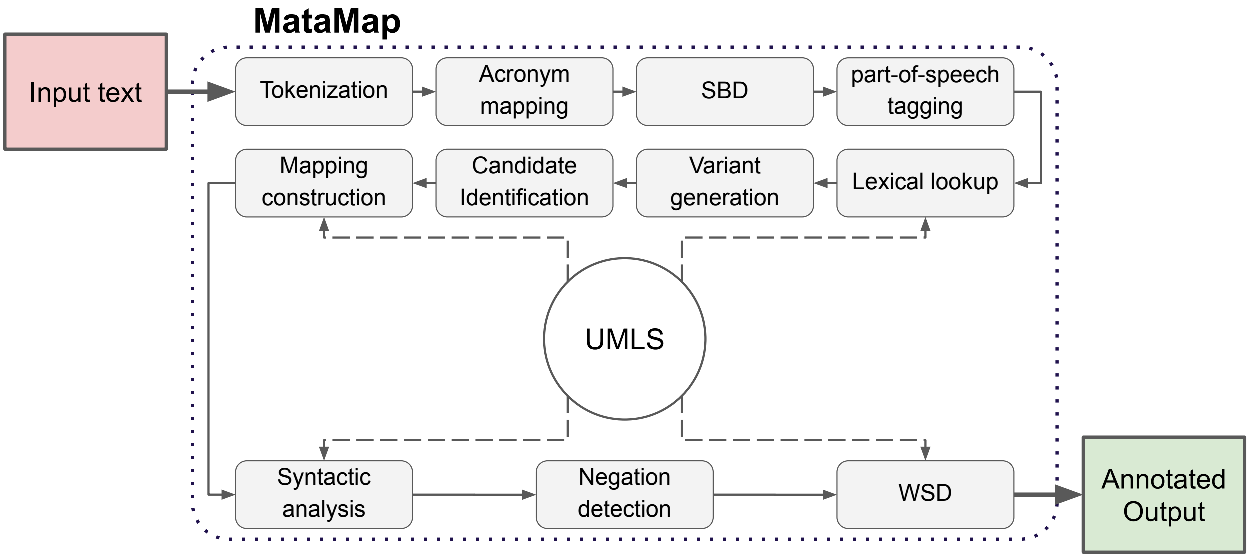
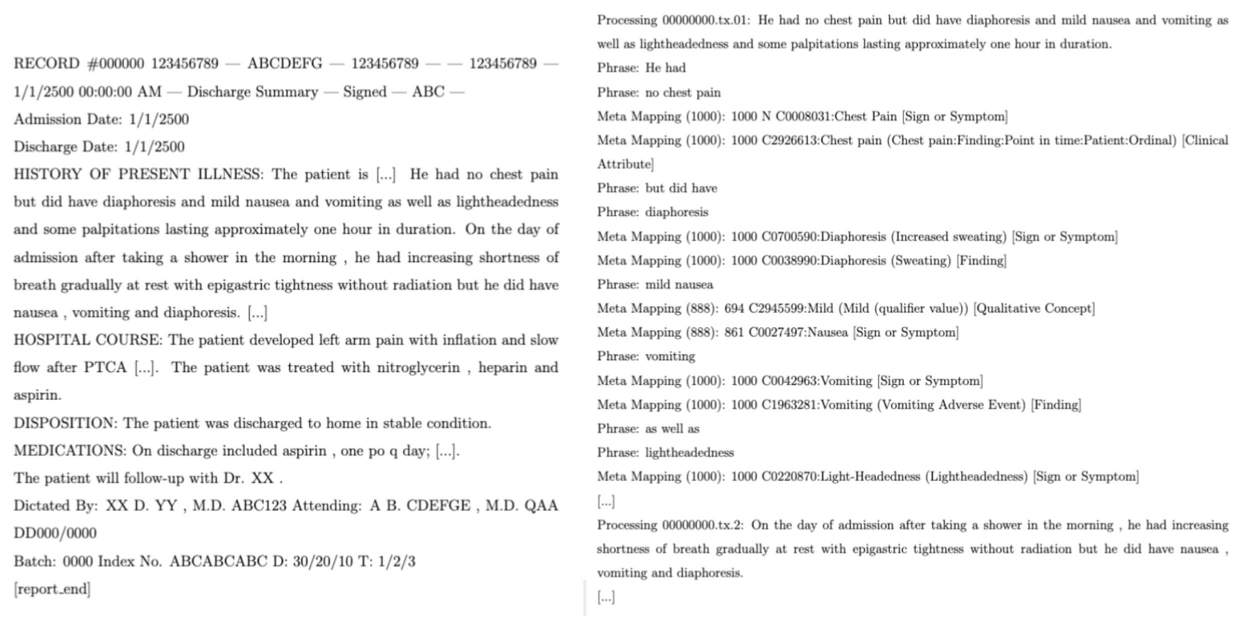


Figure 2.2: MetaMap text processing system. SBP, sentence boundary disambiguation; WSD word sense disambiguation. Figure is created based on [18, 21]

it might take a few minutes to process a two-page consultation note.

2.2 Radiomics

RT relies heavily on radiologic images. They have been utilized by clinicians as noninvasive tools for tumor assessment and treatment planning for decades, providing biological and functional information about tumors and the microenvironment around them. Thanks to recent advances in medical image processing techniques, radiography images have advanced well beyond simplistic tumor viewing tools. Image processing (texture analysis) techniques made it possible to extract hundreds of quantitative and minable imaging phenotype data (called radiomic features) from radiography images. Lambin et al. [25, 26] introduced radiomics in 2012, and thousands of studies have subsequently demonstrated itspotential application in point of care clinical decision support systems [27] and precision medicine [28]. Radiomics has been shown to be an effective tool for evaluating therapy efficiency and responsiveness [29], assisting with early diagnosis [30], and predicting treatment outcomes [31].



RECORD #000000 123456789 — ABCDEFG — 123456789 — — 123456789 — 1/1/2500 00:00:00 AM — Discharge Summary — Signed — ABC —

Admission Date: 1/1/2500

Discharge Date: 1/1/2500

HISTORY OF PRESENT ILLNESS: The patient is [...] He had no chest pain but did have diaphoresis and mild nausea and vomiting as well as lightheadedness and some palpitations lasting approximately one hour in duration. On the day of admission after taking a shower in the morning , he had increasing shortness of breath gradually at rest with epigastric tightness without radiation but he did have nausea , vomiting and diaphoresis. [...]

HOSPITAL COURSE: The patient developed left arm pain with inflation and slow flow after PTCA [...]. The patient was treated with nitroglycerin , heparin and aspirin.

DISPOSITION: The patient was discharged to home in stable condition.

MEDICATIONS: On discharge included aspirin , one po q day; [...].

The patient will follow-up with Dr. XX .

Dictated By: XX D. YY , M.D. ABC123 Attending: A B. CDEFGE , M.D. QAA DD000/0000

Batch: 0000 Index No. ABCABCABC D: 30/20/10 T: 1/2/3

[report_end]

```
Processing 00000000.tx.01: He had no chest pain but did have diaphoresis and mild nausea and vomiting as well as lightheadedness and some palpitations lasting approximately one hour in duration.
Phrase: He had
Phrase: no chest pain
Meta Mapping (1000): 1000 N C0008031:Chest Pain [Sign or Symptom]
Meta Mapping (1000): 1000 C2926613:Chest pain (Chest pain:Finding:Point in time:Patient:Ordinal) [Clinical Attribute]
Phrase: but did have
Phrase: diaphoresis
Meta Mapping (1000): 1000 C0700590:Diaphoresis (Increased sweating) [Sign or Symptom]
Meta Mapping (1000): 1000 C0038990:Diaphoresis (Sweating) [Finding]
Phrase: mild nausea
Meta Mapping (888): 694 C2945599:Mild (Mild (qualifier value)) [Qualitative Concept]
Meta Mapping (888): 861 C0027497:Nausea [Sign or Symptom]
Phrase: vomiting
Meta Mapping (1000): 1000 C0042963:Vomiting [Sign or Symptom]
Meta Mapping (1000): 1000 C1963281:Vomiting (Vomiting Adverse Event) [Finding]
Phrase: as well as
Phrase: lightheadedness
Meta Mapping (1000): 1000 C0220870:Light-Headedness (Lightheadedness) [Sign or Symptom]
[...]
Processing 00000000.tx.2: On the day of admission after taking a shower in the morning , he had increasing shortness of breath gradually at rest with epigastric tightness without radiation but he did have nausea , vomiting and diaphoresis.
[...]
```

Figure 2.3: An example of an unprocessed note (left) and a MetaMap-processed note (right). For better visualization, some parts of the text have been eliminated and replaced with "[...]".

Our focus in this thesis project was on the application of radiomics in palliative RT cancer pain detection.

2.2.1 Radiomics workflow

Figure 2.4 illustrates the six essential steps in the training phase of a typical radiomics-ML study for outcome prediction: (1) image acquisition, (2) manual or semi-automated segmentation of Regions Of Interest (ROIs) on patient images, (3) feature extraction from the segmented ROIs, (4) Feature Selection (FS), (5) building of a classification model to correlate extracted features to each patient's endpoint outcome data, and (6) performance evaluation.

While, first three steps are exclusive to radiomics-based research, steps four through six are standard in any ML process.

This section goes through the first four steps described above, as well as their clinical challenges and applications. Sections 2.3 and 2.4 are devoted to delving deeper into the last

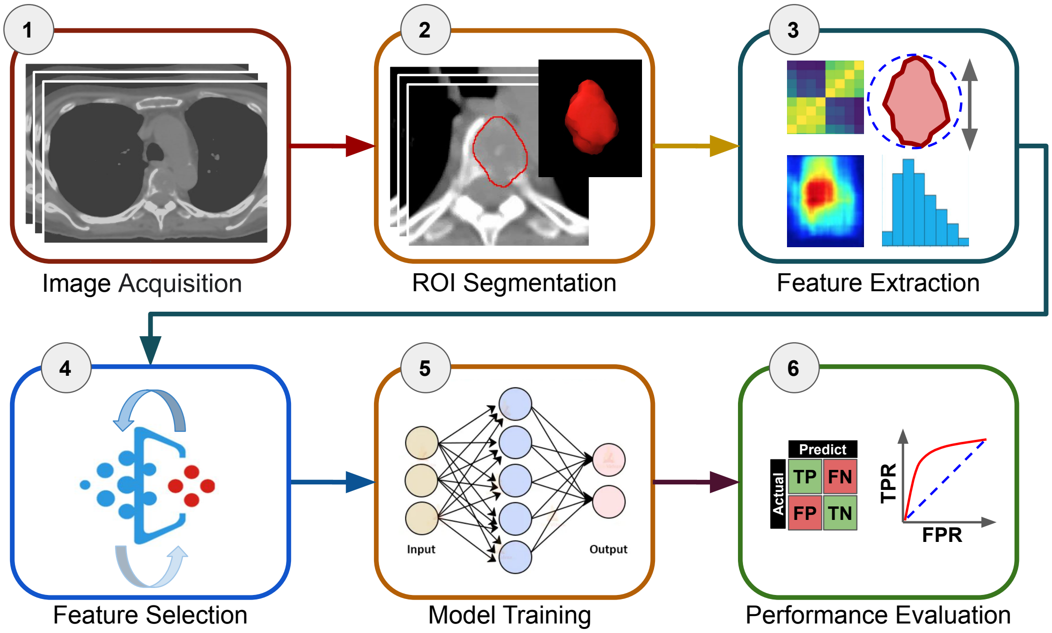


Figure 2.4: A typical workflow of a radiomics-based study. Following the acquisition of the images and the segmentation of the tumor, radiomic features are retrieved with the use of standard libraries. Then, statistical modeling incorporating Feature Selection (FS) and ML is used for outcome prediction, disease categorization, patient clustering, or risk stratification. Using an independent test set, the model's ability to achieve the desired performance is then tested. ROI: region of interest.

two steps.

2.2.2 Image acquisition

Any 2D or 3D medical image can be utilized as a source of radiomic features. However, practically all radiomics models are based on one or more of the three main 3D image acquisition modalities: Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Positron Emission Tomography (PET). This thesis focuses on the use of CT images to extract radiomic features because CT images are often the only source of imaging for palliative RT patients.

The accuracy and reproducibility of radiomic features are determined by the quality and consistency of image acquisition procedures. Changes in image acquisition parameters (like tube voltage, tube current, and slice thickness) can cause variations in the image texture. Prior to radiomic feature extraction, intensity normalization is often used to standardize pixel values and eliminate data acquisition-induced image texture fluctuations. The pixel intensities in CT scans are proportional to tissue attenuation. As a result, by converting pixel values to Hounsfield Units (HU), data acquisition variations can be minimized. When it comes to imaging modalities where pixel values do not have any physical meaning, histogram matching is commonly utilized for image normalization.

Hounsfield Units conversion

CT scans are conducted by rotating X-ray fan beams around the object of interest (see Figure 2.5). As an X-ray beam passes through the body, it is attenuated by tissues according to their electron densities.

The pixel intensity (I) at each angle is determined by the mean attenuation of the tissues that the X-rays pass through and can be described mathematically as follows:

$$I = I_0 \times e^{-\int_0^s \mu(s')ds'} \tag{2.1}$$

where I_0 is the background intensity and $\mu(s')$ is the attenuation of the tissue at location s' along the path s. After the X-ray source completes one full rotation, the CT computer uses a mathematical back-projection algorithm to produce a 2D image slice. CT attenuation values can be represented as HU on a linear density scale. Water is given a value of 0 on the HU scale and all other CT values are calculated using the following expression:

$$HU = 1000 \times \frac{\mu - \mu_{water}}{\mu_{water} - \mu_{air}} \tag{2.2}$$

where μ_{water}, μ_{air}, and μ are the CT linear attenuation coefficients of water, air, and tissue, respectively. The approximate HU values for tissues commonly detected on CT scans are listed in Table 2.1.

Contrast enhancement is one application of HU, which is used to highlight specific structures in an image. It can also be used to generate masks to filter specific structures. For example, HU > 0 removes pixels associated with fat and air.

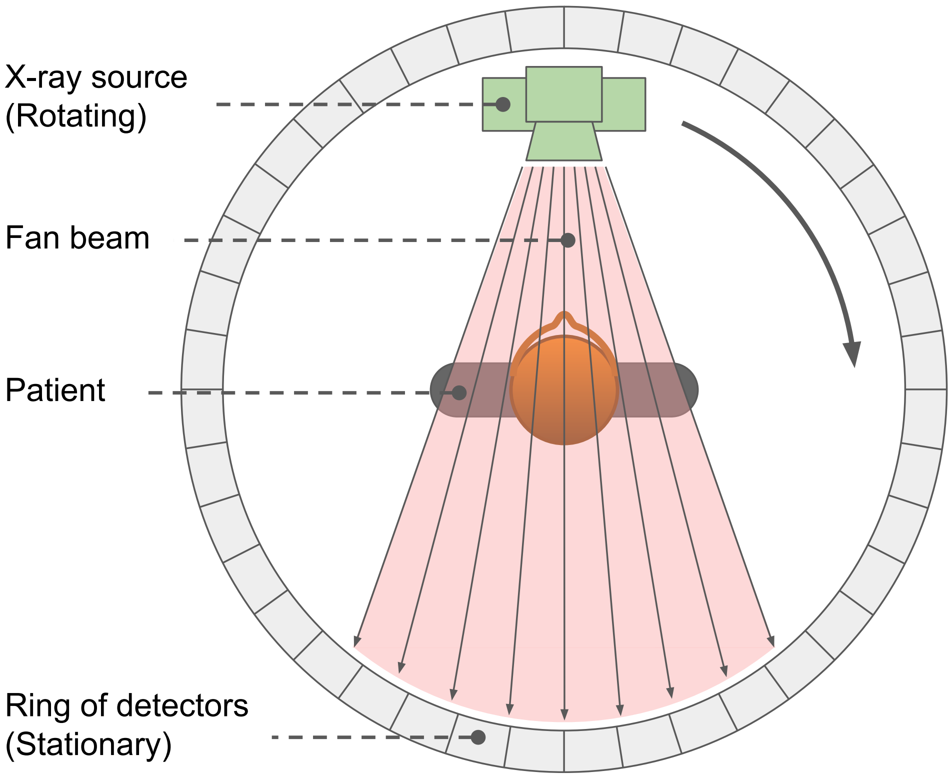


Figure 2.5: Basic principles of a fan beam CT scanner. Figure is generated based on [32].

Substance	Hounsfield Unit
Air	-1000
Lung	-400 → -600
Fat	-60 → -100
Water	0
Soft tissue	40 → 80
Bone	400 → 3000

Table 2.1: Approximate HU values for common tissues in a chest CT scan [33].

2.2.3 Segmentation of regions of interest

The image voxels (3D pixels) that are contained within the tumor ROI are used to calculate radiomic features. Therefore, ROI segmentation is another critical stage that has a direct impact on the quality and reproducibility of radiomics-based investigations. For radiomics research, ROI segmentation is usually done manually or semi-automatically. The most prevalent form of ROI segmentation is manual segmentation by trained physicians. Although manual segmentation is more reliable and benefits from expert knowledge, it has the disadvantage of being time-consuming and exhibiting high intraobserver variability [34, 35].

While fully automated segmentation of brain tumor from multiparametric MRI images has been recently developed [36–38], to the best of our knowledge, there is currently no CT-based fully automated segmentation technique that can detect tumor locations and properly distinguish them from adjacent tissues. Therefore, some manual intervention is required to assure the tumor location and accuracy of segmentation. Semi-automated segmentation combines the benefits of both human intervention and software automation, making it a desirable option for radiomics studies.

Computer-aided segmentation algorithms

There are three types of computer-aided segmentation algorithms: threshold-based, texture-based, and ML-based. The most basic algorithms are threshold-based algorithms, which are commonly employed as a starting point for subsequent segmentation techniques. The most widely utilized threshold-based ROI segmentation criteria are gray level and histogram thresholds [39]. Texture-based segmentation algorithms, such as boundary detection, region expanding, and homogeneity mapping, are typically less generalizable than other segmentation algorithms. In recent years, ML-based segmentation methods like supervised pattern recognition and neural network classifiers have gained popularity [40]. In Chapter 4, we demonstrate that threshold-based ROI segmentation may be sufficient for palliative-intent radiomics modeling of BM pain.

2.2.4 Radiomic features extraction

Radiomic features are imaging phenotype data that are derived mathematically from ROIs in radiography images. These macroscopic features provide insight into disease processes at the molecular level and have relationships with tumor diagnostics and prognosis. Despite the growing interest in radiomics, the majority of published radiomics models are not yet repeatable or generalizable or in clinical use [41, 42]. This is due to the lack of established definitions of radiomic features with verifiable reference values as well as inconsistent implementation of the image processing algorithms required to compute features [43]. In 2016, the Image Biomarker Standardization Initiative (IBSI) was established to create radiomic feature definitions as well as an image processing scheme for calculating standardized radiomic features from images [44]. Since then, most radiomics software packages have been updated to comply with IBSI [45, 46]. ISBI-compliant software include IBEX [47], PyRadiomics [48], RaCaT [49], SERA [50], and ROdiomiX [51]. Radiomic features can be calculated on either the original image or a derived image generated by applying one of various filters to the original image. In this thesis, we used PyRadiomics and extracted 107 standard radiomic features from the original CT images, which are divided into the following categories:

- First Order (19 features),
- Shape-based (26 features),
- Gray Level Co-occurrence Matrix (24 features),
- Gray Level Run Length Matrix (16 features),
- Gray Level Size Zone Matrix (16 features),
- Neighboring Gray Tone Difference Matrix (5 features), and
- Gray Level Dependence Matrix (14 features).

Feature definitions are available in the PyRadiomics documentation [48] (https://pyradiomics.readthedocs.io/en/latest/index.html).

Feature-Label space

Extracted features can be presented in the form of a matrix, $\mathbf{X}$, which consists of n rows (patients in our case) and p columns (radiomic features in our case),

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{pmatrix} \tag{2.3}$$

in which $x_{i,j}$ is the numerical value of the imaging feature j, $(j = 1, 2, ..., p)$ calculated for patient, i, $(i = 1, 2, ..., n)$. In this notation, each row is a vector presenting imaging features for a given patient i;

$$\mathbf{x}_i = \begin{pmatrix} x_{i,1} & x_{i,2} & \dots & x_{i,p} \end{pmatrix} \tag{2.4}$$

Each column is a vector presenting the numerical values of a given feature, j, over all patients.

$$\mathbf{x}_j = \begin{pmatrix} x_{1,j} \\ x_{2,j} \\ \vdots \\ x_{n,j} \end{pmatrix} \tag{2.5}$$

The goal of a typical radiomics-based outcome prediction model is to find a correlation between the extracted radiomic features and the desired tumor outcome. For example, one of the goals of this thesis is to construct a radiomics-based model that can distinguish between painful and painless BM lesions. In this case, we can encode outcomes as binary labels, with 1 indicating a patient with “pain” and 0 indicating a patient with “no pain”. Then, outcomes can be represented as a vector, $\mathbf{y}$, with y_i representing the outcome label for patient i $(y_i \in \{0, 1\})$.

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \tag{2.6}$$

Given examples of $\{x_i, y_i\}$ $(i = 1, 2, \ldots, n)$, am ML model can determine the optimum function, f, that satisfies

$$\mathbf{y} \approx f(\mathbf{x}). \tag{2.7}$$

Section 2.3 will go over some of the ML models, as well as the metrics that each uses to determine the "best" f.

2.2.5 Re-sampling techniques

Because of the nature of medical datasets, it is common to experience a significant imbalance in the number of samples for each outcome label class. In our BM cancer patient data collection, for instance, 86 % of patients reported "pain" $(y = 1)$ while only 14 % reported "no pain" $(y = 0)$. In these kinds of situation, ML models are normally biased in favor of the majority class. This is due to the fact that they disregard the relative distribution of each class in favor of maximizing the overall accuracy. One strategy to address the class imbalance problem is to perform a re-sampling of the training dataset. Re-sampling balances the quantity of samples in each label class, by either ignoring samples from the majority class (undersampling), or duplicating samples from the minority class, (oversampling). Four main re-sampling techniques that we will examine in this thesis are random under-sampling, TomekLinks undersampling [52], Random over-sampling, and Synthetic Minority Oversampling Technique (SMOTE) [53]. Random under-sampling randomly deletes samples in the majority class, and Random over-sampling randomly duplicates samples in the minority class. The TomekLinks technique deletes samples from the majority class that are in closest distance to samples in the minority class. SMOTE generates synthetic samples by averaging the features from k $(k = 5$ in our case) neighbouring samples of randomly selected samples in the minority class.

2.2.6 Feature selection techniques

Radiomics calculates hundreds of features from images, most of which are redundant or irrelevant to the particular outcome. FS techniques are used to identify the most unique, reproducible, and predictive subset of radiomics features. Unfortunately, there is rarely a single best-performing strategy for FS. The same FS technique can result in models performing differently across various studies [54].

In this thesis, we investigated the effects of numerous widely used FS techniques, including Variance Threshold (VT) [55], Principal Component Analysis (PCA) [56], Fast Independent Component Analysis (ICA) [57], Least Absolute Shrinkage and Selection Operator logistic regression algorithm (LASSO) [58, 59], Decision-Tree-Based (TREE) FS [60], and Recursive Feature Elimination with Cross-Validation (RFECV) [61]. Minimum redundancy maximum relevance (MRMRe) [62], mutual information [63], and Kendall rank correlation [64] are some of the less widely utilized FS techniques that we did not include in this thesis. The FS techniques we used are explained in the following subsections.

Variance threshold

VT is a simple baseline thresholding technique for FS. For a feature space $\mathbf{X}$ with n as the number of samples and p as the number of features (as defined by Eq. 2.3), the variance of feature j is defined as

$$S_j^2 = \frac{\Sigma_i^n (x_{ij} - \mu_j)^2}{n-1} \tag{2.8}$$

where μ_j is the mean value of feature j over all samples. The variance of a feature is zero when it has the same value (x_{ij}) across all samples. Therefore, a zero-variance threshold (VT0) filters out constant features. A near zero VT removes features with variance below a set threshold. For instance, a near zero VT of 0.2 removes features with variance smaller than 0.2.

Principal component analysis

PCA is one of the most commonly used dimensionality reduction approaches. PCA is used in FS to project the feature space to a new coordinate system based on the variance of feature values. As a result, the first coordinate (called the first principal component) is formed by the projection of the data with the greatest variation. The second component is made up of the second-largest variance. A schematic representation of a 2D feature space with its principal components is shown in Figure 2.6.

For the feature space $\mathbf{X}$ (Eq. 2.3), the PCA decomposition ($\mathbf{T}$) of $\mathbf{X}$ is $\mathbf{T} = \mathbf{XW}$, where

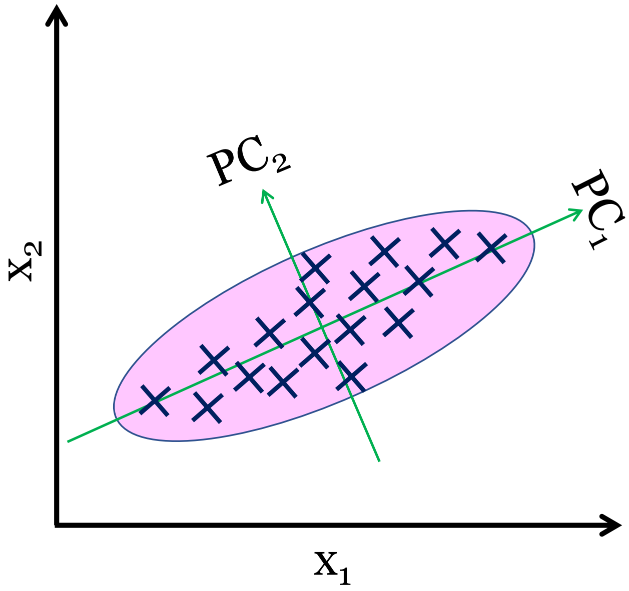


Figure 2.6: A schematic 2D feature space representation with its principal components. PCA is first introduced in [56].

W is the weight vector obtained from the maximum variance expressed as

$$\mathbf{w}_{(\mathbf{k})} = \underset{\|W\|=1}{argmax} \left\{ \|\hat{\mathbf{X}}_k \mathbf{w}\|^2 \right\} \tag{2.9}$$

in which $\hat{\mathbf{X}}_1 = \mathbf{X}$ and $\hat{\mathbf{X}}_k = \mathbf{X} - \sum_{s=1}^{k-1} \mathbf{X}\mathbf{w}_{(s)}\mathbf{w}_{(s)}^{\mathsf{T}}$.

$\mathbf{T} = \mathbf{X}\mathbf{W}$ transforms a feature space $\mathbf{X}$ from an initial space of p variables to a new space with p principal components where components are sorted based on their variance. This new space allows us to keep the first k principal components and reduce the dimension of the feature space. Hence, $\mathbf{T}_L = \mathbf{X}\mathbf{W}_L$ returns a reduced feature space with L features and n samples.

Independent component analysis

The goal of ICA is to find L statistically independent features (also referred to as components) that maximize non-Gaussianity. One of the most widely used ICA algorithms is Fast ICA. Fast ICA uses a nonquadratic nonlinear function $f(u)$, its first derivative $g(u)$, and its second derivative $g'(u)$ to measure non-Gaussianity as

$$f(u) = -e^{-u^2/2}$$

$$g(u) = ue^{-u^2/2}$$

$$g'(u) = (1 - u^2)e^{-u^2/2}$$

To implement Fast ICA, the first step is to center each feature j $(j = 1, 2, ..., p)$ for each sample i $(i = 1, 2, ..., n)$ in the feature space $(\mathbf{X})$, as

$$x_{ij} \leftarrow x_{ij} - \frac{1}{p}\Sigma_{j'}^{p} x_{ij'} \tag{2.10}$$

The centered matrix $\mathbf{X}$ is then given a linear transformation (called whitening) to make its components uncorrelated and with a variance of one. The eigenvalue decomposition of the covariance matrix can be used to perform the whitening of the centered data as

$$\mathbf{X} \leftarrow \mathbf{D}^{-1/2}\mathbf{E}^{\mathsf{T}}\mathbf{X} \tag{2.11}$$

where $\mathbf{E}$ is the eigenvector matrix and $\mathbf{D}$ is the diagonal matrix of eigenvalues. Finally, L linearly independent components are obtained by an iterative procedure as

```
for $j$ in 1 to $L$:
    $w_j$ = Random vector of length $n$
    while $w_j$ changes:
        $w_j = 1/p * \mathbf{X} * g(w_j * \mathbf{X})$
```

The output is a reduced matrix, with n samples and L independent components.

Least absolute shrinkage and selection operator logistic regression

LASSO is a regression analysis approach based on variable selection and regularization. The regularization is accomplished by adding a “penalty” term to the best linear fit produced from the trained data to achieve a lower variance. The penalty is the absolute value of the coefficient of the best fit line (β_j) to the training data. The goal of LASSO regression is to find β_j values to minimize the following expression:

$$\Sigma_{i=1}^{n}(y_i - \Sigma_j x_{ij}\beta_j)^2 + \alpha\Sigma_{i=1}^{p}|\beta_j| \tag{2.12}$$

where the first term is the sum of squared residuals over the number of samples ($i = 1, 2, ..., n$) and the second term is the scaled sum of the absolute value of the magnitude of coefficients over the number of features ($j = 1, 2, ..., p$). α is called the least-squares penalty and defines the number of features to be eliminated. The bigger the α, the more the variance is reduced and, therefore, the more features are eliminated. $\alpha = 0$ reduces the expression to rigid regression, which generates the orthogonal projection of the features with no feature being eliminated.

Decision-tree-based feature selection

Impurity-based feature importance can be computed using a tree-based classifier, which can then be used to filter out irrelevant features. Impurity is defined as the number of times a feature is used in a node of a classifier, weighted by the number of samples split by that note. Features having a high impurity level are considered to be more important [65].

2.3 Machine learning classifiers

Depending on the availability of desired outputs available, ML algorithms are typically classified into two major categories: unsupervised learning and supervised learning. Unsupervised learning is employed when there are no desired outputs, therefore, the goal of these algorithms is to discover patterns in the input data and produce outputs (referred to as labels). On the other hand, supervised learning is used when there is a set of desired outputs. In this thesis we will be using supervised ML algorithms since we have the desired labels. The goal of a supervised ML model is to find a function, $f(\mathbf{X})$, that maps the feature space, $\mathbf{X}$, (Eq. 2.3) to the label vector, y, (Eq. 2.6) as closely as possible. Depending on the algorithm, there are different ways to define the function f, and there are also different ways to map it to y. For example, a linear regression model finds the best fit line ($f = a\mathbf{X} + b$) by minimizing the squared residual:

$$E(f) = (y - f(\mathbf{X}))^2 \tag{2.13}$$

Many factors influence the success of a particular classifier on a given task, including sample size, the quality of acquired data, and the dimensionality of the feature space. An algorithm that performs well on one task may not perform well on another. Therefore, there is no one-size-fits-all learning classifier that can solve all supervised learning problems. As a result, it is common in each study to evaluate multiple classifiers and discover the best performing ones. In this thesis, we investigated the performance of several widely-used ML classifiers, including Support Vector Machine (SVM) [66], Gaussian Naive Bayes (NB) [67], k-Nearest Neighbors (kNN) [68], Quadratic Discriminant Analysis (QDA) [69], Gaussian Process Regression (GPR) [70], Decision Tree (DT) [71], Random Forest (RF) [72], Bagging [72], AdaBoost [72], and Neural Networks (NNet) [73]. These ML classifiers are explained in the following subsections.

2.3.1 Support Vector Machine

SVM is one of the most robust, effective, memory-efficient, and widely used ML algorithms. The SVM's purpose is to find a $p-1$ dimensional hyperplane (where p is the number of features) that maximizes the separation between samples with different labels. The separation can be done by linear classifiers (L-SVM) or by non-linear kernels (such as the radial basis function kernel). Schematics of decision boundaries with linear and non-linear kernels are presented in Figures 2.7 and 2.8.

A linear hyperplane can be defined as $wx - b = 0$, where x is the feature space, w is the slope (normal vector to the hyperplane), and b is intercept of the hyperplane. The algorithm computes the vectors between the hyperplane and each data point. These vectors are called the support vectors. The sum of the lengths of the support vectors is called the margin. The goal is to find w and b such that the margin is maximized. The maximum-margin hyperplane can be achieved by minimizing the cost equation:

$$\lambda \parallel w \parallel^2 + \left[\frac{1}{n}\Sigma_{i=1}^{n} max(0, 1 - y_i(w^T x_i - b))\right] \tag{2.14}$$

The parameter λ defines the trade-off between increasing the margin size and ensuring that

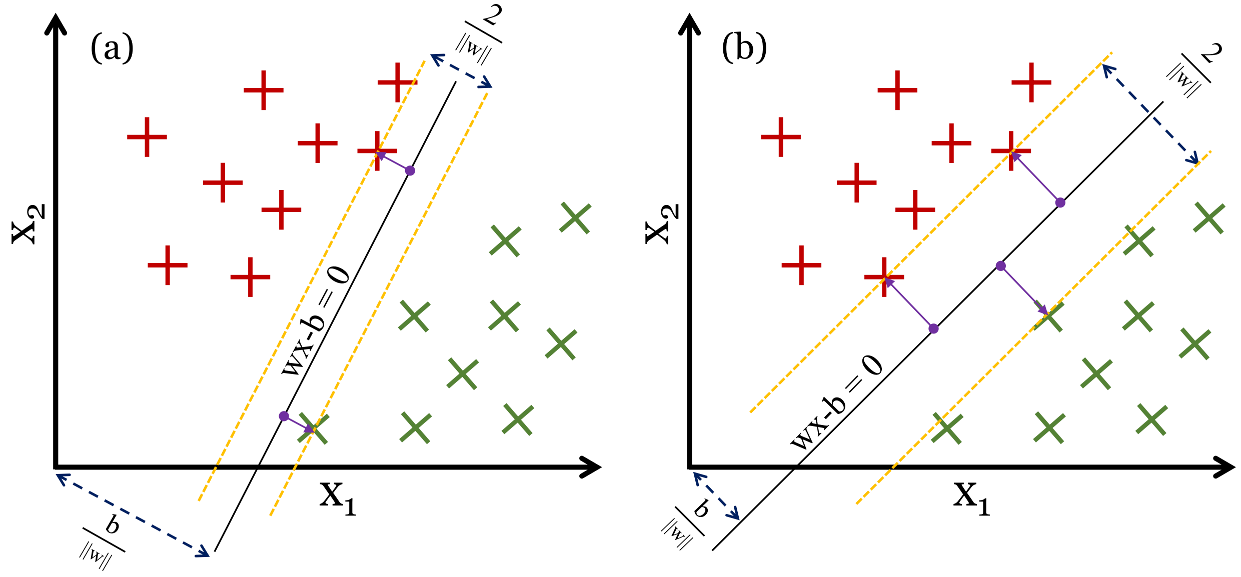


Figure 2.7: A schematic 2D feature space with (a) an example on a linear hyperplane, and (b) a hyperplane that maximizes the margin between two classes. The support vectors are plotted as violet vectors and red and green are used to present two classes.

the samples fall on the correct side of the margin. Figure 2.7 shows a schematic of a hyperplane with small (a) and maximal (b) margins.

Kernel functions can be used to create non-linear decision boundaries for features that are not linearly separable. Non-linear feature spaces are mapped to linear spaces using kernels. Figure 2.8 illustrates an example of a non-linear feature space that can be transformed to a linear space with the help of a radial kernel.

The Radial Basis Function (RBF) kernel is one of the most widely used kernels for feature clustering in SVM. The RBF kernel is defined as

$$K(x, x') = exp(-\gamma \parallel x - x' \parallel^2), \tag{2.15}$$

where $x - x'$ is the distance between all pairs of features, and $\gamma > 0$ is a parameter defining the importance of the neighboring pairs when calculating the support vector for each point in the feature space. A very large γ causes the decision boundary to be formed around every sample with no influence from the neighboring points (this is called over-fitting). On the

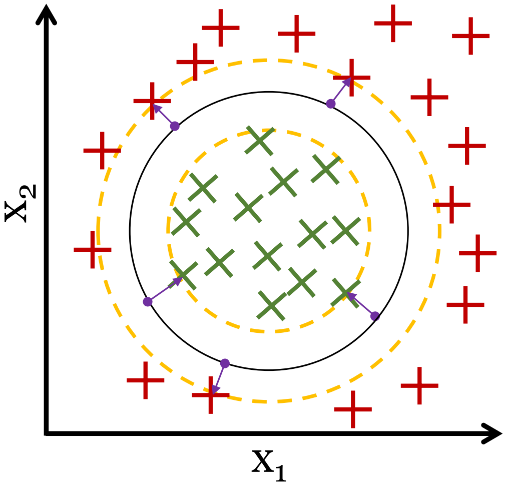


Figure 2.8: A schematic 2D feature space with an example on a non-linear (in this case, radial) hyperplane. The support vectors are plotted in violet.

other hand, a very small γ reduces the decision boundary to a linear hyperplane ($K \rightarrow 1$). The optimum value of gamma is normally found iteratively.

2.3.2 Gaussian Naive Bayes

NB is a simple and scalable ML technique based on Bayes' theorem [74] and the premise that features are equally important and are highly independent of each other. Bayes' theorem is a mathematical formula for determining the likelihood of one event occurring after another has occurred. This is known as a posterior probability, and it is written as follows:

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}, \tag{2.16}$$

where the conditional probability $P(B|A)$ is the probability of event B occurring given that A is true. $P(A)$ and $P(B)$ are the probabilities of events A and B, respectively.

In the continuous feature space, the conditional probability of independent variables can be modeled by a Gaussian distribution:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{Z_y^2}{2}\right), \tag{2.17}$$

in which μ and σ are the mean and standard deviation of the points within each class y, and Z_y (called the z-score) is the normalized distance between each data point x_i and the mean of each class:

$$Z_y = \frac{x_i - \mu_y}{\sigma_y} \tag{2.18}$$

Figure 2.9 depicts an example of two distributions and their conditional probabilities for a given point x_i. A NB model fits the best Gaussian distribution to each of the classes during the training phase. Then, a new point x_i is assigned to a given class based on the highest conditional probability.

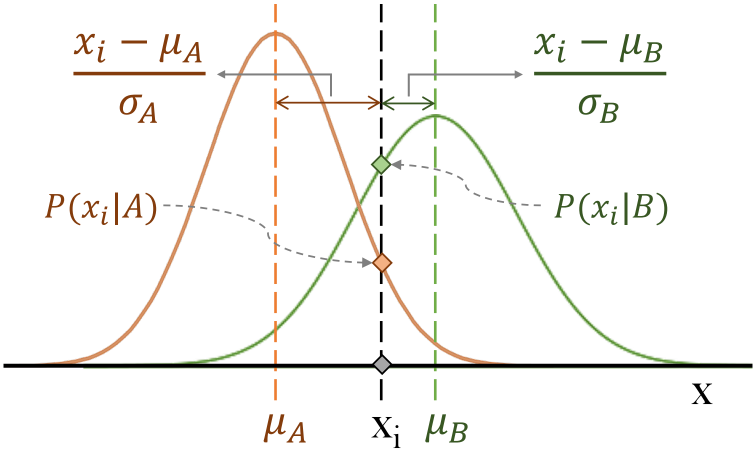

$\frac{x_i - \mu_A}{\sigma_A}$
$\frac{x_i - \mu_B}{\sigma_B}$
$P(x_i|A)$
$P(x_i|B)$
X
μ_A
$\mathrm{X_i}$
μ_B

Figure 2.9: An example of two Gaussian distributions and their conditional probabilities for a given point x_i. $P(x_i|A)$ and $P(x_i|B)$ are the likelihoods of obtaining x_i if x_i belongs to the class A and B, respectively. The z-scores for each distribution are presented in the graph.

2.3.3 Gaussian Process Regression

GPR is an ML process that is also based on posterior probability in the form of a Gaussian distribution (Eq. 2.17). Therefore, probability for each class y can be expressed as:

$$P(y = k|x) = \frac{P(x|y = k).P(y = k)}{P(x)} = \frac{P(x|y = k).P(y = k)}{\sum_l P(x|y = l).P(y = l)}. \tag{2.19}$$

However, unlike NB, GPR assumes that features are correlated with each other. The correlation is modeled by a multivariate normal distribution:

$$P(x|y = k, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} exp\left(-\frac{1}{2}(x - \mu_k)^t \Sigma_k^{-1}(x - \mu_k)\right), \tag{2.20}$$

where Σ is a covariance kernel defining the correlation between features. One of the most commonly used kernels in GPR is RBF (See equation 2.15).

$$\Sigma = cov(x_i, x_j) = exp\left(-\frac{(x_i - x_j)^2}{2}\right) \tag{2.21}$$

GPR assumes samples of the same class are closer to each other in feature space. The drawback of the GPR is that it loses efficiency in high dimensional spaces as shown in Chapter 4.

2.3.4 Quadratic Discriminant Analysis

Similar to the NB and GPR, QDA is also driven by the posterior probability (Eq. 2.19) for each class y as:

$$P(y = k|x) = \frac{P(x|y = k).P(y = k)}{P(x)} = \frac{P(x|y = k).P(y = k)}{\sum_l P(x|y = l).P(y = l)}. \tag{2.22}$$

QDA also assumes that features are correlated with each other and the correlation can be modeled by a multivariate normal distribution as Eq. 2.20, in which the covariance kernel (Σ) defined as:

$$\Sigma_k = \frac{1}{n-1} X_k^T X_k \tag{2.23}$$

The decision boundary is defined by the logarithm of the posterior probability (log-posterior for short):

$$\begin{aligned} \log P(y = k|x) &= \log P(x|y = k + \log P(y = k)) + Constant \\ &= -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x - \mu_k)^t \Sigma_k^{-1}(x - \mu_k) \\ &+ \log P(y = k) + Constant \end{aligned} \tag{2.24}$$

An unlabeled sample is assigned to a class that maximizes the log-posterior. QDA algorithms use singular value decomposition to calculate log-posterior values. To avoid the explicit calculation of Σ^{-1} the solver directly computes the coefficients which makes it faster than GPR.

2.3.5 k-nearest neighbor

kNN is another well-known supervised classifier. Similar to GPR, kNN is also based on the assumption that members of the same class are closer to each other in feature space. For a given unlabeled sample, x_0, the kNN algorithm first finds k (where k is a user-defined parameter) samples from the training set that are closest to x_0 based on their Euclidean distance. Then, within these k samples, it assigns x_0 to the class with the largest population. An example of a kNN classifier is illustrated in Figure 2.10, where the unlabeled gray point is classified as either green or red, depending on the majority of its neighbors. It is assigned to the green class if we consider the five closest neighbors ($k = 5$; dotted circle), and to the red class if we consider ten closest neighbors ($k = 10$; dashed circle).

The drawback of simple kNN is the "majority voting" when the number of samples in both classes is not equal. To avoid this problem, the generalized kNN uses the inverse distance between points as well as $1/k$ as a weighing factor when calculating the conditional probability for class j.

$$P(Y = j|X = x_0) = \frac{1}{k} \sum_{i \in N} (w_i \times I(y_{i=j})) \,, \tag{2.25}$$

where $I = 0$ when $y_i \neq j$ and $w_i = 1/(x_0 - x_i)$.

2.3.6 Decision Trees

A DT is a classification algorithm that works based on the yes-or-no answers that can be extracted from training data. A DT, illustrated in Figure 2.11, consists of nodes (questions that are driven by data), branches (yes-or-no answers to the questions), and leaves (the class labels that are assigned to data points based on the answers to corresponding questions). For instance, a DT node might be $Age > 65$? condition. Passing via this node, samples with $Age > 65$ are routed to the left (*yes*) branch, while all other samples are routed to the right

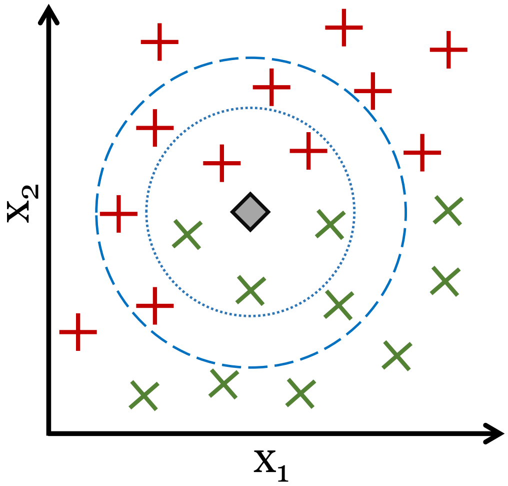


Figure 2.10: Representation of the kNN classification algorithm. The new (unlabeled) point is assigned to the class with the largest number of samples among its k nearest neighbours. In this example, considering five nearest neighbours ($k = 5$) the unlabeled point (grey) must be allocated to the class green, which has the majority of samples (three out of five). Considering $k = 10$, the point must be allocated to class red, which contains six out of ten samples.

(*no*) branch.

A DT is constructed by finding the best feature and condition (the best question) in each tree's node to split the data set into branches. The splitting and optimization are done iteratively based on a series of rules that are extracted from the features. This process is called recursive partitioning. Recursive partitioning ends when all the samples in a node belong to the same class or further iteration does not result in better classification results. The algorithm starts from the root node and finds the best feature that can split the data into two groups. A variety of measures (such as true positive rate, variance, impurity, or information gain) can be used to evaluate the quality of the split and to determine the "best" feature. One basic measure is the True Positive Rate (TPR),

$$TPR = \frac{TP}{TP + FN},$$

where TP and FP are the total true positives and false positives in each subset, respectively.

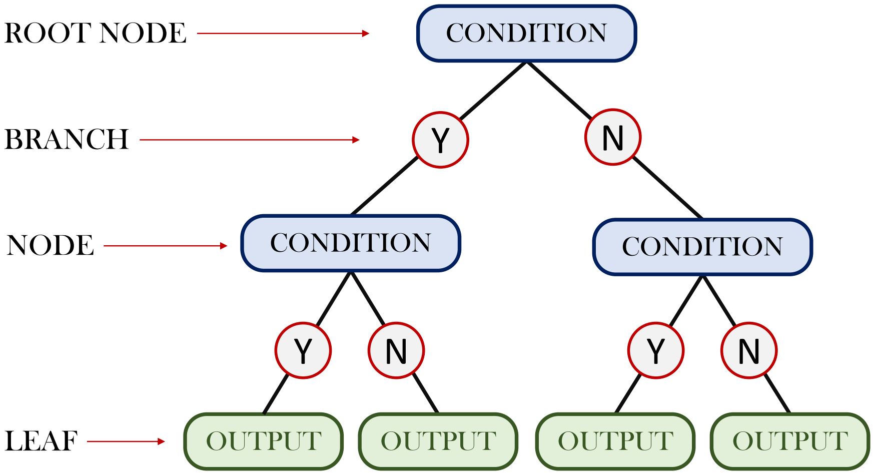


Figure 2.11: A decision tree classifier diagram. Each node (blue box) specifies a condition based on a certain attribute. Each branch is a response to the condition. Leaf nodes (green boxes) represent class labels.

DT divides the data into potential subsets based on each feature on each node and calculates the TPR for each subset. Then, based on the TPR values, the model ranks the features and selects the feature with the highest TPR measure for that specific node. The performance of the DT classifier is highly dependent on the evaluation measure. As a result, the majority of recent DT classifiers use more robust measures such as "information gain". Information gain is defined as a change in the uncertainty of a random variable before and after classification. The uncertainty of a random variable before classification is defined by the entropy equation:

$$H(X) = -\sum_i P(x_i) log(P(x_i)) \tag{2.26}$$

The uncertainty of variable X after learning (assigning value to Y) is reduced. This is measured by conditional entropy:

$$H(X|Y) = -\sum_i P(x_i) \sum_i P(x_i|y_i) log(P(x_i|y_i)) \tag{2.27}$$

Information gain, then, is defined as

$$IG(X|Y) = H(X|Y) - H(X) \tag{2.28}$$

In each node, the DT classifier chooses a feature that maximizes the information gain.

2.3.7 Random Forest

In radiomics studies, RF is one of the most frequently utilized classification methods. It belongs to the category of ensemble learning methods. Ensemble methods employ multiple learning algorithms to improve the predictive performance of the classification. RF is made up of an ensemble of trained decision trees. Each tree classifies data independently, and labels are assigned by average across all of the decision trees. Since RF employs an ensemble of unbiased DT classifiers, it produces accurate predictions across a wide range of data.

2.3.8 Bagging

Bootstrap aggregation, also known as bagging, is an ensemble learning method that is often used to reduce variance when working with a noisy dataset. Bagging is the process of generating bootstrapped datasets from a training set. A bootstrapped set is generated by randomly selecting data from the training set and duplicating part of the points. Thus, it has the same number of samples as the training set but can contain zero, one, or more of each sample. Once bootstrapped sets are generated, each set is used to train an instance of an ensemble model (such as a DT). Finally, the classification probabilities of individual assessments are aggregated to calculate the overall label of a sample.

2.3.9 Adaptive Boosting

AdaBoost, short for “Adaptive Boosting”, is an ensemble learning method that uses the weighted outputs of multiple classifiers to represent the final output of the “boosted” classifier. AdaBoost is called “adaptive” since it modifies the outputs of classifiers in favour of examples that were incorrectly categorised by earlier classifiers.

The AdaBoost error function is defined (compare to Eq. 2.13) as

$$E(f) = e^{-y(\mathbf{X})f(\mathbf{X})} \tag{2.29}$$

In this definition, for a given sample x_i, as $-y(x_i)f(x_i)$ grows larger, the error grows at an exponential rate. Thus, this results in outliers being given higher importance.

2.3.10 Neural Networks

NNets are among the most complex ML techniques that attempt to replicate the human brain by learning from training samples to complete a given task. A basic architecture of a NNet is presented in Figure 2.12. A typical NNet is made up of thousands of processing nodes that are closely coupled and stacked in layers. It is possible for a single node to have several connections to other nodes on both the preceding layer below it (from which it receives data) and the succeeding layer above it (to which it transmits data). The first layer (called the input layer) is made up of p nodes, each of which corresponds to a single feature x_j. The last layer, which is called the output layer, usually has the same number of nodes as there are labels in the data set. All of the layers in between are called hidden layers. The number of hidden layers depends on the complexity of the problem, the size of the training dataset, and the number of classes therein.

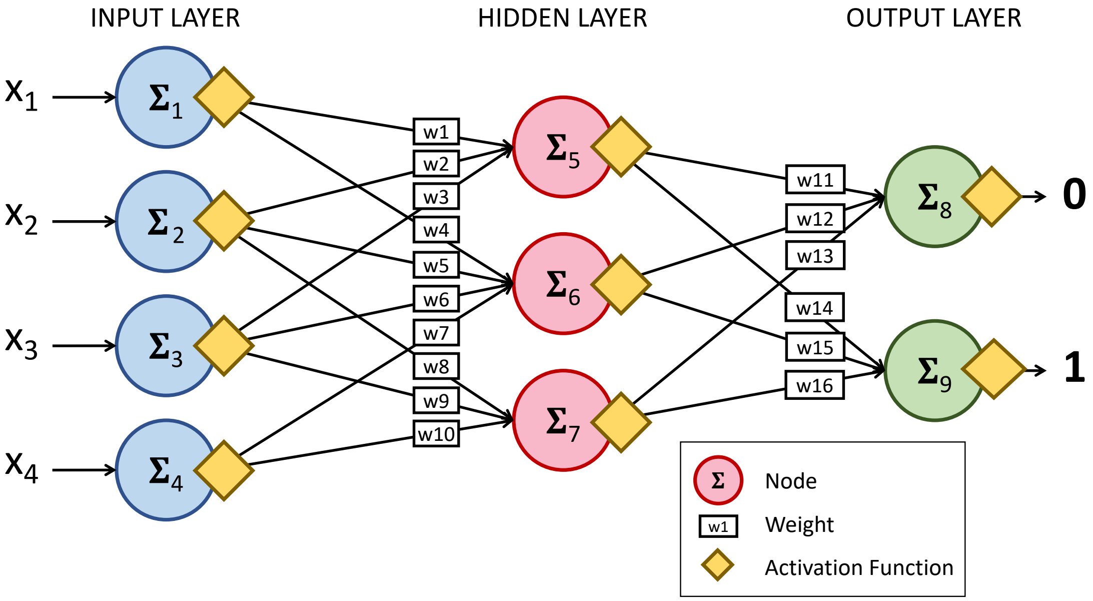


Figure 2.12: Representation of a simple neural network.

In a NNet, each node gives each incoming connection a numerical value, called a "weight". During normal operation of the network, each connection brings the node a unique piece of information in the form of a number, which is then multiplied by the connection's weight. A single number is then obtained by summing all of the individual weights. For example, in

Figure 2.12, the value on node 5, can be calculated as

$$\Sigma_5 = w_1 x_1 + w_2 x_2 + w_3 x_3$$

If the resulting number, Σ, on each node is less than a certain threshold, the node doesn't send any value to the next layer. If the value is greater than the threshold, the node sends the value (the sum of the weighted inputs) to the next layer via all of its outgoing connections. The threshold is defined by an activation function.

There are two widely use activation functions for NNet classifiers, Sigmoid and ReLU (Rectified Linear Unit) [75]. The Sigmoid function is defined as

$$\phi(\Sigma) = \frac{1}{1 + e^{-\Sigma}} \ , \tag{2.30}$$

and ReLU is defined as

$$\phi(\Sigma) = \max(0, \Sigma). \tag{2.31}$$

ReLU has been shown to be faster and have better gradient propagation compared to sigmoidal activation functions [76].

2.4 Performance evaluation

After training an ML model, the final step is to test its performance against a set of ground truth data. This section defines and discusses statistical tools that we used throughout this thesis to evaluate performance of various segments of our pipelines.

2.4.1 k-fold Cross-Validation

It is typical practice in ML tasks to divide the initial data set into training and test data sets. The training set is used to train the model, whereas the test set is utilised to provide an objective evaluation. In situations where the initial data set is small, splitting it into training and test sets may result in skewed data sets. Moreover, training and testing a model on a single small data set can significantly diminish its stability and accuracy. Cross-validation allows us to overcome this issue by training and evaluating the model on multiple subsets of an initial data set.

K-fold cross-validation divides the initial sample into a k sub samples of equal sizes. Then, k-1 of the sub samples are used to train the model, while keeping the remaining sub sample as validation (test) data. Then, the procedure of cross-validation is repeated k times until each of the k sub sets is utilised once as the validation set. This process is illustrated in Figure 2.13. The performance of the model is reported based on the average of all iterations.

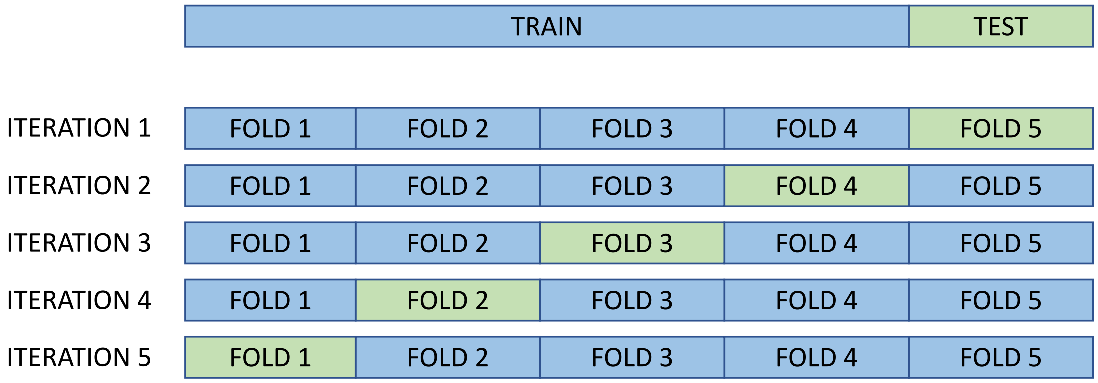


Figure 2.13: A schematic representation of a five-fold (k-fold) cross validation.

2.4.2 Confusion Matrix

A confusion matrix is a type of table used to visualise the performance of a supervised ML classifier. Each row represents the number of samples in a given actual class (ground truth), whereas each column represents the number of the samples assigned to each class by the classifier (predicted class). For example, assume a binary classifier that distinguishes between painful and painless BM (as we will see in chapter 5). We can label painful samples as class 1 (positive) and painless samples as class 0 (negative).

In this example, there are four possible outcomes when comparing a sample's predicted class to its actual class. One, true positive (TP), is the case when sample's actual class is positive (i.e., painful) and the model correctly identified it as belonging to the positive class. Two, false negative (FN), occurs when the sample actually belongs to the positive (painful) class, but the model mistakenly assigned it to the negative (painful) class. Third, false positive (FP), occurs when the sample's actual class is negative (painless), but the model mistakenly classified it as positive (painful). Fourth, true negative (TN), occurs when a sample's actual class is negative (painless) and it is correctly classified as negative by the model. Therefore, the confusion matrix of this classifier is a two-by-two table as shown in Table 2.2. An example of a confusion matrix for a multi-class classifier is presented in Table 3.4.

		Predicted Label	
		POSITIVE	NEGATIVE
True Label	POSITIVE	TP	FN
	NEGATIVE	FP	TN

Table 2.2: The confusion matrix for a binary (two-label) classifier contains 2 correctly-predicted labels and 2 incorrectly-predicted labels. true positive (TP), and true negative (TN), are numbers of sentences that are correctly predicted and false negative (FN) and false positive (FP), are numbers of mislabeled sentences.

Elements of the confusion matrix are utilised to generate four fundamental statistical measures (accuracy, sensitivity, specificity, and F1 score) for evaluating the performance of an ML classifier [77].

Accuracy is the overall measure of the performance. It measures how well the predicted values are close to their true values. It is defined as the total number of correctly classified samples (both TP and TN) divided by the total number of samples as shown in Eq. 2.32.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{2.32}$$

Precision (also known as the positive predictive rate) is the rate of correct prediction within a given predicted class. It is defined as the number of correctly classified samples in each class divided by the total number of predicted samples in that class (both correctly and incorrectly classified) as shown in Eq. 2.33.

$$Precision = \frac{TP}{TP + FP} \tag{2.33}$$

Recall, also known as the sensitivity or TPR, is the number of the correctly detected terms in each class divided by the total number of the true cases in that class (both correctly classified by algorithm and missed). Sensitivity is defined as,

$$Sensitivity = TPR = \frac{TP}{TP + FN} \tag{2.34}$$

Finally, F1-Score is calculated for each class using Eq. 2.35 as the weighted average of precision and recall.

$$F1{-}Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{2.35}$$

2.4.3 Receiver operating characteristic curve

A Receiver Operating Characteristic (ROC) curve is a plot that represents how well the model can separate two classes. It is a plot of sensitivity (TPR) (Eq. 2.34) against False positive Rate (FPR). FPR is defined as

$$FPR = \frac{FP}{FP + TN} \tag{2.36}$$

For example, assume we have a classifier that produced a decision boundary as Figure 2.14 that best separates positive (green) and negative (red) samples. Even with a best classifier, it is common for some samples to fall on the wrong side of the decision boundary. In this example two positive samples and two negative samples are classified incorrectly which results in $FPR = 0.2$ and $TPR = 0.8$. Shifting the decision boundary to the left leads to the correct classification of more positive samples (increase in TPR) at the expense of the incorrect classification of more negative samples (increase in FPR). In contrast, moving the decision border to the right reduces both TPR and FPR.

The ROC curve illustrates how moving the decision boundary affects FPR and TPR.

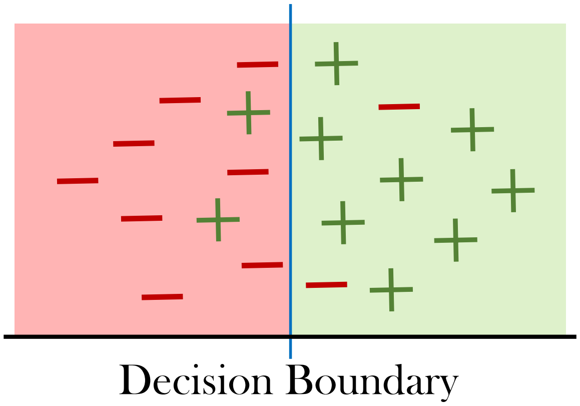


Figure 2.14: An example of a simple decision boundary that best divides samples into positive (green) and negative (red) classes. Even with a best decision boundary, it is common for some samples to be incorrectly classified.

Figure 1 shows the ROC curve for the classifier shown in Figure 2.15. The Area Under the Receiver Operating Characteristic Curve (ROC-AUC) is a metric that measures a model's ability to correctly distinguish between two classes. ROC-AUC = 0.5 (blue dashed line in Figure 2.15) indicates a completely random model, whereas ROC-AUC = 1 indicates a model with perfect prediction ability.

ROC curves and ROC-AUC values are used to evaluate the performance of the classifiers that we developed throughout this thesis. Also, we report accuracy, precision, Sensitivity, and F-1 score of all of our models.

2.4.4 Precision-recall curve

A precision-recall curve is another frequently used metric in applied machine learning for evaluating binary classification models. The precision-recall curve illustrates the tradeoff between precision and recall at various thresholds. This is especially beneficial when there is an imbalance between the two classes' observations.

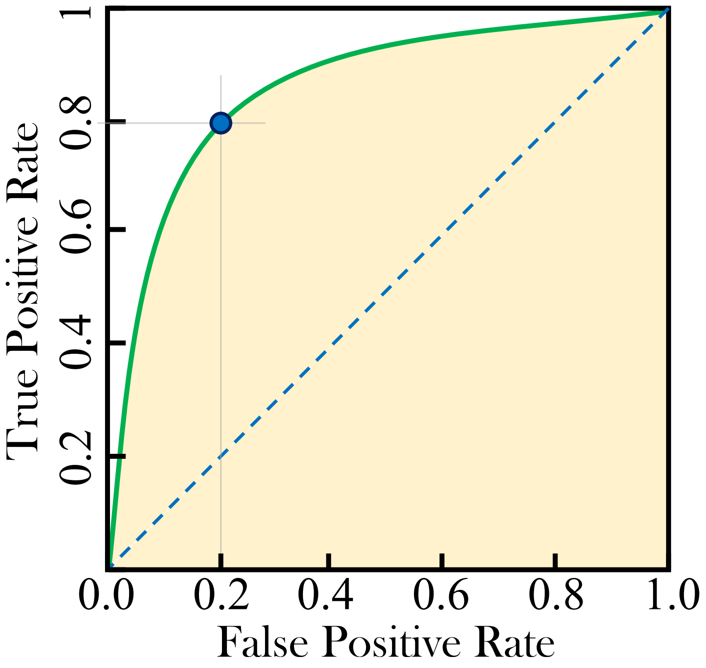


Figure 2.15: A Receiver Operating Characteristic (ROC) curve (green line) is constructed by comparing True positive Rate (TPR) against False positive Rate (FPR) at various decision thresholds for the example binary classifier shown in Figure 2.14. The optimal decision boundary yielded FPR=0.2 and TPR=0.8 (blue dot).

2.4.5 Reliability measurement

Reliability is defined as the capability to replicate measurements [78]. Several well-known mathematical and statistical measures of reliability exist. For instance, the intraclass correlation coefficient (ICC) is a statistical method commonly used to assess interrater reliability and measurement reproducibility [79, 80]. ICC has been used to measure feature reliability in radiomics investigations [81]. Fleiss' kappa is another statistical measure used to determine how well a set number of raters agree on the ratings or classifications of a set number of items [82].

In this thesis, Fleiss' kappa is used to assess the interobserver agreement in the extraction of gold standard labels from clinical notes.

Bibliography

[1] T. B. Murdoch, A. S. Detsky, The inevitable application of big data to health care, JAMA 309 (13) (2013) 1351–1352. doi:10.1001/JAMA.2013.393.
URL https://pubmed.ncbi.nlm.nih.gov/23549579/

[2] T. Elizabeth Workman, J. M. Stoddart, Rethinking information delivery: using a natural language processing application for point-of-care data discovery, Journal of the Medical Library Association : JMLA 100 (2) (2012) 113. doi:10.3163/1536-5050.100.2.009.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3324802/

[3] W. W. Yim, M. Yetisgen, W. P. Harris, W. K. Sharon, Natural Language Processing in Oncology: A Review, JAMA oncology 2 (6) (2016) 797–804. doi:10.1001/JAMAONCOL.2016.0213.
URL https://pubmed.ncbi.nlm.nih.gov/27124593/

[4] T. Zhang, A. M. Schoene, S. Ji, S. Ananiadou, Natural language processing applied to mental illness detection: a narrative review, npj Digital Medicine 2022 5:1 5 (1) (2022) 1–13. doi:10.1038/s41746-022-00589-7.
URL https://www.nature.com/articles/s41746-022-00589-7

[5] S. F. Sung, C. H. Chen, R. C. Pan, Y. H. Hu, J. S. Jeng, Natural Language Processing Enhances Prediction of Functional Outcome After Acute Ischemic Stroke, Journal of the American Heart Association 10 (24) (2021) 23486. doi:10.1161/JAHA.121.023486.
URL https://www.ahajournals.org/doi/abs/10.1161/JAHA.121.023486

[6] S. Velupillai, H. Suominen, M. Liakata, A. Roberts, A. D. Shah, K. Morley, D. Osborn, J. Hayes, R. Stewart, J. Downs, W. Chapman, R. Dutta, Using clinical Natural

Language Processing for health outcomes research: Overview and actionable suggestions for future advances, Journal of Biomedical Informatics 88 (2018) 11–19. doi:10.1016/J.JBI.2018.10.005.

[7] H. J. Dai, C. H. Su, Y. Q. Lee, Y. C. Zhang, C. K. Wang, C. J. Kuo, C. S. Wu, Deep Learning-Based Natural Language Processing for Screening Psychiatric Patients, Frontiers in Psychiatry 11 (2021) 1557. doi:10.3389/FPSYT.2020.533949/BIBTEX.

[8] T. A. Koleck, C. Dreisbach, P. E. Bourne, S. Bakken, Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review, J Am Med Inform Assoc. 26 (4) (2019) 364–379. doi:10.1093/jamia/ocy173. URL www.covidence.org

[9] W. K. Tan, S. Hassanpour, P. J. Heagerty, S. D. Rundell, P. Suri, H. T. Huhdanpaa, K. James, D. S. Carrell, C. P. Langlotz, N. L. Organ, E. N. Meier, K. J. Sherman, D. F. Kallmes, P. H. Luetmer, B. Griffith, D. R. Nerenz, J. G. Jarvik, Comparison of Natural Language Processing Rules-based and Machine-learning Systems to Identify Lumbar Spine Imaging Findings Related to Low Back Pain, Academic Radiology 25 (11) (2018) 1422–1432. doi:10.1016/j.acra.2018.03.008.

[10] R. M. Cronin, D. Fabbri, J. C. Denny, S. T. Rosenbloom, G. P. Jackson, A comparison of rule-based and machine learning approaches for classifying patient portal messages, International Journal of Medical Informatics 105 (2017) 110–120. doi:10.1016/J.IJMEDINF.2017.06.004.

[11] F. Mathew, H. Wang, L. Montgomery, J. Kildea, Natural language processing and machine learning to assist radiation oncology incident learning, Journal of Applied Clinical Medical Physics 22 (11) (2021) 172–184. doi:10.1002/ACM2.13437. URL https://onlinelibrary.wiley.com/doi/full/10.1002/acm2.13437https://onlinelibrary.wiley.com/doi/abs/10.1002/acm2.13437https://aapm.onlinelibrary.wiley.com/doi/10.1002/acm2.13437

[12] A. R. Aronson, T. C. Rindflesch, A. C. Browne, Exploiting a Large Thesaurus for Information Retrieval, RIAO; Intelligent Multimedia Information Retrieval Systems and Management 1 (1994) 197–216. doi:10.5555/2856823.2856842.

[13] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology., Nucleic acids research 32 (Database issue) (2004) 267–70. doi:10.1093/nar/gkh061.
URL http://www.ncbi.nlm.nih.gov/pubmed/14681409

[14] F. B. Putra, A. Arman Yusuf, H. Yulianus, Y. P. Pratama, D. Salma Humairra, U. Erifani, D. K. Basuki, S. Sukaridhoto, R. P. Nourma Budiarti, Identification of Symptoms Based on Natural Language Processing (NLP) for Disease Diagnosis Based on International Classification of Diseases and Related Health Problems (ICD-11), IES 2019 - International Electronics Symposium: The Role of Techno-Intelligence in Creating an Open Energy System Towards Energy Democracy, Proceedings (2019) 1–5doi:10.1109/ELECSYM.2019.8901644.

[15] R. A. Côté, S. Robboy, Progress in medical information management. Systematized nomenclature of medicine (SNOMED), JAMA 243 (8) (1980) 756–762. doi:10.1001/JAMA.1980.03300340032015.
URL https://pubmed.ncbi.nlm.nih.gov/6986000/

[16] F. B. ROGERS, Medical subject headings, Bull Med Libr Assoc. 51 (1) (1963) 114–120.
URL https://pubmed.ncbi.nlm.nih.gov/13982385/

[17] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, Gene Ontology: tool for the unification of biology, Nature genetics 25 (1) (2000) 25. doi:10.1038/75556.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3037419/

[18] A. R. Aronson, F. M. Lang, An overview of MetaMap: historical perspective and recent advances, Journal of the American Medical Informatics Association : JAMIA 17 (3) (2010) 229. doi:10.1136/JAMIA.2009.002733.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2995713/

[19] A. R. Aronson, a. nih gov, Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program., Proceedings of the AMIA Symposium (2001)

17.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243666/

[20] J. Mork, A. Aronson, D. Demner-Fushman, 12 years on - Is the NLM medical text indexer still useful and relevant?, Journal of Biomedical Semantics 8 (1) (2017) 1–10. doi:10.1186/S13326-017-0113-5/FIGURES/6.
URL https://jbiomedsem.biomedcentral.com/articles/10.1186/s13326-017-0113-5

[21] A. T. McCray, A. R. Aronson, A. C. Browne, T. C. Rindflesch, A. Razi, S. Srinivasan, UMLS knowledge for biomedical language processing., Bulletin of the Medical Library Association 81 (2) (1993) 184.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC225761/

[22] A. R. Aronson, MetaMap: Mapping Text to the UMLS Metathesaurus (2006).

[23] A. R. Aronson, MetaMap Evaluation (2001).

[24] S. M. Humphrey, W. J. Rogers, H. Kilicoglu, D. Demner-Fushman, T. C. Rindflesch, Word Sense Disambiguation by Selecting the Best Semantic Type Based on Journal Descriptor Indexing: Preliminary Experiment, Journal of the American Society for Information Science and Technology (Print) 57 (1) (2006) 96. doi:10.1002/ASI.20257.
URL /pmc/articles/PMC2771948//pmc/articles/PMC2771948/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC2771948/

[25] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. Van Stiphout, P. Granton, C. M. Zegers, R. Gillies, R. Boellard, A. Dekker, H. J. Aerts, Radiomics: extracting more information from medical images using advanced feature analysis, European journal of cancer (Oxford, England : 1990) 48 (4) (2012) 441–446. doi:10.1016/J.EJCA.2011.11.036.
URL https://pubmed.ncbi.nlm.nih.gov/22257792/

[26] P. Lambin, R. T. Leijenaar, T. M. Deist, J. Peerlings, E. E. De Jong, J. Van Timmeren, S. Sanduleanu, R. T. Larue, A. J. Even, A. Jochems, Y. Van Wijk, H. Woodruff, J. Van Soest, T. Lustberg, E. Roelofs, W. Van Elmpt, A. Dekker, F. M. Mottaghy, J. E.

Wildberger, S. Walsh, Radiomics: the bridge between medical imaging and personalized medicine, Nature Reviews Clinical Oncology 2017 14:12 14 (12) (2017) 749–762. doi:10.1038/nrclinonc.2017.141.
URL https://www.nature.com/articles/nrclinonc.2017.141

[27] P. Lambin, J. Zindler, B. G. Vanneste, L. V. De Voorde, D. Eekers, I. Compter, K. M. Panth, J. Peerlings, R. T. Larue, T. M. Deist, A. Jochems, T. Lustberg, J. van Soest, E. E. de Jong, A. J. Even, B. Reymen, N. Rekers, M. van Gisbergen, E. Roelofs, S. Carvalho, R. T. Leijenaar, C. M. Zegers, M. Jacobs, J. van Timmeren, P. Brouwers, J. A. Lal, L. Dubois, A. Yaromina, E. J. Van Limbergen, M. Berbee, W. van Elmpt, C. Oberije, B. Ramaekers, A. Dekker, L. J. Boersma, F. Hoebers, K. M. Smits, A. J. Berlanga, S. Walsh, Decision support systems for personalized and participative radiation oncology, Advanced drug delivery reviews 109 (2017) 131–153. doi:10.1016/J.ADDR.2016.01.006.
URL https://pubmed.ncbi.nlm.nih.gov/26774327/

[28] B. Gao, D. Dong, H. Zhang, Z. Liu, S. Payabvash, B. T. Chen, Editorial: Radiomics Advances Precision Medicine, Frontiers in Oncology 12 (2022) 489. doi:10.3389/FONC.2022.853948/BIBTEX.

[29] Z. Liu, S. Wang, D. Dong, J. Wei, C. Fang, X. Zhou, K. Sun, L. Li, B. Li, M. Wang, J. Tian, The Applications of Radiomics in Precision Diagnosis and Treatment of Oncology: Opportunities and Challenges, Theranostics 9 (5) (2019) 1303. doi:10.7150/THNO.30309.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6401507/

[30] R. J. Gillies, M. B. Schabath, Radiomics improves cancer screening and early detection, Cancer Epidemiology Biomarkers and Prevention 29 (12) (2020) 2556–2567. doi:10.1158/1055-9965.EPI-20-0075/70438/AM/RADIOMICS-IMPROVES-CANCER-SCREENING-AND-EARLY.
URL https://aacrjournals.org/cebp/article/29/12/2556/71998/Radiomics-Improves-Cancer-Screening-and-Early

[31] K. Bera, N. Braman, A. Gupta, V. Velcheti, A. Madabhushi, Predicting cancer outcomes with radiomics and artificial intelligence in radiology, Nature reviews. Clinical oncology

19 (2) (2022) 132. doi:10.1038/S41571-021-00560-7.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9034765/

[32] A. M. Luke, K. Shetty, S. Satish, K. Kilaru, Comparison of Spiral Computed Tomography and Cone-Beam Computed Tomography (2014).

[33] M. Lev, R. Gonzalez, CT Angiography and CT Perfusion Imaging, Brain Mapping: The Methods (2002) 427–484doi:10.1016/B978-012693019-1/50019-8.

[34] S. K. Vinod, M. Min, M. G. Jameson, L. C. Holloway, A review of interventions to reduce inter-observer variability in volume delineation in radiation oncology, Journal of medical imaging and radiation oncology 60 (3) (2016) 393–406. doi:10.1111/1754-9485.12462.
URL https://pubmed.ncbi.nlm.nih.gov/27170216/

[35] H. M. Patrick, L. Souhami, J. Kildea, Reduction of inter-observer contouring variability in daily clinical practice through a retrospective, evidence-based intervention, https://doi.org/10.1080/0284186X.2020.1825801 60 (2) (2020) 229–236. doi:10.1080/0284186X.2020.1825801.
URL https://www.tandfonline.com/doi/abs/10.1080/0284186X.2020.1825801

[36] R. R. Savjani, M. Lauria, S. Bose, J. Deng, Y. Yuan, V. Andrearczyk, Automated Tumor Segmentation in Radiotherapy, Seminars in Radiation Oncology 32 (4) (2022) 319–329. doi:10.1016/J.SEMRADONC.2022.06.002.

[37] M. Lin, S. Momin, Y. Lei, H. Wang, W. J. Curran, T. Liu, X. Yang, Fully automated segmentation of brain tumor from multiparametric MRI using 3D context deep supervised U-Net, Medical physics 48 (8) (2021) 4365–4374. doi:10.1002/MP.15032.
URL https://pubmed.ncbi.nlm.nih.gov/34101845/

[38] C. G. B. Yogananda, B. R. Shah, M. Vejdani-Jahromi, S. S. Nalawade, G. K. Murugesan, F. F. Yu, M. C. Pinho, B. C. Wagner, K. E. Emblem, A. Bjørnerud, B. Fei, A. J. Madhuranthakam, J. A. Maldjian, A Fully Automated Deep Learning Network for Brain Tumor Segmentation, Tomography 6 (2) (2020) 186. doi:10.18383/J.TOM.2019.00026.

URL /pmc/articles/PMC7289260//pmc/articles/PMC7289260/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7289260/

[39] N. Sharma, A. K. Ray, K. K. Shukla, S. Sharma, S. Pradhan, A. Srivastva, L. Aggarwal, Automated medical image segmentation techniques, Journal of Medical Physics / Association of Medical Physicists of India 35 (1) (2010) 3. doi:10.4103/0971-6203.58777.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2825001/

[40] D. L. Pham, C. Xu, J. L. Prince, Current Methods in Medical Image Segmentation1, Annual Review of Biomedical Engineering 2 (2000) (2000) 315–337. doi:10.1146/ANNUREV.BIOENG.2.1.315.
URL https://www.annualreviews.org/doi/abs/10.1146/annurev.bioeng.2.1.315

[41] J. E. Park, S. Y. Park, H. J. Kim, H. S. Kim, Reproducibility and Generalizability in Radiomics Modeling: Possible Strategies in Radiologic and Statistical Perspectives, Korean Journal of Radiology 20 (7) (2019) 1124. doi:10.3348/KJR.2018.0070.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6609433/

[42] Q. Qiu, J. Duan, Z. Duan, X. Meng, C. Ma, J. Zhu, J. Lu, T. Liu, Y. Yin, Reproducibility and non-redundancy of radiomic features extracted from arterial phase CT scans in hepatocellular carcinoma patients: Impact of tumor segmentation variability, Quantitative Imaging in Medicine and Surgery 9 (3) (2019) 453–464. doi:10.21037/qims.2019.03.02.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6462568/

[43] I. Fornacon-Wood, H. Mistry, C. J. Ackermann, F. Blackhall, A. McPartlin, C. Faivre-Finn, G. J. Price, J. P. O'Connor, Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform, European radiology 30 (11) (2020) 6241–6250. doi:10.1007/S00330-020-06957-9.
URL https://pubmed.ncbi.nlm.nih.gov/32483644/

[44] A. Zwanenburg, S. Leger, M. Vallières, S. Löck, Image biomarker standardisation initiative, Radiology 295 (2) (2016) 328–338. doi:10.1148/radiol.2020191145.
URL http://dx.doi.org/10.1148/radiol.2020191145

[45] J. J. Foy, K. R. Robinson, H. Li, M. L. Giger, H. Al-Hallaq, I. Samuel G. Armato, Variation in algorithm implementation across radiomics software, Journal of Medical Imaging 5 (4) (2018) 1. doi:10.1117/1.JMI.5.4.044505.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6278846/

[46] A. Bettinelli, M. Branchini, F. De Monte, A. Scaggion, M. Paiusco, Technical Note: An IBEX adaption toward image biomarker standardization, Medical physics 47 (3) (2020) 1167–1173. doi:10.1002/MP.13956.
URL https://pubmed.ncbi.nlm.nih.gov/31830303/

[47] L. Zhang, D. V. Fried, X. J. Fave, L. A. Hunter, J. Yang, L. E. Court, IBEX: an open infrastructure software platform to facilitate collaborative work in radiomics, Medical physics 42 (3) (3 2015). doi:10.1118/1.4908210.
URL https://pubmed.ncbi.nlm.nih.gov/25735289/

[48] J. J. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J. C. Fillion-Robin, S. Pieper, H. J. Aerts, Computational radiomics system to decode the radiographic phenotype, Cancer Research 77 (21) (2017) e104–e107. doi:10.1158/0008-5472.CAN-17-0339/SUPPLEMENTARY-VIDEO-S2.
URL https://aacrjournals.org/cancerres/article/77/21/e104/662617/Computational-Radiomics-System-to-Decode-the

[49] E. Pfaehler, A. Zwanenburg, J. R. de Jong, R. Boellaard, RaCaT: An open source and easy to use radiomics calculator tool, PLoS ONE 14 (2) (2 2019). doi:10.1371/JOURNAL.PONE.0212223.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6382170/

[50] S. Ashrafinia, Quantitative Nuclear Medicine Imaging Using Advanced Image Reconstruction and Radiomics, ProQuest Dissertations and Theses (March) (2019) 294.
URL https://jscholarship.library.jhu.edu/handle/1774.2/61551

[51] H. Bagher-Ebadian, I. J. Chetty, Technical Note: ROdiomiX: A validated software for radiomics analysis of medical images in radiation oncology, Medical physics 48 (1) (2021) 354–365. doi:10.1002/MP.14590.
URL https://pubmed.ncbi.nlm.nih.gov/33169367/

[52] I. Tomek, TWO MODIFICATIONS OF CNN., IEEE Transactions on Systems, Man and Cybernetics SMC-6 (11) (1976) 769–772. doi:10.1109/TSMC.1976.4309452.

[53] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, Journal Of Artificial Intelligence Research 16 (2002) 321–357. doi:10.1613/jair.953.
URL https://arxiv.org/abs/1106.1813v1

[54] A. Demircioğlu, Evaluation of the dependence of radiomic features on the machine learning model, Insights into Imaging 13 (1) (2022) 1–11. doi:10.1186/S13244-022-01170-2/FIGURES/5.
URL https://insightsimaging.springeropen.com/articles/10.1186/s13244-022-01170-2

[55] S. Das, U. Mert Cakmak, Hands-On Automated Machine Learning : a beginner's guide to building automated machine learning systems using AutoML and Python., 1st Edition, Packt Publishing, 2018.

[56] K. Pearson, LIII. On lines and planes of closest fit to systems of points in space, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2 (11) (1901) 559–572. doi:10.1080/14786440109462720.
URL https://zenodo.org/record/1430636

[57] A. Hyvärinen, E. Oja, Independent component analysis: algorithms and applications, Neural Networks 13 (4-5) (2000) 411–430. doi:10.1016/S0893-6080(00)00026-5.

[58] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, D. Gorinevsky, An Interior-Point Method for Large-ScalèScalè 1-Regularized Least Squares, IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING 1 (4) (2007). doi:10.1109/JSTSP.2007.910971.

[59] R. Tibshirani, Regression shrinkage and selection via the lasso: a retrospective, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73 (3) (2011) 273–282. doi:10.1111/J.1467-9868.2011.00771.X.
URL https://onlinelibrary.wiley.com/doi/full/10.1111/j.1467-9868.2011.00771.x

[60] L. Breiman, Random Forests, Machine Learning 2001 45:1 45 (1) (2001) 5–32. doi:10.1023/A:1010933404324.
URL https://link.springer.com/article/10.1023/A:1010933404324

[61] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene Selection for Cancer Classification using Support Vector Machines, Machine Learning 2002 46:1 46 (1) (2002) 389–422. doi:10.1023/A:1012487302797.
URL https://link.springer.com/article/10.1023/A:1012487302797

[62] M. Radovic, M. Ghalwash, N. Filipovic, Z. Obradovic, Minimum redundancy maximum relevance feature selection approach for temporal gene expression data, BMC Bioinformatics 18 (1) (2017) 1–14. doi:10.1186/S12859-016-1423-9/FIGURES/6.
URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1423-9

[63] A. Kraskov, H. Stögbauer, P. Grassberger, Estimating mutual information, Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics 69 (6) (2004) 16. doi:10.1103/PHYSREVE.69.066138/FIGURES/20/MEDIUM.
URL https://journals.aps.org/pre/abstract/10.1103/PhysRevE.69.066138

[64] M. G. KENDALL, A NEW MEASURE OF RANK CORRELATION, Biometrika 30 (1-2) (1938) 81–93. doi:10.1093/BIOMET/30.1-2.81.
URL https://academic.oup.com/biomet/article/30/1-2/81/176907

[65] E. Scornet, Trees, forests, and impurity-based variable importance (1 2020). doi:10.48550/arxiv.2001.04295.
URL https://arxiv.org/abs/2001.04295v3

[66] 1.4. Support Vector Machines — scikit-learn 1.0.1 documentation.
URL https://scikit-learn.org/stable/modules/svm.html

[67] 1.9. Naive Bayes — scikit-learn 1.0.1 documentation.
URL https://scikit-learn.org/stable/modules/naive_bayes.html

[68] 1.6. Nearest Neighbors — scikit-learn 1.0.1 documentation.
URL https://scikit-learn.org/stable/modules/neighbors.html

[69] 1.2. Linear and Quadratic Discriminant Analysis — scikit-learn 1.0.1 documentation.
URL https://scikit-learn.org/stable/modules/lda_qda.html

[70] 1.7. Gaussian Processes — scikit-learn 1.0.1 documentation.
URL https://scikit-learn.org/stable/modules/gaussian_process.html

[71] 1.10. Decision Trees — scikit-learn 1.0.1 documentation.
URL https://scikit-learn.org/stable/modules/tree.html

[72] 1.11. Ensemble methods — scikit-learn 1.0.1 documentation.
URL https://scikit-learn.org/stable/modules/ensemble.html

[73] 1.17. Neural network models (supervised) — scikit-learn 1.0.1 documentation.
URL https://scikit-learn.org/stable/modules/neural_networks_supervised.html

[74] A. Kolmogorov, Foundations of the theory of probability (English translation of the 1933 version) (1950).

[75] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning - whole book, Nature 521 (7553) (2016) 800.
URL http://goodfeli.github.io/dlbook/%0Ahttp://dx.doi.org/10.1038/nature14539

[76] X. Glorot, A. Bordes, Y. Bengio, Deep Sparse Rectifier Neural Networks (2011).

[77] A. Tharwat, Classification assessment methods, Applied Computing and Informatics 17 (1) (2018) 168–192. doi:10.1016/J.ACI.2018.08.003/FULL/PDF.

[78] L. E. Daly, G. J. G. J. Bourke, G. J. G. J. Bourke, Interpretation and uses of medical statistics (2000) 568.

[79] P. E. Shrout, J. L. Fleiss, Intraclass correlations: Uses in assessing rater reliability, Psychological Bulletin 86 (2) (1979) 420–428. doi:10.1037/0033-2909.86.2.420.

[80] J. J. Bartko, The intraclass correlation coefficient as a measure of reliability., Psychological reports 19 (1) (1966) 3–11. doi:10.2466/PR0.1966.19.1.3.

[81] C. Xue, J. Yuan, G. G. Lo, A. T. Chang, D. M. Poon, O. L. Wong, Y. Zhou, W. C. Chu, Radiomics feature reliability assessed by intraclass correlation coefficient: a systematic review, Quantitative Imaging in Medicine and Surgery 11 (10) (2021) 4431. doi:10.21037/QIMS-21-86.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8408801/

[82] J. L. Fleiss, Measuring nominal scale agreement among many raters, Psychological Bulletin 76 (5) (1971) 378–382. doi:10.1037/H0031619.

Chapter 3

Development of a generalizable natural language processing pipeline to extract physician-reported pain from clinical reports: generated using publicly-available datasets and tested on institutional clinical reports for cancer patients with bone metastases

Hossein Naseri, Kamran Kafi, Sonia Skamene, Marwan Tolba, Mame Daro Faye, Paul Ramia, Julia Khriguian, and John Kildea

Article published in: *J Biomed Inform.* 2021 Aug; 120:103864. doi: https://doi.org/10.1016/j.jbi.2021.103864

3.1 Preface

This chapter describes the study that was undertaken to meet objective 1 of this thesis: Construction of an NLP pipeline to extract pain scores from the consultation notes of patients. In this study, a generalizable NLP pipeline was built utilizing methods introduced in Chapter 2 to process unstructured text and extract sentence-level pain scores from patients' clinical notes. Then, these sentence-level pain score were averaged to define note-level physician-reported pain scores. Publicly-available clinical notes were used for training and the pipeline's generalizability was validated by analyzing retrospectively-collected radiation oncology consultation notes of cancer patients with BM at our centre and extracting physician-reported pain scores from them.

3.2 Abstract

Objective The majority of cancer patients suffer from severe pain at the advanced stage of their illness. In most cases, cancer pain is underestimated by clinical staff and is not properly managed until it reaches a critical stage. Therefore, detecting and addressing cancer pain early can potentially improve the quality of life of cancer patients.

The objective of this research project was to develop a generalizable Natural Language Processing (NLP) pipeline to find and classify physician-reported pain in the radiation oncology consultation notes of cancer patients with bone metastases.

Materials and Methods: The texts of 1,249 publicly-available hospital discharge notes in the i2b2 database were used as a training and validation set. The MetaMap and NegEx algorithms were implemented for medical terms extraction. Sets of NLP rules were developed to score pain terms in each note. By averaging pain scores, each note was assigned to one of the three Verbally-Declared Pain (VDP) labels, including no pain, pain, and no mention of pain. Without further training, the generalizability of our pipeline in scoring individual pain terms was tested independently using 30 hospital discharge notes from the MIMIC-III database and 30 consultation notes of cancer patients with bone metastasis from our institution's radiation oncology electronic health record. Finally, 150 notes from our institution were used to assess the pipeline's performance at assigning VDP.

Results: Our NLP pipeline successfully detected and quantified pain in the i2b2 summary notes with 93% overall precision and 92% overall recall. Testing on the MIMIC-III database achieved precision and recall of 91% and 86% respectively. The pipeline successfully detected pain with 89% precision and 82% recall on our institutional radiation oncology corpus. Finally, our pipeline assigned a VDP to each note in our institutional corpus with 84% and 82% precision and recall, respectively.

Conclusion: Our NLP pipeline enables the detection and classification of physician-reported pain in our radiation oncology corpus.This portable and ready-to-use pipeline can be used to automatically extract and classify physician-reported pain from clinical notes where the pain is not otherwise documented through structured data entry.

3.3 Introduction

Two-thirds of cancer patients with advanced metastatic disease experience pain [1], and nearly 50% of these patients identify pain as a significant problem that deteriorates their quality of life [2, 3]. Pain can also induce stress that may suppress the immune system. For instance, it has been demonstrated that pain in metastatic patients can suppress the natural killer cells that control tumor growth and metastasis [4]. Because of these issues, several organizations such as the World Health Organization (WHO) and the American Pain Society recommend that physicians properly document pain in Electronic Health Records (EHRs) to facilitate best practice pain management, follow up, and quality assurance [1, 5–7].

Consultation notes in EHRs represent a wealth of useful information on patients' health and outcomes. But, due to their largely unstructured nature and typically non-standardized formatting, extracting useful information from these unstructured free-text documents efficiently, is a challenging task [8]. This may result in consultation notes being ignored or not optimally used in clinical cancer management and outcomes research.

One potential approach to meet this challenge is to adopt NLP pipeline to parse consultation notes. This approach is the subject of our presently-reported study with a focus on pain mentions.

3.3.1 NLP for pain assessment

Natural Language Processing (NLP) is a branch of Computer Science that utilizes statistical functions and computational algorithms to analyze unstructured free text and extract quantitative information from it [9]. Algorithms can be trained to process large corpora of clinical narratives and extract relevant biomedical information from them. To extract biomedical concepts from clinical texts, one approach is to use pre-trained NLP models such as bidirectional encoder representations from transformers (BERT) [10]. Another approach is to combine the NLP technique with structured databases of clinical terminologies. Such structured databases are designed to categorize and classify medical terms and clinical information into standardized tables with a unique code for each medical concept.

There are several well-known databases of clinical terminologies in-use worldwide. The International Classification of Diseases (ICD) [11] is one of them, maintained by the WHO. ICD-11 is the latest available update of the ICD database. The SNOMED CT is the next one that has encoded over 340,000 multilingual clinical healthcare terminologies [12]. This database is maintained by the SNOMED International association. The Unified Medical Language System (UMLS) [13] is another database maintained by the US National Library of Medicine (NLM). The UMLS provides standard codes for thousands of biomedical concepts and it includes both the ICD and SNOMED CT vocabularies [14]. The NLM also provides the MetaMap NLP tool [15, 16] to extract biomedical concepts from clinical notes and map them to the UMLS database. MetaMap, which is widely utilized in medical NLP applications [17, 18], has built-in libraries for sentence segmentation, concept tokenization and abbreviation/acronym identification [19]. MetaMap uses the NegEx [20] negation detection algorithm to determine whether mentions of medical terms in the corpus were negated. NegEx has a superior performance in negation detection compared to other algorithms [21].

NLP techniques have been used for medical keyword searches, classification of diagnoses, and extraction of cancer phenotype and symptom-related information from clinical notes [22–28]. In some studies, NLP has been used to extract mentions of chest pain and back pain [29, 30]. NLP has also been deployed to identify and classify chronic pain [31, 32], and to extract cancer-related pain scores [33]. Eisman et al. [34] successfully implemented the pre-trained

BERT model to extract angina symptoms from patients' clinical notes. Bui and Zeng [35] developed an NLP algorithm using regular expression analyzes to extract pain terminologies from clinical texts. Then, the authors classified each note into "pain" and "no pain" groups using supervised ML method. However, their algorithm was limited to explicit indications of "pain" and did not achieve accuracy higher than 79% in identifying and assigning pain scores. Heintzelman et al. [33] developed a more robust rules-based NLP technique to process clinical notes and detect all pain terms and their severity scores in each note in their cancer dataset with an accuracy of 96%. Then, for each note they considered the pain term with the maximum severity as the "pain index" and used it to evaluate the correlation between the cancer pain severity and survival rate in metastatic prostate cancer patients. However, upon testing on a publicly-available hospital discharge summary corpus, the accuracy of their NLP algorithm dropped to 64%. Also, the authors of Ref. [33] found that their algorithm needed to be trained on the new pain description patterns that they found in the publicly-available corpora. The authors argued that this lack of generalizability was attributed to more complex hypothetical wordings and past tense descriptions in publicly-available corpora compared to cancer data sets. It has been shown that more generalizable text classification models can be achieved by exploiting word embedding techniques [36, 37]. In study by Tao et al. [38], integration of the GloVe word embedding resulted in a significant performance improvement in the generalizability of extracting prescription information (medication names, dosages and frequencies) from clinical notes. Testing on the i2b2 dataset, authors showed that F-1 score of their algorithm increased from 0.78 to 0.83 when they integrated GloVe word embedding.

The objective of our study was to develop a generalizable (i.e. dataset independent) NLP pipeline to retrospectively process patients' medical notes and identify all pain terms and their severity scores in each note and assign a single Verbally-Declared Pain (VDP) to each note, representing the overall pain of the patient at the time of the consultation. For each note, our VDP was obtained by averaging over the pain scores detected in the note. For generalizability, unlike Heintzelman et al. [33], we first trained our pipeline on a publicly-available dataset, and afterward applied our trained pipeline on another publicly-available dataset and on our institutional radiation oncology dataset. Moreover, motivated by the findings of Tao et al. [38], and in order to provide a more generalizable solution, we used distributed word vectorization methods and word similarity features (GloVe word embedding). Also, unlike Heintzelman et al. [33], that used pain term with the highest pain-score as their pain index, we averaged the pain scores to assign a VDP to each note. We showed that these methods enabled building a database-independent pipeline to identify pain description patterns, exclude irrelevant mentions of pain, and calculate the physician-reported VDP at the time of the hospital visit. This is important as it now allows pain to be reliably extracted from radiation oncology consultation notes in a way that can facilitate further pain-related studies.

The pain-related terms used in this paper are defined in Table 3.1.

Term	Definition
Pain terms	The pain-related medical terms that were collected in Table 3.2. Each note might contain multiple pain terms.
Pain concepts	The UMLS medical concepts that which were obtained by mapping the pain terms to the UMLS metathesaurus (Table 3.19). Multiple pain concepts might be mapped to one pain term.
Pain score	A pain term in a phrase that explicitly indicates an experience (score 1) or denial (score 0) of pain at the time of the hospital visit. Pain terms that were not related to the time of the visit were assigned as irrelevant pain. (See Fig. 3.4) Each note might contain multiple pain scores.
VDP	A three-point Verbally-Declared Pain (VDP) (no mention of pain, pain, no pain) that was assigned to each note by averaging valid pain scores. (See Section 3.4.3)

Table 3.1: Definitions of the terms used in this paper

3.4 Materials and Methods

3.4.1 Corpora

In this study we used three independent corpora to develop and test our NLP pipeline: (i) 1,249 discharge summaries from the Informatics for Integrating Biology & the Bedside (i2b2) #1A Smoking challenge database [39, 40], (ii) 30 discharge summaries from the Medical Information Mart for Intensive Care III (MIMIC-III) database [41], and (iii) 788 consultation notes from the EHRs of 462 metastatic cancer patients previously treated at our institution. Consultation notes for metastatic cancer patients from our institution were extracted from the ARIA database for Radiation Oncology (Varian Medical Systems, Palo

Alto, CA). All patients in our institutional corpus received palliative radiotherapy for a secondary malignant neoplasm of bone at our cancer centre between January 2016 and September 2019. The textual data from our institutional corpus were extracted from Microsoft Word (.doc) documents using the Python textract package [42].

Detailed descriptions of the three corpora are presented in Appendix 3.9.1, and details of the number of characters and words in each corpus are presented in Table 3.15.

All three corpora had similar mean numbers of characters per clinical note (between 7,000 to 9,000, which is equivalent to two or three pages of single-spaced text).

As presented in Figure 3.2, of the 1,249 i2b2 notes, 1,099 randomly-selected notes were used for concept extraction and training the NLP pipeline, 120 notes (4 sets of 30 notes) were used for validation, and the remaining 30 randomly-selected notes were reserved for testing. In each iteration of training, we did a performance evaluation on one set (30 notes) from the validation corpus. The test corpus was used for final performance evaluation once the pipeline was completely developed. Later, 30 notes from the MIMIC-III and 30 notes from ARIA corpora were used for testing of the generalizability of the fully-developed NLP pipeline. It should be noted that the MIMIC-III and ARIA corpora were not used in any of the iterations of the training and validation. Another set of 150 notes from ARIA corpora were used for testing the performance of our NLP pipeline in assigning a VDP label to each note. Cochran's [43] sample size formula was used to determine the confidence interval of the selected sample sizes, as presented in Section 3.9.4, in the Supplementary Information.

3.4.2 Preparation of the validation and test corpora

The notes from the validation corpus were annotated by developers and were used to evaluate the performance of our NLP pipeline in four iterations of the training. The final performance of the pipeline to detect and score pain was evaluated against an expert-annotated (gold-standard) test corpus from each dataset (Figure 3.2).

We extracted all the sentences from each set in the validation and test corpus. The sentences from validation set 1, set 2 and set 3 (2,310, 2,332, 2,075 sentences, respectively) were manually annotated by the primary developer. Validation set 4 (1,012 sentences) was manually annotated by an independent developer. The sentences from the i2b2 (2,361 sentences) and MIMIC-III (2,717 sentences) test corpora were manually annotated by an

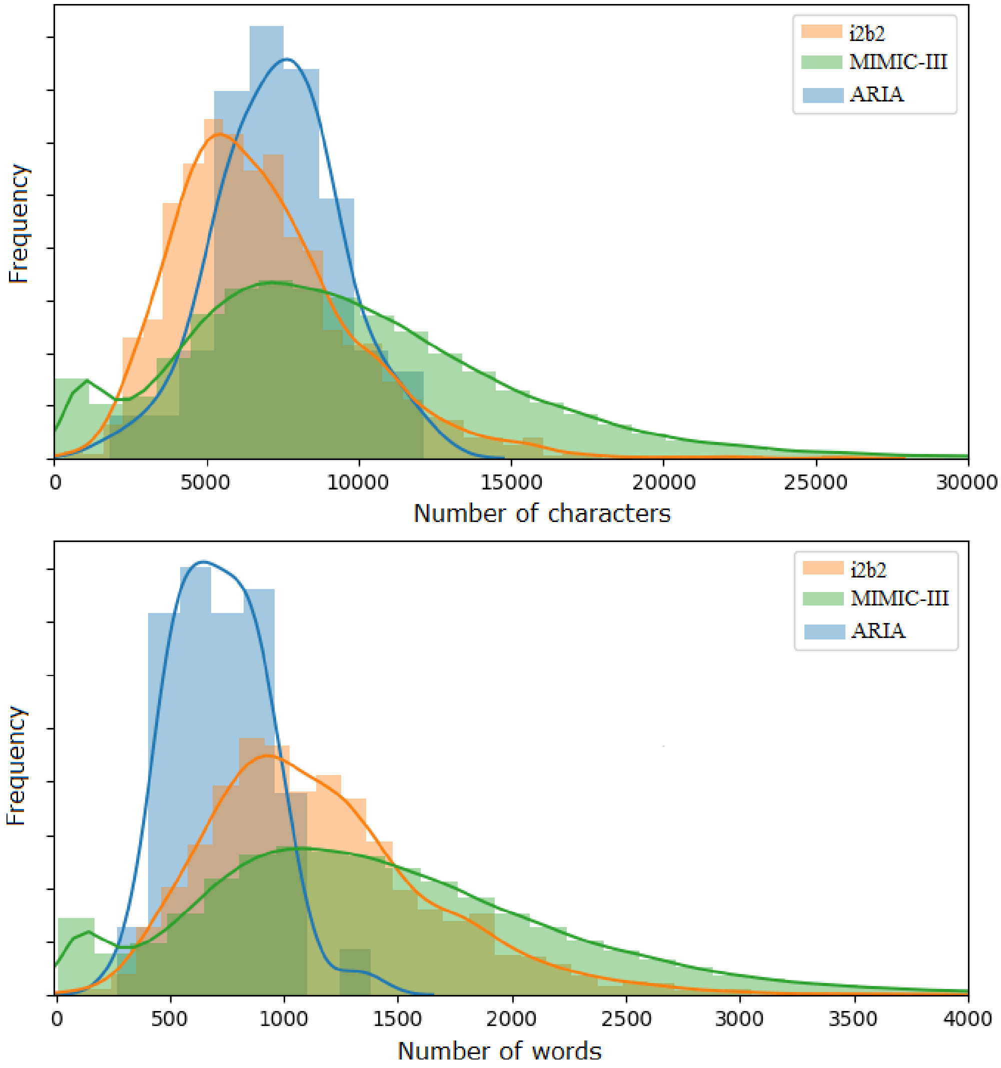


Figure 3.1: Normalized distributions of the number of characters (top panel) and number of the words (bottom panel) in the i2b2 (shown by orange color), MIMIC-III (green), and ARIA (blue) corpora.

MD physician. The sentences from the ARIA test corpus (1,132 sentences) were manually annotated by a radiation oncologist at our institution. The selected sample size resulted in 95% confidence level with less than 1% margin of error (the sample size calculation is presented in Section 3.9.4 in the Supplementary Information).

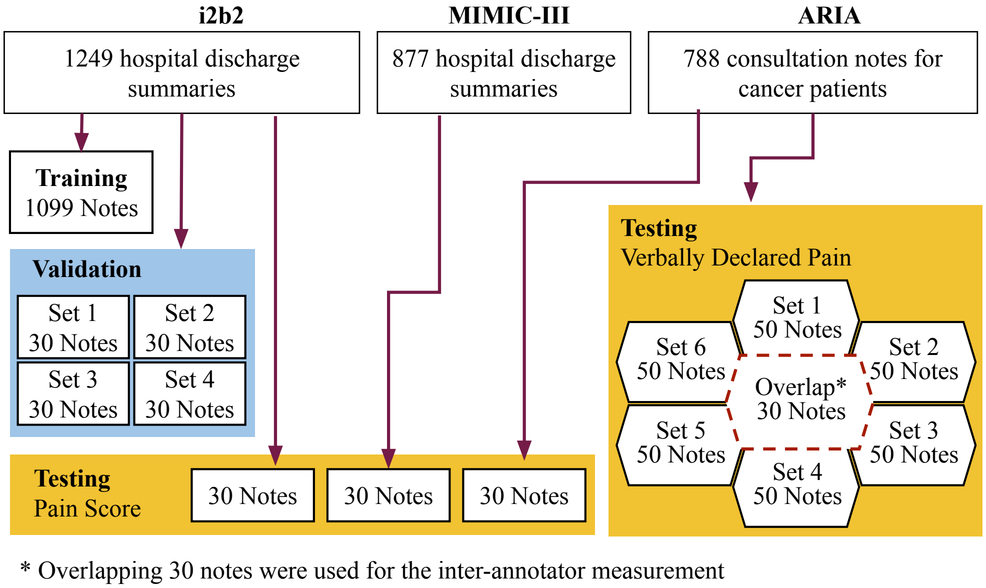


Figure 3.2: The corpora used in this paper. From 1,249 hospital discharge summaries form i2b2 corpora, 1,099 notes were used for concept extraction and training of our NLP pipeline, 120 annotated notes were used for validation of our NLP pipeline in four iterations and 30 notes were reserved for testing of the fully-developed pipeline. The MIMIC-III and ARIA corpora were used only for testing of our NLP pipeline (these two corpora were never used for training). 150 notes from the ARIA corpora were used for testing the Verbally-Declared Pain (VDP) classification method.

Following Heintzelman et al.'s [33] example, sentences from the test sets were annotated by our NLP pipeline. The domain experts (MD physician, radiation oncologist) were then asked to compare their manually-annotated sentences against the NLP annotation results to produce the gold-standard test sets. The rational for this step was to ensure that the experts did not accidentally miss or mislabel any pain term.

To evaluate the accuracy of our VDP classification method, another independent set of 150 notes from the ARIA corpus was annotated by six annotators (one oncologist, one medical physicist, and four oncology residents). Each annotator was asked to annotate a set of 50 notes consisting of 20 unique notes and 30 notes that were shared among all six annotators. These 30 notes were used to report inter-annotator agreement using Fleiss' kappa statistical measure [44]. Each annotator was asked to review each note and assign it to one of the five-grade verbal rating scales; no mention of pain (when pain was not reported in the note or pain was not reflecting the current state of the illness), no pain (if the pain was explicitly denied), mild (pain score 1-3), moderate (pain score 4-6) and severe (pain score 7-10). However, since we found that pain scores were not consistently documented in the radiation oncology consultation notes, which led to poor kappa measures for inter-annotator agreement, we instead defined a three-grade VDP incorporating 'no mention of pain', 'no pain', and 'pain' (by grouping mild, moderate and severe pain scales as 'pain'). The 150 VDP-annotated notes provided a gold-standard for evaluation of the accuracy of our VDP classification method. The selected sample size resulted in 0.026 standard error within a 95% confidence interval. The detailed sample size calculation can be found in Section 3.9.4 of the Supplementary Information.

Because the aim of this project was classifying cancer pain in radiation oncology clinical notes, the accuracy of our VDP classification method was only tested on the ARIA corpus. Given the effort required, we did not ask the radiation oncologists to spend their time annotating i2b2 and MIMIC-III corpora.

3.4.3 Pain detection pipeline

Our pain detection pipeline consisted of three parts: (1) an NLP pipeline to extract all UMLS medical concepts from the text documents, (2) a rules-based classifier to identify pain terms and extract valid pain scores, and (3) a method to calculate an average pain intensity and assign a physician-reported VDP to each note. The terms used in this paper are defined in Table 3.1.

Step 1: UMLS medical concept extraction

A flowchart describing our medical concept extraction pipeline, is provided in Figure 3.3.

The NLP algorithm was constructed in Python 3.7 using the spaCy toolkit [45]. The

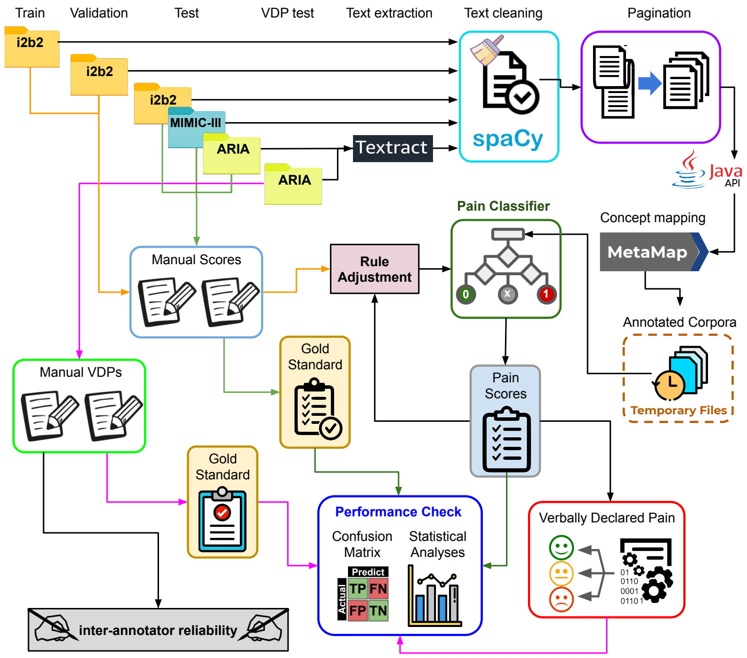


Figure 3.3: Our pipeline for medical concept extraction using MetaMap and NegEx. Text from each clinical note was exported as a text document. The Python spaCy package was used for the NLP of patients' consultation notes for text cleaning. The cleaned medical notes were divided into discrete pages (pagination) and passed to the MetaMap and NegEx algorithms via a Java API [16] for the medical name entity tagging and negation detection, respectively. Then, the processed corpora were passed to the pain classifier (Fig. 3.4) to extract the pain scores. Selection rules were adjusted by evaluating extracted pain scores against the manually annotated pain scores. Finally, the extracted pain scores were stored in the database for VDP calculation, statistical analyses, and performance evaluation.

MetaMap-14 [15, 16] engine was installed on our Ubuntu server and accessed from our custom-written Python code using its Java API. We have made our NLP pipeline and the annotation tool publicly available on GitHub [46].

As shown in Figure 3.3, clinical notes were read by our custom-written Python scripts [47] for pre-processing. Pre-processing was performed using the Python spaCy package to remove white spaces, special characters, and to convert all characters to lowercase. We also used a custom-built lookup table to map pain-related medical acronyms (including “cp”: chest pain,“lbp”: lower back pain, and “akp” : anterior knee pain). Our pipeline did not handle spelling errors. However, in our training and validation we did not see any mislabeling due to spelling errors. After pre-processing, larger documents were divided into discrete pages with a maximum character limit of 8,000 to fit the character limit of MetaMap’s batch processing software. Truncated notes were passed page-by-page to MetaMap via MetaMap’s Java API. MetaMap compiled each file as a ‘freetext’ and segmented it into ‘sentences’. Then, each sentence was processed phrase-by-phrase and was mapped to all possible UMLS concepts. Metamap also, assigned a confidence score for each concept indicating how much each UMLS concept was relevant to the phrase [46]. The NegEx [20] algorithm inside MetaMap was used for negation detection to determine whether mentions of pain terms in the corpus were negated.

Each phrase, together with its assigned clinical concepts, their negation statuses, confidence scores, and ICD codes were stored in a temporary text file. Then, these temporary files were read and the clinical concepts from all phrases of a note were concatenated into a single text file. A sample annotated text is presented in Table 3.14 in the Supplementary Information. Finally, the program read the processed temporary files phrase by phrase and identified all medical concepts with the ‘signs and symptoms’ UMLS tag. If multiple medical concepts mapped to a phrase, the program selected the concept with the highest confidence score. The program also extracted medical concepts with a UMLS ‘pharmacologic substance’ semantic tag in order to identify drug-related phrases. These tags were used to remove drug-related phrases such as “take Tylenol for your back pain”. All identified clinical concepts were organized into a data table together with ICD concept IDs, UMLS confidence scores, and negation indices. These data tables were passed to the pain classifier for pain analysis.

Step 2: Pain classification

Our rules-based classifier for detecting pain scores is presented in Figure 3.4. A lookup table containing Heintzelman et al.’s [33] 66 pain-related medical terminologies was used to determine which ‘signs and symptoms’, detected by the program, were pain-related (Table 3.2).

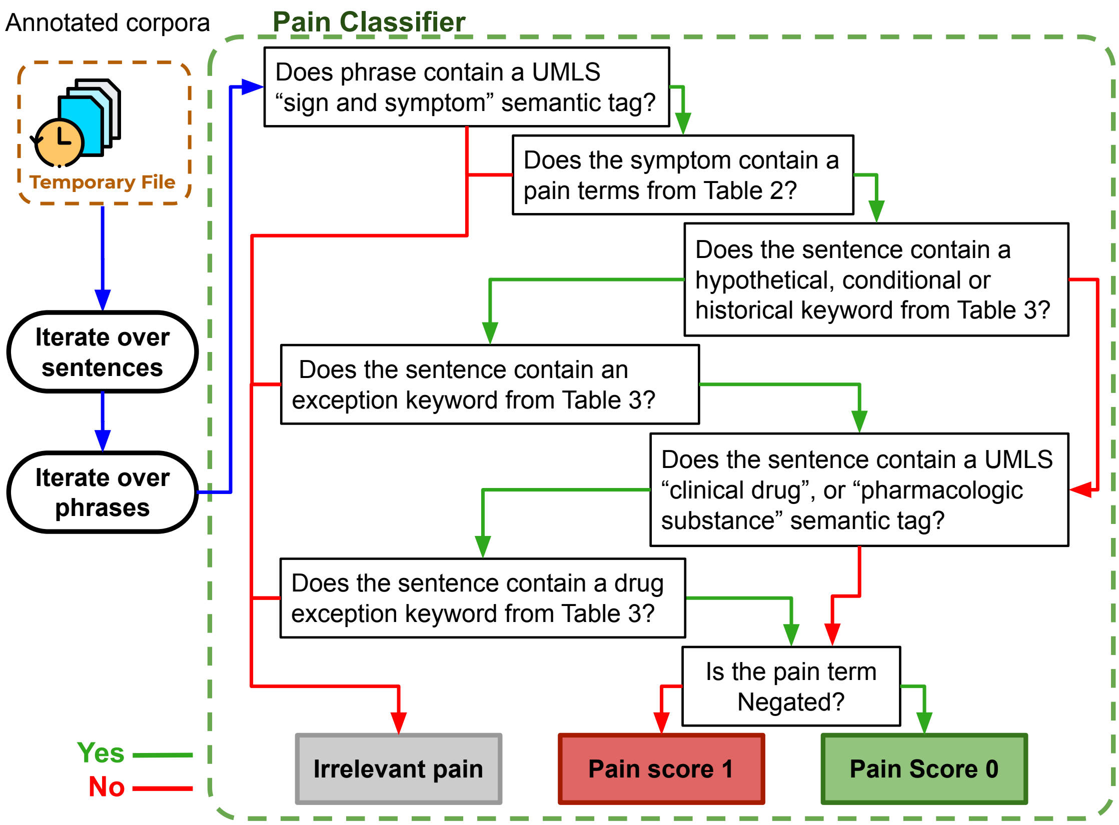

Annotated corpora
Temporary File
Iterate over sentences
Iterate over phrases
Pain Classifier
Does phrase contain a UMLS "sign and symptom" semantic tag?
Does the symptom contain a pain terms from Table 2?
Does the sentence contain a hypothetical, conditional or historical keyword from Table 3?
Does the sentence contain an exception keyword from Table 3?
Does the sentence contain a UMLS "clinical drug", or "pharmacologic substance" semantic tag?
Does the sentence contain a drug exception keyword from Table 3?
Is the pain term Negated?
Irrelevant pain
Pain score 1
Pain Score 0
Yes
No

Figure 3.4: Our NLP pain classification pipeline to extract the physician-reported pain scores from patients' clinical notes. Annotated files were processed phrase by phrase to filter UMLS 'signs and symptoms' tags and identify pain-related biomedical concepts according to Table 3.2. Then, sets of rules were developed to remove hypothetical, historical and drug related mentions of the pain and keep the pain term associated to the state of the pain at the time of the hospital visit. Finally, a pain score was assigned to the detected pain term based on the negation status of the phrase.

In order to obtain the pain score at the time of the consultation/ hospitalization, we excluded irrelevant mentions of pain. For example, we excluded mentions of pain when the patient talked about the history of pain that was not actually presented at the time of the consultation/hospital visit. For this purpose, we trained our pipeline in four iterations by

ache
aching
angina
arthralgia
arthrodynia
burning
cephalalgia
cephalgia
cephalodynia
cervicalgia
cervicodynia
claudication
coccyalgia
coccydynia
coccygalgia
coccygodynia
coccyodynia
coccyxdynia
coxalgia
cp
cramp
discomfort
dolor
dorsalgia
dorsodynia
dysuria
esophagodynia
glossalgia
glossalgias
glossodynia
glossodynias
gonalgia
inguinodynia
lbp
low back syndrome
lumbago
lumbalgia
meralgia
metatarsalgia
muscle weakness
myalgia
myalgias
myodynia
myosalgia
neuralgia
neuralgias
odynophagia
orchialgia
orchidalgia
orchidodynia
osteodynia
otalgia
pain
pancreatalgia
postherpetic
neuralgia
pressure
proctalgia
rectalgia
retrosternal
scapulalgia
scapulodynia
sciatica
sore
tender
tightness

Table 3.2: Pain-related medical terminologies taken from Heintzelman et. al. [33]. These definitions were used to determine pain-related 'signs and symptoms' in the clinical notes by our pain classifier.

manually auditing 5,138 randomly-selected sentences from the training corpora:

1) By randomly examining the training corpora, we created a lookup table containing regular expressions related to conditional, hypothetical, and historical terms. These regular expressions were used to search and exclude any pain term used in a conditional, hypothetical, or historical context (Table 3.3). We used the first validation set to evaluate the performance of the NLP pipeline in correctly detecting valid pain terms.

2) By examining the training corpora, we created a lookup table containing regular expressions describing current events or ongoing situations such as 'present', 'where', and 'control'. This table (called exceptions) is used to avoid the removal of pain terms related to the current state of the illness. Improvements in the performance of our NLP pipeline was evaluated using the validation set 2.

3) We used the Global Vectors Word Representation (GloVe) algorithm [48] to generate semantic embedding vectors for all keywords (regular expressions) in the above-mentioned lookup tables. Then, for each keyword, we found five nearest GloVe words in semantic space and added them into the corresponding lookup tables (Table 3.3). The validation set 3 was used to check the performance of the NLP pipeline at this iteration.

4) We removed pain terms associated with pain medications by excluding phrases containing the UMLS 'pharmacologic substance' and 'clinical drug' semantic type. Also, by randomly examining the training corpus, we created an exception lookup table to avoid removing pain terms associated with non-pain-related or ambiguous pharmacologic substances like 'dob', 'his', and 'lead'. We used the validation set 4 to evaluate how this iteration improved the the performance of the NLP pipeline.

In each iteration of the training, depending on the performance of our NLP pipeline on the validation corpora, we either added more keywords to each of the four lookup tables (Table 3.3) or removed some keywords from the tables. For example, the keyword 'since' was initially in the conditionals lookup table. But after iteration 1, we moved this keyword to the exceptions lookup table, because, we found that most of the sentences with the keyword 'since' were indicating an ongoing event. We added the keyword 'p.r.n' to the conditionals lookup table, since we found that sentences that includes the 'p.r.n' keyword were most likely talking about a prescription drug. Another example was ambiguous drug names. For instance, we found that the MetaMap classified keyword 'his' as Histidine [Pharmacologic Substance] and keyword 'dob' as Dimethoxybromoamphetamine [Pharmacologic Substance]. We added both these terms to the exception look up table.

Once satisfied with the training and validation, we did no more development on our NLP pipeline and used gold-standard corpora to evaluate the final performance of the NLP pipeline. Table 3 contains the final versions of the lookup tables. Our NLP pipeline is available as open-source in Ref. [46]

As illustrated in Figure 3.4, after passing through the selection rules each phrase was assigned to one of the three scores: valid mention of experienced pain (pain score = 1), valid explicit denial of pain (pain score = 0), and no/irrelevant mention of pain (score = nan) by our pipeline. The third label was primarily used for NLP performance evaluation. Examples of NLP extracted pain scores from i2b2 corpora are provided in Tables 3.5 and 3.6.

Step 3: VDP classification method

Valid pain scores were averaged for each note using Eq. 3.1 to obtain the average pain intensity at the time of the consultation.

Conditionals	If, whether, when, in case of, in case, as needed, return,
Hypothetical	might, would, could, should, seek, as needed, call, return, possibly, possible, please, because of, p.r.n.
Historical	History, historical, in the past, previous, before, previously, in the last, prior, recent years
Exceptions	since, present, current, now, where, because of, prevent, manage, diagnosis, control, found, lasted, treated, resolved, comfort, diagnosis, severe, worsening, aggravated, diffuse, severity, increased, score, high, mild, moderate
Drug mentioned	clinical drug, pharmacologic substance
Drug Exceptions	f-, his, lead, histidine, prevent, wake, level, helium, dob

Table 3.3: The lookup tables were formed by examining the training corpus and using the GloVe semantic embedding system. These tables were used to exclude phrases with conditional, hypothetical, historical, and drug-related mentions of pain, and to keep sentences with mentions of the patient's current state of pain in our analysis.

$$Average\ Pain\ Intensity = \frac{\sum(score\ 1\ pains) - \sum(score\ 0\ pains)}{\sum(score\ 1\ pains) + \sum(score\ 0\ pains)} \tag{3.1}$$

To the best of our knowledge, there are no clinical guidelines to assign a VDP score for overall pain [49]. Our rationale for using a weighted average was to take into account the effect of the number of pain mentions. Also, using a weighted average made it easier for us to map average intensity to VDP. Such a weighted averaging has been previously proposed in the literature for the evaluation of multi-site pain [49–51].

A weighted average pain intensity can range from -1 (when 100% of the valid pain mentions were negated) to 1 (if none of the valid pain mentions was negated). We grouped the average pain intensities in two VDPs by setting the intensity threshold at zero as 'no pain' ($average\ pain\ intensity \leq 0$), and 'pain' ($average\ pain\ intensity \geq 0$). We used the Receiver Operating Characteristic (ROC) curve and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) value [52] to examine the performance of a VDP assignment at various intensity thresholds.

3.4.4 Assessment of the pipeline's performance

Annotated notes from the validation corpora were used to check and tune our NLP pipeline at each iteration of the training. The gold-standard corpora, explained earlier, were used to check the performance of our fully developed NLP pipeline. Confusion matrices were produced to compare the pipeline's performance against expert-annotated gold-standard corpora. To avoid bias, NLP developers were kept blinded to the test corpora throughout the entire process.

The confusion matrix for our three-label pain classifier is a 3 × 3 matrix, as presented in Table 3.4. This matrix includes 3 TRUE labels for correctly-scored sentences, and 6 FALSE labels for incorrectly-scored sentences (more details are provided in section 3.9.5).

		Predicted Label		
		Score 1	Score 0	Irrelevant
True Label	Score 1	T_{PP}	F_{PN}	T_{PI}
	Score 0	F_{NP}	T_{NN}	F_{NI}
	Irrelevant	F_{IP}	F_{IN}	T_{II}

Table 3.4: The confusion matrix for the three-label pain classifiers contains 3 correctly-predicted labels and 6 incorrectly-predicted labels. T_{PP}, T_{NN}, and T_{II} are numbers of sentences that are correctly predicted as score 1, score 0, and irreverent pains, respectively. F_{NP}, F_{PN}, F_{PI}, F_{NI}, F_{IP}, and F_{IN} are numbers of mislabeled sentences.

We evaluated the performance of our NLP pipeline for pain scoring and VDP assignment by calculating the precision, recall, and F1-score (F1) from the confusion matrices [52].

3.5 Results

3.5.1 Pain classifier

We tested our NLP pipeline's ability to extract pain terms from notes in the i2b2, MIMIC-III, and ARIA corpora. By processing all the available corpora, we found 19,851, 12,071, and 1,883 suggested pain concepts, respectively. Note that these pain concepts include all the UMLS concepts that were extracted from the clinical notes. This means that multiple pain concepts may have been mapped to one phrase as described earlier.

The result of our rule-based pain detection pipeline (shown in Figure 3.4) for detecting the pain score is presented in Table 3.7. Using the UMLS confidence score to remove duplicate concepts, we obtained uniquely-mapped experienced pain terms and explicitly denied pain terms from the i2b2, MIMIC-III, and ARIA corpora. Finally, by removing conditional, hypothetical, and drug-related pain terms, we obtained 2,845, 1,682, and 2,013 relevant terms presenting the pain score 1 as well as 1,540, 1,427, and 559 score 0 pain terms in the i2b2, MIMIC-III, and ARIA corpora, respectively. Table 3.5 contains a few example sentences from i2b2 corpora in which pain scores were correctly labeled by our NLP pipeline. Examples of pain terms that were not labeled correctly by our NLP pipeline are provided in Table 3.6.

On averaging over the pain scores in each note using Eq. 3.1, we obtained the VDP at the time of consultation/hospitalization in the three corpora. Distribution of the VDP is presented in Table 3.8. Based on our VDP calculations, we found that pain was not documented in 22% of the cancer notes, 13% of our cancer patients denied the experience of pain and at least 65% of cancer patients experienced some level of pain. These results were in agreement with the results reported in the other papers [53].

3.5.2 inter-annotator agreement

Inter-annotator agreement among 6 annotators in assigning notes to a 5-grade pain scale is provided in Table 3.18 (Supplementary Information). We calculated Fleiss' kappa measure and obtained a moderate agreement among 6 annotators ($\kappa = 0.43$). This indicated that pain scores were not sufficiently documented in the consultation notes. Therefore, we instead defined a 3-grade pain scale (called VDP status) by merging 'mild', 'moderate' and

Sentence	**Manual pain score**	**NLP pain score**
he states the feeling returned and then persisted, took a 2nd nitro but it only decreased the pain to a [**2192-2-16**].	1	1
per notes, her abdominal exam was significant for epitastric and right upper quadrant tenderness;	1	1
the patient took one sublingual nitro at home with some relief , but the pain came back as she walking around her home looking for her hospital identification care.	1	1
He had no chest pain but did have diaphoresis and mild nausea and vomiting as well as lightheadedness and some palpitations lasting approximately one hour in duration. ia.	0	0
he had no further episodes of chest pain while in the hospital.	0	0
patient denies shortness of breath , chest pressure , or syncope.	0	0
he denies fevers or chills, shortness of breath or abdominal pain.	0	0
in july , 1989 , he developed chest pain and suffered an inferior myocardial infarction.	-	-
one week prior to admission , the patient had chest pain , which was quickly relieved by one sublingual nitroglycerin.	-	-
morphine 15 mg tablet sustained release sig: one (1) tablet sustained release po every 4-6 hours as needed for pain.	-	-
if you develop chest pain, nausea, vomiting, throat tightness, clamminess or shortness of breath, call your pcp or go to the emergency room.	-	-

Table 3.5: Examples of the sentences from i2b2 corpora that were labeled correctly.

'severe pain' assignments into a single category as 'pain'. We measured the inter-annotator agreement again and we obtained substantial agreement between six annotators in assigning VDP with Fleiss' kappa measure of $\kappa = 0.66$.

3.5.3 Performance of the pain classifier

The confusion matrices, generated by comparing NLP-extracted pain scores against expert-annotated gold-standard from each corpus, are presented in Table 3.9. Based on these confusion matrices, we calculated precision, recall and F1-score. These results are summarized in Table 3.10. To compare the performance of our NLP pipeline with the prior studies, we provided the performances of the pain extraction NLP algorithms presented by Heintzelman et al. [33] and Bui and Zeng-Treitler [35].

Sentence	**Manual pain score**	**NLP pain score**
she refused any consultation at this time by the [*** ****] hospital pain service.	-	1
his left groin was not accessed given his c/o left leg pain post surgery 2 months ago.	-	0
the surgical sites were without any exudate or signs of infection and his tenderness in his right upper extremity was markedly decreased.	1	0
in the ambulance , the patient continued to have the pain and she received one more sublingual nitroglycerin and nasal cannula oxygen.	1	-
the patients abdominal pain could be related to intestinal angina.	1	-
asa , o2 , bb , 1 inch of nitropaste for elev bpof note , pt c/o pain on the r mid-lower back which has been present x 1 wk , reproducible on light palpation.	1	-
history of present illness: 74 y/o female with pmh significant for copd, cad, and hypertension admitted to [**hospital1 18**] on [**6-14**] to the surgery service with two days of epigastric and right upper quadrant pain.	1	-
she does however complain of some urinary frequency (on lasix) in the last few days with out any dysuria or urgency.	0	-

Table 3.6: Examples of the mislabeled sentences from i2b2 corpora.

	i2b2 % (n=4385)	MIMIC-III % (n=3109)	ARIA % (n=2572)
Score 1 pain	64.9	54.1	78.3
Score 0 pain	35.1	45.9	21.7

Table 3.7: The frequency of score 0 and score 1 pain terms labeled by the NLP pipeline in each of the three corpora. Total number of valid pain terms are provided inside the brackets.

3.5.4 Performance of the VDP classifier

The performance of the VDP classification method was evaluated using the 3-grade VDP gold-standard corpora. A 3 × 3 confusion matrix was formed for the three-grade VDP, as explained in the section 3.4.4. Table 3.11 shows the confusion matrix for NLP extracted VDP. The ROC curve is plotted in Figure 3.5 for various intensity thresholds. The ROC-AUC is calculated to be 0.86.

Of the 150 notes selected for the performance evaluation, 14 notes did not have any valid

	i2b2	MIMIC-III	ARIA
Pain	706 (56.5%)	442 (50.4%)	511 (64.9%)
No pain	305 (24.4%)	262 (29.9%)	104 (13.2%)
No mention of pain	238 (19.1%)	173 (19.7 %)	173 (21.9%)

Table 3.8: Verbally-Declared Pain (VDP) at the time of the consultation using all available notes from each corpora. The VDP was obtained by averaging over all pain scores in each note. Percentile values are presented inside the brackets.

		Predicted Label		
	i2b2	Pain score 1	Pain score 0	Irrelevant
True Label	Pain score 1	78	1	11
	Pain score 0	0	22	1
	Irrelevant	5	1	2241

		Predicted Label		
	MIMIC-III	Pain score 1	Pain score 0	Irrelevant
True Label	Pain score 1	51	1	6
	Pain score 0	0	15	3
	Irrelevant	3	1	2635

		Predicted Label		
	ARIA	Pain score 1	Pain score 0	Irrelevant
True Label	Pain score 1	70	1	13
	Pain score 0	1	24	5
	Irrelevant	10	1	1007

Table 3.9: Following the approach described in Table 3.4, for each corpus a three-class confusion matrix was obtained. The name of the corresponding corpus is mentioned in the top left cell of the matrix.

mention of pain (no mention of pain), 112 notes had ‘pain’, and 24 had ‘no pain’ (denied pain) VDP. Among the 112 notes with the mentions of experienced pain, our VDP extraction method correctly classified 104 of them while five were misclassified as no pain and the other three were misclassified as no mention of pain. Among the 24 notes with no pain VDP, our pipeline correctly classified 16 of them and incorrectly labeled seven as pain and one as no mention of pain.

Based on these results, we calculated the precision, recall, and F1-score for the VDP extraction method that are shown in Table 3.12. We achieved 92%, 76%, and 75% precision

Author	Pain Score 1			Pain Score 0		
	P	R	F	P	R	F
Present Study (i2b2)	94.0	86.7	90.2	91.7	95.7	93.6
Present Study (MIMIC-III)	94.4	88.0	91.1	88.2	83.3	85.7
Present Study (ARIA)	86.4	83.3	84.8	92.3	80.0	85.7
Heintzelman et al. [33] [a]	86	95	90	82	95	88
Bui and Zeng-Treitler [35] [b]	73.2	56.6	63.8	78.8	74.2	76.4

[a] Calculated based on the manual annotation of 111 pain mentions that were extracted from 30 discharge summaries from i2b2 database.

[b] Calculated based on manual annotation of 702 pain mentions that were extracted from 100 documents from the US Department of Veterans Affairs' (VA) electronic medical records.

Table 3.10: The precision (P), recall (R) and F1-score (F) of the pain detection pipeline calculated based on the confusion matrices presented in Table 3.9. The performances of the NLP pipelines from prior studies are provided for a comparison.

		Predicted VDP		
	ARIA	Pain	No pain	No mention of pain
True VDP	Pain	104	5	3
	No pain	7	16	1
	No mention of pain	2	0	12

Table 3.11: Following the approach described in section 3.4.4, a three-point VDP confusion matrix was formed based on the manual audit of 120 randomly-selected notes from the ARIA corpora.

ARIA	Precision	Recall	F1-score
Pain	92.0%	92.9%	92.4%
No pain	76.2%	66.7%	71.4%
No mention of pain	75.0%	85.7%	80.4%

Table 3.12: The precision, recall and F1-score of the VDP extraction method has been calculated using Table 3.11.

in classifying the notes into the 'pain', 'no pain', and 'no mention of pain' VDP, respectively.

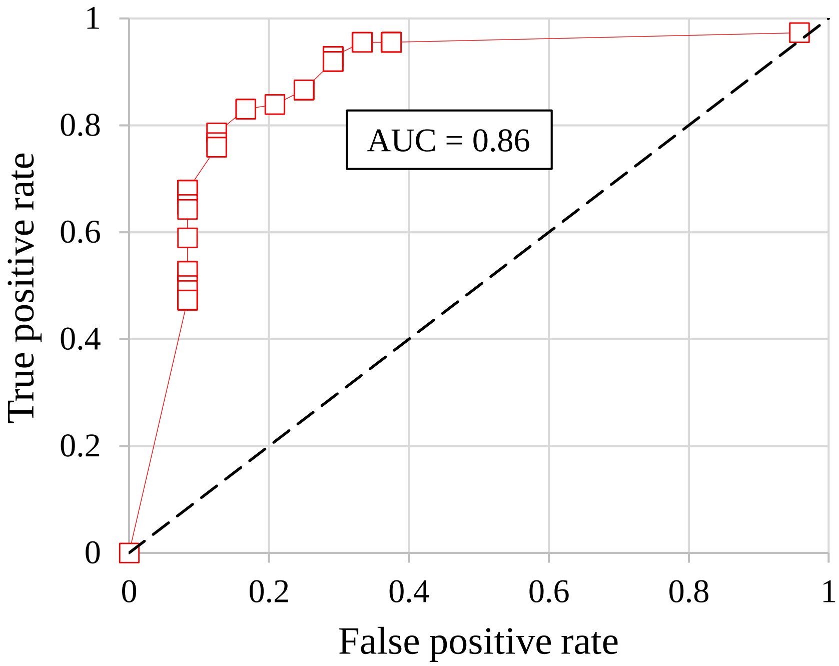


Figure 3.5: The ROC curve was generated to investigate the performance of a VDP classification method at various intensity thresholds. The ROC-AUC is calculated to be 0.86.

3.6 Discussion

3.6.1 Quality of corpora

Comparing the number of words and sentences in Figure 3.1, we found that the consultation notes from the ARIA corpus contained noticeably fewer words and sentences compared to the discharge summaries from the i2b2 and MIMIC-III corpora. Since notes from the i2b2 and MIMIC-III corpora were pre-processed and de-identified for public use, they contained more broken sentences. Nonetheless, we found that the distribution of the length of words and sentences were similar across all three corpora. Therefore, the similarity of the datasets was not very affected by the pre-processing and de-identification steps. This suggests that notes from various resources are similar enough to be used together in a study such as this.

3.6.2 Distribution of the pain terms in the notes

Distribution of the pain terms in the notes from three corpora, presented in Table 3.19 in the Supplementary Information, revealed that pain distribution from the ARIA corpus was notably different from the other two corpora. As expected, the ARIA corpus included only patients with bone metastases,hence, there were more mentions of bone-related pain terms such as back pain and pelvic pain. We also observed that almost 58% of the experienced pain was reported as generic pain without specifying the pain site in the ARIA corpus while this was only 34% and 38% in the other two corpora. We suspect that it was because the consultation notes in ARIA were prepared by radiation oncologists who solely examined cancer patients, while discharge summaries were prepared by general physicians who visited patients with a variety of conditions.

Comparing the experienced pain with the total pain mentions, we detected that pain was experienced in 65% and 54% of the cases in the i2b2 and MIMIC-III corpora respectively, while this number increased to 78% in the ARIA corpus. Again, we assume that the explanation for this might be due to the nature of these three corpora with i2b2 and MIMIC-III containing notes for patients visiting general hospitals while our ARIA database included exclusively notes for cancer patients with bone metastases. Remarkably, the 78% experienced pain for metastatic cancer patients agrees with the results reported in several other studies [54, 55].

3.6.3 Accuracy of the pain score measurements

Performance of our NLP pipeline was evaluated using the gold-standard test sets explained in the section 3.4.4. As presented in table 3.10, our pipeline outperformed prior pain detection pipelines developed by Heintzelman et al. [33] and Bui and Zeng-Treitler [35].

Once we fully trained and tested our pipeline using the i2b2 training corpus, we examined the generalizability of our NLP pain detection pipeline using independent corpora from MIMIC-III and ARIA. Note that our NLP pipeline was used on the MIMIC-III and ARIA corpora without further training on these corpora. The precision of our NLP pipeline in detecting score 1 pain did not change when we applied it to the MIMIC-III corpora. However, it dropped to 86% when we applied our pipeline to the ARIA corpora. The reason for having more mislabeled score 1 pain in the ARIA corpus can be attributed to the difference in the corpus type. The i2b2 and MIMIC-III corpora were general hospital discharge summary notes, while the ARIA corpus comprises radiation oncology consultation notes. As stated previously, up to 50% reduction in the precision is commonly expected when moving from public corpora to private corpora. Therefore, a 12% drop in the precision of our NLP pipeline was reasonable. This suggests that NLP pipelines that are trained on one type of documents (i.e. hospital discharge summaries in this case) can be successfully transferred to analyze patients' other clinical notes (such as cancer consultation notes in this study).

The precision in detecting score 0 pain reduced from 92% to 88% when the MIMIC-III corpus was analyzed. The decrease in precision might be as a result of more diverse negation terms in the MIMIC-III, which includes notes from more diverse sources compared to the i2b2 database. The precision of our pipeline in detecting score 0 pain terms was 92% when analyzing the ARIA corpus. The main reason for such a high precision was because of better sentence segmentation in ARIA corpus compared to i2b2 and MIMIC-III corpora. Both the i2b2 and MIMIC-III were de-identified corpora with a lot of broken sentences. Therefore, it was much harder for our NLP pipeline to detect negation (score 0 pain terms). Examples of mislabeled pain terms are presented in Table 3.6.

The recall parameter provided more information about the behavior of our NLP system. Recall was the measure of how well our pipeline correctly identified all true labels. In the i2b2 and MIMIC-III corpora, we achieved 87% and 88% recall in detecting score 1 pain, respectively. The recall decreased to 83% for the ARIA corpus. As shown in Table 3.9, in the ARIA corpus, a notable number of the score 1 pain was assigned as irrelevant pain. We believe this noticeable mislabeling were related to the pain terms that were describing patient's previous experience of having pain. As expected, most of cancer patients had a history of long term chronic pain which presumably made it difficult for our pipeline to separate them from pain at the time of the consultation.

The recall values for detecting all mentions of score 0 pain were 96%, 83%, and 80% for the i2b2, MIMIC-III, and ARIA corpora, respectively. We believe that this variation in the recall values of score 0 pain was partially due to the layout of the notes in each corpus. For example, in the i2b2 corpus that was used to train our NLP pipeline, each note had a separate section for prescription drugs. Therefore, the drug-related pain terms could be filtered much easier than in the MIMIC-III corpus in which the prescription drugs were mentioned within the notes in an unstructured format. It should be noted that , as we explained in section 3.9.1, we did not cut any segment of the notes in any of the corpora to assure the generalizability of our pipeline.

Having fewer score 0 pain terms might also influence the calculated recall values. Table 3.9 shows that there were only 23, 18 and 30 score 0 pain terms in i2b2, MIMIC-III and ARIA validation corpora, respectively. This means that any mislabeled score 0 pain, introduced a large uncertainty to the recall values.

The overall performance of our NLP pipeline on various corpora was also evaluated using F1-scores. The F1-score did not vary much among the three corpora. F1-score of score 1 pain only decreased from 0.90 in i2b2 to 0.85 in ARIA corpus. Similarly, F1-score of score 0 pain changed from 0.94 in i2b2 to 0.86 in ARIA corpus.

3.6.4 Accuracy of the VDP extraction

Based on the ROC curve with an ROC-AUC value of 0.86 (Figure 3.5), our VDP extraction method had good performance in distinguishing between patients with and without pain. As presented in Table 3.12, our VDP extraction method successfully detected 'pain' with 92.0% precision and 92.9% recall. However, it showed fundamental limitations in detecting 'no pain', with 76.2% precision and 66.7% recall. The main reason for such a high recall and low precision in detecting no pain VDP was that ARIA was an imbalanced corpus, where the classes were not represented equally (i.e. there was $\sim$ 5x more experienced pain than no pain cases, as shown in Table 3.11.)

In addition, investigating the notes in the i2b2 training set, we noticed that when patients reported pain at multiple sites in their body, our classification method was not able to extract VDP precisely. Our method of measuring VDP was confounded by the reality of the notes of metastatic cancer patients, because, for these patients, it is expected to have multiple pain sites with different pain scores in each site.

One possible solution is to add functionality to obtain pain severity from patients' consultation notes by analyzing the pain assessment terminologies (such as severe, mild, controlled) and by capturing numerical pain scores for each identified pain site directly from the consultation notes.

3.7 Conclusion

Our database-independent NLP pipeline, trained using i2b2 hospital discharge summary corpora, was successfully implemented to detect and classify pain from the publicly-available MIMIC-III hospital discharge summary corpus, and our institutional radiation oncology ARIA consultation note database for cancer patients with bone metastases. The pipeline's performance was evaluated against physician-annotated gold standard corpora. Our pipeline achieved a precision and a recall of 89% and 82% in detecting physician-reported pain, respectively, demonstrating successful and state-of-the-art extraction and classification of pain from radiation oncology clinical notes. It also automatically assigned a VDP for each clinical note with 84% and 80% overall precision and recall.

An important and intended application of our NLP tool is that it can be used to reliably extract physician-reported cancer pain from clinical notes in radiation oncology, where the pain is not otherwise documented through structured data entry. Having access to this database-independent NLP pain-extraction pipeline will facilitate further informatics and data-mining studies in radiation oncology that require access to pain information that is typically very difficult to obtain.

3.8 Acknowledgement

Funding for this research was provided by the startup grant of Dr. John Kildea at Research Institute of the McGill University Health Centre (RI-MUHC), The RI-MUHC studentship award, Ruth and Alex Dworkin scholarship from the McGill University Faculty of Medicine, and graduate excellence award from the McGill University Department of Physics. The authors would like to thank Ms. Haley Patrick for the manual audit of our validation sets. We thank Mr. Farzin Khosrow-Khavar for his help with sample size evaluation and statistical analysis. We also thank Dr. Marc David for his clinical support.

Bibliography

[1] W. H. Organization, WHO guidelines for the pharmacological and radiotherapeutic management of cancer pain in adults and adolescents, World Health Organization, 2018. URL www.who.int/http://www.ncbi.nlm.nih.gov/pubmed/30776210

[2] M. G. Nayak, A. George, M. S. Vidyasagar, S. Mathew, S. Nayak, B. S. Nayak, Y. N. Shashidhara, A. Kamath, Quality of life among cancer patients, Indian Journal of Palliative Care 23 (4) (2017) 445–450. doi:10.4103/IJPC.IJPC{_}82{_}17.

[3] L. S. Simon, Relieving pain in America: a blueprint for transforming prevention, care, education, and research, Journal of Pain & Palliative Care Pharmacotherapy 26 (2) (2012) 197–198. doi:10.3109/15360288.2012.678473.

[4] G. G. Page, S. Ben-Eliyahu, The immune-suppressive nature of pain., Seminars in oncology nursing 13 (1) (1997) 10–15. doi:10.1016/S0749-2081(97)80044-7.

[5] D. B. Gordon, J. L. Dahl, C. Miaskowski, B. McCarberg, K. H. Todd, J. A. Paice, A. G. Lipman, M. Bookbinder, S. H. Sanders, D. C. Turk, D. B. Carr, American Pain Society Recommendations for Improving the Quality of Acute and Cancer Pain Management: American Pain Society Quality of Care Task Force, Archives of Internal Medicine 165 (14) (2005) 1574–1580. doi:10.1001/archinte.165.14.1574.

[6] M. P. Cadogan, J. F. Schnelle, N. R. Al-Sammarrai, N. Yamamoto-Mitani, G. Cabrera, D. Osterweil, S. F. Simmons, A standardized quality assessment system to evaluate pain detection and management in the nursing home, Journal of the American Medical Directors Association 6 (1) (2005) 1–9. doi:10.1016/j.jamda.2004.12.002.

[7] T. J. Keay, The mind-set of pain assessment (1 2005). doi:10.1016/j.jamda.2004.12.011.

[8] L. Ohno-Machado, Realizing the full potential of electronic health records: the role of natural language processing, Journal of the American Medical Informatics Association 18 (5) (2011) 539–539. doi:10.1136/amiajnl-2011-000501.

[9] S. Bird, E. Klein, E. Loper, Natural Language Processing with Python, 1st Edition, OŔeilly Media, Inc., 2009.

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[11] WHO, International Classification of Diseases, 11th Revision (ICD-11), WHO (2019).
URL http://www.who.int/classifications/icd/en/

[12] S. International, SNOMED CT January 2020 International Edition - SNOMED International Release notes - SNOMED International Release Management - SNOMED Confluence ([online]).
URL https://confluence.ihtsdotools.org/display/RMT

[13] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology., Nucleic acids research 32 (Database issue) (2004) 267–70. doi:10.1093/nar/gkh061.

[14] Bethesda (MD), UMLS ® Reference Manual, National Library of Medicine (US), 2009.
URL https://www.ncbi.nlm.nih.gov/books/NBK9676/

[15] A. R. Aronson, F. M. Lang, An overview of MetaMap: Historical perspective and recent advances, Journal of the American Medical Informatics Association 17 (3) (2010) 229–236. doi:10.1136/jamia.2009.002733.

[16] D. Demner-Fushman, W. J. Rogers, A. R. Aronson, MetaMap Lite: an evaluation of a new Java implementation of MetaMap, Journal of the American Medical Informatics Association 24 (4) (2017) 841–844. doi:10.1093/jamia/ocw177.

[17] J. Zhang, X. Long, T. Suel, Performance of compressed inverted list caching in search engines, in: Proceeding of the 17th International Conference on World Wide Web 2008, WWW'08, ACM Press, New York, New York, USA, 2008, pp. 387–396. doi:10.1145/1367497.1367550.

[18] L. M. Simon, S. Karg, A. J. Westermann, M. Engel, A. H. A. Elbehery, B. Hense, M. Heinig, L. Deng, F. J. Theis, MetaMap: an atlas of metatranscriptomic reads in human disease-related RNA-seq data, GigaScience 7 (6) (06 2018). doi:10.1093/gigascience/giy070.

[19] R. Reátegui, S. Ratté, Comparison of metamap and ctakes for entity extraction in clinical notes, BMC Medical Informatics and Decision Making 18 (9 2018). doi:10.1186/s12911-018-0654-2.

[20] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, B. G. Buchanan, A simple algorithm for identifying negated findings and diseases in discharge summaries, Journal of Biomedical Informatics 34 (5) (2001) 301–310. doi:10.1006/jbin.2001.1029.

[21] S. Wu, T. Miller, J. Masanz, M. Coarr, S. Halgrim, D. Carrell, C. Clark, Negations Not Solved: Generalizability Versus Optimizability in Clinical Natural Language Processing, PLoS ONE 9 (11) (2014) e112774. doi:10.1371/journal.pone.0112774.

[22] Z. Zeng, Y. Deng, X. Li, T. Naumann, Y. Luo, Natural Language Processing for EHR-Based Computational Phenotyping, IEEE/ACM Transactions on Computational Biology and Bioinformatics 16 (1) (2019) 139–153. doi:10.1109/TCBB.2018.2849968.

[23] X. Wang, A. Chused, N. Elhadad, C. Friedman, M. Markatou, Automated knowledge acquisition from clinical narrative reports., AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium 2008 (2008) 783–787.

[24] I. V. Haller, C. M. Renier, M. Juusola, P. Hitz, W. Steffen, M. J. Asmus, T. Craig, J. Mardekian, E. T. Masters, T. E. Elliott, Enhancing Risk Assessment in Patients

Receiving Chronic Opioid Analgesic Therapy Using Natural Language Processing, Pain Medicine 18 (10) (2016) 1952–1960. doi:10.1093/pm/pnw283.

[25] T. A. Koleck, C. Dreisbach, P. E. Bourne, S. Bakken, Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review, Journal of the American Medical Informatics Association 26 (4) (2019) 364–379. doi:10.1093/jamia/ocy173.

[26] A. Hardjojo, A. Gunachandran, L. Pang, M. R. B. Abdullah, W. Wah, J. W. C. Chong, E. H. Goh, S. H. Teo, G. Lim, M. L. Lee, W. Hsu, V. Lee, M. I.-C. Chen, F. Wong, J. S. K. Phang, Validation of a Natural Language Processing Algorithm for Detecting Infectious Disease Symptoms in Primary Care Electronic Medical Records in Singapore, JMIR Medical Informatics 6 (2) (2018) e36. doi:10.2196/medinform.8204.

[27] G. K. Savova, E. Tseytlin, S. Finan, M. Castine, T. Miller, O. Medvedeva, D. Harris, H. Hochheiser, C. Lin, G. Chavan, R. S. Jacobson, DeepPhe: A natural language processing system for extracting cancer phenotypes from clinical records, Cancer Research 77 (21) (2017) e115–e118. doi:10.1158/0008-5472.CAN-17-0615.

[28] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, J. F. Hurdle, Extracting information from textual documents in the electronic health record: a review of recent research., Yearbook of medical informatics (2008) 128–144doi:10.1055/s-0038-1638592.

[29] S. S. Pakhomov, H. Hemingway, S. A. Weston, S. J. Jacobsen, R. Rodeheffer, V. L. Roger, Epidemiology of angina pectoris: Role of natural language processing of the medical record, American Heart Journal 153 (4) (2007) 666–673. doi:10.1016/j.ahj.2006.12.022.

[30] W. K. Tan, S. Hassanpour, P. J. Heagerty, S. D. Rundell, P. Suri, H. T. Huhdanpaa, K. James, D. S. Carrell, C. P. Langlotz, N. L. Organ, E. N. Meier, K. J. Sherman, D. F. Kallmes, P. H. Luetmer, B. Griffith, D. R. Nerenz, J. G. Jarvik, Comparison of Natural Language Processing Rules-based and Machine-learning Systems to Identify Lumbar Spine Imaging Findings Related to Low Back Pain, Academic Radiology 25 (11) (2018) 1422–1432. doi:10.1016/j.acra.2018.03.008.

[31] T. Y. Tian, I. Zlateva, D. R. Anderson, Using electronic health records data to identify patients with chronic pain in a primary care setting, Journal of the American Medical Informatics Association 20 (E2) (2013) e275. doi:10.1136/amiajnl-2013-001856.

[32] S. J. Fodeh, D. Finch, L. Bouayad, S. L. Luther, H. Ling, R. D. Kerns, C. Brandt, Classifying clinical notes with pain assessment using machine learning, Medical and Biological Engineering and Computing 56 (7) (2018) 1285–1292. doi:10.1007/s11517-017-1772-1.

[33] N. H. Heintzelman, R. J. Taylor, L. Simonsen, R. Lustig, D. Anderko, J. A. Haythornthwaite, L. C. Childs, G. S. Bova, Longitudinal analysis of pain in patients with metastatic prostate cancer using natural language processing of medical record text, Journal of the American Medical Informatics Association 20 (5) (2013) 898–905. doi:10.1136/amiajnl-2012-001076.

[34] A. S. Eisman, N. R. Shah, C. Eickhoff, G. Zerveas, E. S. Chen, W.-C. Wu, I. N. Sarkar, Extracting angina symptoms from clinical notes using pre-trained transformer architectures (2020). arXiv:2010.05757.

[35] D. D. A. Bui, Q. Zeng-Treitler, Learning regular expressions for clinical text classification, Journal of the American Medical Informatics Association 21 (5) (2014) 850–857. doi:10.1136/amiajnl-2013-002411.

[36] V. Major, A. Surkis, Y. Aphinyanaphongs, Utility of General and Specific Word Embeddings for Classifying Translational Stages of Research, AMIA. Annual Symposium proceedings. AMIA Symposium 2018 (2018) 1405–1414.

[37] Y. Si, J. Wang, H. Xu, K. Roberts, Enhancing clinical concept extraction with contextual embeddings, Journal of the American Medical Informatics Association 26 (11) (2019) 1297–1304. doi:10.1093/jamia/ocz096.

[38] C. Tao, M. Filannino, O. Uzuner, Prescription extraction using crfs and word embeddings, Journal of Biomedical Informatics 72 (2017) 60 – 66. doi:https://doi.org/10.1016/j.jbi.2017.07.002.

[39] D. T. Heinze, M. L. Morsch, B. C. Potter, R. E. Sheffer, Medical i2b2 NLP Smoking Challenge: The A-Life System Architecture and Methodology, Journal of the American Medical Informatics Association 15 (1) (2008) 40–43. doi:10.1197/jamia.M2438.

[40] O. Uzuner, Y. Luo, P. Szolovits, Evaluating the State-of-the-Art in Automatic De-identification, Journal of the American Medical Informatics Association 14 (5) (2007) 550–563. doi:10.1197/jamia.M2444.

[41] A. E. Johnson, T. J. Pollard, L. Shen, L. W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R. G. Mark, MIMIC-III, a freely accessible critical care database, Scientific Data 3 (1) (2016) 1–9. doi:10.1038/sdata.2016.35.

[42] D. Malmgren, textract Documentation Release 1.1.0 (2014).
URL https://textract.readthedocs.io/en/stable/

[43] W. G. Cochran, Sampling Techniques, 3rd Edition., John Wiley, 1977.

[44] J. L. Fleiss, J. Cohen, The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability, Educational and Psychological Measurement 33 (1973) 613–619. doi:10.1177/001316447303300309.

[45] spaCy, · Industrial-strength Natural Language Processing in Python ([online]).
URL https://spacy.io/

[46] H. Naseri, textractor; tools for pain scoring. (2021). doi:DOI:10.5281/zenodo.4649625.

[47] Intro to data structures - pandas 1.0.5 documentation ([online]).

[48] J. Pennington, R. Socher, C. D. Manning, GloVe: Global Vectors for Word Representation ([online]).
URL https://nlp.stanford.edu/pubs/glove.pdf

[49] M. S. Wallace, J. North, E. J. Grigsby, L. Kapural, M. R. Sanapati, S. G. Smith, C. Willoughby, P. J. McIntyre, S. P. Cohen, R. M. Rosenthal, S. Ahmed, R. Vallejo, F. M. Ahadian, T. L. Yearwood, A. W. Burton, E. J. Frankoski, J. Shetake, S. Lin, B. Hershey, B. Rogers, N. Mekel-Bobrov, An integrated quantitative index for measuring

chronic multisite pain: The multiple areas of pain (map) study, Pain Medicine (United States) 19 (2018) 1425–1435. doi:10.1093/pm/pnx325.
URL https://pubmed.ncbi.nlm.nih.gov/29474648/

[50] S. D. Rundell, K. V. Patel, M. A. Krook, P. J. Heagerty, P. Suri, J. L. Friedly, J. A. Turner, R. A. Deyo, Z. Bauer, D. R. Nerenz, A. L. Avins, S. S. Nedeljkovic, J. G. Jarvik, Multi-site pain is associated with long-term patient-reported outcomes in older adults with persistent back pain, Pain Medicine (United States) 20 (2019) 1898–1906. doi:10.1093/pm/pny270.
URL https://pubmed.ncbi.nlm.nih.gov/30615144/

[51] M. P. Jensen, C. Tomé-Pires, E. Solé, M. Racine, E. Castarlenas, R. de la Vega, J. Miró, Assessment of pain intensity in clinical trials: Individual ratings vs composite scores, Pain Medicine (United States) 16 (2015) 141–148. doi:10.1111/pme.12588.
URL https://pubmed.ncbi.nlm.nih.gov/25280226/

[52] A. Tharwat, Classification assessment methods, Applied Computing and Informatics 17 (1) (2018) 168–192. doi:10.1016/j.aci.2018.08.003.

[53] M. Kuchuk, C. L. Addison, M. Clemons, I. Kuchuk, P. Wheatley-Price, Incidence and consequences of bone metastases in lung cancer patients, Journal of Bone Oncology 2 (1) (2013) 22–29. doi:10.1016/j.jbo.2012.12.004.

[54] A. Tsuya, T. Kurata, K. Tamura, M. Fukuoka, Skeletal metastases in non-small cell lung cancer: A retrospective study, Lung Cancer 57 (2007) 229–232. doi:10.1016/j.lungcan.2007.03.013.

[55] M. Kuchuk, C. L. Addison, M. Clemons, I. Kuchuk, P. Wheatley-Price, Incidence and consequences of bone metastases in lung cancer patients, Journal of bone oncology 2 (1) (2013) 22–29. doi:10.1016/j.jbo.2012.12.004.

3.9 Appendix: Supplementary Information

3.9.1 i2b2 discharge summaries as the training and test corpora:

The publicly-available de-identified i2b2 discharge summaries were used for primary training and testing of our NLP algorithm. The selected corpora, which included 1,249 discharge summary records, were collected from ‘Partners Healthcare hospitals and physicians network’ [1]. Unfortunately, there was no structured information about the diversity of the providers of the notes. Each record was a de-identified text file (.txt format) [2]. Clinical notes varied between 600 to 23,000 characters in length, with an average of 6886 characters per note. The minimum, maximum, mean, median, and standard deviation of the distributions of the number of characters and words in the corpora are presented in Table 3.15. The normalized distribution of the number of characters per note is plotted in Figure 3.1. As it can be seen in the sample note that presented in Table 3.13, each note included ‘Admission Date’ and ‘Discharge date’, and was organized into several sections with a heading for each. Headings were all written in capital letters, ending with a colon. Sections and section headings were not consistent in the entire corpora. This meant that not only was it possible for each note to have a different set of sections, but also, it was possible for section headings to change from one note to another. For example, a section including the results of the physical examination was labeled as ‘PHYSICAL EXAMINATION:’ in one note, and ‘PHYSICAL EXAMINATION ON ADMISSION:’ in another note. Table 3.16 includes the list of section headings for a sample note. To ensure the generalizability of our algorithm (to be a database-independent tool), we excluded heading information, and used only the descriptive text information for the analysis with our NLP pipeline. From the 1,249 notes, we randomly-selected 150 notes for validation and testing and used the remaining 1099 notes for training our NLP algorithm in different iterations of its development.

RECORD #000000 123456789 — ABCDEFG — 123456789 — — 123456789 — 1/1/2500 00:00:00 AM — Discharge Summary — Signed — ABC —
Admission Date: 1/1/2500
Discharge Date: 1/1/2500
HISTORY OF PRESENT ILLNESS: The patient is [...] He had no chest pain but did have diaphoresis and mild nausea and vomiting as well as lightheadedness and some palpitations lasting approximately one hour in duration. On the day of admission after taking a shower in the morning , he had increasing shortness of breath gradually at rest with epigastric tightness without radiation but he did have nausea , vomiting and diaphoresis. [...]
HOSPITAL COURSE: The patient developed left arm pain with inflation and slow flow after PTCA [...]. The patient was treated with nitroglycerin , heparin and aspirin.
DISPOSITION: The patient was discharged to home in stable condition.
MEDICATIONS: On discharge included aspirin , one po q day; [...].
The patient will follow-up with Dr. XX .
Dictated By: XX D. YY , M.D. ABC123 Attending: A B. CDEFGE , M.D. QAA DD000/0000
Batch: 0000 Index No. ABCABCABC D: 30/20/10 T: 1/2/3
[report_end]

Table 3.13: A sample de-identified hospital discharge summary from i2b2 database.

```
Processing 00000000.tx.01: He had no chest pain but did have diaphoresis and mild nausea and vomiting as well as lightheadedness and some palpitations lasting approximately one hour in duration.
Phrase: He had
Phrase: no chest pain
Meta Mapping (1000): 1000 N C0008031:Chest Pain [Sign or Symptom]
Meta Mapping (1000): 1000 C2926613:Chest pain (Chest pain:Finding:Point in time:Patient:Ordinal) [Clinical Attribute]
Phrase: but did have
Phrase: diaphoresis
Meta Mapping (1000): 1000 C0700590:Diaphoresis (Increased sweating) [Sign or Symptom]
Meta Mapping (1000): 1000 C0038990:Diaphoresis (Sweating) [Finding]
Phrase: mild nausea
Meta Mapping (888): 694 C2945599:Mild (Mild (qualifier value)) [Qualitative Concept]
Meta Mapping (888): 861 C0027497:Nausea [Sign or Symptom]
Phrase: vomiting
Meta Mapping (1000): 1000 C0042963:Vomiting [Sign or Symptom]
Meta Mapping (1000): 1000 C1963281:Vomiting (Vomiting Adverse Event) [Finding]
Phrase: as well as
Phrase: lightheadedness
Meta Mapping (1000): 1000 C0220870:Light-Headedness (Lightheadedness) [Sign or Symptom]
[...]
Processing 00000000.tx.2: On the day of admission after taking a shower in the morning , he had increasing shortness of breath gradually at rest with epigastric tightness without radiation but he did have nausea , vomiting and diaphoresis.
[...]
```

Table 3.14: An example of the annotated text file exported using MetaMap. Each detected medical concept associated with a confidence score, negation status, UMLS id, description and a UMLS semantic type.

Number of characters in the notes.

	minimum	maximum	mean	median	standard deviation
ARIA	1822	12132	7382	7344	1992
i2b2	318	25875	6882	6414	2984
MIMIC-III	54	55728	9619	8878	5540

Number of words in the notes.

	minimum	maximum	mean	median	standard dev
ARIA	266	1381	719	705	207
i2b2	16	4283	1168	1082	510
MIMIC-III	9	7980	1435	1328	828

Table 3.15: The minimum, maximum, mean, median, and standard deviation of the distributions of the number of characters and words, in the notes from the ARIA, i2b2 and MIMIC-III data sets.

Section Heading	
1- RECORD #:	2- ADMISSION DIAGNOSIS:
3- ALLERGIES:	4- DISCHARGE DIAGNOSES:
5- DISPOSITION:	6- HISTORY OF PRESENT ILLNESS
7- HOSPITAL COURSE:	8- MEDICATIONS ON DISCHARGE:
9- LABORATORY DATA:	10- MEDICATIONS ON ADMISSION:
11- PAST MEDICAL HISTORY:	12- PHYSICAL EXAMINATION:

Table 3.16: Each note in the i2b2 database discharge summary was structured into several sections with the following headings. Each heading started with a new line and ended with a colon as shown here.

3.9.2 MIMIC-III clinical documents as the validation data set:

The publicly accessible de-identified health data from MIMIC-III database were used to validate our NLP algorithm [3]. MIMIC-III included a total of 2,083,180 clinical text documents in various categories, as specified in the first line of each document. Table 3.17 shows the number of notes in each category. We found 59,652 notes that were categorized as‘discharge summary’ and only these were used in our study. The normalized distribution of the number of characters per discharge summary is plotted in Figure 3.1, overlaid on the distribution observed in the i2b2 database.

For the purpose of this research, the MIMIC-III database was only used for validation of our NLP algorithm. Therefore, we randomly-selected 30 notes to test the accuracy of the pain detection algorithm. Similar to the i2b2 corpus, MIMIC-III notes were also structured into sections with varying headings, and we used the entire contents of each note in the NLP algorithm for pain extraction.

Category	# of notes	Category	# of notes
1- Nursing/other	822,497	9- Nutrition	9,418
2- Radiology	522,279	10- General	8,301
3- Nursing	223,556	11- Rehab Services	5,431
4- ECG	209,051	12- Social Work	2,670
5- Physician	141,624	13- Case Management	967
6- Discharge summary	59,652	14- Pharmacy	103
7- Echo	45,794	15- Consult	98
8- Respiratory	31,739	**Total**	**2,083,180**

Table 3.17: Number of notes in each category in the MIMIC-III database which included 2,083,180 clinical text documents, arranged into 15 categories.

3.9.3 Institutional radiation oncology consultation notes as the metastatic cancer corpora for cancer pain study:

From the VARIAN Radiation oncology (Varian Medical Systems, Palo Alto, CA) ARIA database at our institution, we searched for the patients that received palliative radiotherapy for a secondary malignant neoplasm of bone between January 2016 and September 2019. From the total of 462 patients who fall within the search criteria we extracted a total of 788 Microsoft word document named as consultation notes. The plain text was extracted from Microsoft Word (.doc) documents using the Python textract package [4].

The normalized distribution of the number of characters per consultation note is displayed in Figure 3.1 along with the i2b2 and MIMIC-III data sets.

We processed notes from all three corpora and found 358 unique pain-related UMLS concepts. These pain concepts are summarized in Table 3.19 together with the frequency of the mentions of each concept.

3.9.4 Determining test set size

By auditing our training corpora, we found that pain is indicated in a 5% (p=0.05) of the sentences. We used Cochran's sample size formula [5] to determine the minimum sample size required to evaluate the study. To ensure that the pain-score detection is precise within a 95% confidence level ($Z_{1-\alpha/2} = 1.96$), and a 1% margin of error (e=0.01), the minimum sample size was determined as,

$$N = p(1-p)\left(\frac{Z_{1-\alpha/2}}{e}\right)^2 = 0.05 \times 0.95 \times \left(\frac{1.96}{0.01}\right)^2 = 1824. \tag{3.2}$$

Therefore, we analyzed about 2000 sentences in each of our validation and the test sets satisfied the minimum sample size requirement.

150 notes ($N = 150$) from the ARIA dataset were used to test our VDP classification method. Cochran's formula was used to determine the confidence level for the selected sample size. According to Table 3.8, probabilities of finding a note in the ARIA corpus with the 'pain' VDP was 64.9% ($p_{pain} = 0.65$) and with the 'no pain' VDP was 13.2% ($p_{no\ pain} = 0.13$). Allowing a 95% confidence interval ($Z_{1-\alpha/2} = 1.96$), the standard error was calculated as:

$$SE = \sqrt{\frac{p(1-p)}{N}} = 0.05 \times \sqrt{\frac{0.13(1-0.13)}{150}} = 0.026 \tag{3.3}$$

This sample size gave us a standard error of 0.027.

3.9.5 Three-label confusion matrix

The confusion matrix for our three-label pain classifier is a 3 × 3 matrix, as presented in Table 3.4. This matrix includes three TRUE labels for correctly assigned labels, and six FALSE labels for incorrectly assigned labels. The TRUE assignments are: T_{PP} (A phrase was correctly labeled as score 1 pain), T_{NN} (A phrase was correctly labeled as score 0 pain), and T_{II} (A phrase was correctly labeled as irreverent pain). The six FALSE labels are as follows: F_{NP} (Score 0 pain was labeled as Score 1 pain), F_{PN} (Score 1 pain was labeled as Score 0 pain), F_{PI} (Score 1 pain was labeled as irrelevant pain), F_{NI} (Score 0 pain was labeled as irrelevant pain), F_{IP} (irrelevant pain was labeled as Score 1 pain), and F_{IN} (irrelevant pain was labeled as Score 0 pain).

We evaluated the performance of our NLP algorithm at detecting pain by calculating the precision, recall, and F1-score (F1) from the confusion matrices [6]. Precision is the measure of how well our model labeled the score 0, score 1, and irrelevant pain terms. Precision is defined as the number of occasions that a given label was assigned correctly correctly divided by the total number of assigned labels. The recall (also known as the sensitivity) is the number of the correctly labeled pain terms divided by the total number of true cases. Finally, $F1\ Score$ is calculated for each class as the weighted average of precision and recall.

The performance of our NLP algorithm in assigning VDP was evaluated using a manual audit of 150 randomly-selected consultation notes from the ARIA corpus that were reserved for this purpose (Figure 3.2). Inter-annotator agreement calculated based on the 30 notes that scored by all 6 annotators as provided in table 3.18. Similar to what we presented in Table 3.4, a 3 × 3 confusion matrix was constructed for VDP, including ‘no mention of pain’, ‘no pain’, and ‘pain’.

Note ID	No mention on pain	No pain		Pain		NLP exported API
			Mild	Moderate	Severe	
1	0	0	0	1	5	0.82
2	0	1	0	1	4	0.71
3	0	0	1	1	4	1.00
4	0	3	1	1	1	0.67
5	0	0	0	0	6	1.00
6	0	0	0	2	4	0.71
7	1	5	0	0	0	-1.00
8	0	0	1	5	0	1.00
9	1	5	0	0	0	-1.00
10	0	0	2	1	3	0.50
11	0	0	0	4	2	1.00
12	0	6	0	0	0	-1.00
13	0	0	1	5	0	-0.33
14	0	0	1	5	0	1.00
15	0	0	1	5	0	1.00
16	0	0	0	4	2	1.00
17	0	6	0	0	0	-1.00
18	0	0	3	2	1	0.75
19	0	0	0	3	3	1.00
20	0	4	2	0	0	0.25
21	0	0	2	4	0	1.00
22	0	0	2	4	0	0.75
23	0	0	6	0	0	1.00
24	0	0	2	4	0	-0.14
25	0	1	4	1	0	0.00
26	5	1	0	0	0	
27	0	0	1	2	3	0.71
28	0	0	0	5	1	0.33
29	0	1	4	1	0	1.00
30	0	6	0	0	0	-1.00

Table 3.18: Inter-annotator agreement between six annotators that labeled verbal pain score in 30 notes. The labels are presented in the columns, while the notes are presented in the rows. Each cell lists the number of annotators who assigned the indicated label (column) note to the indicated note (row). We defined VDP by grouping 'mild', 'moderate' and 'severe' pain scales as 'pain'. The NLP exported average pain intensity (API) is presented in the last column. API was calculated using Eq.3.1. As explained in section 3.4.3, API was used in our VDP classification method as; VDP = 'no pain' (if API < 0), and VDP = 'pain' (if API ≥ 0).

Table 3.19: List of 358 unique pain concepts that were extracted from all the notes from three corpora using MetaMap. The frequency of mentions of each concept is provided in the table.

	Pain concept	i2b2 Frequency (%) (n= 40,787)	MIMIC-III Frequency (%) (n= 13,300)	ARIA Frequency (%) (n= 20,377)
1	abdominal and pelvic pain	0.00	0.01	0.00
2	abdominal angina	0.13	0.02	0.00
3	abdominal cramps	0.17	0.02	0.01
4	abdominal discomfort	0.18	0.17	0.09
5	abdominal pain	2.45	7.85	0.73
6	abdominal pain through to back	0.01	0.00	0.00
7	abdominal tenderness	0.04	0.12	0.08
8	abdominal wind pain	0.01	0.02	0.00
9	absence of pain sensation	0.05	0.07	0.00
10	ache	0.28	0.06	0.37
11	acromioclavicular joint pain	0.00	0.00	0.03
12	acute abdominal pain	0.02	0.03	0.00
13	acute back pain	0.00	0.00	0.10
14	acute chest pain	0.04	0.03	0.00
15	acute headache	0.00	0.00	0.04
16	acute low back pain	0.00	0.00	0.01
17	acute onset pain	0.23	0.17	1.25
18	acute thoracic back pain	0.02	0.00	0.00
19	after pains	0.03	0.18	0.03
20	anal pain	0.00	0.00	0.01
21	angina equivalent	0.32	0.05	0.00
22	angina pectoris	5.50	0.78	0.00
23	angina symptom	0.46	0.13	0.00
24	ankle pain	0.12	0.18	0.04
25	anterior chest wall pain	0.04	0.00	0.17
26	anterior pleuritic pain	0.00	0.00	0.21
27	arm discomfort	0.01	0.06	0.04
28	arm pain	0.13	0.16	0.12
29	arthralgia	0.87	0.85	0.03
30	arthritic pains	0.03	0.01	0.00
31	arthritis pain	0.00	0.00	0.03
32	atypical chest pain	1.60	0.12	0.00
33	back discomfort	0.00	0.00	0.07
34	back pain	0.91	0.95	5.98
35	back pain mid back	0.00	0.05	1.12
36	back pain with radiation	0.01	0.00	0.10
37	back pain, severe	0.03	0.00	0.00
38	back tenderness	0.00	0.03	0.12
39	bilateral headache	0.01	0.02	0.00
40	bladder pain	0.01	0.02	0.00
41	bodily pain	0.00	0.00	0.04
42	body ache	0.00	0.00	0.04
43	body pain	0.13	0.02	0.00
44	bone pain	0.02	0.02	1.36
45	bone tenderness	0.02	0.00	0.43
46	breakthrough pain	0.14	0.23	0.00
47	breast discomfort	0.00	0.00	0.03
48	breast tenderness	0.03	0.00	0.03
49	burning epigastric pain	0.01	0.02	0.00
50	burning feet	0.02	0.00	0.00
51	burning sensation	0.40	0.33	0.41
52	burning sensation of skin	0.00	0.00	0.01
53	bursal pain	0.01	0.00	0.00
54	cachexia	0.11	0.60	0.31
55	calf tenderness	0.41	0.09	0.00
56	cancer pain	0.00	0.00	0.83
57	cardiac pain	0.16	0.00	0.00
58	catch - finding of sensory dimension of pain	0.07	0.10	0.04
59	central pain	0.00	0.02	0.00
60	chest burning	0.16	0.99	0.00
61	chest burning pain of	0.04	0.21	0.00
62	chest discomfort	1.06	0.59	0.07
63	chest pain	17.88	9.25	2.06
64	chest pain angina	0.08	0.03	0.00
65	chest pain at rest	0.14	0.00	0.00
66	chest pain on breathing	0.00	0.12	0.00
67	chest pain on exertion	0.09	0.14	0.00
68	chest pain, sharp	0.06	0.02	0.00
69	chest pressure	1.27	0.50	0.00
70	chest tightness	0.77	0.09	0.00
71	chest tightness or pressure	0.01	0.00	0.00
72	chest wall pain	0.07	0.03	0.22
73	chronic abdominal pain	0.11	0.72	0.00
74	chronic back pain	0.08	0.48	0.47
75	chronic chest pain	0.01	0.02	0.00
76	chronic pain	0.00	0.00	0.18
77	chronic pelvic pain of female	0.03	0.03	0.00
78	clavicle pain	0.00	0.00	0.04
79	cramping sensation quality	0.07	0.06	0.01
80	crushing chest pain	0.02	0.01	0.00
81	deep pain	0.01	0.01	0.01
82	deltoid pain	0.01	0.00	0.00
83	diffuse abdominal pain	0.05	0.29	0.00
84	diffuse pain	0.06	0.13	0.62
85	discomfort	0.00	0.00	3.57
86	discomfort rectal	0.00	0.08	0.00
87	dull chest pain	0.02	0.00	0.00
88	dull pain	0.01	0.00	0.18
89	dysuria	0.80	1.56	0.20
90	ear tenderness	0.00	0.02	0.00
91	earache	0.05	0.46	0.05
92	epigastric burning	0.06	0.01	0.00
93	epigastric discomfort	0.14	0.00	0.03
94	epigastric pain	1.11	0.63	0.05
95	epigastric tenderness	0.05	0.07	0.00
96	esophageal chest pain	0.01	0.00	0.00
97	excruciating pain	0.01	0.00	0.04
98	exercise-induced angina	0.47	0.30	0.00
99	eye pain	0.04	0.08	0.08
100	eye pain, severe	0.00	0.01	0.00
101	facial pain	0.01	0.11	0.04
102	flank pain	0.14	0.14	0.00

Continued on next page

Table 3.19 – *Continued from previous page*

	Pain concept	i2b2	MIMIC-III	ARIA
		Frequency (%)	Frequency (%)	Frequency (%)
		(n= 40,787)	(n= 13,300)	(n= 20,377)
103	foot pain	0.42	0.51	0.03
104	frequent headaches	0.01	0.01	0.00
105	frontal headache	0.02	0.02	0.04
106	gastrointestinal pain	0.17	0.00	0.00
107	generalized abdominal pain	0.00	0.02	0.00
108	generalized aches and pains	0.07	0.02	0.03
109	generalized chest pain	0.00	0.00	0.01
110	great toe pain	0.01	0.00	0.00
111	groin discomfort	0.01	0.00	0.01
112	groin tenderness	0.00	0.00	0.01
113	hand pain	0.13	0.00	0.00
114	head pressure	0.00	0.00	0.01
115	head pressure sensation	0.02	0.00	0.00
116	headache	3.40	2.90	2.19
117	headache associated with sexual activity	0.01	0.00	0.00
118	headache fullness	0.01	0.00	0.00
119	headache persistent	0.01	0.01	0.00
120	headache recurrent	0.01	0.01	0.01
121	headache severe	0.04	0.21	0.05
122	headache worsening	0.00	0.00	0.03
123	heel pain	0.02	0.00	0.00
124	hernia pain	0.01	0.00	0.00
125	hip joint pain	0.00	0.00	1.71
126	hip pain	0.56	0.23	1.65
127	incisional pain	0.06	0.25	0.04
128	inguinal pain	0.19	0.03	0.59
129	injection site pain	0.02	0.00	0.00
130	intermittent abdominal pain	0.01	0.05	0.00
131	intermittent headache	0.00	0.02	0.01
132	intermittent pain	0.35	0.11	0.14
133	intestinal pain	0.01	0.00	0.03
134	ischemic pain	0.01	0.00	0.00
135	ischial tuberosity tenderness	0.01	0.00	0.00
136	jaw pain	0.25	0.02	0.08
137	joint tenderness	0.05	0.00	0.09
138	knee pain	0.94	0.39	0.03
139	left flank pain	0.01	0.00	0.03
140	left lower quadrant pain	0.06	0.15	0.25
141	left sided abdominal pain	0.00	0.02	0.00
142	left sided chest pain	0.25	0.09	0.10
143	left upper quadrant pain	0.02	0.05	0.00
144	leg discomfort	0.00	0.04	0.00
145	liver tender	0.00	0.00	0.01
146	localized pain	0.00	0.02	0.04
147	low back pain	0.45	0.31	1.90
148	lower abdominal pain	0.02	0.03	0.00
149	lower extremity pain	0.25	0.06	0.05
150	lower ribs pain	0.00	0.06	0.00
151	lumbo-sacral pain	0.00	0.00	0.03
152	malaise	0.68	0.87	5.79
153	mandibular pain	0.00	0.00	0.04
154	mastodynia	5.88	2.35	0.12
155	mastodynia of bilateral breasts	0.00	0.00	0.01
156	mastodynia of left breast	0.00	0.00	0.01
157	mastodynia of right breast	0.00	0.00	0.03
158	mechanical pain	0.00	0.32	0.14
159	metastatic bone pain	0.00	0.00	0.49
160	mild pain	0.43	0.37	0.00
161	miserable pain	0.01	0.00	0.00
162	moderate pain	0.01	0.04	0.13
163	morning headache	0.01	0.00	0.00
164	muscle cramp	0.12	0.11	0.09
165	muscle cramps in leg	0.03	0.02	0.00
166	muscle cramps in the calf	0.00	0.01	0.00
167	muscle tenderness	0.01	0.00	0.00
168	musculoskeletal chest pain	0.03	0.00	0.00
169	myalgia	0.38	2.79	0.10
170	nausea or abdominal pain	0.01	0.02	0.00
171	neck cramps	0.00	0.03	0.00
172	neck discomfort	0.02	0.01	0.03
173	neck pain	0.19	0.25	0.60
174	neck tightness	0.01	0.00	0.00
175	nerve pain	0.00	0.02	0.05
176	neuralgia	0.12	0.27	0.34
177	neurological pain	0.00	0.00	0.05
178	night pain	0.01	0.19	0.03
179	non-cardiac chest pain	0.48	0.01	0.00
180	nonspecific abdominal pain	0.01	0.00	0.00
181	occipital headache	0.02	0.02	0.05
182	oral pain	0.11	1.24	0.00
183	other chest pain	0.02	0.00	0.00
184	pain	24.21	20.70	40.00
185	pain aggravated	0.24	0.10	0.55
186	pain and tenderness	0.32	0.02	0.00
187	pain around eye	0.00	0.00	0.04
188	pain characterized by provoking factor	0.00	0.00	0.01
189	pain during injection	0.00	0.13	0.00
190	pain from metastases	0.00	0.02	0.08
191	pain in axilla	0.01	0.00	0.05
192	pain in body part	0.00	0.00	0.01
193	pain in buttock	0.02	0.02	0.14
194	pain in calf	0.21	0.18	0.00
195	pain in cervical spine	0.00	0.00	0.09
196	pain in cheek	0.00	0.02	0.00
197	pain in elbow	0.07	0.02	0.00
198	pain in esophagus (finding	0.01	0.01	0.00
199	pain in femur	0.00	0.00	0.05
200	pain in finger	0.14	0.00	0.00
201	pain in forearm	0.01	0.00	0.00
202	pain in left arm	0.35	0.23	0.07
203	pain in left foot	0.02	0.00	0.00

Continued on next page

Table 3.19 – *Continued from previous page*

	Pain concept	i2b2 Frequency (%) (n= 40,787)	MIMIC-III Frequency (%) (n= 13,300)	ARIA Frequency (%) (n= 20,377)
204	pain in left hand	0.01	0.00	0.00
205	pain in left hip	0.01	0.00	0.00
206	pain in left knee	0.01	0.00	0.00
207	pain in left leg	0.15	0.26	0.00
208	pain in left lower limb	0.00	0.00	0.54
209	pain in left shoulder	0.07	0.05	0.00
210	pain in limb	0.13	0.18	0.00
211	pain in limb, lower leg	0.01	0.03	0.21
212	pain in lower limb	2.00	0.43	0.91
213	pain in lumbar spine	0.00	0.03	0.08
214	pain in right arm	0.04	0.10	0.18
215	pain in right elbow	0.00	0.00	0.01
216	pain in right hand	0.01	0.00	0.00
217	pain in right hip	0.40	0.00	0.00
218	pain in right hip joint	0.00	0.00	1.74
219	pain in right knee	0.07	0.00	0.00
220	pain in right leg	0.37	0.02	0.00
221	pain in right lower limb	0.00	0.00	0.49
222	pain in right lower limb nos	0.00	0.00	0.08
223	pain in right shoulder	0.14	0.18	0.00
224	pain in scrotum	0.01	0.00	0.00
225	pain in spine	0.01	0.00	1.77
226	pain in the coccyx	0.00	0.00	0.03
227	pain in thoracic spine	0.00	0.00	1.57
228	pain in thumb	0.02	0.00	0.00
229	pain in toe	0.01	0.03	0.00
230	pain in wrist	0.06	0.02	0.00
231	pain lower ribs	0.00	0.00	0.04
232	pain neck/shoulder	0.00	0.05	0.00
233	pain of digit	0.11	0.00	0.00
234	pain of ear structure	0.00	0.00	0.01
235	pain of front foot of quadraped	0.01	0.00	0.00
236	pain of left elbow joint	0.00	0.00	0.10
237	pain of left hip joint	0.00	0.00	0.12
238	pain of left shoulder joint	0.00	0.00	0.08
239	pain of left thigh	0.00	0.00	0.20
240	pain of oral cavity structure	0.08	1.24	0.00
241	pain of right arm only	0.04	0.02	0.00
242	pain of right forearm	0.00	0.00	0.01
243	pain of right shoulder joint	0.00	0.00	0.68
244	pain of right thigh	0.00	0.00	0.09
245	pain of skin	0.01	0.00	0.00
246	pain radiating to jaw	0.04	0.00	0.00
247	pain radiating to left arm	0.12	0.00	0.00
248	pain radiating to left leg	0.00	0.00	0.03
249	pain radiating to left shoulder	0.01	0.00	0.00
250	pain radiating to neck	0.02	0.00	0.00
251	pain radiating to right arm	0.03	0.00	0.00
252	pain radiating to right leg	0.01	0.00	0.00
253	pain radiating to right shoulder	0.01	0.00	0.00
254	pain uncontrolled	0.04	0.00	0.00
255	pain with eating	0.01	0.02	0.03
256	pain, burning	0.15	0.00	0.03
257	pain, intractable	0.00	0.03	0.00
258	pain, migratory	0.00	0.00	0.03
259	pain, postoperative	0.27	0.04	0.00
260	pain, referred	0.00	0.02	0.51
261	painful paresthesias	0.01	0.00	0.00
262	painless hematuria	0.00	0.01	0.00
263	pelvic pain	0.05	0.06	1.71
264	pelvic pain female	0.04	0.05	0.77
265	peripheral neuropathic pain	0.00	0.01	0.00
266	persistent mastalgia	0.02	0.02	0.00
267	pleuritic pain	0.23	0.21	0.07
268	posterior cervical pain	0.01	0.05	0.08
269	post-procedural pain	0.01	0.10	0.00
270	precordial pain	0.02	0.00	0.00
271	pubic pain	0.06	0.00	0.31
272	radiating back pain	0.00	0.00	0.26
273	radiating chest pain	0.06	0.13	0.00
274	radiating pain	0.06	0.02	0.71
275	radicular pain	0.01	0.06	0.14
276	rebound tenderness	0.06	0.44	0.01
277	rectal pain	0.00	0.21	0.01
278	recurrent abdominal pain	0.02	0.01	0.00
279	recurrent chest pains	0.00	0.00	0.01
280	recurrent low back pain	0.00	0.00	0.03
281	renal angle tenderness	0.32	0.39	0.00
282	renal pain	0.02	0.00	0.00
283	rest pain	0.41	0.08	0.01
284	retrosternal pain	2.34	0.49	0.00
285	rib pain	0.04	0.12	0.58
286	rib tenderness, lower	0.00	0.00	0.03
287	right flank pain	0.00	0.09	0.08
288	right lower quadrant pain	0.22	0.05	0.03
289	right sided abdominal pain	0.01	0.21	0.04
290	right sided chest pain	0.13	0.09	0.24
291	right upper quadrant abdominal tenderness	0.02	0.00	0.00
292	right upper quadrant pain	0.26	0.35	0.41
293	sacral pain	0.01	0.00	0.25
294	sacroiliac pain	0.00	0.00	0.12
295	scalding pain on urination	0.02	0.03	0.00
296	scalp tenderness	0.02	0.00	0.00
297	scapulalgia	0.17	0.00	0.46
298	sciatic nerve pain	0.00	0.02	0.00
299	sciatica	0.75	0.10	0.31
300	sciatica, bilateral	0.00	0.00	0.01
301	scrotal tenderness	0.01	0.00	0.00
302	sensory discomfort	0.68	0.87	0.00
303	severe back pain	0.00	0.00	0.54
304	severe low backache	0.00	0.00	0.03
305	severe pain	0.62	0.69	0.00

Continued on next page

Table 3.19 – *Continued from previous page*

	Pain concept	i2b2	MIMIC-III	ARIA
		Frequency (%)	Frequency (%)	Frequency (%)
		(n= 40,787)	(n= 13,300)	(n= 20,377)
306	sharp chest pain	0.00	0.00	0.01
307	sharp headache	0.00	0.01	0.00
308	sharp pain	0.11	0.06	0.12
309	shooting pain	0.00	0.10	0.03
310	shoulder discomfort	0.02	0.05	0.12
311	shoulder pain	0.62	0.88	0.24
312	shoulder tenderness	0.03	0.00	0.00
313	side pain	0.03	0.01	0.10
314	sinus headache	0.01	0.03	0.00
315	sinus pain	0.02	0.02	0.00
316	sinus pressure	0.01	0.03	0.00
317	skin tenderness	0.00	0.00	0.03
318	sore mouth	0.00	0.01	0.00
319	sore skin	0.04	0.02	0.01
320	sore throat	0.35	0.56	0.04
321	sore to touch	4.44	22.95	0.01
322	spleen pain	0.00	0.00	0.01
323	stabbing pain	0.03	0.06	0.01
324	stomach ache	0.01	0.02	0.00
325	subcostal pain	0.01	0.00	0.00
326	subtalar joint pain	0.02	0.00	0.00
327	superficial pain	0.01	0.19	0.00
328	suprapubic pain	0.00	0.07	0.07
329	swallowing painful	0.06	0.07	0.18
330	tender mouth	0.00	0.00	0.03
331	tenderness of gums	0.04	0.00	0.00
332	tenderness of tendon	0.01	0.00	0.00
333	tenderness of upper limb	0.00	0.02	0.00
334	thigh pain	0.06	0.04	0.26
335	thigh pain anterior	0.01	0.00	0.31
336	thoracic back pain	0.01	0.01	0.00
337	throbbing headache	0.00	0.01	0.00
338	throbbing pain	0.01	0.14	0.00
339	tibia pain	0.00	0.00	0.05
340	tightness in arm	0.03	0.00	0.00
341	tightness sensation	0.00	0.00	0.01
342	toothache	0.06	0.05	0.00
343	total body pain syndrome	0.02	0.01	0.00
344	transplant pain	0.04	0.00	0.00
345	typical angina	0.42	0.01	0.00
346	umbilical pain	0.00	0.01	0.00
347	unbearable pain	0.00	0.00	0.01
348	uncontrolled pain	0.00	0.00	0.16
349	upper abdominal pain	0.01	0.05	0.00
350	upper back pain	0.01	0.03	0.51
351	upper chest pain	0.03	0.05	0.16
352	upset stomach	0.04	0.00	0.01
353	vascular pain	0.01	0.00	0.00
354	vertex headache	0.00	0.01	0.00
355	very mild pain	0.00	0.00	0.03
356	walking pain	0.03	0.00	0.03
357	wound pain	0.03	0.00	0.00
358	wound tenderness	0.00	0.06	0.00

Continued on next page

Bibliography

[1] O. Uzuner, Y. Luo, P. Szolovits, Evaluating the State-of-the-Art in Automatic De-identification, Journal of the American Medical Informatics Association 14 (5) (2007) 550–563. doi:10.1197/jamia.M2444.

[2] O. Uzuner, T. C. Sibanda, Y. Luo, P. Szolovits, A de-identifier for medical discharge summaries, Artificial Intelligence in Medicine 42 (1) (2008) 13–35. doi:10.1016/j.artmed.2007.10.001.

[3] A. E. Johnson, T. J. Pollard, L. Shen, L. W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R. G. Mark, MIMIC-III, a freely accessible critical care database, Scientific Data 3 (1) (2016) 1–9. doi:10.1038/sdata.2016.35.

[4] D. Malmgren, textract Documentation Release 1.1.0 (2014).
URL https://textract.readthedocs.io/en/stable/

[5] W. G. Cochran, Sampling Techniques, 3rd Edition., John Wiley, 1977.

[6] A. Tharwat, Classification assessment methods, Applied Computing and Informatics 17 (1) (2018) 168–192. doi:10.1016/j.aci.2018.08.003.

Chapter 4

Radiomics-based machine learning models to distinguish between metastatic and healthy bone using lesion-center-based geometric regions of interest

Hossein Naseri, Sonia Skamene, Marwan Tolba, Mame Daro Faye, Paul Ramia, Julia Khriguian, Haley Patrick, Aixa X. Andrade H., Marc David and John Kildea

Article published in: *Sci Rep.* **12**, 9866 (2022). https://doi.org/10.1038/s41598-022-13379-8

4.1 Preface

This chapter describes how we met the second thesis objective: Construction of a radiomics pipeline to extract BM lesion features from radiographic images. In this study, a radiomics pipeline was built to extract BM lesion features from patients' radiography images. To enable rapid labeling of BM lesions, a novel lesion-center-based geometric ROI methodology was introduced. Then, utilizing lesion-center-based geometric ROIs, a pipeline was developed to process simulation-CT images and obtain radiomics imaging phenotypes. The pipeline was tested for its ability to differentiate between healthy and metastatic bone lesions.

4.2 Abstract

Radiomics-based Machine Learning (ML) classifiers have shown potential for detecting Bone Metastases (BM) and for evaluating BM response to radiotherapy (RT). However, current radiomics models require large datasets of images with expert-segmented 3D Regions Of Interest (ROIs). Full ROI segmentation is time consuming and oncologists often outline just RT treatment fields in clinical practice. This presents a challenge for real-world radiomics research. As such, a method that simplifies BM identification but does not compromise the power of radiomics is needed.

The objective of this study was to investigate the feasibility of radiomics models for BM detection using lesion-center-based geometric ROIs. The planning-CT images of 170 patients with non-metastatic lung cancer and 189 patients with spinal BM were used. The point locations of 631 BM and 674 healthy bone (HB) regions were identified by experts. ROIs with various geometric shapes were centered and automatically delineated on the identified locations, and 107 radiomics features were extracted. Various feature selection methods and ML classifiers were evaluated.

Our point-based radiomics pipeline was successful in differentiating BM from HB. Lesion-center-based segmentation approach greatly simplifies the process of preparing images for use in radiomics studies and avoids the bottleneck of full ROIs segmentation.

4.3 Introduction

In recent years, radiomics-based Machine Learning (ML) classifiers have shown great potential for use in the early detection of Bone Metastases (BM) and in assessing response of BM to radiotherapy (RT) [1–20]. However, in order to be clinically acceptable, radiomics models must be trained on large data sets of real-world images. This is challenging as the full 3D segmentation of BM on planning-CT images is time-consuming for radiation oncologists in the clinical context. Often, in the interest of time and given the low doses used in palliative RT, radiation oncologists only delineate treatment field boundaries when treating BM, and they do not fully contour individual BM lesions. As a result, most of the published BM radiomics studies to date were trained and tested with relatively small sample sizes (see Table 4.1), which diminishes their generalizability and their applicability to clinical RT planning. Motivated by the need for large real-world BM data sets, the objective of this study was to determine if a radiomics model can be trained to distinguish BM from healthy bone (HB) using BM lesions denoted as points on planning-CT images rather than using full 3D segmentation.

4.3.1 Radiomics for metastases detection

Radiomics is an automated feature generation method for the extraction of hundreds of quantitative phenotype (radiomics features) from radiology images [21, 22]. ML algorithms can be trained to find relationships between radiomics features and cancer outcomes if provided with sufficient and appropriate data. There are three main steps in the training phase of a typical radiomics study. These include: (1) manual or semi-automated segmentation of Regions Of Interest (ROIs) on patients' images, (2) feature extraction from the segmented ROIs, and (3) generation of a statistical or ML model to correlate extracted features to each patient's endpoint data such as their cancer outcome or other clinically-measured biomarkers [8].

In addition to the need for adequate sample sizes, which is the main motivation behind this study, a radiomics model must overcome two important challenges in order to be reliable in a clinical context. First, it must be clinically reproducible. This is challenging because different radiomics studies use different subsets of radiomics features to achieve optimal models. The variations in published feature selection approaches make radiomics models less clinically reproducible [23, 24]. Therefore, to achieve a clinically-reliable radiomics model, it is important to study and account for the effect of the variation in Feature Selection (FS) methods [25–27].

Depending on the endpoint of interest, various ML classifiers may be used in a radiomics model. Support vector machine, Bayesian network, multivariate logistic regression, k-nearest neighbor, decision trees, random forests, neural network, and convolutional neural networks are among the ML classifiers that are most commonly used in radiomics-based ML models [8–20]. The feasibility of using radiomics-based ML models to distinguish between benign and malignant bone lesions has been reported in previous studies [1–4, 6, 7]. The main details of these studies are summarized in Table 4.1.

Ref	Sample size*	Imaging Modality	ROI	labels	Classifier	Performance (ROC-AUC, A, P, R)†
[1]	36	PET/CT	Manual	benign and metastatic	RF	0.95, 0.88, 0.88, 0.89
[2]	74	Diagnostic-CT	Semi-automated‡	benign and malignant	RF	0.90, 0.92, 0.92, 0.91
[3]	75	PET/CT	Manual	responded and metastatic	kNN	0.76, 0.74, 0.74, 0.74
[4]	100	Diagnostic-CT	Semi-automated‡	benign and malignant	SVM	0.86, -, 0.85, 0.88
[5]	103	Dual-Energy CT	Semi-automated‡	benign and malignant	RF	0.79, 0.78, 0.72, 0.79
[6]	177	CT	Manual	bone island and metastases	RF	0.96, 0.80, 0.96, 0.86
[7]	206	Diagnostic-CT	Manual	benign and malignant	MLR	0.82, 0.86, 0.93, 0.77

Table 4.1: Radiomics-based ML models reported in the literature for distinguishing bone lesions. *Sample size is the total number of samples. †A: Accuracy, P: precision (Specificity), R: Recall (sensitivity). ‡In the semi-automated segmentation methods an expert was required to check and modify the computer-segmented ROIs slice-by-slice.

The radiomics-based ML models listed in Table 4.1 are not readily applicable to our clinical context, palliative RT for BM, for three reasons. First, they have relatively small sample sizes, an inherent problem for generalizability. Second, they require full 3D lesion segmentation, which is challenging to achieve clinically when planning palliative RT for BM. Finally, they were trained on images acquired using diagnostic-CTs or hybrid imaging modalities, whereas palliative RT planning is mostly done on planning-CT (simulation-CT) images alone.

With the above limitations in mind, in this study, we investigated the feasibility of developing a fast and reliable radiomics-based ML pipeline capable of differentiating between BM and HB in RT planning-CT images of cancer patients using just geometric ROIs centered on expert-identified lesion point locations. We investigated the effect of using ROIs with different sizes and geometric shapes. We also examined the performance of different FS methods and ML classifiers in achieving the optimal BM detection pipeline.

4.4 Materials and Methods

4.4.1 Ethics declarations

This retrospective study was approved by the Research Ethics Board of the McGill University Health Centre, Montreal, Quebec, Canada, with the waiver of informed consent. We confirmed that all research were performed in accordance with the relevant guidelines and regulations.

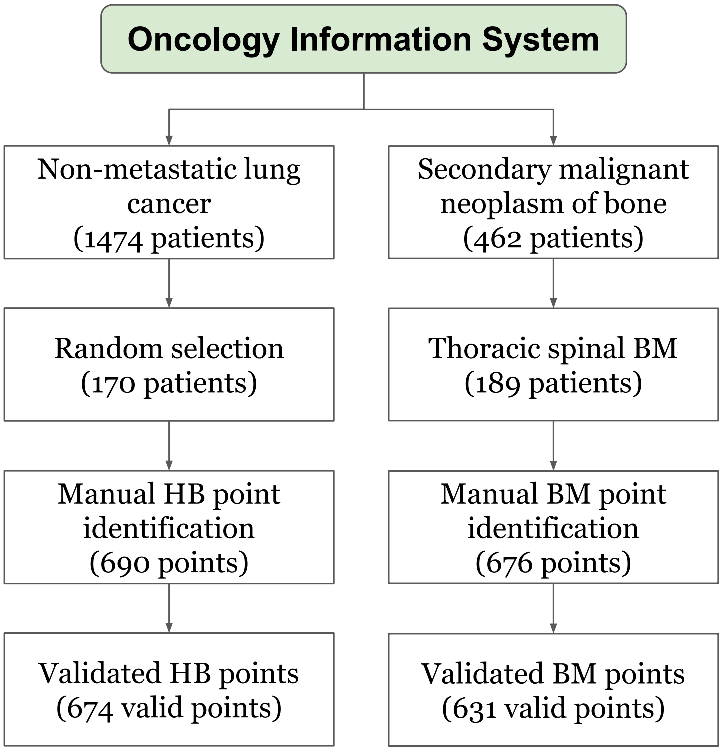

Oncology Information System
Non-metastatic lung cancer (1474 patients)
Secondary malignant neoplasm of bone (462 patients)
Random selection (170 patients)
Thoracic spinal BM (189 patients)
Manual HB point identification (690 points)
Manual BM point identification (676 points)
Validated HB points (674 valid points)
Validated BM points (631 valid points)

Figure 4.1: Flow chart of patient selection.

4.4.2 Patient selection

The planning-CT images of BM and HB patients used in this study were collected from the Oncology Information System at our institution. Our patient selection procedure is presented in Figure 4.1.

BM samples were from patients who received palliative RT for a secondary malignant neoplasm of bone in the thoracic spine between January 2016 and September 2019. HB samples were from individuals who received curative RT for non-metastatic lung cancer (as their CT images covered the same anatomy) during this period.

In total, we found 189 BM patients (96 male and 93 female; mean age± standard deviation (SD), 69 ± 13 y) and 1474 HB patients in our database. To reduce the large imbalance between the number of BM and HB patients, we randomly shuffled the HB sample (in a Microsoft Excel file) and selected the first 170 patients (86 male and 84 female; mean age 71 ± 12 y) to include in our study (see Figure 4.1).

4.4.3 Planning-CT images

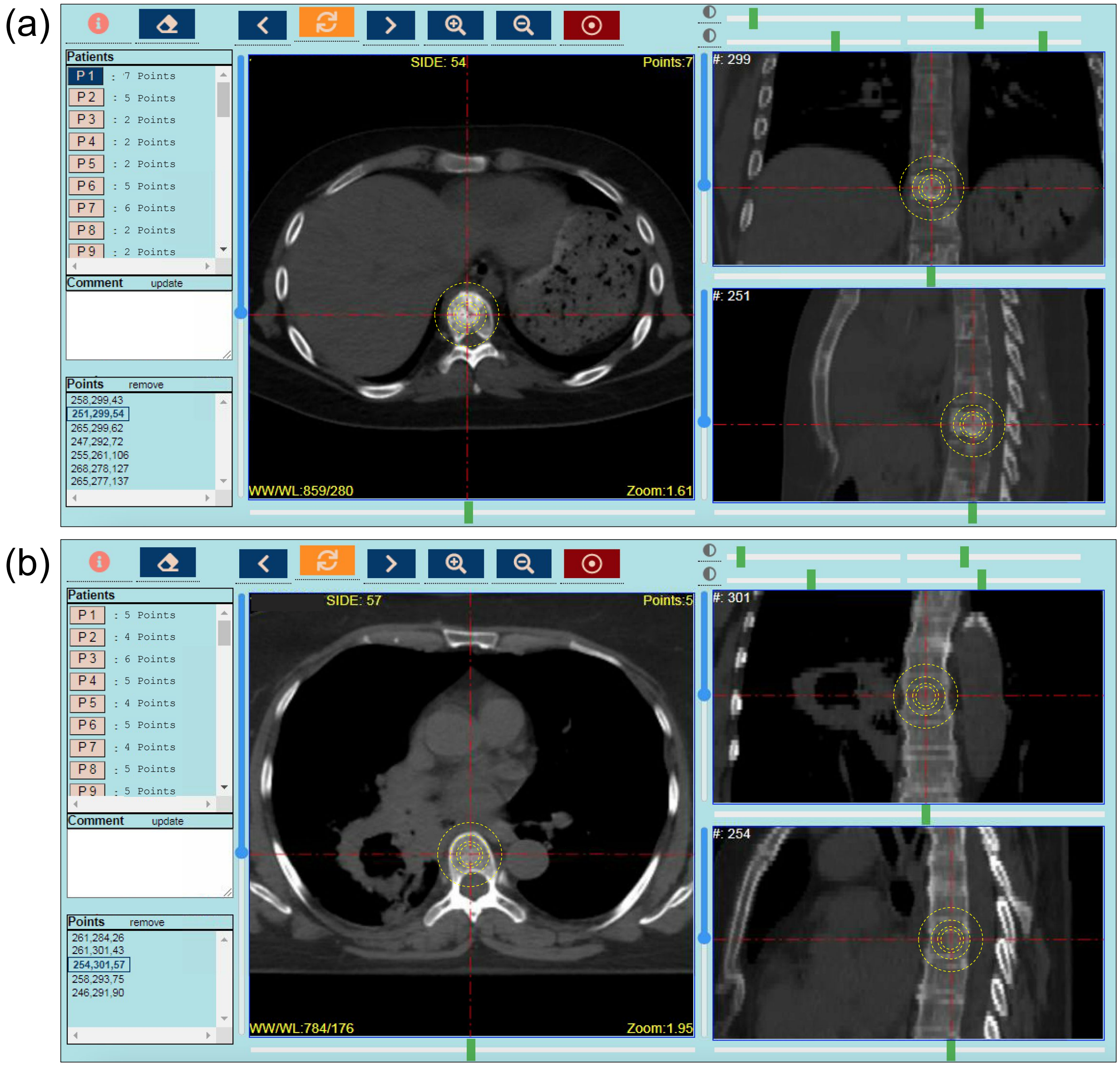
All planning-CT images were generated using one of three Philips' Brilliance Big Bore RT CT scanners at our institution with the acquisition parameters provided in Table 4.2. Planning-CT DICOM files were manually de-identified and exported to a secured hard drive from the Eclipse radiation therapy treatment planning software (Varian Medical Systems, Palo Alto, California), into which they had been previously imported for RT planning.

Tube voltage (kV)	Tube current (mA)	exposure exposure (mm)	Field of view (pixel)	Matrix size (mm)	Slice thickness (mm)	Pixel spacing
120	165-366	240-450	600	512×512	3.0	0.77-1.37

Table 4.2: Planning-CT image acquisition parameters.

4.4.4 Lesion identification

The planning-CT images of the BM patients were randomly divided into five sets using the Python `random.shuffle` module and were loaded into our custom-written 3D DICOM



(a)

Patients

P1 : 7 Points
P2 : 5 Points
P3 : 2 Points
P4 : 2 Points
P5 : 2 Points
P6 : 5 Points
P7 : 6 Points
P8 : 2 Points
P9 : 2 Points

Comment update

Points remove

258,299,43
251,299,54
265,299,62
247,292,72
255,261,106
268,278,127
265,277,137

(b)

Patients

P1 : 5 Points
P2 : 4 Points
P3 : 6 Points
P4 : 5 Points
P5 : 4 Points
P6 : 5 Points
P7 : 4 Points
P8 : 5 Points
P9 : 5 Points

Comment update

Points remove

261,284,26
261,301,43
254,301,57
258,293,75
246,291,90

Figure 4.2: Screenshots of our diCOMBINE 3D lesion labeling web application showing expert-labeled points. (a) A BM lesion, and (b) a HB point. Cross sections of 50 mm, 30 mm, 20 mm, and 15 mm spherical ROIs are visualized with yellow dashed lines on each CT plane.

visualization web application (diCOMBINE [28]) for lesion identifying. diCOMBINE is an open-source software developed by our group using the Python Flask [29] framework for DICOM 3D visualization and lesion point location labeling. The center points of BM lesions were labeled by an expert team comprising one staff radiation oncologist and four radiation oncology fellows. Each expert was asked to label BM center points in one of the five data sets, and a peer expert was tasked with reviewing them and validating the labels. A total of 631 validated BM center points were thus identified in the BM data set. Similarly, the planning-CT images of the HB patients were randomly divided into three sets and were loaded into diCOMBINE for HB labeling. One staff medical physicist and two medical physics graduate students were asked to identify HB points in one of the data sets each. When identifying HB points, the medical physicists were instructed to avoid non-metastatic skeletal complications (such as surgically-treated bone lesions). An average of four HB points were identified in each planning-CT image. Then, we asked each physicist to independently review and confirm the HB points that one of their peers had labeled. A total of 674 validated HB points were identified in this way. Screenshots of our diCOMBINE 3D lesion labeling web application are presented in Figure 4.2. These BM and HB points were used as center points for our automated ROIs delineation.

4.4.5 Delineation of regions of interest

ROIs were automatically delineated in the planning-CT images using geometric shapes centered on the expert-identified point locations. We used four spherical (SP) and five cylindrical along the z-axis (CY) ROIs of various sizes. The characteristics of the ROIs used are specified in Table 4.3. The size ranges were defined to extend from the size of a large bone lesion (∼15 mm) [30] to the maximum size of a spinal vertebra (∼50 mm) [31, 32].

4.4.6 Radiomics feature extraction

The pydicom package (https://pydicom.github.io/pydicom/stable/) was used to read DICOM CT images and normalize pixel data to Hounsfield Units. Then, the normalized CT slices were stored as 3D raster data using the pynrrd package (version 0.4.2) (https://pypi.org/project/pynrrd/0.4.2/). The pynrrd package was also used to generate 3D binary masks from each of the nine ROIs listed in Table 4.3. Finally, The open-source

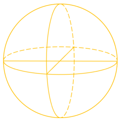
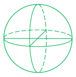
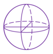
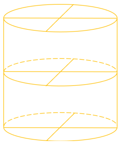
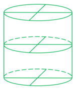
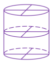
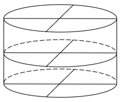
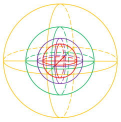
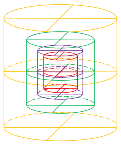
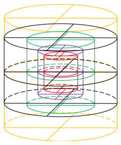
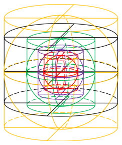
	Spherical				
Abbreviation	SP50	SP30	SP20	SP15	
Diameter (mm)	50	30	20	15	
					
	Cylindrical along the z-axis				
Abbreviation	CY50	CY30	CY20	CY15	CY5030
Width (mm) × Height (mm)	50×50	30×30	20×20	15×15	50×30
					
	Ensemble				
Abbreviation	E4SP	E4CY	E5CY	E9SC	
ROIs	SP50 +SP30 +SP20 +SP15	CY50 +CY30 +CY20 +CY15	CY50 +CY30 +CY20 +CY15 +CY5030	SP50+SP30 +SP20+SP15 +CY50+CY30 +CY20+CY15 +CY5030	
					

Table 4.3: The characteristics of the ROIs used in this study. ROIs from the planning-CT images were segmented using cylindrical and spherical ROIs with various sizes around the expert-labeled BM and HB points.

PyRadiomics package (version 3.0.1) [33] was used to calculate the 3D quantitative radiomics features. For each of the nine ROIs listed in Table 4.3, we extracted 107 radiomics features from each of the planning-CT images. We did not apply any filters prior to feature extraction. These 107 features include 18 First Order, 14 Shape, 24 Gray Level Co-occurrence Matrix (GLCM), 16 Gray Level Size Zone Matrix (GLSZM), 16 Gray Level Run Length Matrix (GLRLM), 14 Gray Level Dependence Matrix (GLDM), and five

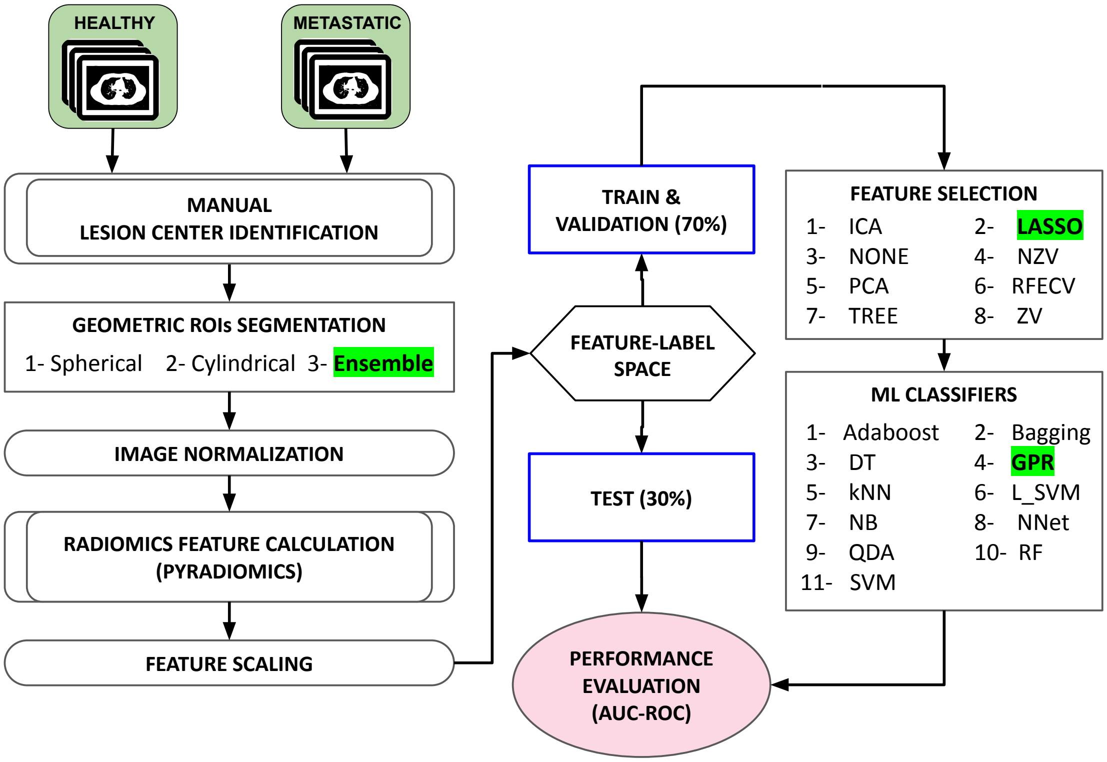


Figure 4.3: The exploration workflow for developing our radiomics-based ML models for classifying metastatic (BM) and healthy (HB) spinal bones. The best performing pipeline, as described in the Results, is highlighted in green.

Neighbouring Gray Tone Difference Matrix (NGTDM) features [34, 35]. We also aggregated radiomics features from multiple ROIs to define four ensemble ROIs, including; 1) E4SP: 428 features extracted from all four spherical ROIs, 2) E4CY: 428 features extracted from the first four cylindrical ROIs, 3) E5CY: 535 features extracted from all five cylindrical ROIs, and 4) E9SC: 963 features extracted from all nine ROIs combined. Our rationale for this approach was that by aggregating features extracted from ROIs with various sizes around the BM centers, we could extract sufficient information about the BMs' shape, size, and other characteristics and distinguish them from HBs using ML classifiers. Similar feature aggregation approaches were used in other studies [36, 37].

4.4.7 Radiomics workflow

Our complete radiomics-based ML workflow is presented in Figure 4.3. After extracting radiomics features for each ROI, we scaled the feature space using z-score normalization [38]. Then, we randomly divided the data set into 70% and 30% stratified training and testing sets, respectively. Each stratified set contained approximately the same BM/HB samples ratio as the initial data set. The training set was used for FS, and ML model development using 5-fold cross-validation [39]. The test set was used for the final performance evaluation. In the present study, we examined the performance of 13 FS methods and 12 ML classifiers as shown in Figure 4.3 and described in the following sections.

4.4.8 Feature selection methods

Radiomics calculates hundreds of features from images and some of them are redundant or are not useful for detecting BM [40]. To identify the most useful radiomics features for differentiating BM and HB, we investigated several supervised and unsupervised FS methods, Principal Component Analysis (PCA) [41], Fast Independent Component Analysis (ICA) [42], zero variance threshold [43] (VT_0), near-zero variance threshold [43], Least Absolute Shrinkage and Selection Operator logistic regression algorithm (LASSO) [44], Recursive Feature Elimination with Cross-Validation (RFECV) r [45], and Decision-Tree-Based (TREE) [46] feature selection. For the LASSO, motivated by Zack et al. [9] we used 20, 24, and 30 features. For the LASSO method, we examined least-squares penalty (α) values of 0.1, 0.5 and 1.0. α controls the stability of the selected features. A LASSO method with a larger α keeps fewer features (the most stable ones) [44]. For near-zero variance, we selected the variance threshold of 0.8 (VT_0.8) as used by Zack et al. [9]. FS techniques were implemented using Python scikit-learn [47] (version 0.24.2) feature selection module (https://scikit-learn.org/stable/modules/feature_selection.html). The performance of these FS methods, along with no FS, was then evaluated using 12 supervised ML classifiers.

4.4.9 Machine learning classifiers

The Python scikit-learn ML package (version 0.20.4) [48] was used to implement our ML classifiers. We used 12 supervised classification models, including the Linear Support Vector Machine [49] (L-SVM), SVM with Radial-basis function kernel [49], Gaussian Naive Bayes (NB) [50], k-Nearest Neighbors (kNN) [51], Quadratic Discriminant Analysis (QDA) [52], Gaussian Process Regression (GPR) [53], Decision Tree (DT) [54], Random Forest (RF) [55], Bagging [55], AdaBoost [55], Neural Networks with stochastic gradient-based solver [56, 57] (NNet) and NNet with Limited-memory Broyden–Fletcher–Goldfarb–Shanno solver [58] (NNet-LBFGS). For both NNet classifiers, we used the rectified linear unit activation function [59] (ReLU).

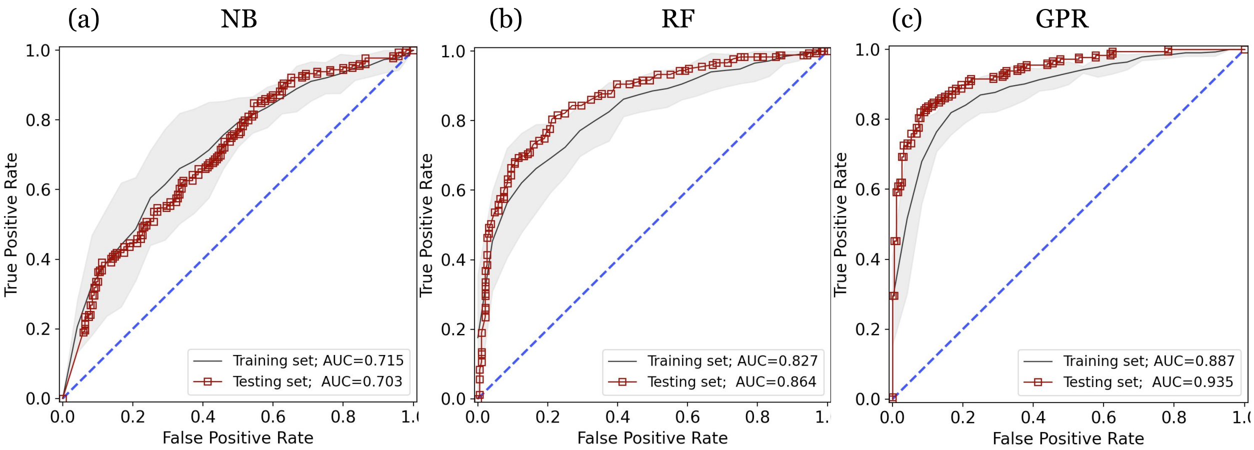
4.4.10 Performance evaluation

The performance of our radiomics-based ML models were measured using the test data set. The standard error of calculations was reported using 5-fold cross-validation on the training data set. We used the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) [60] for performance evaluation. Also, we reported precision and recall for our best-performing pipeline. Matplotlib (version 3.4.3) [61] was used to generate figures.

4.5 Results

4.5.1 Radiomics feature space

A JSON file of the metadata of extracted radiomics features is available in the supplementary dataset in our public repository [62]. The predictive performance of the different FS methods and ML classifiers was evaluated for each ROI on the test set using the ROC-AUC, precision, recall, and F-1 scores. Examples of Receiver Operating Characteristic (ROC) curves are presented in Figure 4.4 for the a) NB (a poor performance), b) RF (a good performance), and c) GPR (the best performance) ML classifiers on the test data set (red squares) and 5-fold validation set (pink lines) using 20 mm spherical ROI (SP20) with no FS. Note that 20 mm SP ROI was selected for visualization purposes throughout this paper for no particular reason. The effect of using the various geometric ROIs will be presented later in this paper. Raw data values, including confusion matrices, ROC graphs, and performance tables (precision, recall, F-1, values on training, validation, and test sets) for all ML classifiers on all ROIs are provided in the output data folder in our public repository [62].

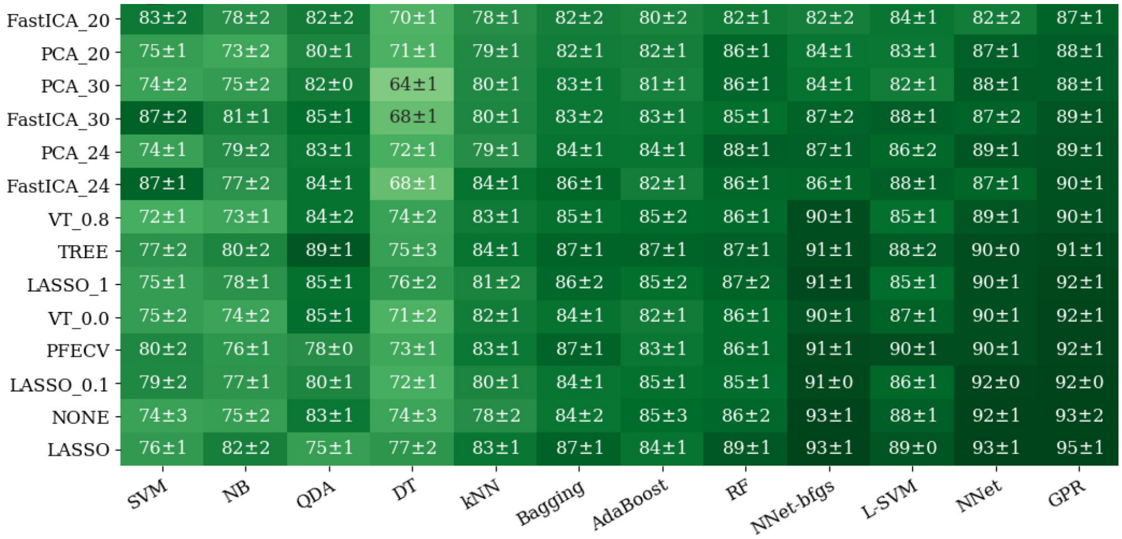


We used 20 mm spherical ROIs (SP20) with no FS for this example. ROC-AUC values are presented in the legends. The 20 mm SP ROI was used for visualization purposes. Full data is available in the supplementary dataset [62].

Figure 4.4: Example ROC curves for our radiomics-based ML models with the a) NB, b) RF, and c) GPR ML classifiers Example ROC curves for our radiomics-based ML models with the a) NB, b) RF, and c) GPR ML classifiers on the training set (black lines) and on the test set (red squares). The gray range represents the mean ROC ± SD of the 5-fold cross-validation used on the training set. Matplotlib (version 3.4.3) (https://pypi.org/project/matplotlib/3.4.3/) is used for visualizing the data.

4.5.2 Effect of feature selection

An example of an ROC-AUC grid for different combinations of ML classifiers and FS methods is presented in Figure 4.5 for the 20 mm SP ROI. As can be seen in Figure 4.5, the best results were achieved by the GPR and NNet classifiers with LASSO FS methods. The RFECV, VT, LASSO, and TREE FS methods outperformed PCA and ICA FS methods. Overall, FS did not have much effect on the performance of the models for the 20 mm ROI. For example,



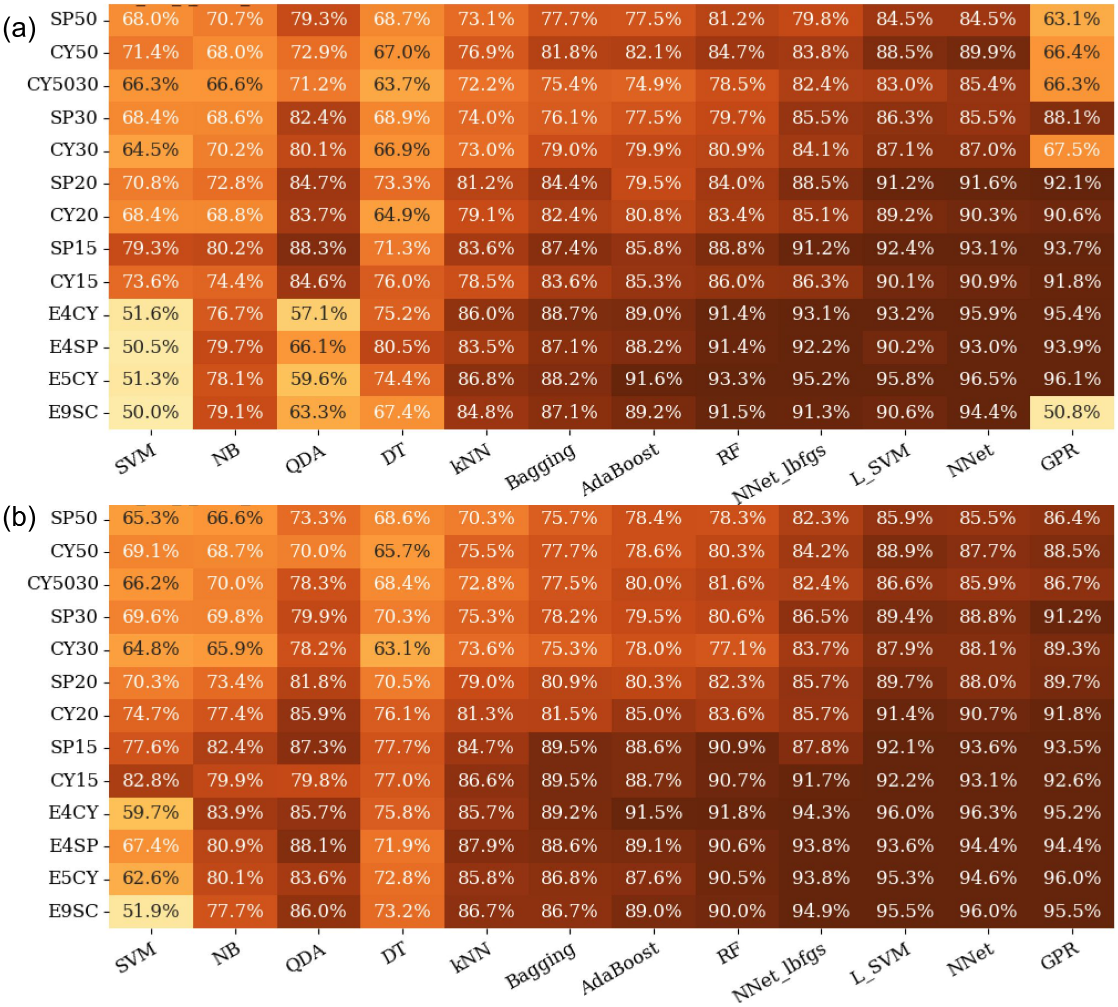
	SVM	NB	QDA	DT	kNN	Bagging	AdaBoost	RF	NNet-bfgs	L-SVM	NNet	GPR
FastICA_20	83±2	78±2	82±2	70±1	78±1	82±2	80±2	82±1	82±2	84±1	82±2	87±1
PCA_20	75±1	73±2	80±1	71±1	79±1	82±1	82±1	86±1	84±1	83±1	87±1	88±1
PCA_30	74±2	75±2	82±0	64±1	80±1	83±1	81±1	86±1	84±1	82±1	88±1	88±1
FastICA_30	87±2	81±1	85±1	68±1	80±1	83±2	83±1	85±1	87±2	88±1	87±2	89±1
PCA_24	74±1	79±2	83±1	72±1	79±1	84±1	84±1	88±1	87±1	86±2	89±1	89±1
FastICA_24	87±1	77±2	84±1	68±1	84±1	86±1	82±1	86±1	86±1	88±1	87±1	90±1
VT_0.8	72±1	73±1	84±2	74±2	83±1	85±1	85±2	86±1	90±1	85±1	89±1	90±1
TREE	77±2	80±2	89±1	75±3	84±1	87±1	87±1	87±1	91±1	88±2	90±0	91±1
LASSO_1	75±1	78±1	85±1	76±2	81±2	86±2	85±2	87±2	91±1	85±1	90±1	92±1
VT_0.0	75±2	74±2	85±1	71±2	82±1	84±1	82±1	86±1	90±1	87±1	90±1	92±1
PFECV	80±2	76±1	78±0	73±1	83±1	87±1	83±1	86±1	91±1	90±1	90±1	92±1
LASSO_0.1	79±2	77±1	80±1	72±1	80±1	84±1	85±1	85±1	91±0	86±1	92±0	92±0
NONE	74±3	75±2	83±1	74±3	78±2	84±2	85±3	86±2	93±1	88±1	92±1	93±2
LASSO	76±1	82±2	75±1	77±2	83±1	87±1	84±1	89±1	93±1	89±0	93±1	95±1

Figure 4.5: The ROC-AUC grid for different ML (x-axis) classifiers and FS methods (y-axis) combinations. The number in front of each PCA or ICA method is the number of selected features used. The number in front of each LASSO method corresponds to the α penalty value (the default value is 0.5). The number in front of each VT method is its variance threshold value. Matplotlib (version 3.4.3) (https://pypi.org/project/matplotlib/3.4.3/) is used for visualizing the data.

for the GPR ML classifier, the performance of our model increased only 2% (from 93% to 95%) with the LASSO method compared to with no FS (NONE).

4.5.3 Effect of geometric ROIs

Two examples of the effect of using geometric ROIs with different sizes and shapes are presented in Figures 4.6. For plot (a), we used no FS. For plot (b), we used the best performing FS method (LASSO). As can be seen in Figure 4.6, the size of the ROI had a significant effect on the performance of our radiomics-based ML models. In general, a smaller ROI resulted in superior performance of models. For example, for the GPR classifier with no FS (the rightmost column of Figure 4.6-a), the ROC-AUC was improved from 86% to 94% when we moved from the SP50 to the SP15 ROI. SP15 resulted in the best overall performance when no FS was used. When we employed FS methods, the ensemble ROIs



(a)

	SVM	NB	QDA	DT	kNN	Bagging	AdaBoost	RF	NNet_lbfgs	L_SVM	NNet	GPR
SP50	68.0%	70.7%	79.3%	68.7%	73.1%	77.7%	77.5%	81.2%	79.8%	84.5%	84.5%	63.1%
CY50	71.4%	68.0%	72.9%	67.0%	76.9%	81.8%	82.1%	84.7%	83.8%	88.5%	89.9%	66.4%
CY5030	66.3%	66.6%	71.2%	63.7%	72.2%	75.4%	74.9%	78.5%	82.4%	83.0%	85.4%	66.3%
SP30	68.4%	68.6%	82.4%	68.9%	74.0%	76.1%	77.5%	79.7%	85.5%	86.3%	85.5%	88.1%
CY30	64.5%	70.2%	80.1%	66.9%	73.0%	79.0%	79.9%	80.9%	84.1%	87.1%	87.0%	67.5%
SP20	70.8%	72.8%	84.7%	73.3%	81.2%	84.4%	79.5%	84.0%	88.5%	91.2%	91.6%	92.1%
CY20	68.4%	68.8%	83.7%	64.9%	79.1%	82.4%	80.8%	83.4%	85.1%	89.2%	90.3%	90.6%
SP15	79.3%	80.2%	88.3%	71.3%	83.6%	87.4%	85.8%	88.8%	91.2%	92.4%	93.1%	93.7%
CY15	73.6%	74.4%	84.6%	76.0%	78.5%	83.6%	85.3%	86.0%	86.3%	90.1%	90.9%	91.8%
E4CY	51.6%	76.7%	57.1%	75.2%	86.0%	88.7%	89.0%	91.4%	93.1%	93.2%	95.9%	95.4%
E4SP	50.5%	79.7%	66.1%	80.5%	83.5%	87.1%	88.2%	91.4%	92.2%	90.2%	93.0%	93.9%
E5CY	51.3%	78.1%	59.6%	74.4%	86.8%	88.2%	91.6%	93.3%	95.2%	95.8%	96.5%	96.1%
E9SC	50.0%	79.1%	63.3%	67.4%	84.8%	87.1%	89.2%	91.5%	91.3%	90.6%	94.4%	50.8%

(b)

	SVM	NB	QDA	DT	kNN	Bagging	AdaBoost	RF	NNet_lbfgs	L_SVM	NNet	GPR
SP50	65.3%	66.6%	73.3%	68.6%	70.3%	75.7%	78.4%	78.3%	82.3%	85.9%	85.5%	86.4%
CY50	69.1%	68.7%	70.0%	65.7%	75.5%	77.7%	78.6%	80.3%	84.2%	88.9%	87.7%	88.5%
CY5030	66.2%	70.0%	78.3%	68.4%	72.8%	77.5%	80.0%	81.6%	82.4%	86.6%	85.9%	86.7%
SP30	69.6%	69.8%	79.9%	70.3%	75.3%	78.2%	79.5%	80.6%	86.5%	89.4%	88.8%	91.2%
CY30	64.8%	65.9%	78.2%	63.1%	73.6%	75.3%	78.0%	77.1%	83.7%	87.9%	88.1%	89.3%
SP20	70.3%	73.4%	81.8%	70.5%	79.0%	80.9%	80.3%	82.3%	85.7%	89.7%	88.0%	89.7%
CY20	74.7%	77.4%	85.9%	76.1%	81.3%	81.5%	85.0%	83.6%	85.7%	91.4%	90.7%	91.8%
SP15	77.6%	82.4%	87.3%	77.7%	84.7%	89.5%	88.6%	90.9%	87.8%	92.1%	93.6%	93.5%
CY15	82.8%	79.9%	79.8%	77.0%	86.6%	89.5%	88.7%	90.7%	91.7%	92.2%	93.1%	92.6%
E4CY	59.7%	83.9%	85.7%	75.8%	85.7%	89.2%	91.5%	91.8%	94.3%	96.0%	96.3%	95.2%
E4SP	67.4%	80.9%	88.1%	71.9%	87.9%	88.6%	89.1%	90.6%	93.8%	93.6%	94.4%	94.4%
E5CY	62.6%	80.1%	83.6%	72.8%	85.8%	86.8%	87.6%	90.5%	93.8%	95.3%	94.6%	96.0%
E9SC	51.9%	77.7%	86.0%	73.2%	86.7%	86.7%	89.0%	90.0%	94.9%	95.5%	96.0%	95.5%

Figure 4.6: The ROC-AUC grid for different combinations of ML classifiers (x-axis) and geometric ROIs (y-axis) with various sizes and shapes (a) with no FS method and (b) with LASSO as the FS method.

(like E4SP and E9SC) out-performed the single-size ROIs. This was most pronounced for the LASSO method, which is presented in Figure 4.6-b. The ROC-AUC grids for other FS methods are provided in the output data folder in our public repository [62].

Comparing Figure 4.6-a and 4.6-b revealed that some ML classifiers (like SVM or GPR) were more sensitive to the use of FS than others (like NNet or RF). Also, we noticed that FS was more important when using large ROIs (such as SP50 or CY50) or ensemble ROIs (such as E4SP or E9SC).

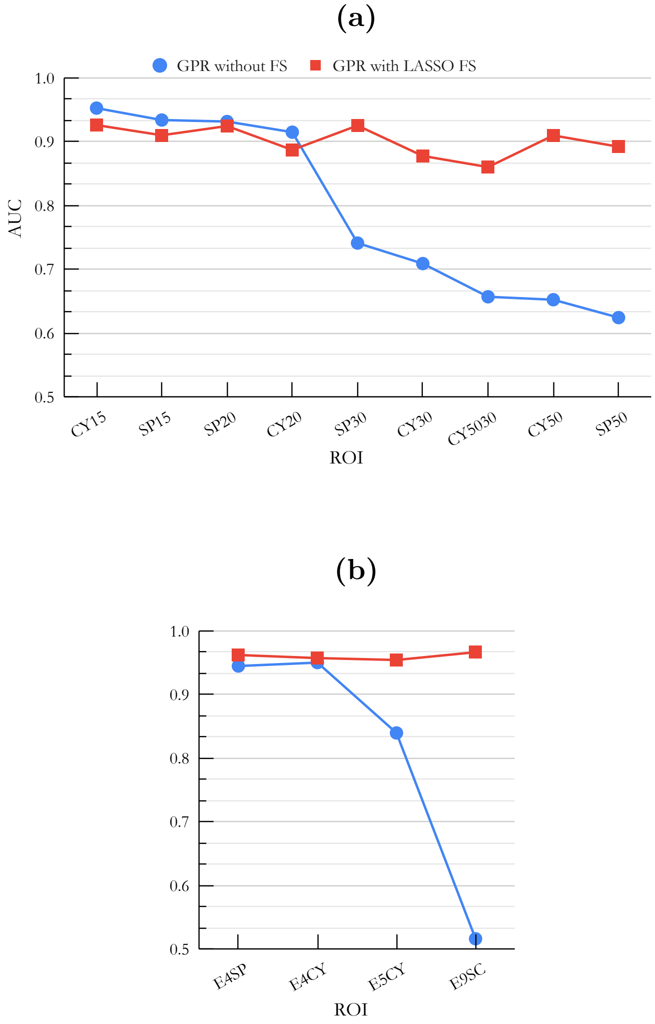
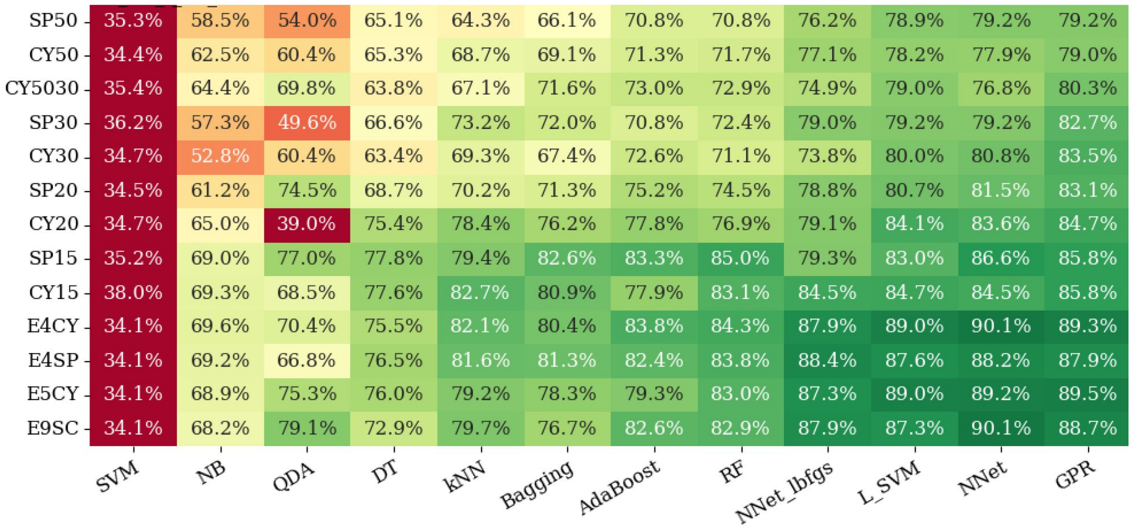


Figure 4.7: An example of ROC-AUCs versus the volume of the ROI for (a) single geometric ROIs and (b) for ensemble ROIs (Ref. Table 4.3). For this graph, we used our best performing ML classifier (GPR), with our best performing FS method (LASSO), and without FS method.

To visualize the effect of the size of ROI on the performance of our models, in Figure 4.7 we show the ROC-AUCs of our best performing ML classifier (GPR) for (a) single geometric

ROIs (sorted by volume), and (b) for ensemble ROIs (sorted by total volume). To show the effect of the use of FS, we plotted the results without FS (blue circles), and with our best-performing FS method (LASSO) (red squares). It can be seen that a smaller ROI resulted in a better performance. Also, FS was more important for larger ROIs (like SP50 and CY50) and ensemble ROIs (like E9SC).



	SVM	NB	QDA	DT	kNN	Bagging	AdaBoost	RF	NNet_lbfgs	L_SVM	NNet	GPR
SP50	35.3%	58.5%	54.0%	65.1%	64.3%	66.1%	70.8%	70.8%	76.2%	78.9%	79.2%	79.2%
CY50	34.4%	62.5%	60.4%	65.3%	68.7%	69.1%	71.3%	71.7%	77.1%	78.2%	77.9%	79.0%
CY5030	35.4%	64.4%	69.8%	63.8%	67.1%	71.6%	73.0%	72.9%	74.9%	79.0%	76.8%	80.3%
SP30	36.2%	57.3%	49.6%	66.6%	73.2%	72.0%	70.8%	72.4%	79.0%	79.2%	79.2%	82.7%
CY30	34.7%	52.8%	60.4%	63.4%	69.3%	67.4%	72.6%	71.1%	73.8%	80.0%	80.8%	83.5%
SP20	34.5%	61.2%	74.5%	68.7%	70.2%	71.3%	75.2%	74.5%	78.8%	80.7%	81.5%	83.1%
CY20	34.7%	65.0%	39.0%	75.4%	78.4%	76.2%	77.8%	76.9%	79.1%	84.1%	83.6%	84.7%
SP15	35.2%	69.0%	77.0%	77.8%	79.4%	82.6%	83.3%	85.0%	79.3%	83.0%	86.6%	85.8%
CY15	38.0%	69.3%	68.5%	77.6%	82.7%	80.9%	77.9%	83.1%	84.5%	84.7%	84.5%	85.8%
E4CY	34.1%	69.6%	70.4%	75.5%	82.1%	80.4%	83.8%	84.3%	87.9%	89.0%	90.1%	89.3%
E4SP	34.1%	69.2%	66.8%	76.5%	81.6%	81.3%	82.4%	83.8%	88.4%	87.6%	88.2%	87.9%
E5CY	34.1%	68.9%	75.3%	76.0%	79.2%	78.3%	79.3%	83.0%	87.3%	89.0%	89.2%	89.5%
E9SC	34.1%	68.2%	79.1%	72.9%	79.7%	76.7%	82.6%	82.9%	87.9%	87.3%	90.1%	88.7%

Figure 4.8: The F-1 score grid for different combinations of ML classifiers (x-axis) and ROIs (y-axis) with different sizes and shapes with LASSO FS method.

The grid of the F-1 scores for the best performing FS method (LASSO) is presented if Figure 4.8. The GPR, NNet, and L-SVM classifiers achieved 0.9 F-1 score in detecting BM using the ensemble ROIs. The ROC-AUC, precision, recall, and F1 score of our best performing pipeline, corresponding to the E9SC ROI, LASSO FS method, and GPR ML classifier, were 96%, 92%, 91%, and 0.9, respectively. The performance of our models for all combinations of FS methods, ML classifiers, and ROIs are provided in the supplementary dataset [62].

4.6 Discussion

In this study, we investigated the feasibility of using a single-point-based geometric ROI to develop a radiomics pipeline to distinguish BM and HB locations in planning-CT images of cancer patients with BM. We investigated various FS methods, and ML classifiers using point-based geometric ROIs with various shapes and sizes.

The time and effort needed for manual 3D segmentation of ROI are significant limitations to achieving large real-world image data sets. This, in turn, hinders the generation of generalizable radiomics-based prognostic ML models [63] for use in the clinic. Another limitation of manual lesion segmentation is inter-observer variability, which has been shown to have a significant impact on the performance and reproducibility of radiomics-based pipelines [64]. Furthermore, manual segmentation tools, designed for radiation therapy treatment planning, intend to load one patient at a time. Therefore, switching between patients is another time-consuming process that slows down the lesion delineation for multiple patients in the research context [65].

Our in-house-developed open-source 3D DICOM visualization and lesion identification tool (diCOMBINE [28]) allowed our collaborating radiation oncologists to quickly review planning-CT images of several hundred patients and efficiently identify 676 BM centers. They found diCOMBINE fast and easy to use, allowing each expert to label around 150 lesions per hour. Based on our experts' anecdotal experience, single-point-based geometric ROI delineation was 10 to 15 times faster than full manual 3D segmentation. These lesion centers were used to generate ROIs automatically. Defining point-based geometric ROIs, instead of full 3D manual segmentation of the ROIs, allowed us to rapidly generate a large sample set, minimize expert imposed uncertainties, and investigate the effect of the size and shape of the ROIs in the performance of our radiomics models. Besides, our point-based radiomics models will allow us to study the feasibility of building an automated BM-identifying pipeline. To the best of our knowledge, no studies on automated BM delineation have been published previously.

Radiomics extracts hundreds of features from an ROI. However, these features are generally highly correlated and contain much noise. Therefore, it is essential to apply proper FS methods to achieve a robust radiomics-based ML pipeline. Among the seven FS methods we examined in this study, we found that PCA and ICA resulted in lower ROC-AUC values than the VT, LASSO, and TREE FS methods. One reason for this difference was that VT, LASSO, and TREE methods automatically defined the optimal number of features, while in PCA and ICA, the number of features was predefined. For highly-correlated features, the optimal number of features (f) is roughly proportional to the square root of the sample size (n) [66]. Accordingly, 30 features would appear to be less than the optimal number of features for our sample size ($f = \sqrt{n} = \sqrt{1305} = 36$). For studies with small sample sizes, such as Zhang et al. [9], that used 112 samples, PCA with 10 features seems to be a suitable FS method. We also noticed that the effect of the FS method depends on the selected ML classifier. For ML classifiers that had built-in FS methods (i.e., RF and NNet), applying FS methods in some cases worsened the overall performance of the model. Inversely, for ML classifiers that did not have built-in FS methods (i.e., GPR), adding FS had a significant effect on the performance of the ML classifier. The effect of the FS method was more significant when working with ensemble ROIs that had many more features. For example, the ROC-AUC value for the GPR ML classifier using the E9SC ROI (963 features) improved from 0.52 to 0.97 when the LASSO FS method was used, as shown in Figure 4.6.

Among the ML methods we examined in this study, we found that GPR, NNet, SVM, and RF resulted in the highest ROC-AUC values and F-1 scores. We showed that the GPR classifier outperformed the NNet classifier for most ROIs. However, for the ensemble ROIs (in which the number of features was large), GPR required a proper FS method (i.e., LASSO). The dimensionality issue of GPR classifiers and their requirement for FS was also discussed in the literature [67, 68].

We found that our radiomics-based ML models performed slightly better on spherical ROIs compared to cylindrical ROIs of similar volumes. More significantly, we found that the smaller ROIs (15 and 20 mm) resulted in better performance compared to the larger ROIs (30 and 50 mm) (Figure 4.6). This might be due to the fact that in larger ROIs there are probably more outlier features captured from bone or organs/tissue surrounding the lesion of interest. Performances of our models did not improve considerably by decreasing the size of ROI below 20 mm, which is roughly the size of a large BM lesion [30]. As can be seen in Figure 4.7, our models performed better on the ensemble ROIs compared to the single ROI when used with FS methods. This could be due to having many features in the ensemble ROIs. For example, the E9SC ROI contains $9 \times 107 = 963$ features. For such a prominent feature space, FS methods become very important.

Although various radiomics pipelines have been previously developed and reported to classify bone lesions, our radiomics-based ML pipeline, reported here, offers several advantages compared to preceding efforts, mainly in the context of palliative radiotherapy planning. First, we, pragmatically, used planning-CT images of cancer patients for extracting radiomics features, whereas prior studies used hybrid modalities or diagnostic-CT images (as listed in Table 4.1). Hybrid modalities allow the development of high-quality prognostic pipelines. However, these pipelines are less clinically applicable in palliative radiotherapy treatment planning for BM, which is often primarily based on a patient's planning-CT scan. Second, all ML classifiers presented in the prior studies were restricted to full 3D segmentation of the lesion volumes. In the real-world clinical workflow for palliative radiotherapy of BM, it is common to use single-slice or lesion-center-based treatment planning with radiation oncologists often defining treatment field limits rather than lesion contours. Therefore, pipelines that require full 3D segmentation of ROI have limited application in real-world palliative radiotherapy [65]. Moreover, 3D segmentation of the ROI is a time-consuming bottleneck that likely compelled all the prior studies to train and test their radiomics pipelines with limited sample sizes. Training on a small sample size diminishes the generalizability and clinical applicability of a radiomics pipeline. In comparison, our point-based pipeline allowed us to avoid the labor-intensive manual segmentation step and train and test our pipeline on a large data set. Finally, in this study, we investigated the effects of FS methods, and ML classifiers in achieving the optimal prognostic model using geometric ROIs. To the best of our knowledge, no prior study performed such a comprehensive optimization.

Our study had some limitations. First, we selected BM and HB from two sets of separate patients. This selection might drive the risk of potential susceptibility to bias if there is a systematic difference between the two sets of images. However, our rationale for using non-metastatic cancer patients to select HBs was to eliminate the possibility of error in labeling HBs by our medical physicists. Second, our collaborating medical physicists could not identify non-metastatic skeletal complications from metastatic bone lesions. Therefore, the non-metastatic skeletal complications (i.e., surgically-removed lesions or bone islands) were ignored when labeling HB points. A solution for this problem would be using pathology data to identify non-metastatic and metastatic lesions but this would significantly increase the required effort. Third, we used a nearly balanced data set of HB and BM patients in this study. However, having an imbalanced sample ratio is common in many real-world radiation oncology outcome data sets [69, 70]. A study with an imbalanced data set is required to evaluate the effect of sample imbalance when building high-performance real-world radiomics-based ML models [71–73]. Forth, while using geometric ROIs significantly simplified the lesion delineation procedure, it ignored some lesion details such as size and shape. One alternative that can be explored as future work is to use deep-learning-based ROI segmentation. Finally, we used single-center planning-CT images from 359 patients in this retrospective study. A multi-center study with a more extensive data set is required to test the generalizability of our radiomics pipeline. Such a big data set would allow us to try more robust deep learning ML classifiers [74, 75] to build an AI tool to scan patients' planning-CT images and identify BM lesions automatically. The present work provides strong motivation to pursue such a multi-center study.

4.7 Conclusion

We demonstrated that our radiomics-based ML models can successfully distinguish between metastatic and healthy bones in planning-CT images using lesion-center-based geometric ROIs. Our results suggest that the GPR classifier with ensemble ROIs is particularly promising for the differentiation of BM and HB. Optimum pipeline performance was obtained using elimination-based FS methods such as LASSO. Our results demonstrate that radiomics features obtained from a lesion-center-based geometrical ROI may be sufficient to train radiomics-based ML classifiers to distinguish between bone lesions when full 3D segmented ROIs are not available. This opens the door to big data artificial intelligence research for cancer patients with BM.

4.8 Data availability

The supporting dataset is provided as a figshare repository [62]. This repository contains three files: 1) "featurespace_metadata.json.zip" file that includes radiomics features extracted from 1273 spinal lesions (healthy or metastatic) from radiotherapy planning-ct images using geometrical Regions Of Interest (ROIs). 2) "output.zip" folder that contains the results of our radiomics-based ML models that were validated and tested using several FS, and ML on single-point-based geometric ROIs with various shapes and sizes. 3) A README.md file that is provided to explain the information about the data structure and file naming patterns.

4.9 Acknowledgements

This research was supported by the startup grant of Dr. John Kildea at Research Institute of the McGill University Health Centre (RI-MUHC), the Ruth and Alex Dworkin scholarship award from the Faculty of Medicine and Health Sciences at McGill University, an RI-MUHC studentship award, a Grad Excellence Award-00293 from the Department of Physics at McGill University, and from the CREATE Responsible Health and Healthcare Data Science (SDRDS) grant of the Natural Sciences and Engineering Research Council. The authors would like to thank Dr. Luc Galarneau for his help with statistical analysis.

4.10 Competing interests

The authors declare no competing interests.

4.11 Author contributions statement

H.N. contributed to the methodology, literature review, software, formal analysis, investigation, visualization, and writing the original draft. S.S. participated in data collection, interpretation, and validation. M.T. participated in data collection, interpretation, and validation. M.F. participated in data collection, interpretation, and validation. P.R. participated in data collection, interpretation, and validation. J.Kh. participated in data collection, interpretation, and validation. H.P. participated in data collection, interpretation, and validation. A.X.A.H. participated in data collection, interpretation, and validation. M.D. participated in conceptualization and methodology. J.Ki. participated in data collection and contributed to the conceptualization, investigation, supervision, funding acquisition, and editing of the original draft. All authors contributed to the review of the paper and approved the final manuscript.

Bibliography

[1] T. Perk, T. Bradshaw, S. Chen, H. J. Im, S. Cho, S. Perlman, G. Liu, R. Jeraj, Automated classification of benign and malignant lesions in 18 F-NaF PET/CT images using machine learning, Physics in medicine and biology 63 (22) (11 2018). doi:10.1088/1361-6560/AAEBD0.
URL https://pubmed.ncbi.nlm.nih.gov/30457118/

[2] M. V. Suhas, A. Mishra, Classification of benign and malignant bone lesions on CT images using random forest, 2016 IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology, RTEICT 2016 - Proceedings (2017) 1807–1810doi:10.1109/RTEICT.2016.7808146.
URL https://manipal.pure.elsevier.com/en/publications/classification-of-benign-and-malignant-bone-lesions-on-ct-images-

[3] E. Acar, A. Leblebici, B. E. Ellidokuz, Y. Başbinar, G. C. Kaya, Machine learning for differentiating metastatic and completely responded sclerotic bone lesion in prostate cancer: A retrospective radiomics study, British Journal of Radiology 92 (1101) (2019). doi:10.1259/bjr.20190286.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6732932/

[4] M. V. Suhas, R. Kumar, Classification of benign and malignant bone lesions on CT imagesusing support vector machine: A comparison of kernel functions, 2016 IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology, RTEICT 2016 - Proceedings (2017) 821–824doi:10.1109/RTEICT.2016.7807941.
URL https://www.researchgate.net/publication/312304564_Classification_

of_benign_and_malignant_bone_lesions_on_CT_imagesusing_support_vector_machine_A_comparison_of_kernel_functions

[5] F. Homayounieh, R. Singh, C. Nitiwarangkul, F. Lades, B. Schmidt, M. Sedlmair, S. Saini, M. K. Kalra, Semiautomatic Segmentation and Radiomics for Dual-Energy CT: A Pilot Study to Differentiate Benign and Malignant Hepatic Lesions, AJR. American journal of roentgenology 215 (2) (8 2020). doi:10.2214/AJR.19.22164.
URL https://pubmed.ncbi.nlm.nih.gov/32406776/

[6] J. H. Hong, J. Y. Jung, A. Jo, Y. Nam, S. Pak, S. Y. Lee, H. Park, S. E. Lee, S. Kim, Development and validation of a radiomics model for differentiating bone islands and osteoblastic bone metastases at abdominal CT, Radiology 299 (3) (2021) 626–632. doi:10.1148/RADIOL.2021203783/ASSET/IMAGES/LARGE/RADIOL.2021203783.VA.JPEG.
URL https://pubs.rsna.org/doi/abs/10.1148/radiol.2021203783

[7] W. Sun, S. Liu, J. Guo, S. Liu, D. Hao, F. Hou, H. Wang, W. Xu, A CT-based radiomics nomogram for distinguishing between benign and malignant bone tumours, Cancer Imaging 21 (1) (2021) 1–10. doi:10.1186/S40644-021-00387-6/FIGURES/4.
URL https://cancerimagingjournal.biomedcentral.com/articles/10.1186/s40644-021-00387-6

[8] A. Vial, D. Stirling, M. Field, M. Ros, C. Ritz, M. Carolan, L. Holloway, A. A. Miller, The role of deep learning and radiomic feature extraction in cancer-specific predictive modelling: a review, Translational Cancer Research 7 (3) (2018). doi:10.21037/21823.
URL https://tcr.amegroups.com/article/view/21823

[9] Y. Zhang, A. Oikonomou, A. Wong, M. A. Haider, F. Khalvati, Radiomics-based Prognosis Analysis for Non-Small Cell Lung Cancer, Scientific Reports 2017 7:1 7 (1) (2017) 1–8. doi:10.1038/srep46349.
URL https://www.nature.com/articles/srep46349

[10] B. Baessler, T. Nestler, D. Pinto dos Santos, P. Paffenholz, V. Zeuch, D. Pfister, D. Maintz, A. Heidenreich, Radiomics allows for detection of benign and malignant histopathology in patients with metastatic testicular germ cell tumors prior to post-chemotherapy retroperitoneal lymph node dissection, European radiology 30 (4) (2020)

2334–2345. doi:10.1007/S00330-019-06495-Z.
URL https://pubmed.ncbi.nlm.nih.gov/31828413/

[11] L. Duron, A. Heraud, F. Charbonneau, M. Zmuda, J. Savatovsky, L. Fournier, A. Lecler, A Magnetic Resonance Imaging Radiomics Signature to Distinguish Benign From Malignant Orbital Lesions, Investigative radiology 56 (3) (2021) 173–180. doi:10.1097/RLI.0000000000000722.
URL https://pubmed.ncbi.nlm.nih.gov/32932375/

[12] B. Laderian, F. S. Ahmed, B. Zhao, J. Wilkerson, L. Dercle, H. Yang, X. Guo, K. Pacak, J. A. Lee, S. E. Bates, J. D. Rivero, L. H. Schwartz, A. T. Fojo, Role of radiomics to differentiate benign from malignant pheochromocytomas and paragangliomas on contrast enhanced CT scans., Journal of Clinical Oncology 37 (15_suppl) (2019) e14596–e14596. doi:10.1200/JCO.2019.37.15{_}SUPPL.E14596.

[13] S. Li, J. Liu, Y. Xiong, P. Pang, P. Lei, H. Zou, M. Zhang, B. Fan, P. Luo, A radiomics approach for automated diagnosis of ovarian neoplasm malignancy in computed tomography, Scientific Reports 2021 11:1 11 (1) (2021) 1–9. doi:10.1038/s41598-021-87775-x.
URL https://www.nature.com/articles/s41598-021-87775-x

[14] P. Yin, N. Mao, H. Chen, C. Sun, S. Wang, X. Liu, N. Hong, Machine and Deep Learning Based Radiomics Models for Preoperative Prediction of Benign and Malignant Sacral Tumors, Frontiers in Oncology 10 (2020) 2235. doi:10.3389/FONC.2020.564725/BIBTEX.

[15] H. Wang, P. Nie, Y. Wang, W. Xu, S. Duan, H. Chen, D. Hao, J. Liu, Radiomics nomogram for differentiating between benign and malignant soft-tissue masses of the extremities, Journal of magnetic resonance imaging : JMRI 51 (1) (2020) 155–163. doi:10.1002/JMRI.26818.
URL https://pubmed.ncbi.nlm.nih.gov/31169956/

[16] J. Wang, X. Liu, D. Dong, J. Song, M. Xu, Y. Zang, J. Tian, Prediction of malignant and benign of lung tumor using a quantitative radiomic method, Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering

in Medicine and Biology Society. Annual International Conference 2016 (2016) 1272–1275. doi:10.1109/EMBC.2016.7590938.
URL https://pubmed.ncbi.nlm.nih.gov/28268557/

[17] L. Zhou, Z. Zhang, Y. C. Chen, Z. Y. Zhao, X. D. Yin, H. B. Jiang, A Deep Learning-Based Radiomics Model for Differentiating Benign and Malignant Renal Tumors, Translational oncology 12 (2) (2019) 292–300. doi:10.1016/J.TRANON.2018.10.012.
URL https://pubmed.ncbi.nlm.nih.gov/30448734/

[18] B. J. Guo, X. He, T. Wang, Y. Lei, W. J. Curran, T. Liu, L. J. Zhang, X. Yang, Benign and malignant thyroid classification using computed tomography radiomics, https://doi.org/10.1117/12.2549087 11314 (2020) 954–961. doi:10.1117/12.2549087.
URL https://doi.org/10.1117/12.2549087

[19] R. Paul, S. Kariev, D. Cherezov, M. B. Schabath, R. J. Gillies, L. O. Hall, D. B. Goldgof, Deep radiomics: deep learning on radiomics texture images, https://doi.org/10.1117/12.2582102 11597 (2021) 8–17. doi:10.1117/12.2582102.
URL https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11597/1159705/Deep-radiomics-deep-learning-on-radiomics-texture-images/10.1117/12.2582102.short

[20] A. Chen, L. Lu, X. Pu, T. Yu, H. Yang, L. H. Schwartz, B. Zhao, CT-based radiomics model for predicting brain metastasis in category T1 lung adenocarcinoma, American Journal of Roentgenology 213 (1) (2019) 134–139. doi:10.2214/AJR.18.20591.
URL https://pubmed.ncbi.nlm.nih.gov/30933649/

[21] M. E. Mayerhoefer, A. Materka, G. Langs, I. Häggström, P. Szczypiński, P. Gibbs, G. Cook, Introduction to Radiomics, Journal of nuclear medicine : official publication, Society of Nuclear Medicine 61 (4) (2020) 488–495. doi:10.2967/JNUMED.118.222893.
URL https://pubmed.ncbi.nlm.nih.gov/32060219/

[22] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. Van Stiphout, P. Granton, C. M. Zegers, R. Gillies, R. Boellard, A. Dekker, H. J. Aerts, Radiomics: extracting more information from medical images using advanced feature analysis, European journal of cancer (Oxford, England : 1990) 48 (4) (2012) 441–446. doi:

10.1016/J.EJCA.2011.11.036.
URL https://pubmed.ncbi.nlm.nih.gov/22257792/

[23] Y. Sugai, N. Kadoya, S. Tanaka, S. Tanabe, M. Umeda, T. Yamamoto, K. Takeda, S. Dobashi, H. Ohashi, K. Takeda, K. Jingu, Impact of feature selection methods and subgroup factors on prognostic analysis with CT-based radiomics in non-small cell lung cancer patients, Radiation Oncology 16 (1) (2021) 1–12. doi:10.1186/S13014-021-01810-9/FIGURES/2.
URL https://ro-journal.biomedcentral.com/articles/10.1186/s13014-021-01810-9

[24] A. Demircioğlu, Measuring the bias of incorrect application of feature selection when using cross-validation in radiomics, Insights into Imaging 2021 12:1 12 (1) (2021) 1–10. doi:10.1186/S13244-021-01115-1.
URL https://insightsimaging.springeropen.com/articles/10.1186/s13244-021-01115-1

[25] P. Yin, N. Mao, C. Zhao, J. Wu, C. Sun, L. Chen, N. Hong, Comparison of radiomics machine-learning classifiers and feature selection for differentiation of sacral chordoma and sacral giant cell tumour based on 3D computed tomography features, European radiology 29 (4) (2019) 1841–1847. doi:10.1007/S00330-018-5730-6.
URL https://pubmed.ncbi.nlm.nih.gov/30280245/

[26] D. A. Delzell, S. Magnuson, T. Peter, M. Smith, B. J. Smith, Machine Learning and Feature Selection Methods for Disease Classification With Application to Lung Cancer Screening Image Data, Frontiers in oncology 9 (12 2019). doi:10.3389/FONC.2019.01393.
URL https://pubmed.ncbi.nlm.nih.gov/31921650/

[27] M. Ligero, G. Torres, C. Sanchez, K. Diaz-Chito, R. Perez, D. Gil, Selection of Radiomics Features based on their Reproducibility, Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference 2019 (2019) 403–408. doi:10.1109/EMBC.2019.8857879.
URL https://pubmed.ncbi.nlm.nih.gov/31945924/

[28] H. Naseri, diCOMBINE: 3D-DICOM Visualization and Lesion Identification Web Application (8 2021). doi:10.5281/ZENODO.5218743.
URL https://zenodo.org/record/5218743

[29] Flask Web Development, 2nd Edition [Book].
URL https://www.oreilly.com/library/view/flask-web-development/9781491991725/

[30] G. Hall, J. Wright, Bone Lesions, Gnepp's Diagnostic Surgical Pathology of the Head and Neck (2021) 689–742doi:10.1016/B978-0-323-53114-6.00008-0.

[31] S. H. Zhou, I. D. McCarthy, A. H. McGregor, R. R. Coombs, S. P. Hughes, Geometrical dimensions of the lower lumbar vertebrae – analysis of data from digitised CT images, European Spine Journal 2000 9:3 9 (3) (2000) 242–248. doi:10.1007/S005860000140.
URL https://link.springer.com/article/10.1007/s005860000140

[32] I. Busscher, J. J. Ploegmakers, G. J. Verkerke, A. G. Veldhuizen, Comparative anatomical dimensions of the complete human and porcine spine, European Spine Journal 19 (7) (2010) 1104–1114. doi:10.1007/S00586-010-1326-9/FIGURES/8.
URL https://link.springer.com/article/10.1007/s00586-010-1326-9

[33] J. J. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J. C. Fillion-Robin, S. Pieper, H. J. Aerts, Computational radiomics system to decode the radiographic phenotype, Cancer Research 77 (21) (2017) e104–e107. doi:10.1158/0008-5472.CAN-17-0339/SUPPLEMENTARY-VIDEO-S2.
URL https://aacrjournals.org/cancerres/article/77/21/e104/662617/Computational-Radiomics-System-to-Decode-the

[34] Radiomic Features — pyradiomics v3.0.1.post9+gdfe2c14 documentation.
URL https://pyradiomics.readthedocs.io/en/latest/features.html

[35] A. Zwanenburg, M. Vallières, M. A. Abdalah, H. J. Aerts, V. Andrearczyk, A. Apte, S. Ashrafinia, S. Bakas, R. J. Beukinga, R. Boellaard, M. Bogowicz, L. Boldrini, I. Buvat, G. J. Cook, C. Davatzikos, A. Depeursinge, M. C. Desseroit, N. Dinapoli, C. V. Dinh, S. Echegaray, I. El Naqa, A. Y. Fedorov, R. Gatta, R. J. Gillies, V. Goh,

M. Götz, M. Guckenberger, S. M. Ha, M. Hatt, F. Isensee, P. Lambin, S. Leger, R. T. Leijenaar, J. Lenkowicz, F. Lippert, A. Losnegård, K. H. Maier-Hein, O. Morin, H. Müller, S. Napel, C. Nioche, F. Orlhac, S. Pati, E. A. Pfaehler, A. Rahmim, A. U. Rao, J. Scherer, M. M. Siddique, N. M. Sijtsema, J. Socarras Fernandez, E. Spezi, R. J. Steenbakkers, S. Tanadini-Lang, D. Thorwarth, E. G. Troost, T. Upadhaya, V. Valentini, L. V. van Dijk, J. van Griethuysen, F. H. van Velden, P. Whybra, C. Richter, S. Löck, The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping, Radiology 295 (2) (2020) 328–338. doi:10.1148/RADIOL.2020191145.
URL https://pubmed.ncbi.nlm.nih.gov/32154773/

[36] P. Fontaine, O. Acosta, J. Castelli, R. De Crevoisier, H. Müller, A. Depeursinge, The importance of feature aggregation in radiomics: a head and neck cancer study, Scientific Reports 2020 10:1 10 (1) (2020) 1–11. doi:10.1038/s41598-020-76310-z.
URL https://www.nature.com/articles/s41598-020-76310-z

[37] K. Wakabayashi, Y. Koide, T. Aoyama, H. Shimizu, R. Miyauchi, H. Tanaka, H. Tachibana, K. Nakamura, T. Kodaira, A predictive model for pain response following radiotherapy for treatment of spinal metastases, Scientific Reports 11 (1) (2021) 1–8. doi:10.1038/s41598-021-92363-0.
URL https://www.nature.com/articles/s41598-021-92363-0

[38] R. Kochendörffer, Kreyszig, E.: Advanced Engineering Mathematics. J. Wiley & Sons, Inc., New York, London 1962. IX + 856 S. 402 Abb. Preis s. 79.—, Biometrische Zeitschrift 7 (2) (1965) 129–130. doi:10.1002/BIMJ.19650070232.
URL https://onlinelibrary.wiley.com/doi/10.1002/bimj.19650070232

[39] Stone M, Cross-Validatory Choice and Assessment of Statistical Predictions, Journal of the Royal Statistical Society. Series B (Methodological) 36 (2) (1974) 111–147. doi:https://doi.org/10.1111/j.2517-6161.1974.tb00994.x.
URL https://www.jstor.org/stable/2984809

[40] S. Rizzo, F. Botta, S. Raimondi, D. Origgi, C. Fanciullo, A. G. Morganti, M. Bellomi, Radiomics: the facts and the challenges of image analysis, European radiology

experimental 2 (1) (12 2018). doi:10.1186/S41747-018-0068-Z.
URL https://pubmed.ncbi.nlm.nih.gov/30426318/

[41] K. P. F.R.S., LIII. On lines and planes of closest fit to systems of points in space, https://doi.org/10.1080/14786440109462720 2 (11) (2010) 559–572. doi:10.1080/14786440109462720.
URL https://www.tandfonline.com/doi/abs/10.1080/14786440109462720

[42] A. Hyvärinen, E. Oja, Independent component analysis: algorithms and applications, Neural Networks 13 (4-5) (2000) 411–430. doi:10.1016/S0893-6080(00)00026-5.

[43] S. Das, U. Mert Cakmak, Hands-On Automated Machine Learning : a beginner's guide to building automated machine learning systems using AutoML and Python., 1st Edition, Packt Publishing, 2018.

[44] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, D. Gorinevsky, An Interior-Point Method for Large-ScalèScalè 1-Regularized Least Squares, IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING 1 (4) (2007). doi:10.1109/JSTSP.2007.910971.

[45] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene Selection for Cancer Classification using Support Vector Machines, Machine Learning 2002 46:1 46 (1) (2002) 389–422. doi:10.1023/A:1012487302797.
URL https://link.springer.com/article/10.1023/A:1012487302797

[46] L. Breiman, Random Forests, Machine Learning 2001 45:1 45 (1) (2001) 5–32. doi:10.1023/A:1010933404324.
URL https://link.springer.com/article/10.1023/A:1010933404324

[47] Fabian Pedregosa, Ga{{\"e}}l Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, {{\'E}}douard Duchesnay, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
URL http://scikit-learn.sourceforge.net.

[48] 1. Supervised learning — scikit-learn 0.20.4 documentation.
URL https://scikit-learn.org/0.20/supervised_learning.html

[49] 1.4. Support Vector Machines — scikit-learn 1.0.1 documentation.
URL https://scikit-learn.org/stable/modules/svm.html

[50] 1.9. Naive Bayes — scikit-learn 1.0.1 documentation.
URL https://scikit-learn.org/stable/modules/naive_bayes.html

[51] 1.6. Nearest Neighbors — scikit-learn 1.0.1 documentation.
URL https://scikit-learn.org/stable/modules/neighbors.html

[52] 1.2. Linear and Quadratic Discriminant Analysis — scikit-learn 1.0.1 documentation.
URL https://scikit-learn.org/stable/modules/lda_qda.html

[53] 1.7. Gaussian Processes — scikit-learn 1.0.1 documentation.
URL https://scikit-learn.org/stable/modules/gaussian_process.html

[54] 1.10. Decision Trees — scikit-learn 1.0.1 documentation.
URL https://scikit-learn.org/stable/modules/tree.html

[55] 1.11. Ensemble methods — scikit-learn 1.0.1 documentation.
URL https://scikit-learn.org/stable/modules/ensemble.html

[56] 1.17. Neural network models (supervised) — scikit-learn 1.0.1 documentation.
URL https://scikit-learn.org/stable/modules/neural_networks_supervised.html

[57] D. P. Kingma, J. L. Ba, Adam: A Method for Stochastic Optimization, 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings (12 2014).
URL https://arxiv.org/abs/1412.6980v9

[58] D. C. Liu, J. Nocedal, On the limited memory BFGS method for large scale optimization, Mathematical Programming 1989 45:1 45 (1) (1989) 503–528. doi:10.1007/BF01589116.
URL https://link.springer.com/article/10.1007/BF01589116

[59] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning - whole book, Nature 521 (7553) (2016) 800.
URL http://goodfeli.github.io/dlbook/%0Ahttp://dx.doi.org/10.1038/nature14539

[60] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters 27 (8) (2006) 861–874. doi:10.1016/J.PATREC.2005.10.010.

[61] J. D. Hunter, Matplotlib: A 2D graphics environment, Computing in Science and Engineering 9 (3) (2007) 90–95. doi:10.1109/MCSE.2007.55.

[62] Hossein Naseri, Sonia Skamene, Marwan Tolba, Mame Daro Faye, Paul Ramia, Julia Khriguian, Haley patrick, Aixa X. Andrade H., Marc David, John Kildea, A radiomics-based machine learning pipeline to distinguish between metastatic and healthy bone using lesion-center-based geometric regions of interest; dataset (3 2022). doi:https://doi.org/10.6084/m9.figshare.19224615.v1.
URL https://figshare.com/articles/dataset/A_radiomics-based_machine_learning_pipeline_to_distinguish_between_metastatic_and_healthy_bone_using_lesion-center-based_geometric_regions_of_interest_dataset/19224615/1

[63] B. Kocak, E. S. Durmaz, E. Ates, O. Kilickesmez, Radiomics with artificial intelligence: a practical guide for beginners, Diagnostic and interventional radiology (Ankara, Turkey) 25 (6) (2019) 485–495. doi:10.5152/DIR.2019.19321.
URL https://pubmed.ncbi.nlm.nih.gov/31650960/

[64] C. Haarburger, G. Müller-Franzes, L. Weninger, C. Kuhl, D. Truhn, D. Merhof, Radiomics feature reproducibility under inter-rater variability in segmentations of CT images, Scientific Reports 2020 10:1 10 (1) (2020) 1–10. doi:10.1038/s41598-020-69534-6.
URL https://www.nature.com/articles/s41598-020-69534-6

[65] B. Kocak, E. S. Durmaz, O. K. Kaya, E. Ates, O. Kilickesmez, Reliability of Single-Slice-Based 2D CT Texture Analysis of Renal Masses: Influence of Intra- and Interobserver Manual Segmentation Variability on Radiomic Feature Reproducibility, AJR. American

journal of roentgenology 213 (2) (2019) 377–383. doi:10.2214/AJR.19.21212.
URL https://pubmed.ncbi.nlm.nih.gov/31063427/

[66] J. Hua, Z. Xiong, J. Lowey, E. Suh, E. R. Dougherty, Optimal number of features as a function of sample size for various classification rules, Bioinformatics (Oxford, England) 21 (8) (2005) 1509–1515. doi:10.1093/BIOINFORMATICS/BTI171.
URL https://pubmed.ncbi.nlm.nih.gov/15572470/

[67] R. Tripathy, I. Bilionis, M. Gonzalez, Gaussian processes with built-in dimensionality reduction: Applications to high-dimensional uncertainty propagation, Journal of Computational Physics 321 (2016) 191–223. doi:10.1016/J.JCP.2016.05.039.

[68] C. E. Rasmussen, C. K. I. Williams, Gaussian Processes for Machine Learning, Gaussian Processes for Machine Learning (11 2005). doi:10.7551/MITPRESS/3206.001.0001.

[69] B. Krawczyk, Learning from imbalanced data: open challenges and future directions, Progress in Artificial Intelligence 5 (4) (2016) 221–232. doi:10.1007/S13748-016-0094-0/TABLES/1.
URL https://link.springer.com/article/10.1007/s13748-016-0094-0

[70] C. Xie, R. Du, J. W. Ho, H. H. Pang, K. W. Chiu, E. Y. Lee, V. Vardhanabhuti, Effect of machine learning re-sampling techniques for imbalanced datasets in 18 F-FDG PET-based radiomics model on prognostication performance in cohorts of head and neck cancer patients, European journal of nuclear medicine and molecular imaging 47 (12) (2020) 2826–2835. doi:10.1007/S00259-020-04756-4.
URL https://pubmed.ncbi.nlm.nih.gov/32253486/

[71] V. Kumar, Y. Gu, S. Basu, A. Berglund, S. A. Eschrich, M. B. Schabath, K. Forster, H. J. Aerts, A. Dekker, D. Fenstermacher, D. B. Goldgof, L. O. Hall, P. Lambin, Y. Balagurunathan, R. A. Gatenby, R. J. Gillies, Radiomics: the process and the challenges, Magnetic resonance imaging 30 (9) (2012) 1234–1248. doi:10.1016/J.MRI.2012.06.010.
URL https://pubmed.ncbi.nlm.nih.gov/22898692/

[72] Y. Sun, A. K. Wong, M. S. Kamel, CLASSIFICATION OF IMBALANCED DATA: A REVIEW, http://dx.doi.org/10.1142/S0218001409007326 23 (4) (2011) 687–719. doi:10.1142/S0218001409007326.

[73] H. He, Y. Ma, Imbalanced learning: Foundations, algorithms, and applications, Imbalanced Learning: Foundations, Algorithms, and Applications (2013) 1–210doi:10.1002/9781118646106.
URL https://onlinelibrary.wiley.com/doi/book/10.1002/9781118646106

[74] J. E. Bibault, P. Giraud, C. Durdux, J. Taieb, A. Berger, R. Coriat, S. Chaussade, B. Dousset, B. Nordlinger, A. Burgun, Deep Learning and Radiomics predict complete response after neo-adjuvant chemoradiation for locally advanced rectal cancer, Scientific Reports 2018 8:1 8 (1) (2018) 1–8. doi:10.1038/s41598-018-30657-6.
URL https://www.nature.com/articles/s41598-018-30657-6

[75] Y. He, I. Pan, B. Bao, K. Halsey, M. Chang, H. Liu, S. Peng, R. A. Sebro, J. Guan, T. Yi, A. T. Delworth, F. Eweje, L. J. States, P. J. Zhang, Z. Zhang, J. Wu, X. Peng, H. X. Bai, Deep learning-based classification of primary bone tumors on radiographs: A preliminary study, eBioMedicine 62 (2020) 103121. doi:10.1016/J.EBIOM.2020.103121.

Chapter 5

A scalable radiomics- and NLP- based machine learning pipeline to distinguish between painful and painless thoracic spinal bone metastases: Algorithm Development and Validation

Hossein Naseri, Sonia Skamene, Marwan Tolba, Mame Daro Faye, Paul Ramia, Julia Khriguian, Marc David and John Kildea

Article accepted to be published in: *JMIR AI* (**in press**). https://doi.org/10.2196/44779

5.1 Preface

This chapter describes the third objective of this thesis: Combining the NLP-quantified pain scores extracted in objective 1 with the radiomic features extracted in objective 2, to develop and evaluate a radiomics-based machine-learning model of pain in patients with BM. In this study, the NLP-quantified patient-level pain scores retrieved using the methodology described in Chapter 3 were combined with the lesion-level radiomic features extracted in Chapter 4 to create a radiomics-based machine-learning model to distinguish between painful and painless BM lesions.

5.2 Abstract

Background The identification of objective pain biomarkers can contribute to an improved understanding of pain, as well as its prognosis and better management. Hence, it has the potential to improve the quality of life of cancer patients. Artificial intelligence can aid in the extraction of objective pain biomarkers for cancer patients with bone metastases.

Purpose To develop and evaluate a scalable Natural Language Processing (NLP) and radiomics-based Machine Learning (ML) pipeline to differentiate between painless and painful Bone Metastases (BM) lesions in simulation-CT images using imaging features (biomarkers) extracted from lesion-centerpoint-based Regions Of Interest (ROIs).

Materials and Methods Patients treated at our comprehensive cancer center who received palliative radiotherapy for thoracic spine BM between January 2016 and September 2019 were included in this retrospective study. Physician-reported pain scores were extracted automatically from radiation oncology consultation notes using an NLP pipeline. BM centerpoints were manually pinpointed on CT images by radiation oncologists. Nested ROIs with various diameters were automatically delineated around these expert-identified BM centerpoints, and radiomics features were extracted from each ROI. The Synthetic Minority Oversampling Technique re-sampling technique, the Least Absolute Shrinkage and Selection Operator logistic regression algorithm (LASSO) feature selection method, and various ML classifiers were evaluated using precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (ROC-AUC).

Results Radiation therapy consultation notes and simulation-CT images of 176 (mean age ± SD, 66 ± 14 y; 95 male) thoracic spine BM patients were used in this study. After BM centerpoint identification, 107 radiomics features were extracted from each spherical ROI using pyradiomics. Data were divided into 70% and 30% training and hold-out test sets, respectively. In the test set, the accuracy, sensitivity, specificity, and ROC-AUC of our best performing model (Neural Network classifier on an ensemble ROI) were 0.82 (132 of 163), 0.59 (16 of 27), 0.85 (116 of 136), and 0.83, respectively.

Conclusion Our NLP and radiomics-based ML pipeline was successful in differentiating between painful and painless BM lesions. It is intrinsically scalable by using NLP to extract pain scores from clinical notes and by requiring just center points to identify BM lesions in CT images.

5.3 Introduction

5.3.1 Overview

Most cancer patients with Bone Metastases (BM) suffer from pain [1] and most receive radiotherapy to control it [2]. But it has been shown that clinicians often underestimate pain [3] and, as a result, many patients with BM receive radiotherapy after their pain has already become debilitating [4].

Although patient-reported outcomes can be used to obtain pain scores directly from patients themselves, the efficacy of these pain scores is limited due to the fact that these ratings are highly qualitative and subjective. Because of this, it is desirable to have pain scoring systems that are more objective. The goal of this study was to explore ways to automatically and objectively quantify pain associated with BMs using CT images.

We hypothesized that tumor features extracted from CT images of BMs contain imaging biomarkers that may be used to objectively identify BM-associated pain. These pain biomarkers may provide the opportunity to develop objective pain-scoring tools to aid in the diagnosis, treatment, understanding, and prognosis of BM pain.

5.4 Background

The search for imaging and non-imaging pain biomarkers has been the focus of numerous studies [5–12]. Various groups have shown how Machine Learning (ML) including Machine Learning (ML) and radiomics, can be used to understand and quantify pain [13–21]. For example, Mashayekhia et al. [22] showed that radiomic features extracted from the pancreas on CT images can help to identify patients with functional abdominal pain. Vedantam et al. [23] explored the viability of employing radiomics features extracted from MRI images to detect pain following percutaneous cordotomy. At least one paper [20] has reported using radiomics to identify painful metastatic lesions in radiographic images. However, we found no reports in the literature of a scalable approach that can be used efficiently on a large set of unlabeled patient data. To the best of our knowledge, our work is the first to combine Natural Language Processing and radiomics to enable an efficient and scalable pain identification pipeline using unstructured data.

A fundamental challenge in developing any AI model for use in medicine is the need to obtain sufficient patient data for training and testing. For example, the dataset used by Wakabayashi et al. [20], in the study that we mentioned earlier, was limited to 69 patients. One limiting factor is obtaining standard patient-reported pain scores for use as ground truth data, and another is obtaining segmented images from which to extract tumor biomarkers. For the work reported in this paper, we overcame the dataset size limitation by employing two novel strategies. First, by combining NLP with radiomics, we quickly mined pain scores from clinical notes and used these NLP-extracted scores to label our radiomics features for supervised learning. Second, by asking our clinical colleagues to pinpoint just the centerpoints of BM lesions in radiotherapy simulation-CT images we maximized the number of lesions identified in the time available. In the medical field, NLP has shown promising results in extracting biomedical information and clinical outcomes such as pain from unstructured text data [24–26]. Moreover, as we reported previously [21], by automatically delineating geometrical regions around BM lesion centerpoints, it is possible to successfully extract radiomics features for robust BM lesion detection. In the present study, we report how our combined radiomics-NLP ML pipeline can successfully identify pain in radiotherapy simulation-CT images of cancer patients with BMs.

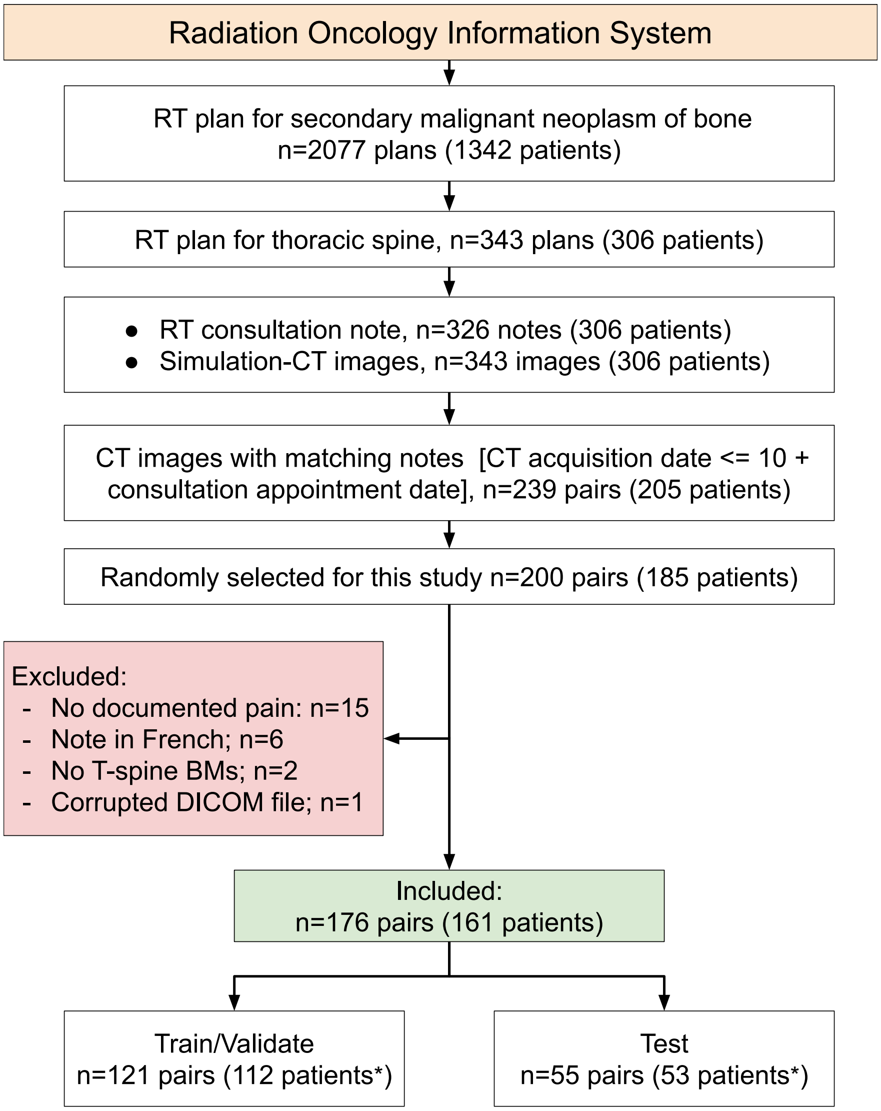
5.5 Methods and Materials

This retrospective study was approved by the Research Ethics Board (REB) of our institution with the waiver of informed consent. We confirm that the entire research was performed in accordance with REB guidelines and regulations.

5.5.1 Data selection

Our patient-selection process is outlined in Figure 5.1. The initial number of 200 pairs of radiation oncology consultation notes and CT images of patients with spinal BM were included in this study based on the minimum sample size calculation as explained in section A.1 of the supplementary information. 120 of the notes and all 200 of the CT images from this study were independently used in two studies we previously reported on [21, 25]. The first of these studies [25] showed the feasibility of extracting pain from consultation notes of cancer patients using NLP. The second [21] demonstrated the feasibility of using lesion-centerpoint based radiomics models to differentiate healthy and metastatic bone lesions in CT scans of patients with BMs. The current study combined the data and results from these two prior studies and expanded upon them to build an NLP- and radiomics-based model to detect pain using the CT scans of patients.

We searched our institution's Oncology Information System for the radiotherapy plans of patients diagnosed with a "secondary malignant neoplasm of bone" between January 2016 and September 2019. From the retrieved list, we selected those that treated thoracic spinal BM. Then, we retrieved the corresponding consultation notes and simulation-CT images. A note-image pair was included if (a) the note was in English, (b) pain was documented, (c) the simulation-CT image was taken up to 10 days post-consultation, and (d) the simulation-CT had BM lesions in the thoracic spine. Patients with multiple but non-overlapping note-image pairs were considered independent samples. Note that we only considered the same patients as new subjects if they had CT scans and associated consultation notes for BM lesions in different areas of their spines. As a result, each BM lesion was included only once in our study. Also, it should be noted that palliative patients normally have their simulation CT scan (for treatment planning) done the same day or within a few days after the consultation, and RT is delivered on the same day or within a few days after treatment planning. To assure that there is no change in the BM lesion structure or pain status, we did not allow more than a 10-day gap between the two. Figure 5.5 in the supporting information displays the distribution of the time interval between the RT consultation and CT acquisition dates.


Radiation Oncology Information System
RT plan for secondary malignant neoplasm of bone
n=2077 plans (1342 patients)
RT plan for thoracic spine, n=343 plans (306 patients)
● RT consultation note, n=326 notes (306 patients)
● Simulation-CT images, n=343 images (306 patients)
CT images with matching notes [CT acquisition date <= 10 + consultation appointment date], n=239 pairs (205 patients)
Randomly selected for this study n=200 pairs (185 patients)
Excluded:
- No documented pain: n=15
- Note in French; n=6
- No T-spine BMs; n=2
- Corrupted DICOM file; n=1
Included:
n=176 pairs (161 patients)
Train/Validate
n=121 pairs (112 patients*)
Test
n=55 pairs (53 patients*)

* Four patients had pairs in both the train and test sets.

Figure 5.1: The patient selection criteria used to obtain the radiotherapy consultation notes and simulation-CT images that formed our training and test datasets. The initial number of 200 note-image pairs included in this study was based on the minimum sample size calculation as explained in Section A.1 in the supplementary information. RT: radiotherapy.

We randomly assigned note-image pairs to the training/cross-validation set (approximately 70%) or the hold-out test set (approximately 30%). We used stratified randomization to preserve the original sample ratio between pain labels in each sample set. In addition, we performed a t-test and a chi-square analysis [27] to ensure that there was no systematic bias in any of our sample sets regarding gender, age, or primary cancer type. Patient demographics are presented in Table 5.1.

5.5.2 NLP-extracted pain labels

Due to the absence of patient-reported pain scores in our oncology information system, we extracted physician-reported pain scores from patients' radiation oncology consultation notes using our previously-reported NLP pipeline [25]. While pain scores were typically reported as part of the "history of the present illness" in our hospital, for the sake of generalizability, we extracted pain scores from the entire note.

Our NLP pipeline first processed the text with MetaMap [28] and mapped it to the UMLS metathesaurus [29] in order to identify pain terminologies and their severity scores. Next, it applied rules to filter out hypothetical, conditional, and historical references to pain in order to focus solely on references to pain at the time of the consultation. Then, it calculated the average pain intensity (API) in each note by averaging the pain scores therein. Finally, it assigned each note a Verbally-Declared Pain (VDP) label, as VDP=‘no pain’ (if $\text{API} < 0$), and VDP=‘pain’ (if $\text{API} > 0$). These pain labels were used to train, validate, and test our radiomics model.

5.5.3 Expert-extracted pain scores

To evaluate the effect of NLP-extracted pain labels on the performance of our pipeline, we also generated best-available ground-truth pain labels using expert-annotated pain scores. To do so, our radiation oncologists used the texTRACTOR [30] pain labeling application to manually read consultation notes and label valid pain scores in our training and test datasets using a 4-grade verbal rating scale (no pain, mild, moderate, severe). A mention of pain was regarded as valid if it reflected the status of pain at the metastatic sites for which treatment was planned at the time of the consultation. Table A1, in the supplementary information, contains all the NLP- and expert-extracted pain scores, and Figure 5.6 illustrates the level

VARIABLE	TRAIN/VALIDATE	TEST	P-VALUE
Number of samples (n)	121	55	
Female	56 (46%)	25 (45%)	
Male	65 (54%)	30 (55%)	
AGE (Mean±SD), years			
Female	63±14	64±12	0.99
Male	67±14	64±13	0.72
p-value	0.2	0.5	
PRIMARY CANCER			0.06
Lung	32 (26%)	20 (36%)	
Breast	23 (19%)	11 (20%)	
Prostate	19 (16%)	5 (9%)	
Multiple Myeloma	8 (7%)	6 (11%)	
Renal Cell Carcinoma	7 (6%)	2 (4%)	
Other and Unknown	64 (53%)	31 (56%)	
BM LESIONS			0.42
Lytic	220 (52%)	76 (47%)	
Blastic	122 (29%)	57 (35%)	
Mix	81 (19%)	30 (18%)	
PAIN LABEL			
Pain	357 (84%)	136 (83%)	
No pain	66 (16%)	27 (17%)	

Table 5.1: Patient demographics in the training and test sets. P-values for numerical values (age) and categorical features (primary cancer site and BM lesion type) are calculated using a two-tailed heteroscedastic t-test and a chi-square test, respectively.

of agreement between them. Due to the quality of the documented pain scores and lack of inter-rater agreement among experts (Fleiss' $\kappa = 0.43$), as explained in [25], we subsequently defined a binary pain score as 'no-pain' and 'pain' in order to establish satisfactory inter-rater agreement ($\kappa = 0.66$) [25]. To create binary ground-truth pain labels comparable to the NLP-extracted labels, we assigned notes scored as 'no pain' to 'no pain' and notes scored as 'mild', 'moderate' and 'severe' pain to 'pain'.

These expert-extracted pain scores were used to measure how well the NLP pipeline works.

5.5.4 Centerpoint identification of BM lesions

BM lesion centerpoints were identified by a team comprising one staff radiation oncologist with 10 years' experience, one radiation oncology fellow, and three 3rd-year radiation oncology residents. Simulation-CT DICOM files were exported from the radiotherapy treatment planning software and de-identified. Then, the CTs were randomly divided into five sets and loaded into the diCOMBINE [31] application for BM lesion centerpoint identification. Our experts were blinded to patients' pain statuses and identities. We requested each to label centerpoints for all visually identifiable BM lesions in all CTs within one of the five sets, and another expert was assigned to validate their labels. A key benefit of this radiomics pipeline [21] is that it does not require full lesion segmentation, making it feasible to engage busy clinicians.

5.5.5 Segmentation of regions of interest

Using our previously-reported methodology [21], we automatically segmented lesion-centerpoint-based nested spherical ROIs. To do this, we first delineated nested spherical (SP) ROIs around the identified BM lesion centerpoints (see Table 5.2, top panel). ROI diameters ranged from 7 mm (3x3 voxels) to 50 mm (average size of the vertebral body [32]). Then, in addition to what was done by Naseri et al. [21], we used Hounsfield units thresholding to exclude fat and air regions from the delineated ROIs . For this, motivated by [33, 34], we applied a threshold to remove voxels with negative Hounsfield units from our ROIs . Hounsfield units less than zero are associated with fat and air [33]. We used OpenCV (version 4.4.0) [35] for Hounsfield units thresholding and applied a Gaussian filter to reduce noise. Then, we used pynrrd (version 0.4.2) [36] to export each ROI as a 3D binary mask and store it as a .nrrd [37] file. Finally, we aggregated these nested ROI masks to form ensemble ROIs . In this study, we examined two contrasting ensemble ROIs as shown in Table 5.2 (bottom panel), one with small size and three layers (EN3) and the other with large size and six layers (EN6). Prior studies [20, 21] have shown that radiomics-based ML models trained on ensemble ROIs have better classification performance compared to single ROI-based models.

	Nested spherical ROIs with Hounsfield units intensity threshold (HU>0)					
Name	SP7	SP10	SP15	SP20	SP30	SP50
Diameter (mm)	7	10	15	20	30	50
	Ensemble ROIs					
Name	EN3			EN6		
Aggrigated ROIs	SP7+SP10+SP15			SP7+SP10+SP15 +SP20+SP30+SP50		

Table 5.2: The characteristics of the spherical and ensemble Regions Of Interest (ROIs) used in this study. HU, Hounsfield units; SP, spherical; EN, ensemble.

5.5.6 Radiomics models

Our radiomics pipeline is illustrated in Figure 5.2. We essentially used our previously-reported pipeline [21] but with our NLP- and expert-extracted pain labels to train and test it. We made one improvement to the pipeline by incorporating Imbalanced-learn (version 0.7.0) [38] as a re-sampling step to account for imbalance (see below).

Radiomics features were extracted from each CT image using masks composed of the ensemble ROIs listed in Table 5.2. Then, the feature space was scaled using z-score normalization [39], and the associated NLP-extracted binary pain labels (pain=1, no pain=0) were incorporated. A single NLP-extracted pain score was assigned to all the lesions extracted from a given paired CT image.

Due to the nature of BM pain [40], there was a large imbalance between the number of painful and painless lesions (493 pain: 93 no pain). Therefore, we used the Synthetic Minority Oversampling Technique (SMOTE) in the training phase as it has been shown

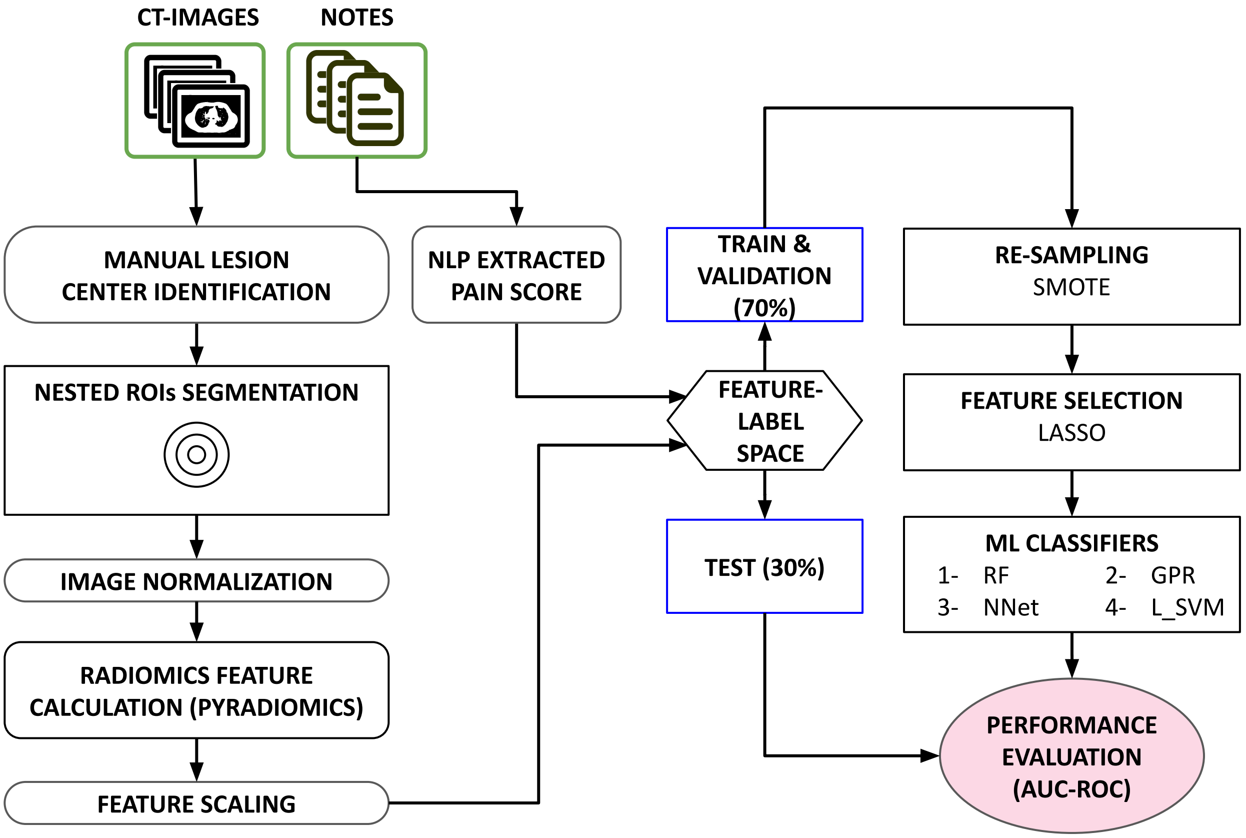


Figure 5.2: The radiomics-based pipeline that we used to select and train a ML model to separate painful and painless BM lesions. Our pipeline is the same as that published by Naseri et al. [21] but using NLP-extracted pain labels and modified to account for sample imbalance.

to be the best-performing re-sampling method for radiomics [41]. We did not apply re-sampling to our test set in order to maintain the original sample imbalance. Then, the Least Absolute Shrinkage and Selection Operator logistic regression algorithm (LASSO) [42] feature selection method was applied to the feature space to remove non-informative features. LASSO is a commonly-used feature selection method in radiomics studies [43, 44]. Finally, we examined the Gaussian Process Regression, Linear Support Vector Machine, Random Forest and Neural Networks classifiers, as they were the best performing ML classifiers in our previous work [21]. We evaluated the performance of our models on the training set using

5-fold cross-validation. Final evaluation was performed on the test set. The Area Under the Receiver Operating Characteristic Curve (ROC-AUC), precision, sensitivity, specificity, and F1-score metrics were used to report the performance of our models on the training and test sets. We also trained and tested our best performing pipeline using the expert-extracted pain scores (best available ground truth) to evaluate the impact of NLP-extracted pain labels.

5.6 Results

5.6.1 Patient demographics

A total of 176 pairs of radiotherapy consultation notes and simulation-CT images of thoracic spinal BM patients were included in this study. As summarized in Table 5.1, 121 sample pairs (mean patient age±SD, female: 63±14y; male: 67±14y; p=0.2, 56 male) were used for training and cross-validation, and 55 sample pairs (mean patient age±SD, female: 64±12y; male 64±13y; p=0.5, 25 male) were used as the test set. The sample selection procedure and data quantities are presented in Figure 5.1. The demographics of the patients in the training and test sets are presented in Table 5.1. The most common primary cancer sites were lung (n=52), breast (n=34), and prostate (n=24).

A total of 586 BM centerpoints were identified by our experts on the training (n=423 lesions) and test (n=163 lesions) datasets. In the training set, 357 (84%) lesions were labeled by the NLP pipeline as painful and 66 lesions were labeled as painless. In the test set, 136 (83%) lesions were identified by the NLP pipeline as painful, and 27 lesions were labeled as painless. This represented a significant but equal imbalance in our training and test sets.

5.6.2 Segmented ROIs

Examples of segmented ROIs with the Hounsfield units threshold applied are presented in Figure 5.3 for painful and painless BMs.

5.6.3 Testing our radiomics models

107 radiomics features were extracted from each of the six nested ROIs . Then, they were aggregated to form feature spaces for the EN3 (with 321 features) and EN6 (with 642

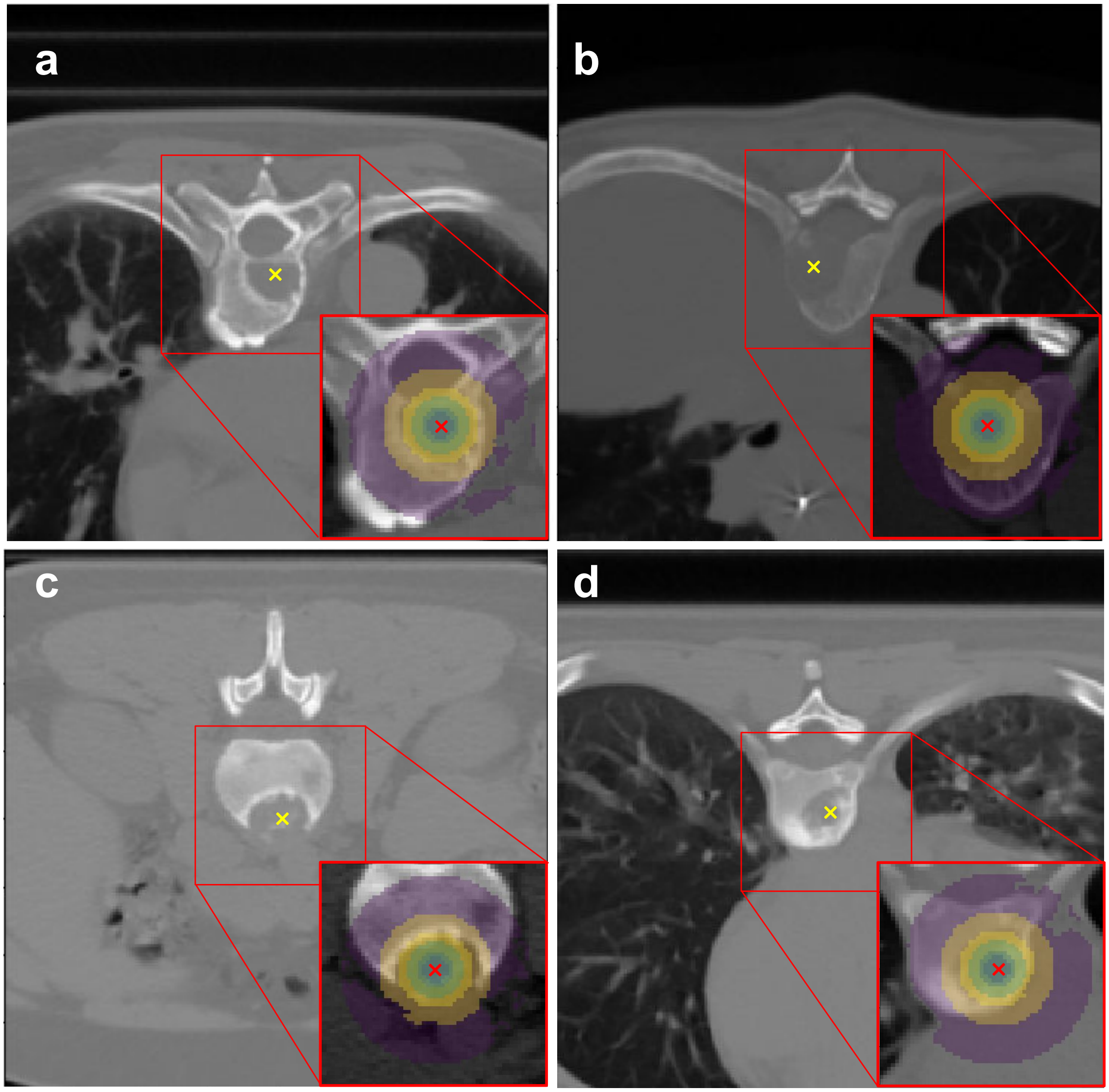


Figure 5.3: Examples of segmented nested spherical Regions Of Interest (ROIs) with the Hounsfield units threshold applied on CT images of patients with painful (a, b) and painless (c, d) bone metastases lesions. Nested ROIs with diameters of 50, 30, 20, 15, 10, and 7 mm are shown in the insets as different hues.

features) ensemble ROIs . Figure 5.4 shows the ROC curve of each model in the training (black lines) and test (red squares) datasets using the EN3 and EN6 ensemble ROIs . On the training set, the gray range represents the mean ROC ± SD of the 5-fold cross-validation. The ROC-AUC and F1-score grids are presented in Table 5.3.

	ROC-AUC grid					F1-score grid			
	Train					Train			
EN3	98.3%	98.1%	84.7%	94.6%	EN3	90.0%	89.9%	79.4%	90.5%
EN6	98.1%	98.3%	89.8%	94.0%	EN6	93.0%	93.0%	84.7%	91.6%
	RF	GPR	L-SVM	NNet		RF	GPR	L-SVM	NNet
	Test					Test			
EN3	67.3%	72.1%	75.2%	73.3%	EN3	60.9%	64.7%	65.4%	63.6%
EN6	74.1%	80.6%	82.4%	82.5%	EN6	63.8%	66.9%	67.4%	69.5%
	RF	GPR	L-SVM	NNet		RF	GPR	L-SVM	NNet

Table 5.3: The ROC-AUCs and F1-scores of our ML classifiers in the training and test datasets using the EN3 and EN6 ensemble ROIs for each of the RF (Random Forest); GPR (Gaussian Process Regression); L-SVM (Linear Support Vector Machine); NNet (Neural Networks) classifiers.

The precision, accuracy, sensitivity, specificity, F1 score, and ROC-AUC values of our best performing pipeline (neural networks with EN6 ROI) are presented in Table 5.4. The performance of this pipeline (trained and tested) on the dataset of expert-extracted pain labels (best-available ground truth) is provided as a quality measurement. The performance of the model from the previously-described prior study by Wakabayashi et al. [20] is also provided for comparison.

5.7 Discussion

Underestimation and under-treatment of cancer pain can significantly diminish cancer patients' quality of life. Accordingly, systems that can objectively measure cancer pain have the potential to improve quality of life. In this study, we created an scalable NLP-radiomics pain identification pipeline. Our pipeline is designed for palliative intent cancer patients undergoing RT therapy, for whom there are typically just two contemporaneous sources of relevant medical information at the time of the treatment: consultation notes and simulation-CT images. We used an NLP pipeline to extract

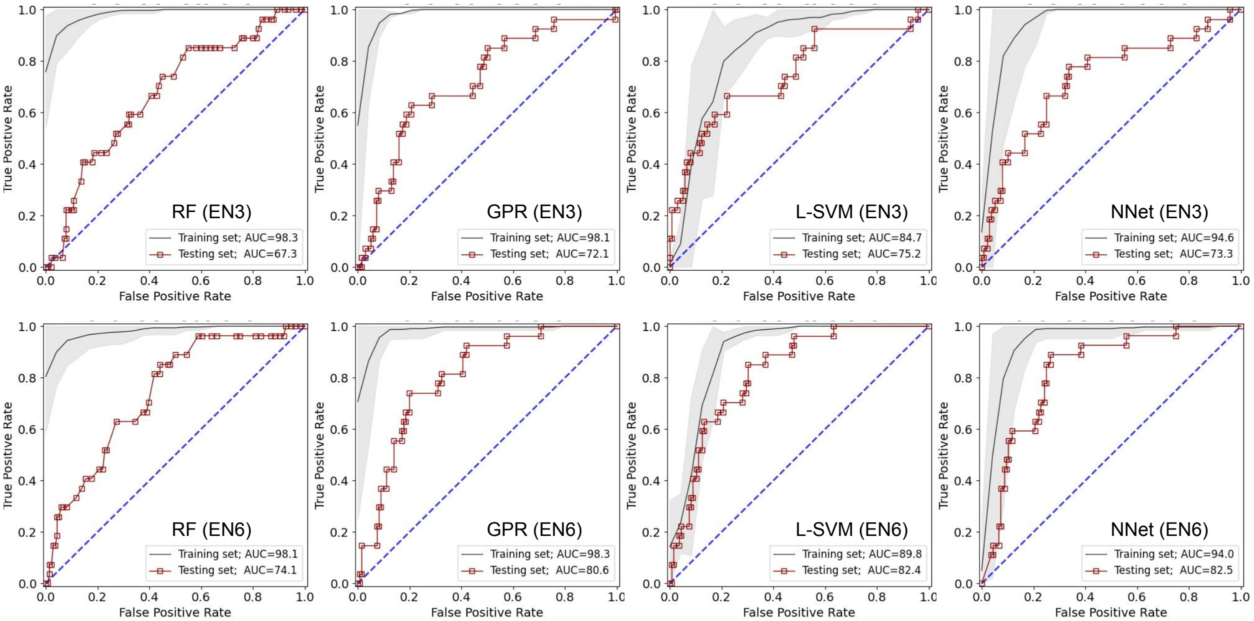


Figure 5.4: ROC curves for our classifiers using three-layer (EN3) (top row) and six-layer (EN6) (bottom row) lesion-centerpoint-based ensemble ROIs in training (black lines) and test (dark red squares) datasets.

physician-reported pain scores from radiotherapy consultation notes. NLP-extracted pain scores are appropriate, whenever structured patient-reported pain scores are unavailable (as is the case for at least 25% to 35% of all cancer patients [20, 45] and for all of the palliative cancer patients treated with RT at our institution at the time the data used in this study). Our lesion-centerpoint-based spherical ROI delineation method significantly sped up the ROI segmentation procedure, enabling us to rapidly delineate BM centerpoints in 176 images for this study. For comparison, the radiomics pipeline that was developed by Wakabayashi et al. [20] required full 3D segmentation of each ROI (69 images).

Due to the unbalanced nature of BM pain, our dataset contained significantly fewer painless samples. In order to better train our models, we applied the SMOTE re-sampling technique to the training set to balance the number of samples with the NLP-extracted 'pain' and 'no pain' labels. We did not apply any re-sampling techniques to our test (hold out) set to maintain the original sample imbalance. Therefore, while our training set was balanced, our test set had five times more 'pain' cases than 'no pain' cases (136 pain versus 27 no pain).

	Accuracy	Precision	Sensitivity	Specificity	F1	ROC-AUC
Current Study (Train)	92.4%	93.2%	92.4%	86.4%	91.6%	94.0%
Current Study (Test)	81.0%	67.9%	59.2%	85.3%	69.5%	82.5%
Current Study (Train), Using Manual Pain Scores	94.2%	94.8%	98.7%	89.7%	94.4%	98.1%
Current Study (Test), Using Manual Pain Scores	83.5%	64.9%	64.7%	85.7%	68.0%	82.3%
Wakabayashi et al. [20] (Train-only)	73.9%		71.0%	86.0%		82.0%

Table 5.4: The performance of our best performing NLP-radiomics pipeline (neural networks with the EN6 ROI) on the training and test sets. The results of the same radiomics model (neural networks with EN6 ROI) when trained and tested using the best-available ground-truth expert-extracted (EE) pain labels, together with the results from a prior study by Wakabayashi et al. [20] are provided for comparison. The reason for having high specificity and low sensitivity in our test set is explained in the Discussion.

This caused a significant change in the pipeline's performance between training and test sets. It has been shown that oversampling improves the overall performance of ML models, but the effect is stronger on the training set due to the inclusion of replicated samples in the cross-validation subsets [46]. Moreover, the imbalance in our test set led to high specificity (ability to properly identify pain instances) and low sensitivity (ability to correctly identify no pain cases) in the performance evaluation. For comparison, the sample imbalance in the study conducted by Wakabayashi et al. was 2:1, resulting in a more balanced relationship between the sensitivity and specificity of their model.

The performance of our pipeline did not improve much when we trained and tested it using expert-extracted pain labels (best-available ground truth). This might be because, in the first experiment, we both trained and tested our pipeline using NLP-extracted pain labels, and in the second experiment, we both trained and tested our pipeline using expert-extracted pain labels. Consequently, after being trained with one set of labels (NLP- or expert-extracted), our pipeline performed well on the test set that was labeled using the same method (NLP or expert). We also demonstrated that our pipeline's performance is comparable to that of Wakabayashi et al., [20] who achieved their results using patient-reported pain labels.

Our pipeline performed significantly better on the EN6 ensemble ROIs compared to the EN3 ROIs . This could be because, in comparison to EN3, our EN6 ensemble ROIs include additional ROIs with sizes of 20, 30, and 50 mm. From visual inspection, we suspect that, in addition to the characteristics of the BM lesion itself, its location (for example, its proximity to the spinal cord) may be a significant contributor to the BM pain. As a result, larger ROIs enable our algorithm to extract characteristics from outside the BM lesion. Wakabayashi et al. also demonstrated the effectiveness of using ROIs outside of the BM lesion.

We are unable to offer a convincing explanation as to why neural networks outperformed random forest and support vector machine classifiers in our analysis. Notwithstanding, it has been demonstrated that neural network classifiers perform better when applied to more difficult problems and larger datasets, while random forest and support vector machine classifiers typically perform well with smaller datasets [44, 47, 48].

Our pipeline was successful in extracting radiomics biomarkers capable of distinguishing between painful and painless BM lesions. These biomarkers potentially provide the opportunity to objectively identify clinical pain-related indicators that may aid in the diagnosis, treatment, and understanding of BM pain.

Our work has several limitations. In the first place, we used data from a single center for this retrospective study. A multicenter study with a larger dataset is necessary to assess the generalizability of our radiomics pipeline for pain quantification. We anticipate that the performance of our NLP-radiomics pipeline will vary based on the pain scoring systems of the cohorts tested. Second, by utilizing lesion-centerpoint-based geometrical ROIs , we ignored lesion characteristics such as size and shape, which may be important in the context of pain. Although we employed Hounsfield units intensity thresholding to preserve some tumor information, we are considering implementing deep-learning-based ROI segmentation in the future as it may better account for full tumor and surrounding tissue characteristics. Lastly, we utilized SMOTE oversampling to address the issue of class imbalance. An alternative solution might be to develop cost-sensitive ML classifiers that account for the cost of misclassifying minority samples [49]. However, there is no clear consensus in the literature on whether cost-sensitive learning outperforms re-sampling [50].A model that can differentiate between painful and painless lesions from medical imaging is a critical component of any possible radiomics-based pain quantification pipeline. The current work not only shows the feasibility of developing a pain quantification tool, but it also removes some of the barriers to its development. As a result, our future work will be to apply our pipeline to patients' past and current CT images and consultation notes in order to develop a longitudinal model of pain. Such a model should take into account not only images (taken before, during, and after delivering RT) but also other internal and external parameters that can influence how pain evolves over time (such as primary cancer type, radiation dose, other treatments, and pain medications). Also it will include patient-reported pain scores to provide more accurate ground-truth pain labels in order to develop a more robust deep learning-based NLP pipeline [26, 51]. This, however, is outside the scope of the current investigation.

In conclusion, we demonstrated that our NLP and radiomics-based ML pipeline can effectively differentiate between painful and painless BM lesions in simulation-CT images using ensemble lesion-centerpoint-based geometrical ROIs . Using NLP-extracted pain labels in conjunction with lesion-centerpoint-based radiomics features is time efficient. This helps to pave the way for the development of quickly-trained and efficient clinical artificial intelligence-based decision-making tools that can objectively measure cancer pain. Such a tool that may help alleviate the burden of pain management and improve the quality of life of patients with BMs.

5.7.1 Acknowledgments

This research was supported by the startup grant of Dr. John Kildea at the Research Institute of the McGill University Health Centre (RI-MUHC), the Ruth and Alex Dworkin scholarship award from the Faculty of Medicine and Health Sciences at McGill University, an RI-MUHC studentship award, a Grad Excellence Award-00293 from the Department of Physics at McGill University, Fonds de recherche du Québec - Santé (FRQS), and by the CREATE Responsible Health and Healthcare Data Science (SDRDS) grant of the Natural Sciences and Engineering Research Council. The authors would like to thank Dr. Luc Galarneau for his help with statistical analysis.

Bibliography

[1] M. H. Van Den Beuken-Van Everdingen, L. M. Hochstenbach, E. A. Joosten, V. C. Tjan-Heijnen, D. J. Janssen, Update on Prevalence of Pain in Patients With Cancer: Systematic Review and Meta-Analysis, Journal of pain and symptom management 51 (6) (2016) 1070–1090. doi:10.1016/J.JPAINSYMMAN.2015.12.340.
URL https://pubmed.ncbi.nlm.nih.gov/27112310/

[2] H. J. Mcquay, S. L. Collins, D. Carroll, R. A. Moore, S. Derry, Radiotherapy for the palliation of painful bone metastases, The Cochrane Database of Systematic Reviews 2013 (11) (11 2013). doi:10.1002/14651858.CD001793.PUB2.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6564087/

[3] S. A. Grossman, Undertreatment of cancer pain: barriers and remedies, Supportive care in cancer : official journal of the Multinational Association of Supportive Care in Cancer 1 (2) (1993) 74–78. doi:10.1007/BF00366899.
URL https://pubmed.ncbi.nlm.nih.gov/7511475/

[4] C. S. Cleeland, N. A. Janjan, C. B. Scott, W. F. Seiferheld, W. J. Curran, Cancer pain management by radiotherapists: a survey of radiation therapy oncology group physicians, International journal of radiation oncology, biology, physics 47 (1) (2000) 203–208. doi:10.1016/S0360-3016(99)00276-X.
URL https://pubmed.ncbi.nlm.nih.gov/10758325/

[5] X. Xu, Y. Huang, Objective Pain Assessment: a Key for the Management of Chronic Pain, F1000Research 9 (2020). doi:10.12688/F1000RESEARCH.20441.1.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6979466/

[6] A. B. Niculescu, H. Le-Niculescu, D. F. Levey, K. Roseberry, K. C. Soe, J. Rogers, F. Khan, T. Jones, S. Judd, M. A. McCormick, A. R. Wessel, A. Williams, S. M. Kurian, F. A. White, Towards precision medicine for pain: diagnostic biomarkers and repurposed drugs, Molecular Psychiatry 2019 24:4 24 (4) (2019) 501–522. doi:10.1038/s41380-018-0345-5.
URL https://www.nature.com/articles/s41380-018-0345-5

[7] M. M. Diaz, J. Caylor, I. Strigo, I. Lerman, B. Henry, E. Lopez, M. S. Wallace, R. J. Ellis, A. N. Simmons, J. R. Keltner, Toward Composite Pain Biomarkers of Neuropathic Pain—Focus on Peripheral Neuropathic Pain, Frontiers in Pain Research 0 (2022) 67. doi:10.3389/FPAIN.2022.869215.

[8] A. Furfari, B. A. Wan, K. Ding, A. Wong, L. Zhu, A. Bezjak, R. Wong, C. F. Wilson, C. DeAngelis, A. Azad, E. Chow, G. S. Charames, Genetic biomarkers associated with pain flare and dexamethasone response following palliative radiotherapy in patients with painful bone metastases, Annals of Palliative Medicine 6 (2) (2018) S240–S247. doi:10.21037/APM.2017.09.04.
URL https://apm.amegroups.com/article/view/16557

[9] J. Gunn, M. M. Hill, B. M. Cotten, T. R. Deer, Observational Study An Analysis of Biomarkers in Patients with Chronic Pain, Pain Physician 23 (1) (2020) 41–49.
URL https://pubmed.ncbi.nlm.nih.gov/32013287/

[10] A. Marchi, R. Vellucci, S. Mameli, A. R. Piredda, G. Finco, Pain Biomarkers, Clinical Drug Investigation 2009 29:1 29 (1) (2012) 41–46. doi:10.2165/0044011-200929001-00006.
URL https://link.springer.com/article/10.2165/0044011-200929001-00006

[11] Y. Ota, M. Connolly, A. Srinivasan, J. Kim, A. A. Capizzano, T. Moritani, Mechanisms and origins of spinal pain: From molecules to anatomy, with diagnostic clues and imaging findings, Radiographics 40 (4) (2020) 1163–1181. doi:10.1148/RG.2020190185/ASSET/IMAGES/LARGE/RG.2020190185.FIG15.JPEG.
URL https://pubs.rsna.org/doi/10.1148/rg.2020190185

[12] I. Tracey, C. J. Woolf, N. A. Andrews, Composite Pain Biomarker Signatures for Objective Assessment and Effective Treatment, Neuron 101 (5) (2019) 783–800. doi:10.1016/J.NEURON.2019.02.019.
URL https://pubmed.ncbi.nlm.nih.gov/30844399/

[13] L. A. Carlson BA, W. Michael Hooten, Pain-Linguistics and Natural Language Processing., Mayo Clinic proceedings. Innovations, Quality & Outcomes 4 (3) (2020) 346–347. doi:10.1016/J.MAYOCPIQO.2020.01.005.
URL https://europepmc.org/article/med/32542226

[14] X. Wang, P. Vp, A. D. Dave, G. Ruaño, J. Kost, Automated Extraction of Pain Symptoms: A Natural Language Approach using Electronic Health Records, Pain Physician 25 (2150-1149) (2022) E245–E246.
URL www.painphysicianjournal.com

[15] P. J. Tighe, B. Sannapaneni, R. B. Fillingim, C. Doyle, M. Kent, B. Shickel, P. Rashidi, Forty-two Million Ways to Describe Pain: Topic Modeling of 200,000 PubMed Pain-Related Abstracts Using Natural Language Processing and Deep Learning-Based Text Generation, Pain medicine (Malden, Mass.) 21 (11) (2020) 3133–3160. doi:10.1093/PM/PNAA061.
URL https://pubmed.ncbi.nlm.nih.gov/32249306/

[16] M. Matsangidou, A. Liampas, M. Pittara, C. S. Pattichi, P. Zis, Machine Learning in Pain Medicine: An Up-To-Date Systematic Review, Pain and Therapy 10 (2) (2021) 1067. doi:10.1007/S40122-021-00324-2.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8586126/

[17] K. I. Neijenhuijs, C. F. Peeters, H. van Weert, P. Cuijpers, I. V. d. Leeuw, Symptom clusters among cancer survivors: what can machine learning techniques tell us?, BMC Medical Research Methodology 21 (1) (2021) 1–12. doi:10.1186/S12874-021-01352-4/FIGURES/2.
URL https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-021-01352-4

[18] J. H. Hong, J. Y. Jung, A. Jo, Y. Nam, S. Pak, S. Y. Lee, H. Park, S. E. Lee, S. Kim, Development and validation of a radiomics model for differentiating bone islands and osteoblastic bone metastases at abdominal CT, Radiology 299 (3) (2021) 626–632. doi:10.1148/RADIOL.2021203783/ASSET/IMAGES/LARGE/RADIOL.2021203783.VA.JPEG.
URL https://pubs.rsna.org/doi/abs/10.1148/radiol.2021203783

[19] W. Sun, S. Liu, J. Guo, S. Liu, D. Hao, F. Hou, H. Wang, W. Xu, A CT-based radiomics nomogram for distinguishing between benign and malignant bone tumours, Cancer Imaging 21 (1) (2021) 1–10. doi:10.1186/S40644-021-00387-6/FIGURES/4.
URL https://cancerimagingjournal.biomedcentral.com/articles/10.1186/s40644-021-00387-6

[20] K. Wakabayashi, Y. Koide, T. Aoyama, H. Shimizu, R. Miyauchi, H. Tanaka, H. Tachibana, K. Nakamura, T. Kodaira, A predictive model for pain response following radiotherapy for treatment of spinal metastases, Scientific Reports 11 (1) (2021) 1–8. doi:10.1038/s41598-021-92363-0.
URL https://www.nature.com/articles/s41598-021-92363-0

[21] H. Naseri, S. Skamene, M. Tolba, M. D. Faye, P. Ramia, J. Khriguian, H. Patrick, A. X. Andrade Hernandez, M. David, J. Kildea, Radiomics-based machine learning models to distinguish between metastatic and healthy bone using lesion-center-based geometric regions of interest, Scientific Reports 12 (1) (2022) 1–13. doi:10.1038/s41598-022-13379-8.
URL https://www.nature.com/articles/s41598-022-13379-8

[22] R. Mashayekhi, V. S. Parekh, M. Faghih, V. K. Singh, M. A. Jacobs, A. Zaheer, Radiomic features of the pancreas on CT imaging accurately differentiate functional abdominal pain, recurrent acute pancreatitis, and chronic pancreatitis, European Journal of Radiology 123 (2020) 108778. doi:10.1016/J.EJRAD.2019.108778.

[23] A. Vedantam, I. Hassan, A. Kotrotsou, A. Hassan, P. O. Zinn, A. Viswanathan, R. R. Colen, Magnetic Resonance-Based Radiomic Analysis of Radiofrequency Lesion Predicts Outcomes After Percutaneous Cordotomy: A Feasibility Study, Operative neurosurgery (Hagerstown, Md.) 18 (6) (2020) 721–727. doi:10.1093/ONS/OPZ288.
URL https://pubmed.ncbi.nlm.nih.gov/31665446/

[24] M. Elbattah, E. Arnaud, M. Gignon, G. Dequen, The role of text analytics in healthcare: A review of recent developments and applications, HEALTHINF 2021 - 14th International Conference on Health Informatics; Part of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2021 (2021) 825–832doi:10.5220/0010414508250832.

[25] H. Naseri, K. Kafi, S. Skamene, M. Tolba, M. D. Faye, P. Ramia, J. Khriguian, J. Kildea, Development of a generalizable natural language processing pipeline to extract physician-reported pain from clinical reports: Generated using publicly-available datasets and tested on institutional clinical reports for cancer patients with bone metastases, Journal of biomedical informatics 120 (8 2021). doi:10.1016/J.JBI.2021.103864.
URL https://pubmed.ncbi.nlm.nih.gov/34265451/

[26] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference 1 (2018) 4171–4186. doi:10.48550/arxiv.1810.04805.
URL https://arxiv.org/abs/1810.04805v2

[27] D. Freedman, R. Pisani, R. Purves, Statistics (1998) 123.
URL https://books.google.com/books/about/Statistics.html?id=mviJQgAACAAJ

[28] A. R. Aronson, a. nih gov, Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program., Proceedings of the AMIA Symposium (2001) 17.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243666/

[29] A. T. McCray, A. R. Aronson, A. C. Browne, T. C. Rindflesch, A. Razi, S. Srinivasan, UMLS knowledge for biomedical language processing., Bulletin of the Medical Library Association 81 (2) (1993) 184.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC225761/

[30] H. Naseri, texTRACTOR: an NLP tool to extract physician-reported pain scores from clinical notes, Zenodo (3 2021). doi:10.5281/ZENODO.4649625.
URL https://zenodo.org/record/4649625

[31] H. Naseri, diCOMBINE: 3D-DICOM Visualization and Lesion Identification Web Application (8 2021). doi:10.5281/ZENODO.5218743.
URL https://zenodo.org/record/5218743

[32] I. Busscher, J. J. Ploegmakers, G. J. Verkerke, A. G. Veldhuizen, Comparative anatomical dimensions of the complete human and porcine spine, European Spine Journal 19 (7) (2010) 1104–1114. doi:10.1007/S00586-010-1326-9/FIGURES/8.
URL https://link.springer.com/article/10.1007/s00586-010-1326-9

[33] H. J. Deglint, R. M. Rangayyan, F. J. Ayres, G. S. Boag, M. K. Zuffo, Three-Dimensional Segmentation of the Tumor in Computed Tomographic Images of Neuroblastoma, Journal of Digital Imaging 20 (1) (2007) 72. doi:10.1007/10278-006-0769-3.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3043888/

[34] A. Ulano, M. A. Bredella, P. Burke, I. Chebib, F. J. Simeone, A. J. Huang, M. Torriani, C. Y. Chang, Distinguishing Untreated Osteoblastic Metastases From Enostoses Using CT Attenuation Measurements, AJR. American journal of roentgenology 207 (2) (2016) 362–368. doi:10.2214/AJR.15.15559.
URL https://pubmed.ncbi.nlm.nih.gov/27101076/

[35] G. Bradski, The OpenCV Library, Dr. Dobb's Journal of Software Tools (2000).

[36] A. Elliott, M. Everts, T. Braun-Jones, R. Court, H. Johnson, I. Norton, J. Warner, A. Ghayoor, H. Meine, P. Fischer, S. Ekström, T. Billah, D. Brown, G. M. Fleishman, J. v. d. Gronde, Nils, K. Leinweber, M. Scheifer, mhe/pynrrd: v0.4.3 Released (4 2022). doi:10.5281/ZENODO.6501810.
URL https://zenodo.org/record/6501810

[37] Teem, nrrd.
URL http://teem.sourceforge.net/nrrd/index.html

[38] G. Lemaˆıtre Lemaˆıtre, F. Nogueira, C. K. Aridas char, Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning, Journal of Machine Learning Research 18 (2017) 1–5. doi:10.5555/3122009.
URL http://jmlr.org/papers/v18/16-365.html.

[39] C. H. Brase, C. P. Brase, Understanding basic statistics (2013) 5.

[40] K. Torvik, J. Hølen, S. Kaasa, O. Kirkevold, A. Holtan, U. Kongsgaard, T. Rustøen, Pain in elderly hospitalized cancer patients with bone metastases in Norway, International journal of palliative nursing 14 (5) (2008) 238–245. doi:10.12968/IJPN.2008.14.5.29491.
URL https://pubmed.ncbi.nlm.nih.gov/18563017/

[41] C. Xie, R. Du, J. W. Ho, H. H. Pang, K. W. Chiu, E. Y. Lee, V. Vardhanabhuti, Effect of machine learning re-sampling techniques for imbalanced datasets in 18 F-FDG PET-based radiomics model on prognostication performance in cohorts of head and neck cancer patients, European journal of nuclear medicine and molecular imaging 47 (12) (2020) 2826–2835. doi:10.1007/S00259-020-04756-4.
URL https://pubmed.ncbi.nlm.nih.gov/32253486/

[42] R. Tibshirani, Regression shrinkage and selection via the lasso: a retrospective, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73 (3) (2011) 273–282. doi:10.1111/J.1467-9868.2011.00771.X.
URL https://onlinelibrary.wiley.com/doi/full/10.1111/j.1467-9868.2011.00771.x

[43] P. Yin, N. Mao, C. Zhao, J. Wu, C. Sun, L. Chen, N. Hong, Comparison of radiomics machine-learning classifiers and feature selection for differentiation of sacral chordoma and sacral giant cell tumour based on 3D computed tomography features, European radiology 29 (4) (2019) 1841–1847. doi:10.1007/S00330-018-5730-6.
URL https://pubmed.ncbi.nlm.nih.gov/30280245/

[44] J. D. Shur, S. J. Doran, S. Kumar, D. Ap Dafydd, K. Downey, J. P. O'connor, N. Papanikolaou, C. Messiou, D. M. Koh, M. R. Orton, Radiomics in oncology: A practical guide, Radiographics 41 (6) (2021) 1717–1732. doi:10.1148/RG.2021210037/

ASSET/IMAGES/LARGE/RG.2021210037TBL6.JPEG.
URL https://pubs.rsna.org/doi/10.1148/rg.2021210037

[45] R. J. Fleischman, D. G. Frazer, M. Daya, J. Jui, C. D. Newgard, Effectiveness and Safety of Fentanyl Compared with Morphine for Out-of-Hospital Analgesia, Prehospital emergency care : official journal of the National Association of EMS Physicians and the National Association of State EMS Directors 14 (2) (2010) 167. doi:10.3109/10903120903572301.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2924527/

[46] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, F. Herrera, Learning from Imbalanced Data Sets, Learning from Imbalanced Data Sets (2018). doi:10.1007/978-3-319-98074-4.

[47] Q. Sun, X. Lin, Y. Zhao, L. Li, K. Yan, D. Liang, D. Sun, Z. C. Li, Deep Learning vs. Radiomics for Predicting Axillary Lymph Node Metastasis of Breast Cancer Using Ultrasound Images: Don't Forget the Peritumoral Region, Frontiers in Oncology 10 (2020) 53. doi:10.3389/FONC.2020.00053/BIBTEX.

[48] C. S. Lisson, C. G. Lisson, M. F. Mezger, D. Wolf, S. A. Schmidt, W. M. Thaiss, E. Tausch, A. J. Beer, S. Stilgenbauer, M. Beer, M. Goetz, Deep Neural Networks and Machine Learning Radiomics Modelling for Prediction of Relapse in Mantle Cell Lymphoma, Cancers 14 (8) (4 2022). doi:10.3390/CANCERS14082008.
URL https://pubmed.ncbi.nlm.nih.gov/35454914/

[49] N. Thai-Nghe, Z. Gantner, L. Schmidt-Thieme, Cost-sensitive learning methods for imbalanced data, Proceedings of the International Joint Conference on Neural Networks (2010). doi:10.1109/IJCNN.2010.5596486.

[50] A. Liu, C. Martin, B. La Cour, J. Ghosh, Effects of Oversampling Versus Cost-Sensitive Learning for Bayesian and SVM Classifiers, Vol. 8, Springer, Boston, MA, 2010. doi:10.1007/978-1-4419-1280-0.

[51] S. Tamang, M. Humbert-Droz, M. Gianfrancesco, Z. Izadi, G. Schmajuk, J. Yazdany, Practical Considerations for Developing Clinical Natural Language Processing Systems

for Population Health Management and Measurement, JMIR Medical Informatics 11 (2023) e37805. doi:10.2196/37805.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9846439/

[52] T. M. F. Smith, W. G. Cochran, Sampling Techniques, Second Edition., Applied Statistics 13 (1) (1964) 54. doi:10.2307/2985224.

5.8 Appendix: Supplemental Information

5.8.1 Sample size calculation

We used Cochran's sample size formula [52] to determine the minimum sample size required to evaluate the performance of the pipeline. An initial audit of our data set showed that the probabilities of finding a patient in our data set with 'pain' was 85% (ppain=0.85) and with 'no pain' was 15% (pno-pain=0.15). To ensure that the pain-score detection is precise within a 95% confidence level ($Z_{1-\alpha/2} = 1.96$), and a 5% margin of error (e=0.05), the minimum sample size was determined as,

$$N = p_{pain} * p_{no-pain} \left(\frac{Z_{1-\alpha/2}}{e} \right)^2 = 0.85 * 0.15 * (1.96/0.05)^(2) = 196$$

Therefore, we included 200 patients in this study to satisfy the minimum sample size requirement.

5.8.2 Time gap between the consultation note and CT acquisition dates

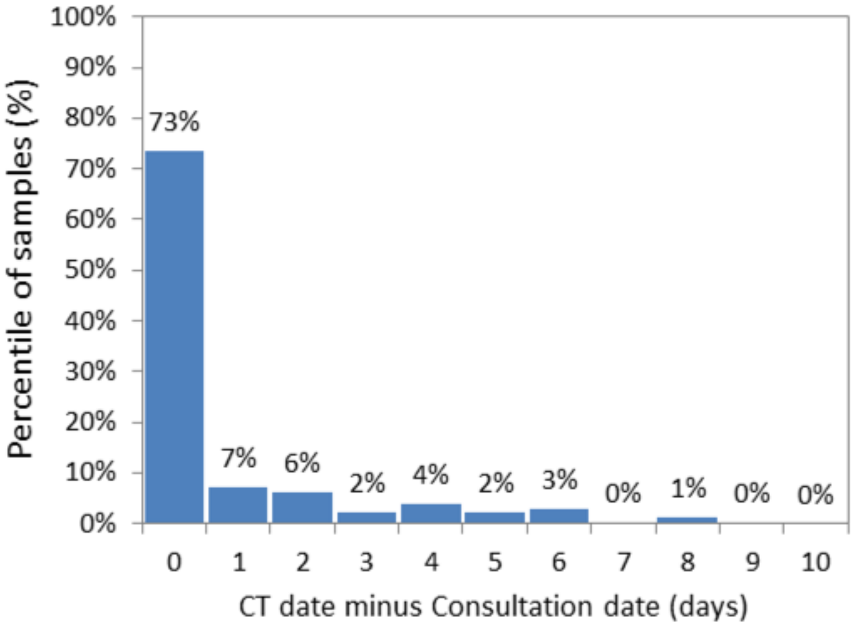


Figure 5.5: The distribution of the time interval between the CT acquisition date and RT consultation date (n = 239 pairs).

5.8.3 Pain labels

NLP-extracted APIs and VDP values, and expert-extracted pain scores in our database are presented in Table 5.5. The box plot comparing the distribution of NLP-extracted API values versus expert-extracted pain scores is shown in Figure 5.6.

id	api	vdp	pain score	id	api	vdp	pain score	id	api	vdp	pain score
p1	-1	no pain	none	p67	1	pain	severe	p133	0.14	pain	moderate
p2	-0.5	no pain	none	p68	1	pain	severe	p134	1	pain	moderate
p3	-0.33	no pain	none	p69	0.5	pain	severe	p135	1	pain	moderate
p4	-0.33	no pain	none	p70	0.5	pain	severe	p136	0.82	pain	moderate
p5	-1	no pain	none	p71	1	pain	severe	p137	1	pain	moderate
p6	-1	no pain	none	p72	0.43	pain	severe	p138	1	pain	moderate
p7	-1	no pain	none	p73	0.33	pain	severe	p139	0.45	pain	moderate
p8	-0.5	no pain	none	p74	1	pain	severe	p140	1	pain	moderate
p9	-1	no pain	none	p75	0.67	pain	severe	p141	0.33	pain	moderate

p10	-0.33	no pain	none	p76	0.2	pain	none	p142	0.47	pain	moderate
p11	-1	no pain	none	p77	1	pain	none	p143	0.6	pain	moderate
p12	-0.5	no pain	none	p78	0	pain	none	p144	1	pain	moderate
p13	-1	no pain	none	p79	0.6	pain	none	p145	0.14	pain	moderate
p14	-1	no pain	none	p80	0.5	pain	none	p146	1	pain	moderate
p15	-1	no pain	na	p81	0	pain	none	p147	0	pain	moderate
p16	-1	no pain	na	p82	1	pain	na	p148	1	pain	moderate
p17	-1	no pain	na	p83	1	pain	na	p149	0.67	pain	mild
p18	-1	no pain	na	p84	1	pain	na	p150	0.6	pain	mild
p19	-0.14	no pain	na	p85	1	pain	na	p151	1	pain	mild
p20	-1	no pain	moderate	p86	1	pain	na	p152	1	pain	mild
p21	-0.43	no pain	moderate	p87	0.5	pain	moderate	p153	0	pain	mild
p22	-0.33	no pain	mild	p88	0	pain	moderate	p154	0.5	pain	mild
p23	-1	no pain	mild	p89	0.87	pain	moderate	p155	0.11	pain	mild
p24	-1	no pain	mild	p90	0.5	pain	moderate	p156	0.33	pain	mild
p25	0.33	pain	severe	p91	0.14	pain	moderate	p157	0.5	pain	mild
p26	1	pain	severe	p92	1	pain	moderate	p158	1	pain	mild
p27	0.67	pain	severe	p93	0.2	pain	moderate	p159	0	pain	mild
p28	0.33	pain	severe	p94	1	pain	moderate	p160	0.45	pain	mild
p29	1	pain	severe	p95	0.2	pain	moderate	p161	1	pain	mild
p30	1	pain	severe	p96	0	pain	moderate	p162	1	pain	mild
p31	0.5	pain	severe	p97	1	pain	moderate	p163	0	pain	mild
p32	1	pain	severe	p98	1	pain	moderate	p164	0	pain	mild
p33	0.33	pain	severe	p99	1	pain	moderate	p165	0.33	pain	mild
p34	0	pain	severe	p100	0.6	pain	moderate	p166	0.14	pain	mild
p35	0.5	pain	severe	p101	0.5	pain	moderate	p167	1	pain	mild
p36	0.67	pain	severe	p102	0.2	pain	moderate	p168	1	pain	mild
p37	0.33	pain	severe	p103	1	pain	moderate	p169	0.33	pain	mild
p38	0.71	pain	severe	p104	1	pain	moderate	p170	0.14	pain	mild
p39	1	pain	severe	p105	1	pain	moderate	p171	0.33	pain	mild
p40	0.75	pain	severe	p106	1	pain	moderate	p172	1	pain	mild
p41	0.67	pain	severe	p107	0.5	pain	moderate	p173	1	pain	mild
p42	1	pain	severe	p108	1	pain	moderate	p174	0.64	pain	mild
p43	1	pain	severe	p109	1	pain	moderate	p175	1	pain	mild

p44	0.5	pain	severe	p110	0.33	pain	moderate	p176	0.67	pain	mild
p45	0.75	pain	severe	p111	1	pain	moderate	p177		na	mild
p46	0.56	pain	severe	p112	1	pain	moderate	p178		na	mild
p47	1	pain	severe	p113	1	pain	moderate	p179		na	mild
p48	1	pain	severe	p114	1	pain	moderate	p180		na	moderate
p49	0.11	pain	severe	p115	0.78	pain	moderate	p181		na	moderate
p50	0.6	pain	severe	p116	0	pain	moderate	p182		na	moderate
p51	0.6	pain	severe	p117	1	pain	moderate	p183		na	moderate
p52	1	pain	severe	p118	1	pain	moderate	p184		na	moderate
p53	0.67	pain	severe	p119	0.5	pain	moderate	p185		na	moderate
p54	1	pain	severe	p120	0.6	pain	moderate	p186		na	moderate
p55	1	pain	severe	p121	1	pain	moderate	p187		na	na
p56	1	pain	severe	p122	1	pain	moderate	p188		na	na
p57	1	pain	severe	p123	0.5	pain	moderate	p189		na	na
p58	0.71	pain	severe	p124	0.33	pain	moderate	p190		na	na
p59	0	pain	severe	p125	0.56	pain	moderate	p191		na	na
p60	0	pain	severe	p126	0.67	pain	moderate	p192		na	na
p61	0.33	pain	severe	p127	1	pain	moderate	p193		na	na
p62	0	pain	severe	p128	0.67	pain	moderate	p194		na	none
p63	0.33	pain	severe	p129	1	pain	moderate	p195		na	severe
p64	1	pain	severe	p130	0.33	pain	moderate	p196		na	severe
p65	0.2	pain	severe	p131	0.33	pain	moderate	p197		na	severe
p66	0.43	pain	severe	p132	1	pain	moderate				

Table 5.5: The performance of our best performing NLP-radiomics pipeline (neural networks with the EN6 ROI) on the training and test sets. The results of the same radiomics model (neural networks with EN6 ROI) when trained and tested using expert-extracted pain labels, together with the results from a prior study by Wakabayashi et al. [20] are provided for comparison. The reason for having high specificity and low sensitivity in our test set is explained in the discussion section.

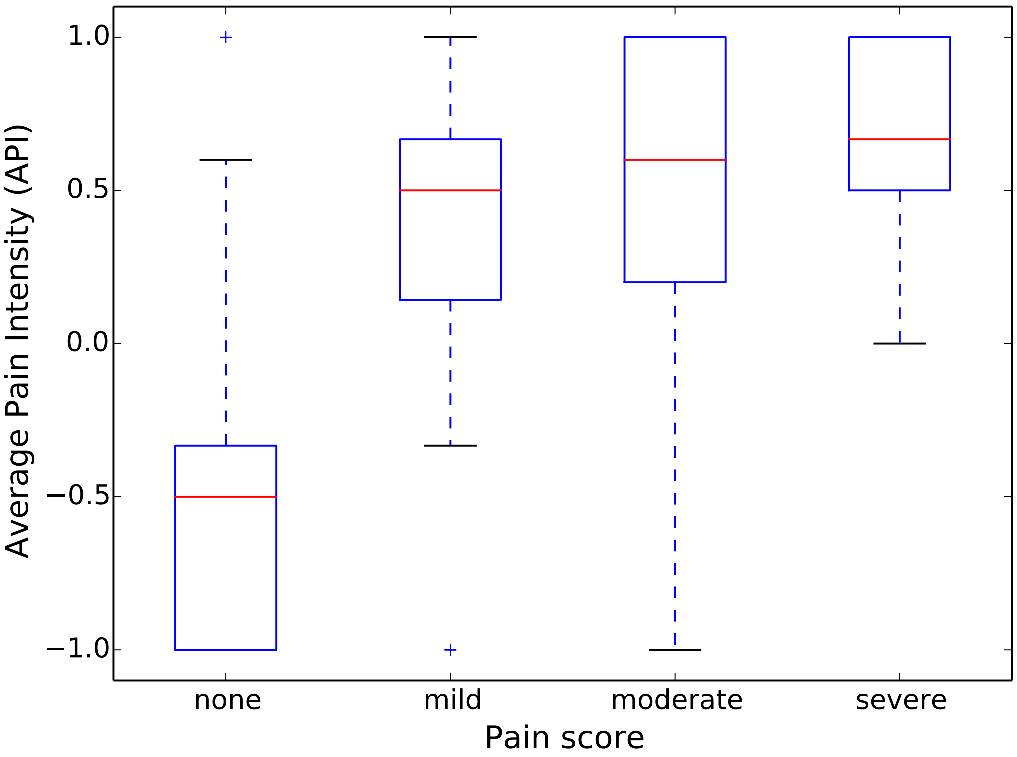


Figure 5.6: Relation between expert-extracted pain scores and NLP-extracted average pain intensities (API). The box plot is generated using the pyplot package.

Chapter 6

Conclusions

6.1 Summary and novelty of work

6.1.1 Objectives

The ability to objectively measure and predict cancer pain has the potential to play a role in personalized care and improve the quality of life for cancer patients. Motivated by this potential, the overarching objective of this thesis project was to develop, implement, and evaluate an AI pipeline to detect pain in the simulation-CT images of cancer patients with BM.

In working towards our overarching objective, we demonstrated that NLP applied to radiation oncology consultation notes and radiomics analysis of simulation-CT scans can be combined to find imaging biomarkers (features) that can be used to identify pain caused by BM. This thesis not only made progress toward objectively detecting BM pain using radiographic images, but it also demonstrated the use of a generalizable and scalable pain detection pipeline that can potentially be applied to different study contexts in the future and, ultimately, translated into the clinic.

This chapter provides a summary of the thesis project, highlighting the main results of each of the three objectives that went into achieving the overarching objective.

6.1.2 Objective 1: Construct an NLP pipeline to extract pain scores from the consultation notes of patients

Patients with BM receiving palliative RT are not always asked to fill out standardized pain screening questionnaires. This was the case at our institution. Accordingly, in this study, we made use of pain scores that physicians recorded in radiation oncology consultation notes. To automatically extract these pain scores, we developed and trained a database-independent NLP pipeline using the publicly-accessible i2b2 and MIMIC-III hospital discharge summary corpora. Then, we used our pipeline to extract pain scores for cancer patients with BM from the consultation notes in our institutional radiation oncology information system.

The performance of our NLP pipeline was evaluated using pain scores in physician-annotated best-available gold standard corpora. We obtained these physician-annotated pain scores with the help of clinicians who used our in-house developed manual pain annotation and scoring tool (texTRACTOR).

Our work demonstrated that pain is poorly documented in consultation notes and that physician-reported pain ratings lack sufficient resolution to extract high resolution numerical or verbal pain scores; yet, they can still be used to extract binary pain scores as “pain” and “no-pain”. We also showed that a generalizable NLP pipeline can be trained on publicly-available data to extract these binary pain labels. Our pipeline successfully extracted and identified physician-reported binary pain labels from our radiation oncology clinical notes, with 80% recall and 84% precision.

6.1.3 Objective 2: Construct a radiomics pipeline to extract BM lesion features from radiographic images of patients

Radiomics-based ML models have demonstrated the potential to detect BM, evaluate BM response to RT, and predict outcomes. However, current radiomics models require large imaging datasets with 3D ROIs that have been segmented by experts. Full ROI segmentation is time-consuming for clinicians, posing a difficulty for large-scale real-world radiomics research. Consequently, a method to facilitate simple BM identification without compromising the efficacy of radiomics is desired. The purpose of this objective was to investigate the viability of constructing a rapid pipeline for radiomics research employing geometric ROIs based on just lesion centerpoints.

To achieve this objective, we created a custom-written 3D DICOM visualization web application (diCOMBINE) for radiation oncologists to quickly identify lesion centerpoints. Then, folowing BM lesion centerpoint identification by our radiation oncologist colleagues, radiomic features were calculated using spherical and cylindrical ROIs automatically delineated around each centerpoint.

We evaluated and demonstrated the efficacy of our centerpoint-based radiomics pipeline for discriminating between healthy and metastatic bone lesions. Using ensemble ROIs, we showed that the GPR, NNet, and L-SVM classifiers achieved an F-1 score of 0.9 in detecting BM. The ROC-AUC, precision, recall, and F1 score of our best performing pipeline, which corresponded to the E9SC ROI, LASSO FS technique, and GPR ML classifier, were 96%, 92%, 91%, and 0.9, respectively. These results are comparable to those of other studies utilizing full 3D segmented ROIs. Our lesion-centerpoint-based segmentation technique significantly simplifies the preparation of images for radiomics research and eliminates the bottleneck of obtaining full 3D ROI segmentation.

6.1.4 Objective 3: Combine the NLP-quantified pain scores extracted in objective 1 with the radiomic features extracted in objective 2, to develop a radiomics-based machine-learning model of pain in patients with BM

In this objective, we combined the tools that we developed in our earlier objectives to evaluate the viability of utilizing radiomics-based ML models trained using NLP-extracted pain scores to assist in the identification of pain using the simulation-CT images of cancer patients with BM.

First, we extracted radiomics features from simulation-CT images of BM patients using the ensemble lesion-centerpoint-based geometrical ROIs introduced in Objective 2 (Chapter 4). Second, we merged these features with pain scores extracted from patient consultation notes using the NLP pipeline we developed for Objective 1 (Chapter 3). Finally, we created a ML model to differentiate between painful and painless BM lesions.

We showed that our NLP- and radiomics-based neural network model was able to differentiate between painful and painless BM lesions on simulation CT scans. The accuracy, sensitivity, specificity, and ROC-AUC scores of our top-performing model (Neural Network classifier on the EN6 ensemble ROI) in the test set were 0.82 (132 of 163), 0.59 (16 of 27), 0.85 (116 of 136) and 0.82, respectively.

Our radiomics-based neural network model that was trained on NLP-extracted pain scores showed comparable diagnostic performance (ROC-AUC, 82.5%) to the same radiomics-based model trained using expert-extracted pain scores (ROC-AUC, 82.3%). It is also consistent with similar work reported in the literature [1] that used patient-reported pain scores and full 3D segmented ROIs (ROC-AUC, 82.0%). However, according to our collaborating radiation oncology experts, data preparation for our pipeline, which uses NLP to extract pain scores and just lesion centerpoints for lesion segmentation, is 15 times more time-efficient due to the need for minimal expert involvement. This paves the way for the development of scalable "Big Data" ML pipelines with semi-automated data curation via NLP and centerpoint-based radiomics.

6.2 Future directions

The following subsections detail possible extensions of the present project to build upon its findings and address some of its limitations.

6.2.1 NLP for extracting pain scores

Our NLP pipeline outputs a single VDP score for each consultation note. As explained in Chapter 3 when we investigated the i2b2 training set, we discovered that our classification algorithm struggled to extract VDP when patients reported pain in multiple sites in their bodies. Patients with BM often have multiple pain sites with different pain scores on each site, which confounds our VDP measurement approach for effective BM pain extraction.

Future work should incorporate a pipeline to obtain pain severity from the consultation notes of patients by extracting numerical pain scores and pain assessment terms (such as severe, mild, and controlled). In addition, the pipeline should attempt to identify and extract pain sites so that pain scores can be localized for patients with multiple cancer sites. The evaluation of such a high-resolution pain scoring pipeline necessitates the collection of patient-reported pain scores via standard pain questionnaires and graphical tools for patients to indicate the location of their pain. Using electronic questionnaires accessible through patient engagement tools such as the Opal patient portal developed by our research group (https://www.opalmedapps.com/) [2], we believe it will be possible to collect location-identified gold-standard pain scores directly from patients.

6.2.2 Automated BM detection pipeline

Future work in constructing an automatic BM detection pipeline should acquire a much larger dataset of images with lesion centerpoint labels to better train a ML model, as explained in Chapter 4. PET scans and pathology reports may be utilized to better identify and classify healthy bones, metastatic lesions, and non-metastatic skeletal complications (e.g., surgically-removed lesions or bone islands). Also, although the use of geometric ROIs considerably simplified the lesion delineation method in our work, it ignored some potentially useful lesion features such as size and shape. Future work should thus consider the use of deep-learning-based ROI segmentation methods.

Such a deep-learning-based application will require training on a large collection of images that generally requires collecting data from multiple centers. A large multicenter dataset would also allow for testing of our radiomics pipeline's generalizability and its broader clinical acceptability.

6.2.3 Identifying imaging biomarkers for pain quantification

Although our results demonstrated that it is possible to use radiomics features to distinguish between painful and painless BM lesions, more research is required to figure out the connections between these radiomics features and subjective sensations of pain. As we explained in Chapter 5, future work should attempt to improve upon our pain detection pipeline by using more reliable pain data. For example, patient-reported pain scores could be used, allowing for more accurate ground-truth pain labeling.

Using such granular pain data, it will be possible to assess the stability, pain dependence, and predictive capability of each radiomics feature. Additionally, it will be possible to examine the impact of the number and type of BM lesions on the intensity of pain. Due to the lack of granularity of the extracted pain scores, these investigations were not feasible for this thesis [3].

Finally, our pipeline may be applied to past and present CT images of patients, as well as consultation notes, in order to develop a longitudinal model of pain.

6.2.4 Pain prediction in clinical radiation oncology

Medical practitioners and patients can benefit from pain prediction models when making treatment choices. An effective pain prediction model should provide quantitative prognostic information to facilitate personalized clinical decision-making, with the end goal of improving patient outcomes, such as quality of life. While outcome prediction models have the potential to greatly improve patient care, as introduced in this thesis, the vast majority of these models have not yet been implemented into clinical practice.

In order for a model to be used in the clinic, it must first be tested and validated on large, multi-center databases that have been standardized and have labels that can be trusted (gold standard data).

The minimum number of samples necessary for clinically verifying an end-to-end outcome prediction model is highly dependent on parameters such as the outcome's characteristics and the quality of the collected data. With many types of data (images, pathology reports, genetic sequencing results, etc.), it may be sufficient to validate a typical outcome prediction model using just a minimum of one thousand patients. However, in hospital settings, it can be difficult to find even this many patients with complete data. The problem stems from the complexity of obtaining patient information from unstructured health records and the lack of standardized data warehouses in hospitals. Obtaining multi-center data is an even greater challenge. The main barriers are differences in standardization among healthcare systems, non-uniform record-keeping processes, and the necessity to preserve patients' privacy and confidentiality.

In recent years, several initiatives have been launched to address these issues. The implementation of unified data repositories and application programming interfaces (API) for medical systems will faciliate access to anonymized patient data within a hospital. The implementation of standardized data transfer protocols such as HL7-FHIR and mCODE will allow for uniform information flow between various health care systems. The development of multi-institution patient portals, such as Opal, that permit patients to access their medical data from various centers, will enable secure and patient-consented acquisition of multi-center data. Finally, the availability of publicly-available multi-center datasets of de-identified patient data, such as the Cancer Imaging Archive and MIMIC-III clinical databases, will allow new AI algorithms to be tested for generalizability and scalability.

Any data included in a ML-based outcome prediction model must be preprocessed to ensure that the data are comprehensive, non-sparse, cleaned, standardized, and correctly labeled. In the majority of existing ML pipelines, preprocessing is performed manually or semi-automatically with minimal or no standardization, resulting in a bottleneck for preparing standard data for large-scale ML models. Some of these challenges can be overcome by using automated data collection pipelines and standardized data preparation techniques. For example, collecting patient-reported pain data using automatically-scheduled standard electronic pain questionnaires sent to the patient using the Opal patient portal will enable access to large-scale non-sparse ground truth data at our institution. Furthermore, the image processing tools created for this thesis adhere to the protocols specified by the Image Biomarker Standardization Initiative (IBSI), enabling multi-center research.

While the work presented in this thesis does not include a clinically applicable model for predicting pain in patients with BM, it does represent a step towards this vision and helps pave the way for the development of rapidly-trained and efficient clinical AI-based decision-making tools that may help reduce the burden of pain management and improve the quality of life for patients with BM.

Bibliography

[1] K. Wakabayashi, Y. Koide, T. Aoyama, H. Shimizu, R. Miyauchi, H. Tanaka, H. Tachibana, K. Nakamura, T. Kodaira, A predictive model for pain response following radiotherapy for treatment of spinal metastases, Scientific Reports 11 (1) (2021) 1–8. doi:10.1038/s41598-021-92363-0.
URL https://www.nature.com/articles/s41598-021-92363-0

[2] J. Kildea, J. Battista, B. Cabral, L. Hendren, D. Herrera, T. Hijal, A. Joseph, Design and Development of a Person-Centered Patient Portal Using Participatory Stakeholder Co-Design, J Med Internet Res 2019;21(2):e11371 https://www.jmir.org/2019/2/e11371 21 (2) (2019) e11371. doi:10.2196/11371.
URL https://www.jmir.org/2019/2/e11371

[3] H. Naseri, S. Skamene, M. Tolba, M. D. Faye, P. Ramia, J. Khriguian, M. David, J. Kildea, A scalable radiomics- and NLP- based machine learning pipeline to distinguish between painful and painless thoracic spinal bone metastases: Algorithm Development and Validation (In press), JMIR AI (12 2022). doi:10.2196/44779.
URL https://preprints.jmir.org/preprint/44779